

Studies in Computational Intelligence 1102

Abdellah Idrissi *Editor*

Modern Artificial Intelligence and Data Science

Tools, Techniques and Systems



Springer

Studies in Computational Intelligence

Volume 1102

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Abdellah Idrissi
Editor

Modern Artificial Intelligence and Data Science

Tools, Techniques and Systems

 Springer

Editor

Abdellah Idrissi
IPSS Team, Artificial Intelligence and Data
Science Group,
Computer Science Department,
Faculty of Science
Mohammed V University in Rabat
Rabat, Morocco

ISSN 1860-949X

ISSN 1860-9503 (electronic)

Studies in Computational Intelligence

ISBN 978-3-031-33308-8

ISBN 978-3-031-33309-5 (eBook)

<https://doi.org/10.1007/978-3-031-33309-5>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Artificial Intelligence (AI) is often all over the news these last few years. Some fear that AI is getting so advanced that people will start losing their jobs and their habits. In fact, one area humans will always be better at is emotional intelligence. Computers can't be emotional, but they can trick us into believing they are.

Yes, we can say that AI is breaking records lately on all fronts. Its catalyst is the abundance of data that is constantly being generated with the various existing technological and material means. Being coupled with data science and applied to different areas of everyday life, AI is usable everywhere and it can be applied and can help enormously in every area. In fact, we have really a special opportunity in terms of transformation. We cannot predict what will happen in the the near future. But what is sure, is that almost all services will have an AI component. There's an entire ecosystem under construction that will bring a new set of products and services. Now, we are witnessing more than another industrial revolution, it is something that exceeds humanity and even life itself.

AI affects all our lives on a daily basis. It influences our choices in a way or another. It is potentially offering solutions and alternatives to many practical issues of everyday life and if well explored it can positively revolutionize the life of the people. Thus, we must approach this discipline to contribute to its better orientation in the service of humanity.

To sum up, over the past few decades, we are undoubtedly witnessing an unprecedented technological revolution. Thus, we live in an era where everything changes very quickly and where everything is connected thanks to very advanced developments both in new technologies in general and in computing in particular. AI, big data, internet of things, cognitive computing, high-performance computing, blockchain, cloud and edge computing, and virtualization, have invaded the world of computing. These new disciplines allow the development of several applications to facilitate human tasks and thus mainly deal with real-world activities which, at a certain time, were relatively difficult to accomplish such as health, education, agriculture, energy, environment, security, cyber security, industry, economy, finance, transport, societal issues, and in general, in all areas we can think of. Our goal, as researchers and the

scientific community, is to guide all these technologies to put them at the service of humanity.

All research chapters published in this book are focused on this large domain. Therefore, this book entitled *Modern Artificial Intelligence and Data Science* is focused on intelligent applications whose main objective is to enrich existing scientific research to enlarge the state of the art and to propose new methods and innovative approaches to deal with various problems in society. In this context, we have tried to explore some important branches of AI and presented several ideas using AI and its extension in solving certain real-life problems, for known and unknown environments. Thus, this book is composed of 26 chapters grouped into 4 topics dealing with AI and its applications, namely, AI and recommendation systems (chapters *TopKws Algorithm in the Map-Reduce Paradigm for Cloud Computing QoS Recommendation System* to *A Bi-LSTM Neural Network to Forecast Stock Market Index*); AI and Advanced Technologies (chapters *Fast Yolo V7 Based CNN for Video Streaming Sea Ship Recognition and Sea Surveillance* to *Comparative Analysis of Skyline Algorithms used to Select Cloud Services Based on QoS*); AI applied to Blockchain and IoT (chapters *A State-of-the-Art Survey on Ransomware Detection using Machine Learning and Deep Learning* to *A Comparative Study of Consensus Algorithms Used in Blockchain and Their Adaptation to the IoT Networks*); and finally AI applied to some Academic and Real Problems (chapters *Dynamics Behavior of Vehicular Traffic Flow in a Scale-Free Complex Network* to *Experimenting with Polymorphic Creativity Support Tools to Support Innovation in Participatory Ideation*).

The first topic entitled *Artificial Intelligence and Recommendation Systems* deals with map-reduce model for Top_KWS recommendation algorithm in cloud computing QoS, a new context-based factorization machines for Context-Aware Recommender Systems (CARS), a review of recommendation systems and their applications in e-learning, a novel courses recommendation system based on Divide & Conquer and Skyline BNL algorithms, an application of deep reinforcement learning solution in financial markets, the use of exascale computing systems to implement scalable solutions for the analysis of massive data and a bi-LSTM neural network to forecast stock market index.

The second topic deals with the interaction between AI and advanced technologies such as machine learning and computer vision. In this context, the readers will discover in these seven chapters a Fast Yolo V7-based CNN for video streaming sea ship recognition and sea surveillance, a graph-based approach for multilingual sentiment analysis, a method to predict blood glucose levels in type 1 diabetes Using LSTM, a machine learning approach to identify sarcasm in social media, a method to predict school dropout using machine learning algorithms, a survival prediction in patients with nasopharyngeal cancer using machine learning techniques, and as last chapter in this topic a Comparative analysis of skyline algorithms used to select cloud services on the basis of QoS.

Part “Artificial Intelligence Applied to Blockchain and IoT” is consecrated for AI applied to blockchain and IoT. The six chapters presented here deal with a survey of ransomware detection studies using machine learning and deep learning techniques, an improvement of the application of blockchain technology for tracking processes

in the supply chain integrated business intelligence, studying consensus mechanisms for blockchain, design and construction of a smart agricultural greenhouse, an implementation and management of a home automation control system (Smart Home), a comparative study of consensus algorithms used in blockchain and their adaptation to the IoT networks.

The fourth and last topic illustrates AI applied to some academic and real problems. In this context, the authors present in these six chapters some studies on the dynamics behavior of vehicular traffic flow in scale-free complex network, a customer journey map discovery approach, an efficient improvement for unmanned aerial vehicle-assisted clustered wireless sensor network data collection, a synergistic fibroblast optimization algorithm for solving knapsack problem, implications of perceived ease of use, perceived usefulness, and self-efficacy on TVETs' acceptance of JAWS as an assistive computer application software, and finally, last but not least effort, deals with an experimenting with polymorphic creativity Support tools to support innovation for participatory ideation.

We consider that this book presents some really interesting advances in the field of AI coupled with data science and their applications. It contributes to their evolution, emergence, and particularly their guidance in the service of the humanity.

We would like to thank all the authors for their interactions, involvements, and interesting contributions.

In addition, we would like to warmly thank and sincerely acknowledge the great efforts of the editors, especially Professor Janusz Kacprzyk, Dr Thomas Ditzinger, Dr Sylvia Schneider, Dr Hemavathy Manivannan, and Dr Manopriya Saravanan for their great help and support and to any person who contribute to promote the Springer Nature Publisher, particularly the Series of Computational Intelligence Studies.

Rabat, Morocco

Prof. Abdellah Idrissi

Contents

Artificial Intelligence and Recommendation Systems	
<i>TopKWS</i> Algorithm in the Map-Reduce Paradigm for Cloud Computing QoS Recommendation System	3
Kaoutar El Handri, Abdellah Idrissi, and Aicha Er-Rafyg	
A New Context-Based Factorization Machines for Context-Aware Recommender Systems	15
Rabie Madani, Abdellah Idrissi, and Abderrahmane Ez-Zahout	
Review of Recommendation Systems and their Applications in E-learning	25
Aicha Er-Rafyg and Abdellah Idrissi	
Improvement of Courses Recommendation System using Divide and Conquer Algorithm	37
Aicha Er-Rafyg, Abdellah Idrissi, and Kaoutar El Handri	
Deep Reinforcement Learning in Financial Markets Context: Review and Open Challenges	49
Youness Boutyour and Abdellah Idrissi	
A Survey of Parallel Computing: Challenges, Methods and Directions	67
Meryem Bouras and Abdellah Idrissi	
A Bi-LSTM Neural Network to Forecast Stock Market Index	83
Zakaria Al Bakkari, Ikram El Azami, and Adil El Makrani	
Artificial Intelligence and Advanced Technologies	
Fast Yolo V7 Based CNN for Video Streaming Sea Ship Recognition and Sea Surveillance	99
Abdelilah Haijoub, Anas Hatim, Mounir Arioua, Slama Hammia, Ahmed Eloualkadi, and Antonio Guerrero-González	

Graph Convolutional Network for Multilingual Sentiment Analysis 111
 El Mahdi Mercha, Houda Benbrahim, and Mohammed Erradi

Predicting Blood Glucose Levels in Type 1 Diabetes Using LSTM 121
 Dounia Nasir, Mohamed Elmehdi Ait Bourkha, Anas Hatim,
 Said Elbeid, Siham Ez-ziymy, and Khalid Zahid

**Sarcasm Detection on Social Media using Machine Learning
 Approach** 137
 Chahrazad Lagrini and Abdellah Idrissi

School Dropout Prediction using Machine Learning Algorithms 147
 Said Ouabou, Abdellah Idrissi, Abdeslam Daoudi,
 and Moulay Ahmed Bekri

**Survival Prediction in Patients with Nasopharyngeal Cancer Using
 some Machine Learning Methods** 159
 Abdellah Idrissi, Hasna Lakrim, and Mehdi Bouskri

**Comparative Analysis of Skyline Algorithms used to Select Cloud
 Services Based on QoS** 169
 El Khammar Imane, Abdellah Idrissi, Mohamed El Ghmary,
 and Kaoutar El Handri

Artificial Intelligence Applied to Blockchain and IoT

**A State-of-the-Art Survey on Ransomware Detection using
 Machine Learning and Deep Learning** 183
 Loubna Moujoud, Meryeme Ayache, and Abdelhamid Belmekki

**Improving the Application of Blockchain Technology for Tracking
 Processes in the Supply Chain Integrated Business Intelligence** 201
 Khadija El Fellah, Adil El Makrani, and Ikram El Azami

Studying Consensus Mechanisms for Blockchain 213
 Hamza El Mezouari and Fouzia Omary

Design and Construction of a Smart Agricultural Greenhouse 225
 Moulay Ahmed Bekri, Abdellah Idrissi, Said Ouabou,
 and Abdeslam Daoudi

**Implementation and Management of a Home Automation Control
 System (Smart Home)** 233
 Abdeslam Daoudi, Abdellah Idrissi, Moulay Ahmed Bekri,
 and Said Ouabou

**A Comparative Study of Consensus Algorithms Used in Blockchain
 and Their Adaptation to the IoT Networks** 245
 Mohamed Aghroud, Mohamed Oualla, Abdeslam Jakimi,
 and Lahcen Elbermi

Artificial Intelligence Applied to Some Academic and Real Problems

Dynamics Behavior of Vehicular Traffic Flow in a Scale-Free Complex Network 261
Siham Lamzabi, Kaoutar El Handri, Marwa Benyoussef, Hamid Ez-Zahraouy, and Abdelilah Benyoussef

Customer Journey Map Discovery Approach 275
Imane El Alama and Hanae Sbai

Unmanned Aerial Vehicle-Assisted Clustered Wireless Sensor Network Data Collection Efficiency Improvement 281
Mohamed Abid, Said El Kafhali, Abdellah Amzil, and Mohamed Hanini

Synergistic Fibroblast Optimization Algorithm for Solving Knapsack Problem 295
T. T. Dhivyaprabha and P. Subashini

Acceptance of Job Access with Speech (Jaws) as an Assistive Computer Application Software 307
Lihle Ndlovu, Anass Bayaga, and Sylvan Blignaut

Experimenting with Polymorphic Creativity Support Tools to Support Innovation in Participatory Ideation 319
Muhammad Mustafa Hassan, Imran Arshad Choudhry, Markku Tukiainen, and Adnan N. Qureshi

About the Editor

Abdellah Idrissi is graduated Ph.D. in Artificial Intelligence. He is currently a member of the IPSS team where he leads a research group on artificial intelligence and its applications. He is the author of four books and co-author of several publications in international journals and conferences. He is also the co-author of two patents and others are pending. He was a guest editor of five special issues in renowned journals. He is a member of the editorial board of several international journals and a member of the TPC of several international conferences. He is the founder and general chair of two International Conferences, namely, “Modern Artificial Intelligence and Data Science Systems (MAIDSS)” and “Modern Intelligent Systems Concepts (MISC)”, and has chaired numerous international conferences and workshops. He has supervised seven doctoral thesis, which have been defended with excellence, and many more are in progress. He is the founder and coordinator of the Master in Artificial Intelligence and Data Science. He is a partner of several national and international projects and was particularly a partner of the MOSAIC project, funded by the European Commission (FP7 612076), in which 12 partners representing 12 different countries participated. He was, in this last project, the leader in the implementation of the Technology Platform in the Maghreb Region.

Artificial Intelligence and Recommendation Systems

*Top*_{KWS} Algorithm in the Map-Reduce Paradigm for Cloud Computing QoS Recommendation System



Kaoutar El Handri, Abdellah Idrissi , and Aicha Er-Rafyg

Abstract With the evolution of Big Data, Cloud computing has become very popular because it enables the efficient use of the IT resources needed to manage these massive amounts of data. As a result, Cloud markets have become very competitive, given their needs and the technological development that increasingly leverages Big Data infrastructures. As the growing number of various services in the cloud is rapidly evolving in the cloud market, selecting the best service from the vast amount of data is a great challenge. This paper presents a new *TopK* Selection Algorithm based on the Map-Reduce paradigm for recommending the cloud QoS. This processing explores a combined System of recommendation called the n-CSRSS (New Cloud Service Recherche and Selection System) that helps the user choose the best services according to their requirements. The experimental results show that our Recommender system can effectively recommend the right combination of Cloud services to consumers.

Keywords Map-reduce · Top-K · Skyline cloud computing QoS · Recommendation system

1 Introduction

These two IT forces currently interest businesses worldwide: Big Data analysis and cloud computing [1]. The former offers the promise of generating rich information that will lead to competitive advantage, innovation, and increased revenue [2]. Cloud

K. El Handri (✉)

LIMIE Laboratory, Higher Institute of Engineering and Business, 393 Rte d'El Jadida, Casablanca, Morocco

e-mail: kaoutar.elhandri@isga.ma

K. El Handri · A. Idrissi · A. Er-Rafyg

IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department, Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco

e-mail: idrissi@um5r.ac.ma

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,

Studies in Computational Intelligence 1102,

https://doi.org/10.1007/978-3-031-33309-5_1

computing can potentially increase business coordination and productivity as an IT service delivery model while improving efficiency and reducing costs. Both continue to evolve. Professionals are no longer satisfied with knowing how to store Big Data but are trying to analyze this data in a way that is relevant to real business needs [3, 4]. The recommendation system is one of the most important tools for interacting with the end-user and analyzing their requirements. However, a recommendation is an action of computing a list of items (Top-K items) that the user will like the most [5, 6]. The calculation of recommendation lists is done by assigning scores for objects according to their popularity or preferences; one of the most used methods in recommendation systems is the Top_K s algorithm. Besides, the Top-k algorithm's interest is finding the K items with the best global score. However, the data volume and their velocity require some efficient parallel solutions [7, 8] that overcame the restricted resources of a centralized aspect. Our motivating application is a decentralization of the Top_{kWS} algorithm [1, 5, 9] by using Hadoop [21] and particularly Map-Reduce framework (Lin et al. 2010) with its recommendation tools. In fact, within the suggested system, we provide cloud service recommendations based on user preferences. The employment of mentioned technologies has provided the MapReduce version of Top_{kWS} , which represents a meaningful role in the previous work [1, 5, 9, 10] to refine the Skyline result. The MCDA tolls were used in the aggregation [1, 11] function by the adapted weighted sum method (WSM). In addition, thanks to this hybridization, we showcase that our Top_{kWS} [5] algorithm outperforms the Fagin algorithm in terms of runtime measurement and the commonly used Metric correlation study in Cloud Service QoS. The parallel model of Top_{kWS} referred to $MRTop_{kWS}$ in this paper. The rest of this paper is organized as follows. In Sect. 2, we discuss related work. Section 3 exposes our contribution using the Map-Reduce paradigm for combining the Top_k algorithm and the MCDA methods. Section 4 evaluates the parallel approach using the Cloud Service Recommendation database. Finally, in Sect. 5, we conclude this work while giving some perspective.

2 Related Works

2.1 The Use of MCDA in Recommender System

MCDA (Multi-criteria Decision Aid) methods are tools that help to find a decision. Furthermore, among these methods, we apply ELECTRE IS and the weighted sum method, which will remain the MCDA approach and algorithms used in this paper. But before talking about these two algorithms, we need to showcase the nature of the concept behind the MCDA methods, in general, that motivates us to combine them with the Recommender system and start to speak about what we call a Multi-criteria recommendation system (MCRS).

2.2 Map-Reduce and Big Data Architecture

Many Big Data architecture is starting to be used, especially the famous Hadoop Platform. Currently, Hadoop is still the leading platform for Big Data. It is used for storing and processing huge volumes of data [12]. Hadoop is an open-source software framework for storing data and running applications on clusters of standard machines. This framework will allow processing of massive data on clusters ranging from one to numerous hundred devices. Its insistence manages the distribution of data in the cluster machines. The Hadoop framework consists of many Open Source components [13], all connected to a set of core modules designed to capture, process, manage, and analyze large volumes of data. These core technologies are: Hadoop Distributed File System (HDFS). This file system supports a conventional hierarchical directory but distributes files to a set of storage nodes on a Hadoop cluster. It is a programming model, and an execution framework for parallel processing applications in batch mode [14, 15]. It is designed to manage many potentially considerable files in a highly distributed environment (up to thousands of servers). It breaks down the processing of an operation (called a “job” at Hadoop) into several steps, two of which are elementary, to optimize parallel data processing. These operations are, respectively, the mapping and the reducing. In contrast, the mapping performs a specific function on each element of the input list: from a file in the form $\{key, value\}_i$, it then generates an output list in the same way shown in Fig. 1. It should be noticed that there is also an operation between Mapping and Reducing called Shuffling, which rearranges the list elements to prepare the Reducing. The Reducing is then processed, giving the final output. Thanks to YARN, Hadoop has seen several improvements; For example, YARN (Yet Another Resource Negotiator) takes care of the scheduling of tasks (jobs) and allocates resources to the cluster to run applications, and arbitrate when there is a conflict of resources. It also monitors the execution of jobs. Hadoop also [16] has expressed a significant change, notably Hadoop 2.0. However, Hadoop 3.0 includes new features such as erasure coding to manage fault tolerance. Hadoop 3.x also reduces storage costs from 200 up to 50%. It has also implemented a new command-line tool known as Disk balancer. Thus, Hadoop 3.x has improved overall performance Table 1 shows the added value of Apache Hadoop3 over Apache Hadoop2. In this paper, we use MapReduce to paralleling our algorithm, which is a central component of the Apache Hadoop framework. We also operate another new component based on a similar principle and technical behind MapReduce. This component is called the Apache Pig [17]. However, Pig is an abstraction above Map-Reduce. It is a tool and platform applied to analyze extensive flow data. It is commonly used with Hadoop; we can perform all data manipulation and operations in Hadoop using Pig. For Pig, we need to manipulate a high-level language called Pig Latin. This language offers various operators with which programmers can develop their functions for reading, writing, analyzing [18], and processing data. In this paper, we use Pig as an enhancement of our Map-Reduce code, thanks also to the advantage given by Pig comparing it with Map-Reduce, which can be shown in the Table 1.

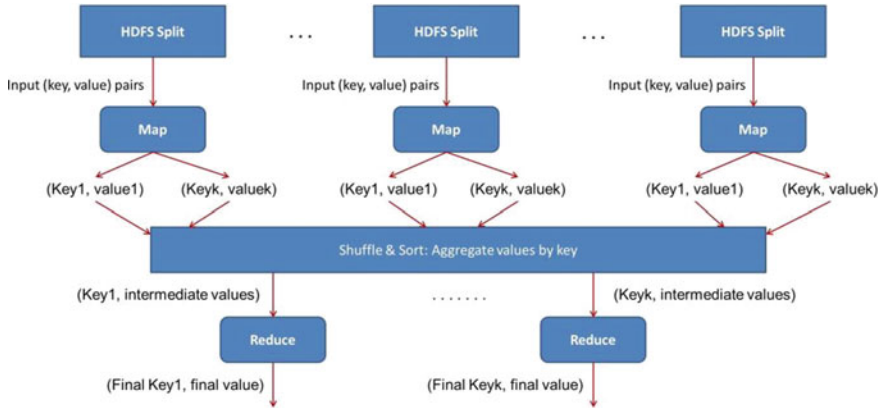


Fig. 1 MapReduce work source: [19]. In the MapReduce framework, the operations consist of two main phases: the Map and Reduce phases. In the Map phase, the input data is divided into smaller chunks, and a map function is applied to each chunk independently to generate intermediate key-value pairs. In the Reduce phase, the intermediate key-value pairs are shuffled and sorted based on the keys, and a reduce function is applied to each unique key to produce the final output. Figure 1 visually depicts the flow and interaction of these operations in the MapReduce framework

Table 1 The added value of Apache Hadoop 3 above Apache Hadoop 2

Attributes	Hadoop2.x	Hadoop3.x
Handling fault-tolerance	Through replication	Through erasure coding
Storage	Consumes 200% in HDFS	Costumes just 50%
Scalability	Limited	Improved
File system	DFS, FTP, and Amazon S3	All features plus Microsoft Azure Data lake File System
Manuel intervention	Not needed	Not needed
Cluster resource management	Handled by YARN	Handled by YARN
Data balancing	User HDFS balancer for this purpose	User intra-data node balancer

3 Our Contribution: $MRTop_{kws}$ Algorithms

Various studies concerned with the parallelism of Top_k algorithms attempt to use the Map-Reduce Framework see Fig. 1, as it is the same state in our solution proposal. However, in this figure, we show that to finalize the implementation of Map-reduce, we take the key-value pair of the reducer and write it in the file by the record writer. By default, it separates the key and the value by a tab and each record by a linefeed character. We can configure it to get a more productive output format. But the final data is still written to the HDFS. In this work, we propose the Algorithms 2 and 3 of parallel Top_{kws} computations based on this paradigm for refining the skyline result, which will be called: $MRTop_{kws}$.

Algorithm 1 Top_{kWS}

```

1:  $Ls$ : input PriorityQueue
2:  $tlArray$ : array of items
3:  $Item$ : input object which will calculate its score function
4:  $k$ : the number of objects returned by the algorithm
5:  $TopList$ : output list of the tuples forming the solution
6: Define  $Ls$  as priority queue based on  $ScoreFunction$ 
7: function  $ComputeTopK$ 
8:  $returned = 0$ 
9: while  $returned < k$  do
10:    $Ls' = \emptyset$   $Ls \in PriorityQueue$ 
11:    $result = Compare(Ls, item)$ 
12:   if  $result < 0$  then
13:     Select from  $Ls$  the object  $item$  with the maximum  $ScoreFunction$ 
14:     Remove the head of this queue or returns null if this queue is empty
15:      $Ls.add(item)$ 
16:     Update  $ScoreFunction(item)$ , and update  $Ls$  accordingly
17:   else  $ScoreFunction(item)$  is completely known
18:     Report( $item, ScoreFunction(item)$ ) and
19:      $returned = returned + 1$ 
20:   end if
21:
22: end while
23: return  $TopList$ 
24: end function

```

This algorithm takes as input a set of values representing Cloud service criteria. To replace them in the form: (key, value) for it to be processed by the Mapper.

Mapper phase. At line 5 of Algorithm 1, a StringTokenizer is used to split the string obtained from a transformation. For each word received, we create a couple whose key is the word, and the value is worth the offset of the line. For each data slicing, Hadoop creates a Map task that will execute the map function developed accordingly. Each data slicing is processed by only one task Map for more details

Table 2 $MRTop_k$ final output while using pig and MapReduce for $k = 10$

Service	Key	Value (score)
Service number1	137,794	8,266,108.90688207
Service number2	70,573	1,374,702.1587389937
Service number3	99,828	333,200.8879954284
Service number4	160,649	238,354.3083541053
Service number5	142,991	174,565.89832607133
Service number6	37,883	74,508.39833356984
Service number7	135,662	67,738.27213285805
Service number8	119,152	31,215.643822761635
Service number9	9118	28,001.2485926331172
Service number10	128,721	16,254.91752235506

Table 3 Dataset used for scalability study

Datasets input values input size		
DS1	3633	200, 6 KB
DS2	50,000	2, 7 MB
DS3	100,000	5, 5 MB
DS4	200,000	11, 1 MB
DS5	500,000	27, 7 MB

about the used datasets in this experiment see the Table 3. It should be noted that it is not the data transported to the program but the reverse. Hadoop will try to find the closest node containing the data to transfer the Map function. It is a Data Locality Optimization. **Reducer phase.** Hadoop will launch the Reduce tasks until all tasks are completed Map. Because the input of the reduce function corresponds to the output of the map functions. Then each task Reduce generates an output file to be stored as shown in Table 2, this time in the HDFS file system (Fig. 2).

Algorithm 2 $MRTop_{kws}$: Map phase

```

1: input:  $D_{si}$ 
2: Output: records of  $DS$ 
3: function  $Map(key, value)$ 
4: //verify normalization of criteria
5:  $String = nextTokenizer(valuetostring)$ 
6: Emit ( $key, value$ )
7: end function

```

Algorithm 3 $MRTop_{kws}$: Reduce phase

```

1: input: A subset of values  $key_1, key_2, \dots$  with the associated sets of  $val_1, val_2, \dots$ 
2: Output:  $Top_kScore$  for value key
3: function  $Reduc(key, value)$  //with key sorted in descending order of scores
4: //and with  $w_{ij}$  the user input and  $,label \in 0, 1$ 
5:  $int\ k = -1$ 
6: for ( $val - i : values$ )
7:    $++k$ 
   ( $val_i : values$ )
8:    $++k$ 
9:   if ( $label[k] == 0$ ) then
10:      $Score+ = w_i[k] * (1/val)$ 
11:   else ( $label[k] != 0$ )
12:      $Score+ = w_i[k] * (val)$ 
13:   else
14:      $Score+ = 0$ 
15:   end if
16: end for
17:  $res.set(Score)$ 
18: Emit( $key, res$ )
19: skyp a objects with score
20: Emit ( $key, Top_kScore$ )
end function

```

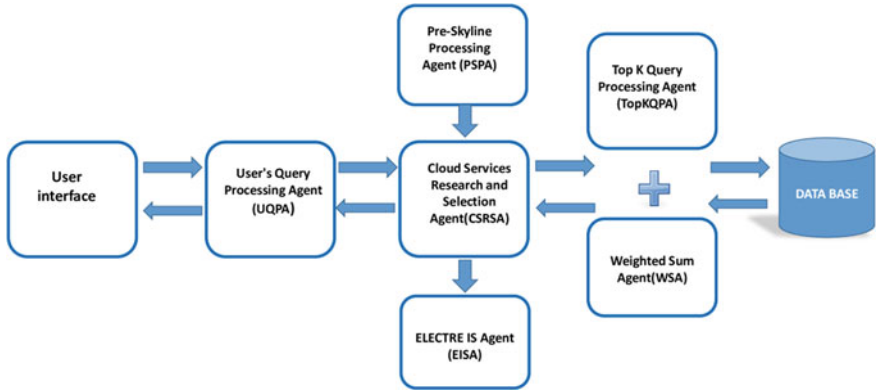


Fig. 2 Top_k and WS Agent using for Skyline refining problem in (CSRSS), extracted from [5]

4 Results and Discussion

After the Reducer calculates the scores Associated with keys and outputs the k pairs with the highest scores, we identify if the criteria are to be minimized or maximized .then calculate the weighting to be assigned to each label that is 0 or 1 and calculate the final score. After applying our score function to have a precise decision about the first k recommended cloud services, we still have to schedule our output and select the services by fixing an integer k. Our goal is still to detect the Top_k Cloud service with a higher degree of importance according to the user choice, which mains the Cloud services with higher scores (Fig. 3).

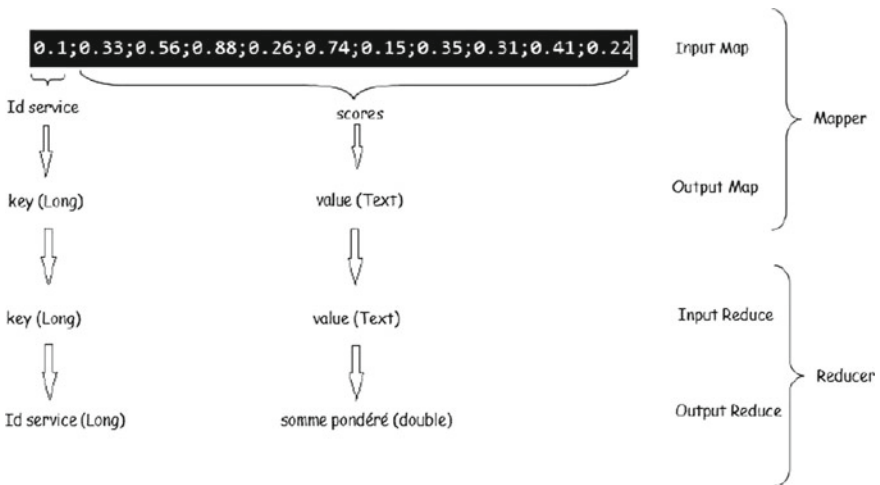


Fig. 3 The weights given by the user assigned to each service in parallel

- Value of type String (it's a line in the file).
- Key: it is the key of type LongWritable (the Id of the service, and it is the line offset).

After applying our weighted sum function to have a precise decision about the first k recommended services, we still have to schedule our output and select the services by setting an integer k . The choice of Pig instead of Map-Reduce was not random, especially in this phase. Still, the data flow, the processing difficulty concerning the values in MapReduce, and the program's complexity were our constraints. Here we will load our Map-Reduce output as a table to be processed with the Pig Latin language thanks to the communication between the two components of the Hadoop platform. Then apply a simple query to order and limit the selection of the K items. Knowing that this query has a highly complex Map Reduce java code behind it. Still, thanks to the Pig Latin, it becomes something straightforward. The output of Map Reduces before using Pig for ordering the score. For this experiment, we use Hadoop 3.0.0.0, with java 8 (Fig. 4).

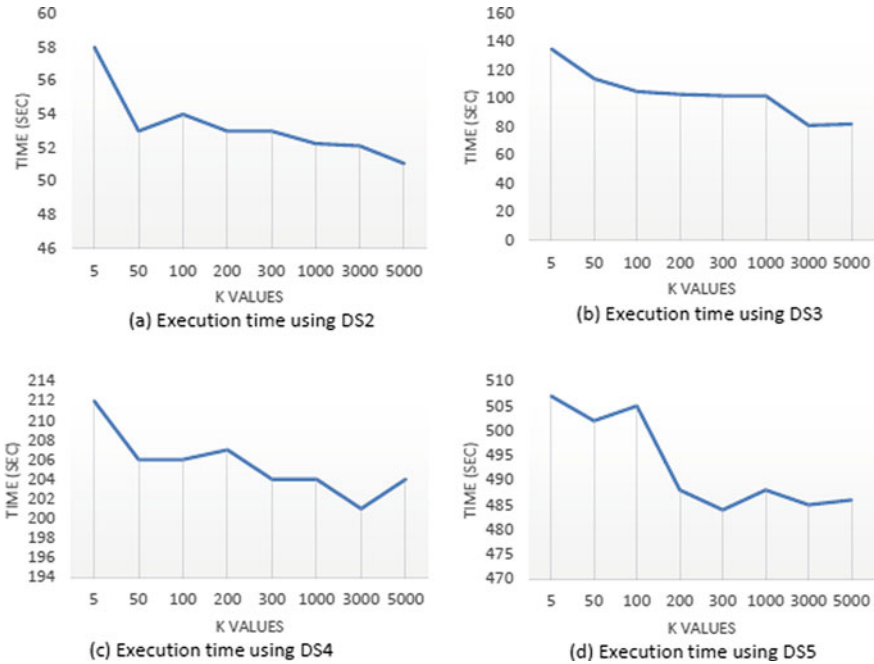


Fig. 4 runtime of $MRTop_k$ for Skyline refining problem in (CSRSS)

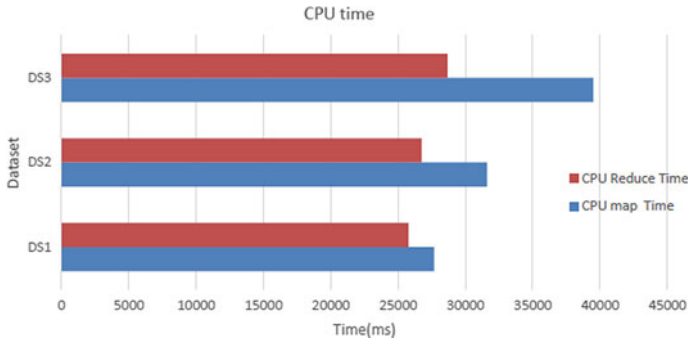


Fig. 5 Runtime of MRT_{Top_k} for Skyline refining problem in (CSRSS)

5 Conclusion

In this work, we tried to model our Top_k algorithm using MapReduce as the main component of the Hadoop Framework. This modeling was done by combining MapReduce and Pig. The latter is also based on the Map-Reduce paradigm. However, Pig ensures a more significant acceleration of 10 times faster than MapReduce. So the addition of Pig gives us high-level data flow processing and simplicity of intercommunication with Hadoop. In the following work, we will try to compare this work with our previous work that was based on Spark and benefit from the technology of the Spark Streaming and Machine learning library [1, 12, 20, 21], compare our algorithm with the distributed version of the Top_k algorithm like the Fagin algorithm and applied the MRT_{Top_k} in other application fields (Fig. 5). Moreover, we also intend to adapt several methods as those presented in [4, 22-32] to this domain.

References

1. K. El Handri, A. Idrissi, système collaboratif d'aide à la décision à base des recommandations multi critères (Sep 03 2020 MA Patent 50776)
2. K. El Handri, A. Idrissi, Efficient topkws algorithm on synthetics and real datasets. Under review in Int. J. Artif. Intell. IJ-II, submitted 14 mai (2020)
3. A.M.S. Osman, A novel big data analytics framework for smart cities. *Futur. Gener. Comput. Syst.* **91**, 620–633 (2019)
4. A. Idrissi, C.M. Li, J.F. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks, in *18th IEEE International Conference on Tools with Artificial Intelligence* (2006)
5. A. Idrissi, K. Elhandri, H. Rehioui, M. Abourezq, Top-k and skyline for cloud services research and selection system. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*. p. 40. ACM (2016)
6. P. Riyaz, S.M. Varghese, A scalable product recommendations using collaborative filtering in hadoop for bigdata. *Procedia Technol.* **24**, 1393–1399 (2016)

7. N. Bloyet, P.F. Marteau, E. Frènod, Étude lexicographique de sous-graphes pour l'élaboration de modèles structures à activité-cas de la chimie organique, in *Extraction et Gestion des Connaissances: Actes de la conférence EGC*, vol. 79 (2019), p. 3
8. K. Meena, J. Sujatha, Reduced time compression in big data using mapreduce approach and hadoop. *J. Med. Syst.* **43**(8), 239 (2019)
9. B.D. dans le Cloud, Big data dans le cloud: des technologies convergentes (2013)
10. Institute, I.T.R.: big data kernel description (2019)
11. D.E. O'Leary, Artificial intelligence and big data. *IEEE Intell. Syst.* **28**(2), 96–99 (2013)
12. K.E. Handri, A. Idrissi, Comparative study of topk based on fagin's algorithm using correlation metrics in cloud computing qos. *Int. J. Internet Technol. Secur. Trans.* **10**(1–2), 143–170 (2020)
13. M. Bakratsas, P. Basaras, D. Katsaros, L. Tassioulas, Hadoop mapreduce performance on ssds: the case of complex network analysis tasks, in *INNS Conference on Big Data* (2016)
14. A. Saxena, A. Chaurasia, N. Kaushik, N. Kaushik, Handling big data using mapreduce over hybrid cloud, in *International Conference on Innovative Computing and Communications* (Springer, 2019), pp. 135–144
15. I. Yaqoob, I.A.T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N.B. Anuar, A.V. Vasi-lakos Big data: from beginning to future. *Int. J. Inf. Manag.* **36**(6), 1231–1247 (2016)
16. E.A. Balas, M. Vernon, F. Magrabi, L.T. Gordon, J. Sexton, et al, Big data clinical research: validity, ethics, and regulation, in *MedInfo* (2015), pp. 448–452
17. K. Tannir, *Optimizing Hadoop for MapReduce*. Packt Publishing Ltd (2014)
18. M. Ojha, K.P. Singh, P. Chakraborty, S. Verma, P.S. Pandey, An empirical study of aggregation operators with pareto dominance in multiobjective genetic algorithm. *IETE J. Res.* 1–11 (2017)
19. N. Khan, M. Alsaqer, H. Shah, G. Badsha, A.A. Abbasi, S. Salehian, The 10 vs, issues and challenges of big data, in *Proceedings of the 2018 International Conference on Big Data and Education*, (2018), pp. 52–56
20. K. El Handri, A. Idrissi, Étude comparative de top-k basée sur l'algorithme de fagin en utilisant des métriques de corrélation dans la qualité de service de cloud computing, in *EGC* (2019), pp. 359–360
21. K. El Handri, A. Idrissi, Parallelization of top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
22. A. Idrissi, K. Elhandri, H. Rehioui M. Abouezq, Top-k and skyline for cloud services research and selection system, in *International conference on Big Data and Advanced Wireless Technologies* (2016)
23. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. arXiv preprint. <https://doi.org/10.48550/arXiv.1709.10372> (2015)
24. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Busi. Intell. Data Mining* (2017)
25. K. EL Handri, A. Idrissi, Efficient Top-kws algorithm on synthetics and real datasets. in *International journal of Artificial Intelligent (IJAI)* (2020)
26. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Tech. Sec. Trans.* **10** (2020)
27. M. Abouezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Archit.* **9**(2–3), 136–148 (2020)
28. M. Abouezq, A. Idrissi, Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comp. Sci. Appl.* **6**(6) (2015)
29. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electr. Eng.* **73**, 289–303 (2018)
30. M. Essadqi, A. Idrissi, A. Amarir, An Effective Oriented Genetic Algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**, 170–179 (2018)

31. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment, in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (2016)
32. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Robot. Intell. Syst.* **14**(3), 65–70 (2020)

A New Context-Based Factorization Machines for Context-Aware Recommender Systems



Rabie Madani , Abdellah Idrissi , and Abderrahmane Ez-Zahout 

Abstract Traditional recommender systems provide recommendations based on two dimensions (item and user), which can impact the quality of recommendation since they only use two entities and neglect other influential factors. Context-aware Recommender systems (CARS) solve this problem by taking in consideration several factors that affect user behavior like contextual factors. In fact, information such as time, location and companion have an influence on users and their choices. Besides item and user information, CARS use extra information (Contextual information) to produce more accurate and personalized recommendations. Factorization Machines (FM) is the most used algorithm in recommendation due to its efficiency, however like every algorithm FM has some drawbacks, one of them is its inability to capture feature interaction strength. In this paper, we proposed a new CARS based on an extension of FM adapted with contextual data and able to capture interaction strength and make the difference between features belonging to different contexts called Context-Based Factorization Machines (CBFM). Experiments show that CBFM improves the accuracy of prediction and realizes competitive performance compared with baselines.

Keywords Recommender systems · Context-aware recommender · Systems factorization machines · Context-based factorization machines

R. Madani (✉) · A. Idrissi · A. Ez-Zahout
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: rabie.madani@um5.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

A. Ez-Zahout
e-mail: abderrahmane.ezzahout@um5s.net.ma

1 Introduction

Today, the use of recommender systems has become crucial in several areas. For instance, RS recommend products to customers in e-commerce, suggest movies and TV-shows to users in streaming platforms and last but not least recommend songs in music apps. Recommender systems, which are information filtering systems, try to understand user behavior by using their past data, then recommend elements that may match with their requirements [1]. RS bring many benefits to companies by helping them to effectively market their products and services, increase their sales and as a result increase the income, and also help users to reduce the time of searching and make it easy for them to find what they need in a short time. Context-aware recommender systems [2] are a special type of Recommender Systems developed to overcome some drawbacks of traditional RS. While traditional RS use items and user's data to provide recommendation, CARS use additional data to generate more accurate predictions. Basically, this approach revolves around the concept of context which represents every factor that influences customer behavior and differs from application to another. For instance, a restaurant recommender system can provide recommendations to customers based on their profiles and restaurant descriptions, the recommendation could be interesting in terms of menus, quality of food and prices, but if this recommendation ignore contextual factors such as customer location and how far is he to the restaurant, in this case it become less attractive and not convenient. One of the biggest problems of RS is sparsity which means a which we can define as a high percentage of empty values contained in the ratings matrix. Many works [3–5] have been proposed to tackle this problem and showed acceptable performance, but the study proposed by Rendle [6] called Factorization Machines surpassed all previous works and showed good performance. However, FM is developed to be used with Collaborative Filtering approach where the data contains few dimensions, and also FM fails to capture interaction strength and processes all features although equally there are not from the same context. The major contributions of this study are:

- It represents a new variant of factorization machines, adapted with high dimensional data.
- It captures interaction strength and processes features from different contexts in a different way.

The rest of this paper is presented as follows. The second section represents the state-of-the-art works. The third section proposes in detail our model. The fourth section discusses the obtained results. And the last section represents a conclusion of the proposed work.

2 Related Works

Lately, many works have tried to incorporate contextual information in the process of generating recommendations, in order to achieve more performance and to develop robust and efficient contextual recommenders. For instance, Adomavicius et al. [7] proposed a multi-dimensional (MD) Recommender system that uses additional contextual data to generate recommendations. Furthermore, they presented an estimation method to identify where MD Recommender system overcomes the traditional approach and the opposite, then select the best choice for recommendation. Baltrunas et al. [8] introduced a Context-Aware RS based on Matrix Factorization algorithm which computes contextual features interactions using extra model parameters. In this work, conduct experiments on three models then choose the best one. The first one assumes that the all contextual dimension has a uniform impact on ratings. The second one supposes that each context has a different impact on ratings.

The last one supposes that all contexts have the same impact. Hariri et al. [9] presented a CARS based on a text mining method to extract hidden contexts from reviews. The model combines extracted data with user rating, then generates a utility function of items. Kramár et al. [10] proposed a framework for selecting context types dynamically using light-weight semantics. They also develop a new approach that uses implicit feedback to extract patterns from user behavior based on their search sessions. Unger et al. [11] represent a latent context-aware RS which aims to automatically learn latent contexts from data collected by mobile sensors using unsupervised Deep Neural Networks. They also proposed a hybrid model that utilizes latent contexts and explicit data to recommend items. Lahlou et al. [12] represented a textual CARS based on Factorization machines to predict ratings. The model considers the entire reviews as contextual data, then uses this data as an input for the prediction algorithm. Livne et al. [13] proposed a latent sequential latent model for Context-aware Recommender system using a deep learning technique. The model uses a long short-term memory to extract latent contexts from contextual data sequences and to learn high order interactions for user/item dimension and also for contextual dimensions. Jeong and Kim [14] introduced a deep learning context aware model which combines autoencoders and non-supervised techniques to extract patterns and predict ratings. The proposed method also tries to address the problem of sparsity which is a classic challenge for existing CARS. Vu and Le [15] represented a multicriteria Context aware approach based on deep neural networks techniques to generate better suggestions for users. The approach uses deep learning models to aggregation function learning and to predict scores. Madani and Ezzahout [22] proposed a new method for contextual information extraction from text. The method uses two components, the first one consists of a customized named entity recognition and BERT model and aims to extract contexts from users' reviews. The second one uses Factorization Machines to predict ratings using outputs of the first component and user/item data. The majority of the aforementioned works are based on Factorization machines or on deep learning to improve the accuracy or to tackle some drawbacks of FM such as nonlinear problems. However, The FM algorithm still suffers from many limitations,

one of them is that it considers all features equal and ignores the fact that features do not interact in the same way with other features from other contexts. In the next section we will present our proposed solution for this problem.

3 Proposed Model

In this work, we intend to use contextual information in addition to traditional data in order to generate personalized recommendations. We propose an extension of FM that is able to deal with extra data and the dimensionality expansion from one hand. On the other hand, we try to address the inability of FM to make the difference between contexts. Figure 1 shows the architecture of our model. The model consists of three layers: the first one is the input layer, the second one is the embedding layer where inputs are transformed to embedding vectors and the third layer is the CFM layer which is the most important layer in this work.

3.1 Factorization Machines

FM is a generic supervised learning technique that transforms real values to low dimensional latent vectors and can be used for regression, ranking and classification tasks. FM trains the model in a linear complexity which makes it well suited for large datasets and also it efficiently handles high sparse data. For these reasons, FM is widely used in prediction and real-world recommendations. The equation of FM

Fig. 1 The architecture of the CBFM model

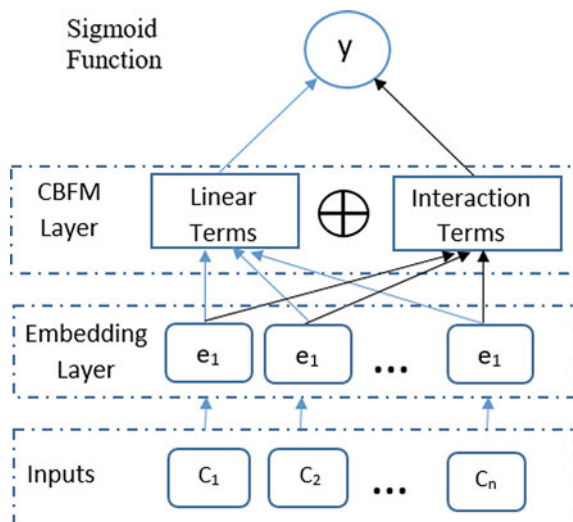
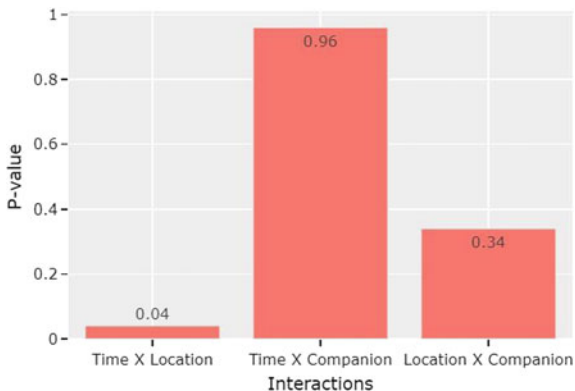


Fig. 2 The interaction between contexts



is defined as:

$$y_{FM}(x) = w_0 + \sum_{i=1}^d w_i x_i + \sum_{i=1}^d \sum_{j=i+1}^d \langle v_i, v_j \rangle x_i x_j \quad (1)$$

where w_0 global bias, w_i the weight for i th feature, $\langle v_i, v_j \rangle$ the interaction between i and j features.

Despite all the advantages of FM, it fails to make the difference between contexts. For instance, the model features interaction between Time, Companion and Location FM uses the same latent vector from the feature “Morning” belonging to Time with features belonging to different contexts (*Companion* $\langle v_{Morning}, v_{Wife} \rangle$, *Location* $\langle v_{Morning}, v_{Loc1} \rangle$). However, the feature interaction strength differs from pair of contexts to another and features do not act in the same way with different features from different contexts. Confirm the validity of this hypothesis, we use ANOVA [16] which is a statistical tool that computes the features interaction strengths between two variables, and verifies the impact of interactions on the outcome. Figure 2 shows that the interaction strength between three contexts differ from pair to another depending on the features participating in interactions and their original contexts. For the first interaction, the result shows a strong interaction between Time and Location since the p-value is less than 0.5. For the second and the third interactions, the obtained values are higher than 0.5, thus the interactions for both pairs are weak. The major conclusion we can extract from this experiment is that the features do not interact in the same way with other features if their original context is different.

3.2 Context Based Factorization Machines (CBFM)

To address this problem, we propose a new variant of FMs called Context Based Factorization machines (CBFM) which can make the difference between contexts during the interaction by using extra weights. The equation of CBFM is defined as

follow:

$$y_{CBFM}(x) = w_0 + \sum_{i=1}^d w_i x_i + \sum_{i=1}^d \sum_{j=i+1}^d \langle v_i w_{c(i)}, v_j w_{c(j)} \rangle x_i x_j \quad (2)$$

where w_0 is a global bias, v_i and v_j latent vectors of i and j , $w_{c(i)}$ and $w_{c(j)} \in \mathbf{R}$ are weights to differentiate context from another.

4 Experiments

This section presents the experimental results. Firstly, we will give a description of the dataset and the implementation. Then, we will compare our model with five baselines and discuss the results.

4.1 Datasets

To verify the effectiveness of our model, we use two datasets: The first one is the Yelp dataset [17] which contains more than 1.9 M users and more than 150 k businesses. The dataset consists of four contextual information extracted from users' reviews (Time, Location, Environmental and Companion). The second one is the Amazon dataset [18] which consists of 82.83 M ratings, 20.98 M users, 9.3 M items and four contexts extracted from reviews. To evaluate the CBFM model we use Mean Square Error (MSE). We can define its equation as:

$$MSE = \frac{1}{N} \sum_{i=1}^n (r_n - \hat{r})^2 \quad (3)$$

4.2 Comparisons

We select six models to compare the performance of the CBFM model:

FM [6]: Factorization machines is the most used algorithm in recommendations. PMF [19]: Probabilistic Matrix Factorization model estimates the latent vectors by using Gaussian distribution. CDL [20]: Collaborative Deep Learning for RS combines a Bayesian model and collaborative filtering to learn latent representation and predict ratings. DeepFM [21]: Deep Factorization Machines combine FMs and deep neural networks to capture high and low order feature interactions. DB-CARS [14]: Deep Learning-Based CARS uses deep learning and autoencoder to predict ratings and extract features. CFM [22]: Contextual Factorization machine is based on the generic FM and contextual data to predict scores.

4.3 Results and Discussion

We use a core i7, 16 GB desktop equipped with tensorflow [23] to implement our solution. The datasets are splitted into train (80%) and test (20%) data. We fix the mini batch at 4096, we use L2 regularization to prevent overfitting, and we also use Adam as an optimization method.

Figure 3 shows the obtained result from Yelp dataset. As we can see, The CBFM model achieves good performance by reaching the lowest value (1.321) of MSE followed by the DB-CARS model which reaches the second lowest value of MSE (1.213), however, FM achieves the highest MSE since the most of models utilized in this experiment are extensions of the original Factorization Machine. To confirm the obtained results using Yelp dataset, we conduct another experiment using the Amazon dataset. As shown in Fig. 4, we obtain the same order of performance which affirms the results obtained in 3. Our model overcomes all six models and achieves a significant improvement in terms of performance.

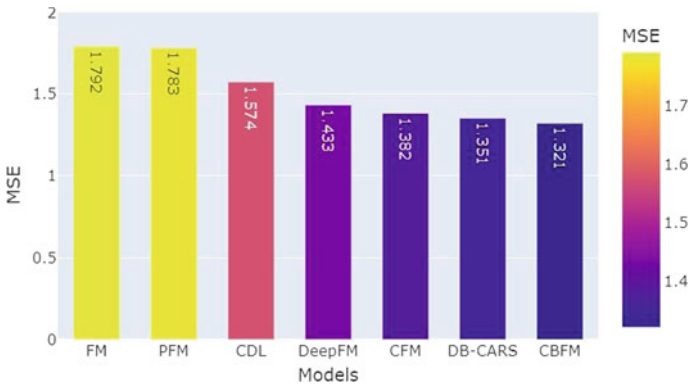


Fig.3 Yelp dataset results

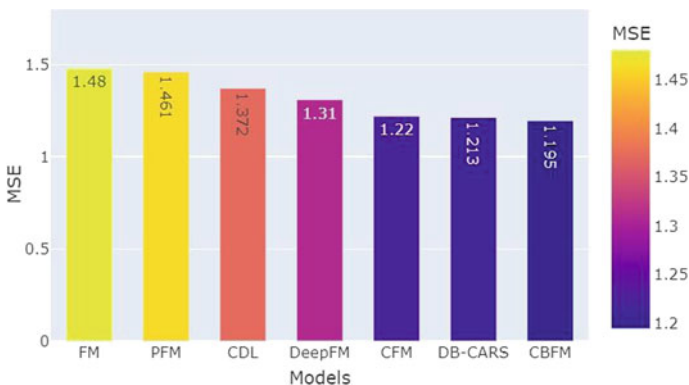


Fig. 4 Amazon dataset results

5 Conclusion

In the paper, we propose a Context-Aware Recommender System based on a new variant of factorization machines called Context-Based Factorization Machines. This variant is able to make the difference between contexts and to take in consideration the fact that features do not interact in the same way with all features. The proposed variant uses additional weights to each feature's latent vector to differentiate a context from another. The performance results show that the CBFM model achieves competitive results compared to six baseline models. For future work, we will try to propose a more efficient model able to capture second order feature interactions and also to solve the inability of FM to solve nonlinear problems. To this end, we can rely on the various works presented in [24–35].

References

1. R. Madani, A. Ez-Zahout, A. Idrissi, An overview of recommender systems in the context of smart cities, in *2020 5th International Conference on Cloud Computing and Artificial Intelligence (CloudTech)*, (2020), pp. 1–9. <https://doi.org/10.1109/CloudTech49835.2020.9365877>
2. G. Adomavicius, B. Mobasher, F. Ricci, A. Tuzhilin, Context-aware recommender systems. *AI Mag.* **32**(3) Art. no. 3 (2011). <https://doi.org/10.1609/aimag.v32i3.2364>
3. Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, (2008), pp. 426–434
4. S. Rendle, L. Schmidt-Thieme, Pairwise interaction tensor factorization for personalized tag recommendation, in *WSDM '10: Proceedings of the third ACM International Conference on Web Search and Data Mining*. (ACM, New York, NY, USA, 2010), pp. 81–90
5. R. Salakhutdinov, A. Mnih, Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. *Int. Conf. Mach. Learn.* **25** (2008)
6. S. Rendle, Factorization machines, in *2010 IEEE International Conference on Data Mining*, (2010), pp. 995–1000. <https://doi.org/10.1109/ICDM.2010.127>
7. G. Adomavicius, et al., Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst. (TOIS)* **23**(1), 103–145 (2005)
8. L. Baltrunas et al., Matrix factorization techniques for context aware recommendation. *Proc. Fifth ACM Conf. Recomm. Syst.* **2011**, 301–304 (2011)
9. N. Hariri, B. Mobasher, R. Burke, Y. Zheng, Context-Aware Recommendation Based On Review Mining (2011)
10. T. Kramár, M. Bieliková (2012) Dynamically selecting an appropriate context type for personalization, in *Proceedings of the Sixth ACM Conference on Recommender Systems* (Dublin, Ireland, 2012), pp. 321–324
11. M. Unger, A. Bar, B. Shapira, L. Rokach, Towards latent context-aware recommendation systems. *Knowl. Based Syst.* **104** (2016)
12. F.Z. Lahlou, H. Benbrahim, I. Kassou, Review aware recommender system: using reviews for context aware recommendation. *IJDAI* **10**(2), 28–50 (2018). <https://doi.org/10.4018/IJDAI.2018070102>
13. A. Livne, M. Unger, B. Shapira, L. Rokach, deep context-aware recommender system utilizing sequential latent context. *arXiv* (2020). <https://doi.org/10.48550/arXiv.1909.03999>.
14. S.-Y. Jeong, Y.-K. Kim, Deep learning-based context-aware recommender system considering contextual features. *Appl. Sci.* **12**(1), Art. no. 1 (2022). <https://doi.org/10.3390/app12010045>

15. S.-L. Vu, Q.-H. Le, A Deep learning based approach for context- aware multi-criteria recommender systems. *csse*, **44**(1) Art. no. 1, (2022). <https://doi.org/10.32604/csse.2023.025897>
16. R.A. Fisher, *Statistical Methods for Research Workers*. Oliver and Boyd (1925)
17. Yelp Dataset. <https://www.yelp.com/dataset>
18. Amazon review data. <https://jmcauley.ucsd.edu/data/amazon/>
19. R. Salakhutdinov, A. Mnih, Probabilistic matrix factorization, in *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA, (2007), pp. 1257–1264
20. H. Wang, N. Wang, D.-Y. Yeung, Collaborative Deep Learning for Recommender Systems. [arXiv:1409.2944](https://arxiv.org/abs/1409.2944) [cs, stat]. <http://arxiv.org/abs/1409.2944>
21. H. Guo, R. Tang, Y. Ye, Z. Li, et X. He, DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv*, (2017). <https://doi.org/10.48550/arXiv.1703.04247>
22. R. Madani, A. Ez-zahout, A review-based context-aware recommender systems: using custom NER and factorization machines. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **13**(3) Art. no. 3, 42/30 (2022). <https://doi.org/10.14569/IJACSA.2022.0130365>
23. M. Abadi, et al., TensorFlow: a system for large-scale machine learning, in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. (USA, 2016), pp. 265–283
24. K. EL Handri, A. Idrissi, Efficient Top-kws algorithm on synthetics and real datasets. in *International journal of Artificial Intelligent (IJAI)*, (2020)
25. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syts. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
26. A. Idrissi, K Elhandri, H. Rehioui, M. Abourezq. Top-k and skyline for cloud services research and selection system. *International Conference on Big Data and Advanced Wireless Technologies* (2016)
27. A. Idrissi, F. Zegrari. A new approach for a better load balancing and a better distribution of resources in cloud computing. *arXiv preprint arXiv: 1709.10372*. (2015)
28. A. Idrissi, C.M. Li, J.F. Myoupo. An algorithm for a constraint optimization problem in mobile ad-hoc networks. 18th IEEE International conference on tools with artificial intelligence. Washington, USA, (2006)
29. H. Rehioui, A. Idrissi. A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data Min.* (2017)
30. M. Abourezq, A. Idrissi, H. Rehioui. An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Archit.* **9**(2–3), 136–148. (2020)
31. M. Abourezq, A. Idrissi. Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comput. Sci. Appl.* **6**(6) (2015)
32. F. Zegrari, A. Idrissi, H. Rehioui. Resource allocation with efficient load balancing in cloud environment. *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*. (2016)
33. F. Zegrari A Idrissi. Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Robot. Intell. Syst.* **14**(3), 65–70 (2020)
34. M. Essadqi, A. Idrissi, A. Amarir. An effective oriented genetic algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**, 170–179 (2018)
35. S. Retal, A. Idrissi. A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electri. Eng.* **73**, 289–303 (2018)

Review of Recommendation Systems and their Applications in E-learning



Aicha Er-Rafyg  and Abdellah Idrissi

Abstract In recent years, we are in front of exponential increasing of the learning resources on the Web. It takes a lot of the learner's time and effort to locate training resources that are appropriate for his or her needs. In this framework, Recommendations Systems are crucial to the field of e-learning. They help the learner to find useful training content that is relevant to their learning needs. In this paper we present the recommendation systems as well as their types, the advantages and drawbacks of each type of recommendation, we will also present some works using recommender systems in e-learning and mention their applied techniques, algorithms, dataset used, and the recommended resources.

Keywords E-learning · Recommendation system · Learning management system · Collaborative filtering · Knowledge-based · Content-based · Hybrid

1 Introduction

E-learning or Online Learning is a web-based teaching system, which allows the dissemination of content as well as communication [1]. The use of computers and electronic communications makes it possible to dispense with classes and to no longer be bound to fixed schedules. Therefore, the transmission of knowledge can be done at any time and from anywhere [2]. Online learning through Learning Management System (LMS) has become a popular and effective way to deliver distance learning to learners of all ages and backgrounds [3]. LMS platforms provide a variety of features and tools that enable instructors to create and manage online courses,

A. Er-Rafyg (✉) · A. Idrissi
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: a.errafyg@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

deliver learning content, track learner progress, and facilitate communication and collaboration between learners and instructors [4].

The benefits presented by e-learning have generated an exponential increase in learning resources on the Web. On the other hand, due to this information overload, online learners are increasingly struggling with the choice of the most appropriate learning materials that are best suited to their learning needs. In order to tackle this issue, it is necessary to rely on technologies capable of filtering available data, based on users' needs. In this regard, a recommendation system presents a powerful mechanism to tackle the information overload problem and to help users in finding relevant content that meets their needs.

This article is structured into five main sections; the following section provides a brief overview of Recommendation Systems (definitions, types, advantages and disadvantages). The third section exposes applications of some recent recommendations system in e-learning. The fourth section compares their techniques, algorithms, used dataset and the recommended resources. The last section concludes the article and highlights some perspectives.

2 Overview of Recommendation Systems

With the advancement of computers, and particularly the Internet, our way of life has undergone significant transformation. Once the user is connected, he or she can perform a variety of tasks as desired, including shopping, searching, consulting the news, learning online, watching movies, etc. [5]. In order to assist users in finding what they want on the web, recommendation systems (RS) have been developed. These systems base their recommendations on a variety of factors and criteria, such as the opinions of a user community, as people are currently experiencing information overload [6].

2.1 *Definitions of Recommendation Systems*

There is a general definition of Robin Burke, but given the range of proposed classifications for referral systems, they can be described in many different ways [7]. Which defines them as follows: "Systems that can provide customized recommendations to guide the user to interesting and useful resources within a large data space". Recommendations systems offer personalized suggestions to users, these systems often rely on Collaborative Filtering (CF) [8]. They made their debut in 1992 with the Tapestry system, which supported CF [9]. Lu et al. [10] define recommendations system as software and techniques for suggesting items to the user [7]. The item is the term that designates what the system will recommend, it can be a product to buy, a movie to watch, a training content, a service and the list is very long.

The foundation of most recommendation systems is the estimation of the user's rating for unrated products coming from him or her. Based on this assessment, which is often based on the ratings this user has made to previous articles, the algorithm suggests the item(s) with the highest estimation [11].

Another definition stated by Rao and Talwar [12] as authors consider the recommendations system software application that uses algorithms to analyze data and provide recommendations to users. These recommendations can be in the form of products, movies, music, articles, etc. A recommender system's objective is to offer users personalized recommendations based on their preferences, previous actions, and other criteria [10].

2.2 Types of Recommendation Systems

It is possible to categorize RS in several ways depending on how referrals are made [7], 8, or the data that was used to propose a product [9]:

- **Demographic recommendation:** are a kind of recommendation system that offers personalized recommendations based on user demographic information including age, gender, location, and other personal characteristics. The system groups users into categories according to their demographic data, and then recommends items to them based on the tastes of other users in the same demographic class [10].
- **Knowledge-based recommendations:** to identify the user's requirements and the best manner to meet those needs, Knowledge-based recommendations uses fundamental knowledge of related and similar items [7]. What should a recommendation system perform, for example, for a car or another comparable product? The response to the question: "Why are they buying a car?" is to learn more about the user's preferences. Is comfort more essential than fuel efficiency? Based on this data, the system can utilize knowledge-based reasoning to determine which goods satisfy the user's requirements and produce a recommendation [11].
- **Content-based recommendation:** is a type of recommender system that generates recommendations depending on the features of the recommended items. The basic idea of this method is that if a user likes a particular item, they are likely to be interested in other items that share similar characteristics or features [13]. For example, if a user gave a movie in the "Comedy" genre a positive review, the system will suggest further comedies [7].
- **Collaborative recommendations:** is the most researched recommendation method that has developed throughout time [14], It has been extensively utilized in useful recommendation systems, particularly those in e-commerce, like those of Amazon and Netflix [15].

This technique makes suggestions to a new user with related interests based on the actions or preferences of a group of users. These approaches are based on the notion that users with similar preferences or previous choices are likely to have

similar interests in the future [16]. The two types of collaborative filtering are memory-based and model-based techniques [7].

Memory-based: Use similarity metrics to find users or products that are similar based on past behavior. Then, these algorithms provide recommendations based on the actions of comparable users or objects. Depending on whether similarity is computed between individuals or items, this method can also be categorized as user-based or item-based collaborative filtering [14].

Model-based: Instead of measuring the correlation between users, this approach measures the correlation between the material. The fundamental concept is to recommend things based on the system's search for items that are comparable to those a user has already selected. In contrast to the user-centric approach, which is relatively memory resource-hungry, preliminary processing is carried out on an evaluation matrix in order to identify related things and thereafter give recommendations in real time [16].

- **Hybrid recommendation:** This approach combines two or more of the above techniques to increase precision and relevance of the recommendations. For example, a hybrid recommender system may use content-based filtering to generate an initial set of recommendations, and then use collaborative filtering to refine the recommendations based on the behavior of other users [7, 14].

2.3 Advantages and Disadvantages of Each Type of RS

As shown in Table 1, each type of recommendation system has its advantages and disadvantages. Each system is suitable for a specific need, e.g. the content-based recommendation will be the good solution in the case where there is a new system that will be put in place, because the cold start problem will not occur. On the other hand, a hybrid recommendation-based system will be more efficient in all cases since it uses the best in combined approaches, although its implementation is complicated.

3 E-learning Recommendation Systems

By assisting the learner in finding training materials that are appropriate to their learning requirements, recommendation systems play a crucial role in the field of e-learning. This section will highlight some works on recommendation system applications in e-learning.

El Mabrouk et al. suggested a hybrid recommendation system for online learning platforms that makes use of data mining. This system seeks to suggest the most appropriate content to a user of an online learning platform and enables users to focus by making content more accessible. Also, because most recommendation systems rely on analyzing learner profiles, the suggested system has been created to consider a number of factors when making recommendations [17].

By taking into account the learner’s expertise, Herath et al. suggested an architecture for a personalized learning recommendation system. In this design, e-assessment, also known as technology enhanced assessment (TEA), is used to determine the learner’s level of understanding. The learning model, domain model, electronic assessment model, and recommendation model make up the four components of the research’s suggested recommendation system [18].

In order to recommend learning resources to learners, Tarus et al. proposed a hybrid recommender system for online learning that combines collaborative filtering (CF) algorithms, sequential pattern mining (SPM) and context awareness. The system incorporates more contextual data about the learner to personalize recommendations, and the SPM algorithm mines web logs to identify the learner’s sequential access

Table 1 Advantages and disadvantages of each recommendation system

	Advantages	Disadvantages
Demographic recommendation (DF)	<ul style="list-style-type: none"> • No need for user evaluation history which is required by collaborative and content-based techniques [9] • Can be implemented quickly and easily [9] 	<ul style="list-style-type: none"> • An effective recommendation requires collecting full demographic information about users which is difficult because they involve privacy issues [9] • A new user and a new article both have a cold-start issue [9]
Knowledge-based recommendations (KF)	<ul style="list-style-type: none"> • These systems emphasize a great deal on knowledge sources that collaborative filtering and content-based filtering methods do not fully utilize [12] • In this type of recommendation we don’t have problem with cold start [12] 	<ul style="list-style-type: none"> • The so-called knowledge acquisition bottleneck affects knowledge-based recommenders in that knowledge engineers must put in a lot of effort to formalize the knowledge held by domain experts into executable representations [12] • Complexity of their design and the difficulty of their implementation makes them less favorable to apply [14]
Content-based recommendation (CB)	<ul style="list-style-type: none"> • They do not require data about other users so don’t have problem with cold start • Are able to suggest products to individuals with particular tastes • Explain the recommended elements (citations of their characteristics) • Are able to recommend new and unpopular items to each user [9] 	<ul style="list-style-type: none"> • For items that are not analyzable by machines humans must manually insert their features [9] • Human involvement is very irrational, expensive, time-consuming, and subjective [9] • The elements are restricted to the functions or initial descriptions that they were given. Because of this restriction, content-based approaches are dependent on properties that are explicitly stated [9]

(continued)

Table 1 (continued)

	Advantages	Disadvantages
Collaborative recommendations (CF)	<ul style="list-style-type: none"> • Are based solely on the judgment of the community of participating users [9] • May be used with almost any kind of item, that is, articles, news, websites, movies, songs, books, jokes,... [9] • Do not require domain knowledge to mark the functionality of elements [9] 	<ul style="list-style-type: none"> • The cold-start problem when a novel item is inserted into the database, no user may recommend the item until another user has reviewed it [9] • A small or larger user community will occasionally contain people with outlandish beliefs or preferences. Seldom do these users get recommendations [9]
Hybrid recommendation	<ul style="list-style-type: none"> • Use the best in combined approaches in order to raise the limitations [14] 	<ul style="list-style-type: none"> • These systems are the most complicated in their design and implementation, but they have better prediction results [14]

habits. The CF algorithm is used to predict the learner's preference for a resource based on their previous interactions with the system and the behavior of other similar learners [19].

Obeid et al. provide a framework for a semantic recommender system that may be used to assist students in choosing the best major and university by employing the semantic web and machine learning approaches. The suggested system is divided into four key components: data collection, ontology, information processing utilizing machine learning, and ontology. Finally, the fourth section compares students' interests, generates recommendations, and saves a file of student recommendations to be used in subsequent procedures [20].

In order to facilitate the suggestion of Learning Objects (LOs) in an e-learning environment, Hinz et Pimenta's method takes into account the reputation of the users making the recommendations. The e-learning environment, the recommendations component, and the reputation mechanism are the three primary elements of the model, which the authors refer to as "e-RecRep." [21].

Dhanda et al. advise a tailored recommendation technique for academic literature to help researchers identify the finest publications for their study areas. Moreover, it uses the information in the document and the individual tastes of the user to provide recommendations [22].

In an online discussion forum, Albatayneh et al. provide a framework built on an inventive recommendation architecture that may suggest engaging messages to learners based on a semantic content-based filtering and their negative ratings [23].

A CBF-CF-GL technique integrating content-based, collaborative filtering, and high learner ratings is suggested by Turnip et al. Given that the hybrid method combining CBF and CF has been successfully applied in numerous prior research for various recommendation problems, a combined method of CBF-CFGL was chosen [24].

An intelligent learning system based on a hybrid recommendation algorithm is proposed by Li et al. It provides students, teachers, and other staff with the multimedia resources they require by using a system for personalized recommendation of teaching resources [25].

Morsomme et al. in [26] proposes a content-based course recommender system for liberal arts education. The system is designed to recommend courses based on the course descriptions and the students' interests.

The authors of the study in [27] proposed a machine learning-based recommender system to improve students' learning experiences. They developed a system that uses different algorithms to predict the most suitable resources for a particular student. The system considers the students' learning style and their preferences to recommend the most relevant content. The authors collected data from students who studied courses in different fields.

The study by Fernandez et al. [28] focused on developing a recommender system to support higher education students in their subject enrollment decision. The authors utilized a hybrid recommendation approach that combined content-based and collaborative filtering techniques. The system was designed to consider several factors, including the student's academic background, interests, and the relevance of the subject to their chosen degree.

Dwivedi et al. in [29] propose an effective trust-aware e-learning recommender system based on the learning styles and knowledge levels of the users. The proposed system is designed to address the challenges of traditional e-learning systems that often provide generic recommendations without considering the individual differences in learning styles and knowledge levels of the users. The system uses a combination of content-based and collaborative filtering techniques to generate personalized recommendations. In addition, it incorporates trust-awareness by analyzing the trust levels of the users towards other users, as well as the content being recommended.

In [30] Bhaskaran et al. proposed a hybrid recommendation system for e-learning applications that combines content-based filtering and collaborative filtering techniques. The system is designed to improve the effectiveness and efficiency of personalized learning by providing recommendations based on the user's past behaviors and the behavior of similar users. The authors used a clustering algorithm to group users based on their learning behavior and interests, and then applied a collaborative filtering approach to recommend resources to users within each cluster. The content-based approach was used to recommend resources that are similar to the resources that a user has already shown interest in.

4 Discussion

The same applications for recommendation systems in e-learning that were published in the previous section are mentioned here. These systems used a variety of techniques to provide recommendations for learning materials. We'll talk about them, mention

their methods, the algorithms that were utilized, the dataset that was used, and the resources that were recommended.

As shown in Table 2, the most used recommendation techniques in the cited works are the hybrid and content-based method.

The adoption of the hybrid method is explained by its principle that allows the combination between two or more methods that are complete and gives better results.

Also these studies demonstrate the potential of recommendation systems in improving the e-learning experience for students by providing personalized and relevant recommendations of learning resources.

The second point that can be noticed from the Table 2 is the diversity of the algorithms used and the resources recommended in the works, which allows us to deduce that this field of study is full of perspectives and challenges.

5 Conclusion and Future Work

In this paper we have presented the recommendation systems as well as their types, the advantages and disadvantages of each type of system. We have also exposed some recent recommendation systems applications in e-learning and compare their techniques, algorithms, dataset used, and the recommended resources.

The works presented in Sect. 3 of this paper show that recommendation systems play a very important role in the field of e-learning. The diversity of techniques and algorithms used to allow for a lot of challenges and perspectives to better help learners to choose the information that meets their needs.

Our perspective is to propose a new architecture of courses recommendation system for e-learning. Based on the learner's preferences we will start by using the Skyline BNL algorithm and assessing the system by varying the number of criteria used to generate the recommendations from 2 to 9 and see the number of courses recommended. As other perspectives, we studied the algorithms developed in [31–42] and we intend to adapt them to improve our approaches.

Table 2 Description of recommendation systems applications

		References
Methods of recommendation	– Hybrid	– [17, 21, 24, 25, 27–30]
	– Content-based	– [17, 22, 23, 26]
	– Collaborative	– [21]
	– Knowledge-based	– [20] (ontology-based)
Algorithms	– ID3 algorithm (Decision tree)	– [17]
	– Sequential pattern mining (SPM)	– [21]
	– k-means	– [20, 25]
	– The pearson coefficient	– [21]
	– High-utility itemset mining (EFIM)	– [22]
	– Vector space model	– [23]
	– Content based algorithms	– [17, 24]
	– Correlation-based algorithms	– [24]
	– Different machine learning algorithms	– [27, 28]
	– Genetic K-means	– [29]
	– Kullback–Leibler distance	– [26]
	– Personalized recommendation algorithm	– [25]
	Dataset	– Learner profil
– File log		– [17]
– Cookies		– [17]
– Historical data		– [17]
– Learner’s knowledge		– [17]
– Learners registered in an LMS		– [21]
– ACL Anthology Network		– [22]
– Database of posted messages		– [23]
– Collected data		– [24, 26, 27]
– Real-world dataset from a public Spanish university		– [28]
– Movie Lens dataset		– [29]
– Educational dataset with 1000 learners		– [30]
– Web learning platform (http://evaluate.guoshi.com/publishg/)		– [25]
Recommended resources		– Content
	– Videos	– [21]
	– Audios	– [21]
	– University choice, orientation in higher education	– [20]
	– Papers/references	– [22]
	– Posted messages in discussion forum	– [23]

References

1. W. A. Cidral, T. Oliveira, M. Di Felice, M. Aparicio, E-learning success determinants: Brazilian empirical study. *Comput. Educ.* (2017). <https://doi.org/10.1016/j.compedu.2017.12.001>
2. E.T. Welsh, C.R. Wanberg, K.G. Brown, M.J. Simmering, E-learning: emerging uses, empirical results and future directions. *Int. J. Train. Dev.* **7**(4), 245–258 (2003). <https://doi.org/10.1046/j.1360-3736.2003.00184.x>
3. P. Mehta, K. Saroha, Recommendation system for learning management system, in *Information and Communication Technology for Sustainable Development*, vol. 10, ed. by D.K. Mishra, M.K. Nayak, et A. Joshi (Singapore: Springer Singapore, 2018), pp. 365–374. https://doi.org/10.1007/978-981-10-3920-1_38
4. D. Moonsamy, I. Govender, Use of blackboard learning management system: an empirical study of staff behavior at a South African University. *Eurasia J. Math., Sci. Technol. Educ.* **14**(7), 3069–3082 (2018). <https://doi.org/10.29333/ejmste/91623>
5. Ticha-Diploˆme de Docteur en Informatique.pdf
6. P. Resnick, H.R. Varian, Recommender systems. *Communications of the Acm* **40**(3), 3 (1997)
7. G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **17**(6), 734–749 (2005). <https://doi.org/10.1109/TKDE.2005.99>
8. M. Balabanović, Y. Shoham, Fab: content-based, collaborative recommendation. *Commun. ACM* **40**(3), 66–72 (1997). <https://doi.org/10.1145/245108.245124>
9. K. Nageswara Rao, Application domain and functional classification of recommender systems—a survey. *DESIDOC J. Libr. Inf. Technol.* **28**(3), 17–35 (2008). <https://doi.org/10.14429/djlit.28.3.174>
10. R. Burke, Hybrid recommender systems: survey and experiments 40, (2002)
11. R. Burke, Integrating Knowledge-Based and Collaborative-Filtering Recommender Systems, p. 4
12. F. Ricci (ed.), *Recommender Systems Handbook*. (Springer, New York, 2011)
13. D. Jannach, Recommender Systems, p. 353
14. N. Mittal, Data, learning and privacy in recommendation systems, p. 137 (2016)
15. J. Bobadilla, F. Serradilla, A. Hernando, Collaborative filtering adapted to recommender systems of e-learning. *Knowl. Based Syst.*, vol. **22**(4), 261–265 (2009). <https://doi.org/10.1016/j.knsys.2009.01.008>
16. M. Ghennane, Le web social et le web sémantique pour la recommandation de ressources pédagogiques (2015)
17. M. El Mabrouk, S. Gaou, M.K. Rtili, Towards an intelligent hybrid recommendation system for e-learning platforms using data mining. *Int. J. Emerg. Technol. Learn. (iJET)* **12**(06), 52 (2017). <https://doi.org/10.3991/ijet.v12i06.6610>
18. D. Herath, L. Jayarathne, An architecture for a personalized learning recommendation on knowledge level of learner **9**(6), 7, (2018)
19. J.K. Tarus, Z. Niu, D. Kalui, A hybrid recommender system for e-learning based on context awareness and sequential pattern mining. *Soft Comput.* **22**(8), 2449–2461 (2018). <https://doi.org/10.1007/s00500-017-2720-6>
20. C. Obeid, I. Lahoud, H. El Khoury, P.-A. Champin, Ontology-based recommender system in higher education, in *Companion of the the Web Conference 2018 on The Web Conference 2018—WWW '18* (Lyon, France, 2018), pp. 1031–1034. <https://doi.org/10.1145/3184558.3191533>
21. V.T. Hinz, M.S. Pimenta, Integrating Reputation to Recommendation Techniques in an e-learning Environment. (2018)
22. M. Dhanda, V. Verma, Personalized recommendation approach for academic literature using high-utility itemset mining technique, in *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, vol. 719, ed. by P.K. Sa, M.N. Sahoo, M. Murugappan, Y. Wu, B. Majhi (Springer Singapore, Singapore, 2018), pp. 247–254. https://doi.org/10.1007/978-981-10-3376-6_27

23. N.A. Albatayneh, I.G. Khairil, F.-F. Chua, Utilizing learners' negative ratings in semantic content-based recommender system for e-learning forum
24. R. Turnip, D. Nurjanah, D.S. Kusumo, Hybrid recommender system for learning material using content-based filtering and collaborative filtering with good learners' rating, in *2017 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)* (Miri, Sarawak, Malaysia, 2017), pp. 61–66. <https://doi.org/10.1109/IC3e.2017.8409239>
25. H. Li, H. Li, S. Zhang, Z. Zhong, J. Cheng, Intelligent learning system based on personalized recommendation technology. *Neural Comput. Appl.* (2018). <https://doi.org/10.1007/s00521-018-3510-5>
26. R. Morsomme, S.V. Alferez, Content-based Course Recommender System for Liberal Arts Education (2019), p. 6
27. N. Yanes, A.M. Mostafa, M. Ezz, S.N. Almuyqil, A Machine Learning-Based Recommender System for Improving Students Learning Experiences. *IEEE Access* **8**, 201218–201235 (2020). <https://doi.org/10.1109/ACCESS.2020.3036336>
28. A.J. Fernandez-Garcia, R. Rodriguez-Echeverria, J.C. Preciado, J.M.C. Manzano, F. Sanchez-Figueroa, Creating a recommender system to support higher education students in the subject enrollment decision. *IEEE Access*, **8**, 189069–189088 (2020). <https://doi.org/10.1109/ACCESS.2020.3031572>
29. P. Dwivedi, K.K. Bharadwaj, Effective trust-aware e-learning recommender system based on learning styles and knowledge levels p. 17, (2021)
30. S. Bhaskaran, R. Marappan, B. Santhi, Design and analysis of a cluster-based intelligent hybrid recommendation system for e-learning applications. *Mathematics* **9**(2), 197 (2021). <https://doi.org/10.3390/math9020197>
31. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
32. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10**, (2020)
33. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Architect.* **9**(2–3), 136–148 (2020)
34. M. Abourezq, A. Idrissi, Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comput. Sci. Appl.* **6**(6), (2015)
35. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment. *Proc. Int. Conf. Big Data Adv. Wirel. Technol.* (2016)
36. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Robot. Intell. Syst.* **14**(3), 65–70 (2020)
37. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data Min.* (2017)
38. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Elect. Eng.* **73**, 289–303 (2018)
39. A. Idrissi, K. Elhandri, H. Rehioui and M. Abourezq, Top-k and Skyline for cloud services research and selection system. *Proc. Int. Conf. Big Data Adv. Wirel. Technol.* (2016)
40. A. Idrissi, C.M. Li, J.F. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks. in *18th IEEE International Conference on Tools with Artificial Intelligence*, Washington, USA (2006)
41. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. *arXiv preprint arXiv: 1709.10372* (2015)
42. M. Essadqi, A. Idrissi, A. Amarir, An effective oriented genetic algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**, 170–179 (2018)

Improvement of Courses Recommendation System using Divide and Conquer Algorithm



Aicha Er-Rafyg , Abdellah Idrissi , and Kaoutar El Handri

Abstract The recommendation of online courses is a crucial tool for learners who want to learn on their own. The issue of recommendations in the educational sector is unique and cannot be dissociated from irrational elements like learner preferences, performance expectations, etc. In order to increase learners' excitement for studying, online course recommendations should aim to match learners' needs with appropriate courses. In this paper, we provide a novel approach to course recommendation for students with multiple preferences. This method achieves the objective of recommendation by generating a list of recommended courses using the Divide and Conquer (D&C) algorithm which makes it possible to divide the database of available courses according to the category chosen by the learner and then use the Skyline Block-Nested Loops (BNL) algorithm to recommend courses that most closely match the preferences of the learner.

Keywords Recommendation system · Skyline algorithm · Block nested loop · Divide and conquer

A. Er-Rafyg (✉) · A. Idrissi · K. El Handri
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: a.errafyg@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

K. El Handri
e-mail: kaoutar.elhandri@isga.ma

K. El Handri
LIMIE Laboratory, Computer Science Department, Higher Institute of Engineering and Business,
Casablanca, Morocco

1 Introduction

The transition to online learning has accelerated in recent years, especially after the COVID-19 pandemic. While online learning offers many benefits, it can also be overwhelming for learners trying to navigate the vast amount of information and options available to them. Today, students and schools all around the world depend on digital learning as a vital resource. This was a whole new type of schooling that many educational institutions had to go through. Today, online learning is used for both academic and extracurricular activities by learners.

The learner would feel overwhelmed when searching for a course that will be well matched to his needs because of the exponential growth of data in the sector of online learning [1]. For this reason, recommender systems have become increasingly important in e-learning, as they help learners find courses and resources that are tailored to their needs and interests [2], as they provide a potent method of combating the issue of information overload and assisting users in locating pertinent material that satisfies their requirements.

This article is structured in seven main sections; the next section presents the background. The third section exposes the problematic statement. In the fourth we can find the research methodology. The fifth section presents some related works. The proposed approach is presented in the sixth section. The last section concludes the article and highlights some perspectives.

2 Background

2.1 Recommendation Systems

We may be given a variety of options in different situations, but we could not have the time, knowledge, or resources to fully consider each option. In these situations, we frequently rely on the recommendations of others to guide our decision-making [3, 4, 13]. Like we are planning a trip to a new city, we may not know which hotel or restaurant to choose. In such cases, we may look at reviews or ratings from other travelers to help us make a decision. Similarly, when we are considering which book to read or which movie to watch, we may rely on the recommendations of friends or family members who have similar interests. By filtering the retrieved information items based on past suggestions or recommendations made by other users concerning those items, recommender systems are tools used in the processes of information

access with the aim of assisting users in their information search operations [5]. The main elements of a recommender systems are the following [3].

- User profiles are regularly employed by recommender systems to collect data on a user's preferences or requirements. Through either explicit or implicit user feedback, these profiles can be created.
- A representation of the items that have all the crucial items' features. Experts in the particular application field typically include this information.
- The method by which recommendations are made. Several methods can be employed. Some of the most commonly used techniques include demographic, knowledge-based, content-based, collaborative (item-based or user-based), demographic, or hybrid methods which integrate numerous strategies to maximize their benefits and minimize their downsides.
- Rating history which are the values that users put into the system when looking over an item or changing a value that has already been given

In any instance, as there are no better or worse approaches than others a priori, the success or failure of a technique to provide recommendations depends mostly on the scope of specific information where they are being applied.

2.2 Divide-and-Conquer Algorithm

Börzsöny et al. were the ones to originally introduce Skyline Operator [6]. It has been extensively developed and used to a variety of research subjects, including multi-criteria decision making and database visualization, data mining [7]. Börzsöny et al. employed the skyline operator to handle challenges involving selecting interesting points from a wide number of points that best fit pre-defined conditions, especially when the requirements are complementary [8].

The Skyline computation is used to extract all the non-dominated points from a large dataset. A non-dominated point is a tuple that is not dominated by any other tuples in the dataset, meaning it is either comparable or better than all other tuples in all aspects, and superior to at least one tuple in at least one aspect. By finding all the non-dominated points, the Skyline algorithm generates a list of tuples that represents the best options in the dataset, based on the criteria used to evaluate them. This can be useful in a variety of applications, such as in decision-making or in recommending products to customers [9]. The classic example of Skyline is illustrated in Fig. 1, Consumers are often looking for the cheapest beachfront hotel, but there is a conflict of interests since hotels are more costly the closer they are to the beach. The ultimate choice is up to the consumers; user A might pay more for a hotel that is closer to the beach, whilst user B would pick a less expensive hotel that is farther away from the beach. The best strategy is to just show consumers hotels that are interesting, that is,

hotels that are not more expensive or farther away from the user than any other hotel. The Skyline is made up of these hotels. The graphical representation of the points that make up a skyline, as seen in Fig. 1, inspired the word “skyline.”

The most effective way to compute the Skyline is to use the algorithm. The benefit of algorithms is that they can compute any Skyline, no matter how many dimensions it has. There might be several algorithms utilized, e.g., INDEX [10], Bitmap [10], the Block-Nested Loops algorithm (BNL) [6], the Divide and Conquer algorithm (D&C) [6], B-Tree [11], and many more.

The main idea behind the Divide and Conquer algorithm in the context of skyline computation. The goal is to efficiently compute the global skyline from a large dataset by breaking it down into smaller, more manageable subsets that can fit into memory.

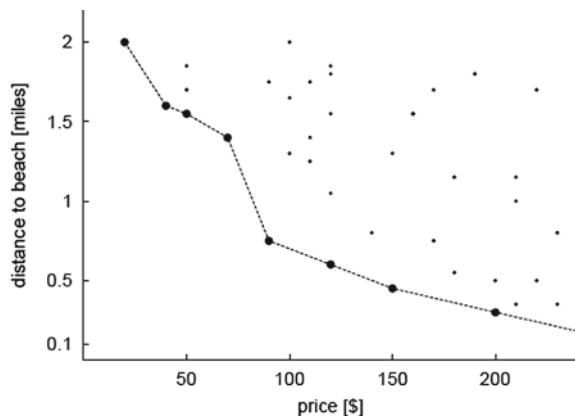
The algorithm works by recursively dividing the dataset into smaller subsets until each subset can be stored in memory. Then, it computes the local skyline for each subset, which consists of all the non-dominated points within that subset.

Once the local skylines have been computed, the algorithm begins to combine them in a series of pairwise merges, building up the global skyline incrementally. At each merge step, the non-dominated points from the two local skylines are compared to identify the new set of non-dominated points for the combined subset.

By iteratively merging the local skylines, the algorithm can efficiently compute the global skyline without having to store the entire dataset in memory at once. This can be particularly useful for large datasets where it is not feasible to compute the skyline using traditional methods [6]. The algorithm works as follows.

- i. Compute the α -quantiles of a set of input points along a certain dimension dp . Create m partitions from the points so that each partition fits in the memory. Let these partitions be $P_1; \dots; P_m$.
- ii. Using any known skyline computation algorithm, compute the skyline S_i of partition P_i , $1 \leq i \leq m$.
- iii. After merging the S_i pairwise, calculate the overall skyline. M -way partitioning is used during the merging procedure so that all sub-partitions can be combined in main memory.

Fig. 1 An illustration of the skyline as it appears in [6]



3 Problematic Statement

Users of e-learning platforms can choose from a wide variety of courses to suit their interests, but this has led to certain unexpected issues. Students may choose a specific course based on the instructor's reputation, the duration of the course, its level of difficulty, or other considerations unrelated to a particular subject area.

In situations where there is a large amount of information available, such as in the case of MOOCs (massive open online courses), locating courses that fit their needs can be difficult for learners. By examining user data, recommendation systems provide a potential remedy to this issue, such as their interests, behavior, and preferences, to make personalized recommendations.

We therefore want a recommendation system that can effectively manage huge preference queries in order to suit the demands of each learner. This system should enable learners to rapidly explore the data space without having to assign individual preference weights to each criterion.

3.1 Research Methodology

The research's conclusions were based on analyses of the literature, examinations of online learning environments, and the difficulties associated with recommending courses that will suit learners' goals while avoiding time waste. On the other hand, in order to overcome these barriers, we propose a new approach to course recommendation for students with multiple preferences. This method achieves the objective of recommendation by generating a list of recommended courses using the Divide and Conquer (DC) algorithm which divides the database of available courses according to the category chosen by the learner and then apply the Skyline Block-Nested Loops Algorithm (BNL) which will suggest the courses that best suit the learner's preferences.

4 Related Work

Recommendation Systems have attracted a lot of attention from researchers in recent years as a substitute to help learners choose from among the many e-learning courses available online those that best suit users' needs. This is a result of e-rising learning's popularity. Many research on this topic have been published in the literature.

For example, Bakhshinategh et al. [12] described how the assessment of the GAs can be utilized to create course recommendations. The main objective of the authors' suggested method is to recommend to students courses that would either enhance their overall competence profile or particular abilities that they wish to enhance. Based on the dates that the students gave, the algorithm operates in a collaborative filtering environment, taking into consideration the time component (i.e. recent student assessments add more value to the recommendation) [12].

Asra et al. in [13] proposed an accurate and practical algorithm for MOOCs named 'Novel Online Recommendation Algorithm for Massive Open Online Courses (NoR-MOOCs), which will utilize rating data of learners for recommendations.

Dahdouh et al. [14] developed a system specifically for the online learning environment that recommends courses. It tries to identify connections between students' course activities using the association rules approach to assist the student in selecting the most suitable learning materials they concentrated on the analysis of historical data from prior course enrollments or log data. The frequent itemsets notion is specifically discussed in the paper to identify the intriguing rules in the transaction database. The authors then utilized the guidelines they had extracted to locate a catalog of courses that were more suited to the learner's tastes and habits. They then put the system into practice using the R programming language and the FP-growth algorithm.

Apaza et al. [15] developed a mechanism for recommending courses based on previous college students' grades. The courses that are currently offered on websites like Coursera, Udacity, Edx, etc. will be recommended by our algorithm. To do this, the following probabilistic topic models are applied. On the one hand, the Latent Dirichlet Allocation (LDA) topic model infers subjects from the course syllabus's provided information. Topics are also taken from a massive open online course (MOOC) syllabus, on the other hand. A content-based recommendation system matches these two sets of topics and grade information in order to suggest to students online courses that are appropriate for them. Initial findings demonstrate the applicability of our strategy.

The work of Perumal et al. [16] aims to produce a list of recommendation alternatives with the greatest anticipation ratings of several important concepts that are prepared for internet visitors to read. This method chooses the important ideas that are of greater interest to internet consumers using the fuzzy family tree similarity algorithm. Empirical analyses show that the suggested method is effective and practical for including the essential ideas in the recommendation list that would otherwise be missed by the traditional tree similarity method. Based on the user key concept rate (UKCR) matrix, the suggestion alternatives, and neighbors arranged in order of semantic and content similarity, anticipatory ratings are calculated.

5 Proposed Approach

The traditional approach of students looking for interesting and pertinent courses is no longer applicable given the increasing quantity of courses offered online. Students think that the website would offer personalized services and will automatically suggest engaging courses to them based on their unique needs and preferences. In our previous work [17], Our recommender system for courses makes use of the Skyline algorithm, which creates a list of courses that is certain to include every course for any set of linear feature requirements. However, this approach has shown its limitations when the learner requires multiple criteria at the same time in making a decision because the number of returned courses is considered as large for a learner who wants to find his required courses. To further refine the final results, we thought of combining Skyline BNL with the Divide and Conquer algorithm (Fig. 2).

From the technological perspective, in the proposed work, courses recommendations based on D&C and Skyline BNL algorithms. The proposed system works as follow:

- The learner will sign in to our platform and make a profile for themselves
- In the system the course will be presented by a profile, The value that is present for each criterion is all that is used to create the course profile.
- The user will choose the desired category and then he expresses his needs by selecting the criteria that must be taken into consideration
- The system first begins to divide the database, by the category chosen by the learner, into sub-databases and then will apply the Skyline BNL algorithm on



Fig. 2 Architecture of the proposed approach

each of the sub-databases in order to recommend courses that meet to the needs of the learner as shown in Algorithm 1.

Algorithm1: D&C
<ul style="list-style-type: none"> - L: input list of tuples - Lo, S: output Lists - p: tuple - Catg: list of categories available in the dataset - i: index of the categorie <p>Fonction step1(L,a):</p> <p>FOR EACH p in L:</p> <p style="padding-left: 2em;">IF p[i] = a :</p> <p style="padding-left: 4em;">Lo = Lo + {p}</p> <p>Return Lo</p> <p>Fonction step2(L,a):</p> <p>FOR EACH i in Catg</p> <p style="padding-left: 2em;">S= step1(L, i)</p> <p style="padding-left: 2em;">ComputeSkyline(S)</p> <p>Return S</p>

6 Experiment Results and Discussion

6.1 Used Datasets

Throughout the tests, we evaluate our approach using real data from Kaggle, which includes 3682 courses from the Udemy platform (<https://www.kaggle.com/datasets/andrewmvd/udemy-courses>) this dataset contains 12 columns, we only used 6 that meet the criteria proposed by our approach.

6.2 Results and Discussion

The results presented in Table 1 and Fig. 3 demonstrate that the new approach, which combines the Skyline BNL algorithm with the Divide and Conquer algorithm, is effective in recommending courses to learners. The approach was evaluated by varying the number of criteria used to generate the recommendations from 2 to 6, and the results showed that the number of courses returned was reduced, and the execution time was also decreased compared to the previous work [17] that used only the Skyline algorithm.

These findings suggest that the combined approach is a more efficient and effective way to generate personalized recommendations for learners, particularly when multiple criteria are involved. By partitioning the courses into smaller subsets based on additional criteria and then applying the Skyline BNL algorithm to each subset, the approach is able to generate a more targeted list of recommendations that match the learner’s specific needs and interests.

Overall, the results suggest that the combination of the Skyline BNL algorithm with the Divide and Conquer algorithm can be a valuable tool for course recommendation systems, particularly in the context of the increasing quantity of courses

Table 1 Comparison of results between our previous work and the new approach

Number of criteria	Approach used	Recommended list size	Response time (s)
2	BNL	14	0,028,019
	D&C + BNL	2	0.032974
3	BNL	22	0,04,503
	D&C + BNL	4	0.022601
4	BNL	32	0,049,032
	D&C + BNL	23	0.024855
5	BNL	50	0,057,038
	D&C + BNL	27	0.029974
6	BNL	84	0,108,073
	D&C + BNL	70	0.026049

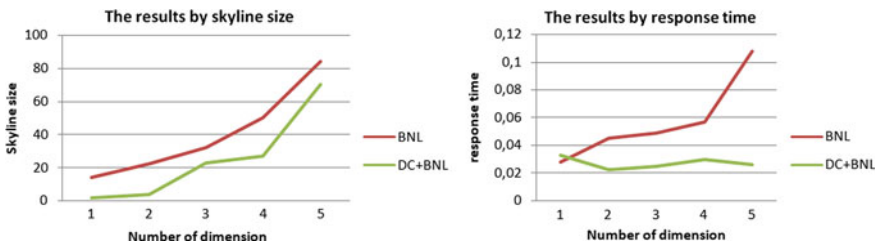


Fig. 3 Graphical presentation of results

offered online, where learners may be overwhelmed by the sheer number of options available.

7 Conclusion and Future Work

In this work we combined the Divide & Conquer and Skyline BNL algorithms to improve the results of courses recommender system proposed in our previous work. The goal was to provide more personalized recommendations while reducing the overwhelming number of courses offered. By reducing the list of recommended courses, the use of these two algorithms resulted in significant improvements in the outcomes of the recommender system. The Divide & Conquer algorithm is a technique that divides a large dataset into smaller subsets for easier processing, while the Skyline BNL algorithm is a technique that selects the best options from a set of candidates based on a set of criteria. By combining these two techniques, the study aimed to provide more accurate and relevant recommendations for learners. Several algorithms can be applied in our work. Specifically, those presented in [18–29] could be suited to better improve our investigations.

References

1. T. Huang, G. Zhan, H. Zhang, H. Yang, MCRS: a Course recommendation system for MOOCs 68, p. 4
2. F. Ricci (ed.), *Recommender Systems Handbook* (Springer, New York, 2011)
3. R. Burke, Hybrid recommender systems: survey and experiments 40, (2002)
4. J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: a survey. *Decis. Support. Syst.* **74**, 12–32 (2015). <https://doi.org/10.1016/j.dss.2015.03.008>
5. M. Balabanović, Y. Shoham, Fab: content-based, collaborative recommendation. *Commun. ACM* **40**(3), 66–72, (1997). <https://doi.org/10.1145/245108.245124>
6. S. Borzsony, D. Kossmann, K. Stocker, The Skyline operator, in *Proceedings 17th International Conference on Data Engineering* (Heidelberg, Germany, 2001), pp. 421–430. <https://doi.org/10.1109/ICDE.2001.914855>
7. H. Du, L. Shao, Y. You, Z. Li, D. Fu, A two phase method for skyline computation, in *Proceedings of 2019 Chinese Intelligent Systems Conference*, vol. 592, ed. by Y. Jia, J. Du, W. Zhang (Springer Singapore, Singapore, 2020), pp. 629–637. https://doi.org/10.1007/978-981-32-9682-4_66
8. M. Abourezq, A. Idrissi, F. Yakine, Routing in wireless Ad Hoc networks using the Skyline operator and an outranking method, in *Proceedings of the International Conference on Internet of things and Cloud Computing—ICC '16* (Cambridge, United Kingdom, 2016), pp. 1–10. <https://doi.org/10.1145/2896387.2900333>
9. M. Abourezq, Cloud service selection using the skyline and multi criteria decision aiding (2017)
10. R. Liu, D. Li, An Efficient Skyline Computation Framework, [arXiv:1908.04083 \[cs\]](https://arxiv.org/abs/1908.04083), août 2019, Consulté le: 20 septembre 2019. [En ligne]. Disponible sur: <http://arxiv.org/abs/1908.04083>
11. A. Idrissi, M. Abourezq, Skyline in cloud Computing **60**, 12 (2005)

12. B. Bakhshinategh, G. Spanakis, O. Zaiane, S. ElAtia, A course recommender system based on graduating attributes, in *Proceedings of the 9th International Conference on Computer Supported Education* (Porto, Portugal, 2017), pp. 347–354. <https://doi.org/10.5220/0006318803470354>
13. A. Khalid, K. Lundqvist, A. Yates, M.A. Ghzanfar, Novel online Recommendation algorithm for Massive Open Online Courses (NoR-MOOCs). *PLoS ONE* **16**(1), Art. no 1 (2021). <https://doi.org/10.1371/journal.pone.0245485>
14. K. Dahdouh, A. Dakkak, L. Oughdir, A. Ibriz, Building an e-learning recommender system using Association Rules techniques and R environment. *Int. J. Inf. Sci.* **2**, 8 (2019)
15. R.G. Apaza, E.V. Cervantes, L.C. Quispe, J.O. Luna, Online Courses Recommendation based on LDA, p. 7
16. S.P. Perumal, K. Arputharaj, G. Sannasi, Fuzzy family tree similarity based effective e-learning recommender system. in *2016 Eighth International Conference on Advanced Computing (ICoAC)* (Chennai, India, 2017), pp. 146–150. <https://doi.org/10.1109/ICoAC.2017.7951760>
17. A. Er-Rafyq, M. Abourezq, A. Idrissi, A. Bouhouch, Courses recommendations using Skyline BNL algorithm, 19
18. A. Idrissi, K. Elhandri, H. Rehioui and M. Abourezq, Top-k and Skyline for cloud services research and selection system. *Proc. Int. Conf. Big Data Adv. Wirel. Technol.* (2016)
19. A. Idrissi, C.M. Li, J.F. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks. in *18th IEEE International Conference on Tools with Artificial Intelligence*, Washington, USA (2006)
20. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. arXiv preprint arXiv: 1709.10372 (2015)
21. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
22. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10**, (2020)
23. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data Min.* (2017)
24. M. Abourezq, A. Idrissi, Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comput. Sci. Appl.* **6**(6), (2015)
25. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Architect.* **9**(2–3), 136–148 (2020)
26. M. Essadqi, A. Idrissi, A. Amarir, An effective oriented genetic algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**, 170–179 (2018)
27. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Elect. Eng.* **73**, 289–303 (2018)
28. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Robot. Intell. Syst.* **14**(3), 65–70 (2020)
29. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment. *Proc. Int. Conf. Big Data Adv. Wirel. Technol.* (2016)

Deep Reinforcement Learning in Financial Markets Context: Review and Open Challenges



Youness Boutyour  and Abdellah Idrissi 

Abstract The advancement of reinforcement learning techniques has increased their application across many industries, including the financial markets. An overview of reinforcement learning's application in the financial markets is given in this article. We start by outlining the fundamental ideas of financial markets and reinforcement learning. We next go over some examples of how reinforcement learning has been used to resolve problems related to market making, algorithmic trading, portfolio allocation, and optimal execution in the financial markets. Then, we examine a selection of twenty-one of the most relevant research papers using reinforcement learning to tackle each of the aforementioned types of financial market problems. By using this selection as a comparative analysis, we were able to examine the results of each research paper and gain a deeper understanding of the challenges. As an outcome, we were able to discern the various methods and architectures used to finally list the research directions that require in-depth examinations.

Keywords Deep Reinforcement Learning · Autonomous Agent · Multi-Agent Reinforcement Learning · Markov Decision Process · Quantitative Finance · Financial Markets

1 Introduction

Many researchers have studied financial market analysis during the last few decades. The complicated mechanisms and varied financial market behaviors make it difficult to use and replicate the results of those simplified environments in practice. It could even lead to unsuccessful strategies. Additionally, several market-related factors

Y. Boutyour (✉) · A. Idrissi
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: youness.boutyour@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,
Studies in Computational Intelligence 1102,
https://doi.org/10.1007/978-3-031-33309-5_5

consideration may result in complicated models that are challenging to compute using conventional stochastic techniques.

Machine learning models are rapidly being used in the financial market owing to advances in artificial intelligence research and improved machine performances. Furthermore, many scientific studies discuss time series prediction, portfolio optimization, and trading decision-making in diverse financial markets.

One of the areas of artificial intelligence that has contributed to the scientific advancement of the use of AI in the financial sector is reinforcement learning (RL). Video games [1], robot control [2], and board games [3] are areas where reinforcement learning has achieved astounding success. Deep neural networks' availability as effective function approximators is the key to Reinforcement Learning's rapid progress. This type of learning aims to learn and automate decision-making oriented towards one or more objectives. To help agents learn how to make the best choices feasible through repeated experiences and interaction with the environment, it trains them how to interact with this environment. Reinforcement Learning is a formal framework that defines the interactions between one agent or multiple agents and an environment expressed by states, actions, and rewards. Many papers have covered the topic of reinforcement learning in finance, but this paper intends to go a step further by covering the most recent studies. This overview offers a methodical foreword to the use of reinforcement learning in the financial market. It is structured as follows: First, we will discuss the foundations of reinforcement learning, starting with Markov decision processes (MDP), which are the basis for modeling problems in reinforcement learning. This modeling allows us to discuss the different learning approaches to list the value-based and policy-based algorithms used in conjunction with ideas and techniques of the deep learning field. Secondly, we will move to understand the financial market and its different problems and characteristics. It will help us to properly model the financial market environments necessary to apply reinforcement learning algorithms. Finally, we will discuss the Open challenges in using reinforcement learning algorithms in those markets and obtain successful strategies and results (Fig. 1).

According to the following function $p(s', r | s, a)$, in an MDP, state S_{t+1} and reward R_{t+1} depend only on the prior state S_t , and action A_t [5]:

$$p(s', r | s, a) = Pr \left\{ R_{t+1} = r, S_{t+1} = s' | S_t = s, A_t = a \right\} \quad (1)$$

The agent's primary objective while interacting with the environment is to maximize returns which are determined by Eq. 2; as a result, the best possible policy,

Fig. 1 Illustration of the interaction between agent and environment [4]



state-value function V^* (Eq. 4), and action-value function Q^* (Eq. 3) are selected. The Bellman Optimality equation for figuring out Q^* is indeed very important [6].

$$G_t = R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 R_{t+3} + \dots \quad (2)$$

$$Q^*(s, a) = E[R_{t+1} + \gamma \max_{a'} Q^*(s', a')] \quad (3)$$

$$V_\pi(s, a) = E[R_{t+1} + \gamma(V_\pi(S_{t+1})) | S_t = s] \quad (4)$$

1.1 Markov Decision Process Formalism

For reinforcement learning, we define a Markov decision process (MDP) as a tuple of (S, A, P_a, R_a, γ) where [7, 8]:

- S is a finite set of permissible states of the environment and s_0 is the initial state of this environment.
- A is a finite set of actions; A_s is the finite set of actions applicable to state s , if different sets of actions apply to different states.
- P_a is the transition Probability Matrix that would change state to a state s' at time $t + 1$ given an action a in the state s at time t . It is expressed by:

$$P_a(s, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$$

- After action a changes state s to state s' , $R_a(s, s')$ is the reward earned.
- The discount factor $\gamma \in [0, 1]$ stands for the difference between the future and present rewards value.

1.2 Types of Reinforcement Learning Algorithms

Three possible functions can be learned by an agent in Reinforcement Learning:

- A value function that forecasts the quality of each state or state/action pair, $v_\pi(s)$ or $Q_\pi(s, a)$,
- A policy, $\pi(s)$ or $\pi(s, a)$, which maps state to action: $a \sim \pi(s)$,
- An environment model $P(s' | s, a)$.

From these functions, three types of RL algorithms can be deduced [9]:

- Value-based algorithms, such as: SARSA [10], DQN [11], DDQN [12], etc.
- Policy-based algorithms, such as: REINFORCE [13].

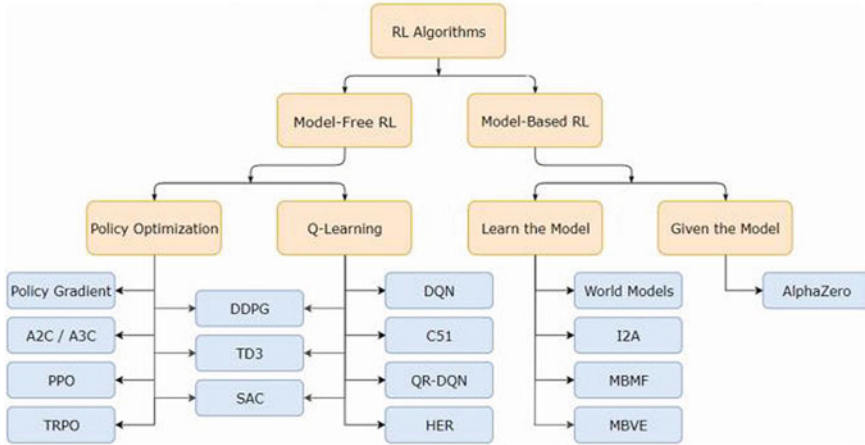


Fig. 2 Taxonomy of Reinforcement Learning Algorithms, this list is not exhaustive [16]

- Model-based algorithms, such as: iLQR [14], MPC, MCTS [9], MBMF, AlphaZero, etc.

The value-Based and Policy-Based models and their derivatives form the so-called model-free algorithms. Despite the fact that adopting model-based approaches may increase sampling efficiency, model-free approaches typically have a simpler implementation and a more reliable tuning procedure [15].

The learning methods used in the RL space are depicted in high-level detail in Fig. 2.

1.3 Policy and Value Functions

The policy is the probability of executing action \mathbf{a} while the agent is in state \mathbf{s} at the time \mathbf{t} . A mapping function is used by the agent to learn how to build correspondences between the state and the actions to maximize the total return. There are two types of policies (See Fig. 3).

- **Deterministic policies:** It has a single, distinct action for each state. In other words, the probability function becomes a straightforward mapping function, with a function's value being 1 for some action $\mathbf{A}_t = \mathbf{a}$ and 0 for all other actions \mathbf{A}_t . The agent learns to only choose the best possible action.
- **Stochastic policies:** The chance that the agent will choose one of the many possible actions for a given state is provided by the probability function $\pi(\mathbf{a}|\mathbf{s})$ (Table 1).

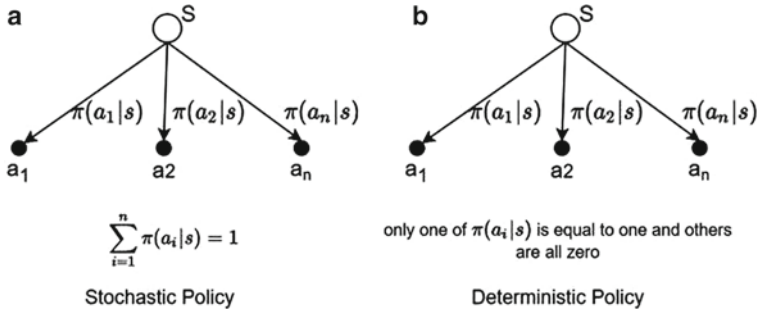


Fig. 3 Types of policies. **a** Stochastic policy **b** Deterministic policy

Table 1 Popular Reinforcement Learning Algorithms

Algorithm	Full name	References
A2C/A3C	Asynchronous advantage actor-critic	Mnih et al. [17]
PPO	Proximal policy optimization	Schulman et al. [2]
SAC	Soft actor-critic	Haarnoja et al. [18]
DDPG	Deep deterministic policy gradient	Lillicrap et al. [19]
TD3	Twin delayed DDPG	Fujimoto et al. [20]
DQN	Deep Q-networks	Mnih et al. [11]

1.4 Multi-Agent Reinforcement Learning

In contrast to single-agent RL, Multi-Agent RL (MARL) uses multiple agents to solve sequential decision problems through a trial-and-error technique [21]. The joint actions of all agents directly influence the state of the environment and the reward received by each agent. The main task of each agent is of course to maximize its cumulative reward. The interaction then becomes not only with the environment but also among the agents. Multi-agent reinforcement learning integrates the principles of reinforcement learning and multi-agent systems [22]. Figure 4 illustrates a multi-agent reinforcement learning schema.

There are three categories of MARL problems according to the behavior of the agents: competitive, cooperative, and mixed. In a competitive behavior, the sum of the agents' rewards is equal to zero. One agent win means a defeat for another. In cooperative behavior, agents work together to maximize a shared reward value. One agent win is a victory for all agents. In contrast, mixed behavior involves both competitive and cooperative agents; some actions may result in a win/win situation, while others may end in a win/loss situation (Fig. 5).

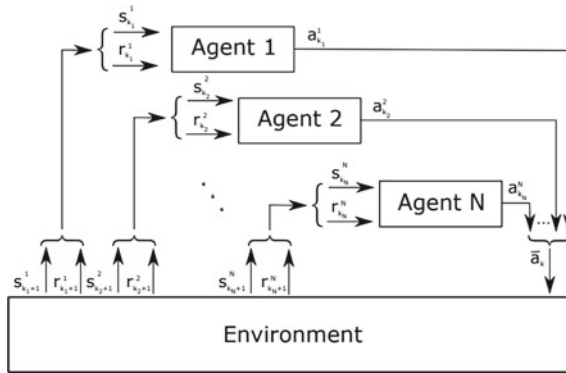


Fig. 4 Multi-agent reinforcement learning schema



Fig. 5 Hide and seek game simulated by cooperative agents in different environments [23]

The multi-objective function optimization is a common task in real-world problems. All agents in a fully cooperative environment share the same reward function.

The term “multi-objective” refers to a problem where each agent’s reward function has different values. The agents involved in decentralized decision-making must express their preferences to one another. Due to this heterogeneity, multi-agent reinforcement learning must incorporate communication protocols, and communication-efficient algorithms must be investigated [24].

2 Financial Markets

Financial markets are vital to the efficient operation of capitalist economies because they distribute resources and provide liquidity for businesses and entrepreneurs. Markets for buyers and sellers facilitate trading in financial holdings. The creation of security products by financial markets results in gains for lenders and investors as well as the availability of funds for those who require additional finance. These markets are created by buying and selling different financial instruments forms, such as stocks, bonds, currencies, and derivatives. They set effective and reasonable prices. Each day, the financial markets trade assets worth trillions of dollars.

Types of Financial Markets

The following list includes the major financial markets that belong to the capital markets.

- **Stock Market:** The stock market is arguably the financial market that is most prevalent. These are places where investors and traders can buy and sell shares listed by companies. Stock markets, often referred to as equities markets, are used by businesses to collect funds through an IPO, after which shares are traded between different buyers and sellers in a market known as the secondary market.
- **Bond Market:** Bonds are securities in which an investor lends money for a predetermined period at an agreed-upon interest rate. An agreement between a lender and borrower specifies the terms of the loan and its repayments. Municipalities, states, sovereign governments, and corporations issue bonds to fund operations and projects [25].
- **Derivatives Market:** A derivative is a contract between two or more parties, the value of which is determined by an agreed-upon underlying financial asset, such as a security, or group of assets, such as an index. Secondary securities known as derivatives take the whole of their value from the price of the original security to which they are tied.
- **Commodities Market:** Commodities include crops, animals, cocoa, oil, meat, gold, and other natural resources, which are traded on the commodities market by traders and investors.
- **Foreign Exchange Market (Forex):** The foreign exchange market, sometimes known as the forex market, is a place where participants can speculate, buy, sell, hedge, and deal in currency pairs. As a result of the liquidity of cash as an asset, the forex market is the most liquid in the entire globe.
- **Cryptocurrency Market:** Cryptocurrencies, or decentralized digital assets built on blockchain technology, like Bitcoin [26] and Ethereum, have been introduced and have grown significantly over the past few years. Today, a variety of independent online cryptocurrency exchanges offer hundreds of cryptocurrency tokens for trading. These exchanges provide traders with access to digital wallets where they can exchange one cryptocurrency for another or fiat money like dollars or euros.

3 Financial Markets Open Challenges for RL

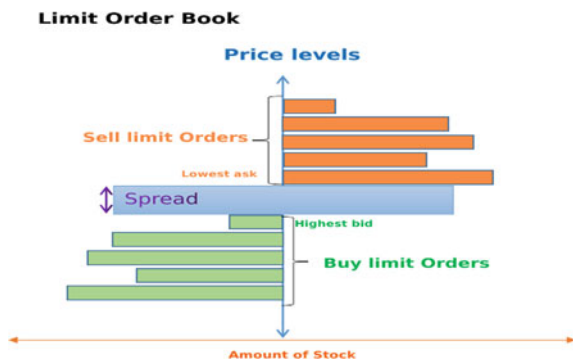
This section will look at the types of financial market-related challenges that fall under the category of problems that can be solved using reinforcement learning.

3.1 Market Making

In order to produce liquidity, market-making is a fundamental trading problem where an agent continuously offers to buy or sell assets. Inventory risk adds to the difficulty of the scenario by raising the possibility of building up a negative position and subsequently losing money. Since the introduction of the electronic limit order book (LOB) (see Fig. 6), It has become more automated because of the task becoming nearly challenging for individuals to do because of the need to manage more data and work on ever-shorter time periods [27]. A market maker's goal is to benefit by earning the spread (difference of ask and bid prices) while amassing an unnecessarily big position instead of profiting by finding the proper price movement direction. Three significant sources of risk are present for market makers:

- **The inventory risk** is the possibility of creating a high negative net inventory, which substantially raises volatility caused by market fluctuations.
- **The adverse selection risk** is the possibility of a directional price movement that overwhelms the market maker's limit orders and prevents a price reversion by the end of the trading period. Because marketers must often dispose of their inventory at the close of each trading session, this might result in a considerable loss.
- **The execution risk** is related to the execution of limit orders within a given time frame. This execution is not assured for sure.

Fig. 6 Illustration of a limit order book



3.2 Portfolio Allocation and Optimization

Asset allocation is an investing strategy that divides up a portfolio's assets following a person's financial status, risk tolerance, investment horizon, and goals to balance risk and reward. When a portfolio has a fixed asset allocation, it stays that way until the investor, or the manager of the portfolio acting on the investor's behalf decides to modify it. A well-known fixed allocation strategy called the "60/40 portfolio" has served as a trustworthy benchmark for investors who are willing to undertake moderate risks. The equal weight method is another well-liked approach to fixed asset allocation. Regardless of outside circumstances, this strategy accords the same importance to every item in the portfolio. To maximize the portfolio's risk-adjusted returns for a certain collection of assets, deep neural networks are utilized to solve the deep reinforcement learning problem that can be used to simulate portfolio management. These neural networks must recognize the market's behavior when reallocating the assets to ensure convergence. The advantage of using multi-asset portfolios is that the diversification of investments allows higher returns per unit of risk than managing a single asset [28].

3.3 Optimal Trade Execution

Institutional investors, banks and hedge funds all face the challenge of executing huge positions in the most advantageous way over a specific negotiating horizon. Rebalancing a portfolio inadvertently runs the risk of causing big negative price swings because other savvy traders may interpret the signal. Investors need to strike a balance between trading hastily and receiving poor execution prices and trading cautiously and becoming vulnerable to unexpected market fluctuations. Agents frequently encounter a wealth of data, such as previous asset values and many other economic factors. The algorithmic trading literature includes a significant section on choosing the optimum trading strategy in the face of all of this data. Researchers have previously proposed a probabilistic approach based on experimental findings, such as an "Ornstein–Uhlenbeck" process also called a stochastic volatility process [29]. They estimated model parameters using historical data, gave performance metrics that they aimed to maximize, and then dealt with the issue analytically using stochastic optimal control approaches. The Almgren–Chriss [30] approach is one of the earliest techniques in this genre, and it assumes that prices move in a Brownian manner.

3.4 Decision-Making in Trading

Algorithmic trading—also known as quantitative trading—can be thought of as the process of automatically generating trading decisions using a set of mathematical rules calculated by a program. The primary benefit of algorithmic trading for markets has already been demonstrated to be the notable increase in liquidity [31]. The use of algorithmic trading strategies is appropriate for a wide range of markets. A trader may invest in commodities futures, trade stocks and shares on stock exchanges, or trade forex. The current growth of cryptocurrencies also presents attractive opportunities. This type of trading is usually based on the price and volume of the asset, patterns, and technical indicators such as the moving average, RSI (Relative Strength Indicator), MACD, etc. An agent is programmed to interact with the financial market by making legal decisions according to the target market. The movement of the market is usually represented by Japanese candlesticks which are formed in a unit of time called Timeframe. Figure 7 shows the movement of a given asset represented by candlesticks. By specifying its State Space, Action Space, and Reward Function, the Partially Observed Markov Decision Process (POMDP) is used to describe the trading problem. The trading environment is a POMDP model, which reflects the nature of trade in the real world. The agent’s responsibility is to make decisions about when, how much, and how to trade, with the typical goal of maximizing profit while taking risk into consideration. In this situation, there are a variety of challenges, including partial observability, a sizable action space, and a rigorous formulation of rewards and training goal [32].

4 Related Works and Discussions

In this section, we will present a brief summary of the recent works done in the area of applying reinforcement learning to financial market challenges. We discuss the most relevant research papers proposing models for each challenge.



Fig. 7 Representation of market movements using Japanese candlesticks

4.1 Market Making

There are few scientific publications in the literature that uses reinforcement learning to solve the Market Maker problem. Table 2 summarizes the results of those reviewed articles. Ganesh et al. [33] used a multi-agent architecture to simulate a dealer market that provides liquidity to the market by continuously offering bid and ask prices. The goal is to maximize the total PnL of the model. To achieve this goal, they used the standard implementation of the PPO algorithm with three formulations for the reward function in a competitive environment. As a result, the Market Maker agent learned to offer the optimal ask and bid prices better than the other models. Alexey Bakshaev [34], on his side, implemented his model based on the Soft Actor-Critic (SAC) algorithm to simulate the Market Maker. He conducted an extensive study comparing different models and use cases. The models showed significant results with better Sharpe ratio and PnL values. For their part, Selser et al. [35] modeled two algorithms based on Deep Q-Network (DQN) and Tabular Q-Learning (TQL) successively. After the agent training phase, experiments showed favorable results for the DQN-based model. This model had the best Sharpe ratio and standard deviation wealth. Like Alexey Bakshaev, Gasperov et al. [36] used the SAC algorithm to model a stochastic control problem of an optimal market maker. The Deep Reinforcement Learning (DRL) agent trained on a Limit Order Book (LOB) model based on the multivariate linear Hawkes process. To compare the DRL model and classical Market-Making techniques, they ran Monte Carlo simulations with the simulator's produced synthetic data. Statistical metrics such as PnL, Mean episodic return, MAP, and Sharpe ratio are used to evaluate and compare generated models. The results of the experiments have shown that the DRL model could play a reliable role as a

Table 2 A small subset of research papers uses DRL in the market-making problem

Algorithm	Goal	Performance	References
PPO	Maximize the total PnL	<ul style="list-style-type: none"> • Dealer market modelization as a multi-agent system • Learning the optimal pricing point 	Ganesh et al. [33]
SAC	Maximize reward expressed as PnL minus the penalty function	Better sharpe ratio and PnL values	Bakshaev [34]
DQN/TQL	Maximize cumulative discounted reward	Best sharpe ratio than benchmarks methods	Selser et al. [35]
SAC	Maximize the expectation wealth and minimize the inventory risk	Favorable sharpe ratio and higher mean PnL value	Gasperov et al. [36]
Decentralized MADDPG)	Maximize profit	<ul style="list-style-type: none"> • Formalize the dealer market as a stochastic differential game • Learn the optimal pricing ask/bid in a competitive environment 	Cont et al. [37]

control agent in a market-maker environment. Cont et al. [37] used a decentralized multi-agent architecture based on the DDPG algorithm to model the interaction of a Market Maker with the market as a stochastic differential game of intensity control with partial information. The study focused on learning bid-ask pricing spreads in a competitive environment with other market makers to maximize profit. The experiments showed that the generated model was able to learn to quote ask and bid prices better than the competitors. It should be noted that the experiments also showed that the number of competing Market Makers influences the results obtained and can generate losses.

4.2 Portfolio Allocation and Optimization

Numerous scientific papers have examined the portfolio optimization problem using reinforcement learning techniques. A subset was studied to provide a summary table of the algorithms used, the reward function employed, the performance obtained, and the results found (See Table 3). Liang et al. [38] used the DDPG, PPO, and Policy-gradient (PG) algorithms to generate three models dealing with the portfolio optimization problem. The experiments applied to the Chinese stock market data. The implementation of algorithms was inspired by the work of Jiang et al. [39] with the difference of using a Deep Residual Network instead of a CNN. Wang [40], on his side, used reinforcement learning to solve the large-scale Markowitz mean–variance (MV) portfolio allocation problem. To describe this problem, the continuous-time exploratory control framework is being used. The implemented models have been tested on S&P500 data and have had a consistent annual return of about 10%. The result far outperforms classical methods and other RL models on a monthly and daily trading basis. “Pretoriu” et al. [41] implemented models based on DDPG, PPO, A2C, and FRONTIER algorithms to tackle the portfolio optimization problem. The main ideas of the study were the inclusion of non-linear terms in the cost expression and

Table 3 A small subset of research papers using DRL in portfolio allocation and optimization problems

Algorithm	Reward	Performance	References
DDPG/PPO/PG	Accumulative portfolio value	Models outperform the UCRP method	Liang et al.
EMV	Mean–variance	Consistent annual returns over 10%	Wang [40]
Frontier (Monte Carlo method)	Discounted future rewards	RL model outperforms mean–variance optimization models	Pretoriu et al. [41]
A3C	Sharpe ratio	Higher average sharpe ratio than classic benchmarks methods	Ahn et al. [42]
Various RL algorithms	Differential return	27.6% annual return	Durall [43]

the introduction of investor preferences as model parameters. In the majority of assessed risk values, the RL models surpassed the standard mean–variance model. Another model in the literature was done by Ahn et al. [42]. The paper used the A3C algorithm taking into account the statistical characteristics of the financial market while avoiding over-optimization by using Monte Carlo simulation data. The model performed better than other classical techniques such as Markowitz, Risk-Budgeting, and Equal Weight. The portfolio return, Sigma, Maximum Draw-down, and Sharpe ratio were all compared. Another study by Durall [43] conducted extensive experiments by implementing nine different models based either on reinforcement learning algorithms or traditional portfolio optimization models such as tangency portfolio, minimum variance portfolio, risk parity, and equal weight. The results confirmed the stability of the classical methods, whereas the RL models were difficult to interpret, especially with significant variations. The PPO and SAC algorithms gave the best results on the bullish and bearish market trends.

4.3 Optimal Execution

Several reinforcement learning methods are applied to address the optimal trade execution problem, such as DQN, A2C, PPO, and DDPG algorithms. The study of scientific papers has shown that the actions taken by intelligent agents are composed of a single action allowed to follow the general trend, to several ones (buy, sell, hold) to take different decisions depending on the targeted problem. The state of the environment usually consists of attributes like the date, the current price of the asset, the spread, the available or desired quantity, and other elements depending on the studied paper. These papers use different functions of rewards: total profit, return, Sharpe ratio, Sortino ratio, PnL, and trading cost. Table 4 summarizes a subset of works done to address the optimal execution problem with reinforcement learning. Hendricks et al. [44] introduce the DQN algorithm in the Almgren-Chriss solution to prove that using reinforcement learning for solving the optimal execution problem can be achieved by sending a market order sequence based on a limit order book. This technique outperforms the static Almgren-Chriss method, the most widely used technique in this area, and performs better in the South African equity market. Ning et al. [45] also employed a modified double DQN model that maximized PnL (Profit and Loss) while minimizing total trading costs. Pan et al. [4446] used an architecture based on the PPO algorithm to maximize return and PnL. They proposed a method consisting of two phases: HALOP-stage1 and HALOP-stage2. With a discretized Gaussian policy, in stage 1, the agent selects a target percentage price change and evaluates a subset of discrete stocks; in stage 2, the agent selects a tick-based stock using a fine-grained Gaussian and softmax policy. Unlike the use of limit orders like Pan [46], Fang [47] used market orders which can lead to the execution of the order volume with disadvantageous prices, especially when the spreads are wide.

Table 4 A small subset of research papers using DRL in optimal execution problem

Algorithm	Goal	Performance	References
DQN	Maximize profit	RL model outperforms the base model	Hendricks et al. [44]
DQN	Maximize PnL	PnL(model) > PnL(TWAP)	Ning et al. [48]
HALOP (based on PPO)	Maximize return/PnL	HALOP's return and PnL are bigger than TWAP, VWAP and ODP ones	Pan et al. [46]
TD	Maximize profit	outperforms TWAP and VWAP strategies	Moallemi et al. [49]
PPO	Minimize the expected total execution cost	Proposed model outperforms TWAP strategy	Fang et al. [47]

4.4 Decision-Making in Trading

Table 5, summarizes the methods and results in a small set of scientific papers that have studied the trading problem using reinforcement learning. Different algorithms were used such as DDQN, A3C, TD3, and others. Ponomarev et al. [50] used a policy-gradient algorithm (A3C) in the reinforcement learning model with return as the reward function. They obtained an annual profit of 110% on the RTS futures dataset. Lucarelli et al. [51] implemented a model with the DDQN algorithm, they use the profit and the Sharpe ratio as a reward function. The models obtained have a higher average cumulative return than the classical trading strategies such as “Buy & Hold” and “Sell & hold”. Based on the same algorithm as [51], Théate et al. [32] implemented a model using daily returns as a reward function. The Sharpe ratio obtained outperformed that of classical trading strategies. Jin [49] experiments with a different method named R3L, which consists in estimating the distribution of cumulative

Table 5 A small subset of research papers using DRL in trading problem

Algorithm	Reward	Performance	References
A3C	Returns	Profitability of 110% per annum including costs	Ponomarev et al. [50]
DDQN	Sharpe ratio/profit	Higher cumulative average return	Lucarelli et al. [51]
DDQN	Daily returns	RL's sharpe ratio outperforms classic trading strategies	Théate et al. [32]
Online transfer learning	Cumulated returns	Total return of 350% over five years	Borrageiro et al. [53]
R3L	Portfolio value	R3L sharpe ratio performs better than classic strategies	Jin [49]
TD3	Portfolio value	2.68 sharpe ratio on unseen data	Kabbani et al. [52]

rewards instead of the expectation by the critic network. The obtained model had a better Sharpe ratio than the classical trading strategies. While Kabbani et al. [52], based their model on the TD3 algorithm, while using the portfolio value as a reward function, the main finding is the high value of the Sharpe ratio on the test data which is 2.68.

5 Conclusion

This paper provides an overview of the use of reinforcement learning in financial markets. The field of RL in financial markets is relatively new but has attracted a lot of attention from researchers in recent years. There is a multitude of open problems in this field, which range from using RL to model financial markets, to proposing new methods and architectures, and optimizing agents for financial market problems. The purpose of this work is to present the state-of-the-art for the different types of challenges that have used RL as an approach in financial markets.

The findings of the research presented in this paper are noteworthy and encourage further investigation in each area. Each open challenge has a set of hyper-parameters and research tracks that need to be discovered, tuned, and optimized. This highlights the ongoing and active nature of the field, with many opportunities for future research and development.

Finally, the paper concludes that deep reinforcement learning can handle most of these problems. This is due to the ability of deep RL to handle high-dimensional, non-linear problems, which are often present in financial markets. We also intend to use some methods presented in [54–65] to confront them on our subject and draw from them conclusions related to this area of finance.

References

1. V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015)
2. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms (2017). arXiv preprint [arXiv:1707.06347](https://arxiv.org/abs/1707.06347)
3. D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016)
4. R.S. Sutton, A.G. Barto, *Introduction to reinforcement learning*, 1st edn. (MIT Press, Cambridge, MA, 1998)
5. R. Bellman, A Markovian decision process. *J. Math. Mech.* **6**(5), 679–684 (1957). [Online]. <http://www.jstor.org/stable/24900506>

6. R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction*, 2nd edn. (A Bradford Book, Cambridge, MA, USA, 2018)
7. K. Arulkumar, M.P. Deisenroth, M. Brundage, A.A. Bharath, Deep reinforcement learning: a brief survey. *IEEE Signal Process. Mag.* **34**(6), 26–38 (2017)
8. D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel et al., A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362**(6419), 1140–1144 (2018)
9. D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **529**(7587), 484–489 (2016)
10. G.A. Rummery, M. Niranjan, *On-line Q-learning using connectionist systems* (University of Cambridge, Tech. rep, 1994)
11. V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing atari with deep reinforcement learning. arXiv e-prints (2013)
12. H. Van Hasselt, Double Q-learning, in *Advances in Neural Information Processing Systems*, vol. 23, ed. by J.D. Lafferty, et al. (Curran Associates, 2010), pp. 2613–2621
13. R.J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**(3–4), 229–256 (1992). ISSN: 0885-6125
14. W. Li, E. Todorov, Iterative Linear quadratic regulator design for non-linear biological movement systems, in *Proceedings of the 1st International Conference on Informatics in Control, Automation and Robotics (ICINCO 1)*, ed. by H. Araújo, A. Vieira, J. Braz, B. Encarnação, M. Carvalho, (INSTICC Press, 2004), pp. 222–229. ISBN: 972-8865-12-0
15. D. Yarats, Y. Zhang, I. Kostrikov, B. Amos, J. Pineau, R. Fergus, Improving sample efficiency in model-free reinforcement learning from images. arXiv e-prints (2020)
16. SpinningUp OpenAI website (2022). [Online]. https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html
17. V. Mnih, A. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, *Asynchronous Methods for Deep Reinforcement Learning* (ICML, 2016)
18. T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor* (ICML, 2018)
19. T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, *Continuous Control with Deep Reinforcement Learning* (ICLR, 2016)
20. S. Fujimoto, H. Hoof, D. Meger, *Addressing Function Approximation Error in Actor-Critic Methods* (ICML, 2018)
21. S. Gronauer, K. Diepold, Multi-agent deep reinforcement learning: a survey. *Artif. Intell. Rev.* **55**, 895–943 (2022). <https://doi.org/10.1007/s10462-021-09996-w>
22. Y. Shoham, K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game Theoretic, and Logical Foundations* (Cambridge University Press, 2008)
23. B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, I. Mordatch, Emergent tool use from multi-agent autotutorials. arXiv preprint [arXiv:1909.07528](https://arxiv.org/abs/1909.07528) (2019)
24. C.F. Hayes, et al. A practical guide to multi-objective reinforcement learning and planning (2021)
25. F. Black, M. Scholes, The pricing of options and corporate liabilities. *J. Polit. Econ.* **81**, 637–654 (1973)
26. S. Nakamoto, Bitcoin: a peer-to-peer electronic cash system (2009)
27. J. Hasbrouck, G. Saar, Low-latency trading. *J. Financ. Mark.* **16**(4), 646–679 (2013)
28. E. Zivot, *Introduction to Computational Finance and Financial Econometrics* (Chapman and Hall Crc, 2017)
29. G.E. Uhlenbeck, L.S. Ornstein, On the theory of Brownian motion. *Phys. Rev.* **36**, 823–841 (1930)
30. R. Almgren, N. Chriss, Optimal execution of portfolio transactions. *J. Risk* **3**, 5–40 (2001)
31. T. Hendershott, C.M. Jones, A.J. Menkveld, Does algorithmic trading improve liquidity? *J. Financ.* **66**, 1–33 (2011)

32. T. Théate, D. Ernst, An application of deep reinforcement learning to algorithmic trading. *Expert Syst. Appl.* **173**, 114632 (2021)
33. S. Ganesh, N. Vadori, M. Xu, H. Zheng, P. Reddy, M. Veloso, Reinforcement learning for market making in a multi-agent dealer market. (2019) [arXiv:1911.05892v1](https://arxiv.org/abs/1911.05892v1)
34. A. Bakshaei, Market-making with reinforcement-learning. (2020). [arXiv:2008.12275v1](https://arxiv.org/abs/2008.12275v1)
35. M. Selser, J. Kreiner, M. Maurette, Optimal market making by reinforcement learning. (2021). [arXiv:2104.04036v1](https://arxiv.org/abs/2104.04036v1)
36. B. Gasperov, Z. Kostanjcar, Market making with signals through deep reinforcement learning. *IEEE Access* **9**(2021). <https://doi.org/10.1109/ACCESS.2021.3074782>
37. R. Cont, W. Xiong, Dynamics of market making algorithms in dealer markets: learning and tacit collusion (2022)
38. Z. Liang, H. Chen, J. Zhu, K. Jiang, Y. Li, Adversarial deep reinforcement learning in portfolio management. (2018). [arXiv:1808.09940v3](https://arxiv.org/abs/1808.09940v3)
39. Z. Jiang, X. Dixing, J. Liang, A deep reinforcement learning framework for the financial portfolio management problem. (2017). [arXiv:1706.10059](https://arxiv.org/abs/1706.10059)
40. H. Wang, Large-scale continuous-time mean-variance portfolio allocation via reinforcement learning. (2019)
41. R. Pretorius, T.L. Zyl, Deep reinforcement learning and convex mean-variance optimisation for portfolio management. *J. IEEE Trans. Artif. Intell.* (2022)
42. J. Ahn, S. Park, J. Kim, J. Lee, Reinforcement learning portfolio manager framework with Monte Carlo simulation (2022)
43. R. Durall, Asset allocation: from markowitz to deep reinforcement learning (2022)
44. D. Hendricks, D. Wilcox, A reinforcement learning extension to the Almgren-Chris framework for optimal trade execution, in *2014 IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFER)*, (IEEE, 2014), pp. 457–464
45. Z. Zhang, S. Zohren, S. Roberts, Deep reinforcement learning for trading. *J. Financ. Data Sci.* **2**, 25–40 (2020)
46. F. Pan, T. Zhang, L. Luo1, J. He1, S. Liu, Learn continuously, act discretely: hybrid action-space reinforcement learning for optimal execution. (2022). [arXiv:2207.11152v1](https://arxiv.org/abs/2207.11152v1)
47. J. Fang, J. Weng, Y. Xiang, X. Zhang, Imitate then transcend: multi-agent optimal execution with dual-window denoise PPO. (2022). [arXiv:2206.10736](https://arxiv.org/abs/2206.10736)
48. B. Ning, F.H.T. Ling, S. Jaimungal, Double deep Q-learning for optimal execution. (2018). [arXiv:1812.06600](https://arxiv.org/abs/1812.06600)
49. B. Jin, An intelligent algorithmic trading based on a risk-return reinforcement learning algorithm. (2022). [arXiv:2208.10707v2](https://arxiv.org/abs/2208.10707v2)
50. E.S. Ponomarev, I.V. Oseledetsa, A.S. Cichocki, Using reinforcement learning in the algorithmic trading problem, mathematical models and computational methods (2019)
51. G. Lucarelli, M. Borrotti, A deep reinforcement learning approach for automated cryptocurrency trading, in *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations* (Crete, Greece, 2019), pp. 247–258
52. T. Kabbani, E. Duman, Deep reinforcement learning approach for trading automation in the stock market. (2022). [arXiv:2208.07165v1](https://arxiv.org/abs/2208.07165v1)
53. G. Borrageioro, N. Firoozye, P. Barucca, The recurrent reinforcement learning crypto agent. *IEEE Access* (2022)
54. A. Idrissi, K. Elhandri, H. Rehioui, M. Abouzeq, Top-k and Skyline for cloud services research and selection system, in *International Conference on Big Data and Advanced Wireless Technologies* (2016)
55. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. *arXiv preprint* [arXiv:1709.10372](https://arxiv.org/abs/1709.10372) (2015)
56. A. Idrissi, C.M. Li, J.F. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks, in *18th IEEE International Conference on Tools with Artificial Intelligence* (2006)
57. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data Min.* (2017)

58. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
59. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10** (2020)
60. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Architect.* **9**(2–3), 136–148 (2020)
61. M. Abourezq, A. Idrissi, Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comput. Sci. Appl.* **6**(6) (2015)
62. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electr. Eng.* **73**, 289–303 (2018)
63. M. Essadqi, A. Idrissi, A. Amarir, An effective oriented genetic algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**, 170–179 (2018)
64. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment, in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (2016)
65. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Rob. Intel. Syst.* **14**(3), 65–70 (2020)
66. G. Jeong, H.Y. Kim, Improving financial trading decisions using deep Q learning: predicting the number of shares, action strategies, and transfer learning. *Expert Syst. Appl.* **117**, 125–138 (2019)
67. S. Bansal, R. Calandra, K. Chua, S. Levine, C. Tomlin, Model-based priors for model-free reinforcement learning (2016)
68. C. Moallemi, M. Wang, A reinforcement learning approach to optimal execution. *Quant. Financ.* **22**(2022)

A Survey of Parallel Computing: Challenges, Methods and Directions



Meryem Bouras and Abdellah Idrissi 

Abstract The processing of massive data in our real world today requires the necessity of high-performance computing systems such as massively parallel machines or the use of the cloud. And with the progression of parallel technologies in the coming years, Exascale computing systems will be used to implement scalable solutions for the analysis of massive data in the fields of science and economics. In order to achieve this, new design and implementation obstacles must be addressed to maximize the computing power of these new HPC systems in running Big Data and machine learning applications. Extreme Data pertains to massive amounts of Big Data that require processing and analysis in (near) real-time, necessitating a large number of memory and computing elements. Extreme Data pertains to massive amounts of Big Data that require processing and analysis in (near) real-time, necessitating a large number of memory and computing elements. Exascale computing systems, which are currently in development, will process and analyze substantial data repositories and continuous data streams, including scientific data generated at rates of hundreds of gigabits-per-seconds, millions of images each day, or billions of social data posts queried in real-time using an in-memory components database. The scale of such applications' data exceeds the capabilities of traditional disks and commercial storage systems, necessitating a vast number of cores to process them. This Research Topic aims to focus on data-intensive algorithms, systems, and applications running on systems composed of up to millions of computing elements, which underpin the Exascale systems, in response to the need for improvements in current concepts and technologies.

Keywords High performance computing (HPC) · Machine learning algorithms · Data processing · Parallel programming

M. Bouras (✉) · A. Idrissi
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: meryem.bouras@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,
Studies in Computational Intelligence 1102,
https://doi.org/10.1007/978-3-031-33309-5_6

1 Introduction

The introduction of parallelism in computers' architecture was a result of the confluence of three elements: the needs of applications, the limits of sequential architectures, and the existence of the property of parallelism in applications [1].

Parallelism is often attached to the execution performance of applications. This term encompasses a variety of concepts depending on the requirements of the application. In fact, whatever the field of application, parallelism can be seen as a solution for two needs: processing power and/or availability.

Processing power covers two main concepts: processing latency and processing throughput.

Latency represents the time required for the execution of a process.

The throughput represents the number of executable processes per unit of time.

These two notions can be independent. Reducing latency is more difficult than increasing throughput.

In the first case (latency), it is a question of fighting against time which depends on technological capacities.

Whereas, in the second case (throughput), if several processes are independent, the increase in the number of resources is enough to execute more processes at the same time.

Processing power also depends on the capacity and organization of computer memory. Some applications require data sets that are larger than the addressability of a sequential computer.

Multiplying the resources that each have their own memory increases the size of the total addressable memory. Some parallel architectures' organizational structures allow them to address more memory than sequential architectures.

Parallelism is the use of a group of processors capable of communicating and cooperating in order to speed up the resolution of a problem [2].

There are three fundamental parameters to take into account:

- The architecture: the type of memory, the interconnection of processors, ...
- The algorithm: the parallel algorithm must be designed taking into account the types of data handled, granularity, dependencies...
- Programming: choosing the appropriate language to express the parallelism depends on the architecture used.

2 Background

2.1 Partitioning

Partitioning (decomposition) consists of decomposing the problem into tasks, then:

- Determine concurrent (independent) tasks.
- Identify dependencies between tasks.

The decomposition into tasks makes the parallelism possible, with the objective of occupying the maximum number of processors.

= The degree of parallelism is given by the number of concurrent tasks at a given instant.

2.2 Granularity

Granularity refers to the amount of computation required to execute (execution time) a task. This execution time must be large compared to the time it takes to schedule it, place it, as well as communicate (send and receive data).

- coarse-grained tasks with many calculations
- Fine-grained tasks with very few calculations
- Finding a compromise based on performance (calculation, data access, system overhead, communications, disk, etc.)

2.3 Scheduling

Equal use of processors while keeping the volume of communications between processors as small as possible [2].

PS: All processors must have the same volume of calculations to execute.

Static scheduling: If scheduling is performed during the program initialization phase.

Dynamic scheduling: If scheduling is performed during program execution.

3 Parallel Architectures

Parallel processing is a form of information processing that exploits concurrent events at several levels: program level, instruction level, and operation level.

- Single processor architecture: A single processor that executes program instructions sequentially.
= The instructions are executed one after the other.
- Parallel architecture: Several processors execute the same program in parallel.
= Independent sequences of instructions are executed simultaneously.

3.1 Levels of Parallelism

- Parallelism internal to the processor (pipeline; calculation unit; cores)
- Parallelism internal to a machine (processors or accelerators)
- Machines connected by a network (computers, clusters)
- Multiple internet-scale connected computing centers (grid).

3.2 Definitions

Definition 1 Parallel computing is the use of multiple computing units (processors, cores, etc.) to solve the same problem (application). It consists of dividing a big problem into small independent ones [1].

Definition 2 Parallel computing is the set of software techniques and hardware that allows the execution of a program using several computing units rather than just one [2, 5].

Definition 3 A “parallel machine” (computer) is one that is made up of multiple processors that can communicate with one another [4].

3.3 Principle

The general principle is [2, 5]:

- Break down the program into tasks (sequences of instructions).
- Look for dependencies between these tasks and identify independent tasks.
- Assign independent tasks to different processors in order to be executed in parallel (simultaneously).

- Manage communications: manage synchronizations and exchanges of information between processors.

Important Note

Keep in Mind:

- data distribution: communicate as little as possible.
- load balancing: ensure that the load is distributed evenly among the processors.
- task granularity: the ability to balance computation and communication.

4 Advantages and Difficulties of Parallelism

4.1 Advantages

Gain in speed: increase the power of calculations, which leads to a reduction in execution times.

Theoretically, with a multiprocessor, we can calculate faster than with a single processor. In the ideal case, with p processors, the execution time is p times faster than with a single processor.

Gain in memory size: Theoretically, with p processors, we have p times more memory, which makes it possible to solve problems of larger size.

Exploiting new computing platforms: exploiting the parallelism available in modern processors (multi-core, multithreading, GPU).

4.2 Difficulties

- Manage task partitioning: identify concurrency (identify independent tasks).
- Manage task scheduling: assignment of tasks to processors.
- Manage memory access and communications.

= Control task dependencies: control information exchange.

5 Sequential Architectures

Single-processor architectures are based on the Von Neumann model (1946) which gave the outline for building an electronic machine which is made up of several functional blocks:

- Memory: contains the program (instructions) and the data.
- control unit: takes care of the sequencing of operations. It loads instructions from memory
- An Arithmetic and Logic Unit: Loaded with instructions to execute
- I/O: A data transfer input/output unit.

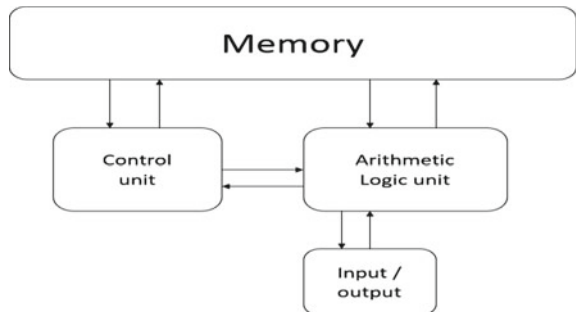
In a single-processor architecture, the instructions are executed sequentially (one after the other) (Fig. 1).

5.1 From Sequential to Parallelism

The introduction of parallelism led to:

- take into account new concepts compared to the sequential model:
 - Competition.
 - Communication.
 - The synchronization.
- reevaluate sequential algorithms to adapt them to parallel environments.
 - Partitioning.
 - Scheduling.
 - Data distribution.

Fig. 1 Von Neumann architecture



Differences

- Sequential programming is based on the Von Neumann model (a unique model).
 - A single execution flow.
 - Only one instruction is carried out at a time.
 - A processor.
 - For parallel programming, there is no single model.
 - Several executions flows.
 - Several instructions were executed simultaneously.
 - Several processors.
- = There are different aspects to consider:
- At the architectural level.
 - At the algorithmic level.
 - At the programming level.

This leads to more difficult programming than sequential programming.

6 Classification of Parallel Architectures

A good knowledge of the type of parallel architecture allows the research and the good design of the solution to the problem to be parallelized. The method of programming is not independent of the machine used.

6.1 Flynn's Classification

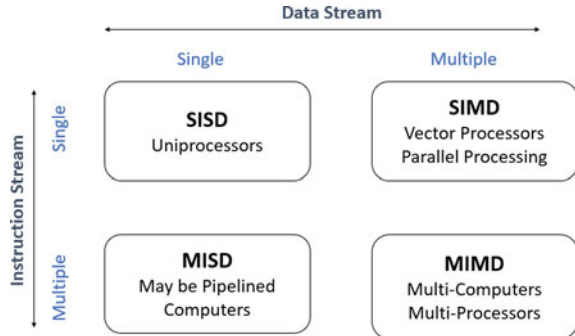
Flynn's Classification is a categorization of parallel computer architectures based on the concurrency in processing sequences, data, or instructions from the perspective of an assembly language programmer [6].

M.J. Flynn introduced this classification, which is based on the number of instructions and data items that are altered at the same time.

An instruction stream is a set of instructions read from memory, while a data stream is the output of actions performed on the data in the processor.

“stream” means the flow of instructions or flow of data.

Fig. 2 Schema of Flynn's classification



Parallel processing can occur in either the data stream, the instruction stream, or in both simultaneously (Fig. 2).

- **SISD** means Single Instruction and Single Data
- **SIMD** means Single Instruction and Multiple Data
- **MISD** means Multiple Instruction and Single Data
- **MIMD** means Multiple Instruction and Multiple Data.

Architecture SISD

Sequential architecture with a single instruction stream (SI) and a single data stream (SD). This is the model of sequential architecture. It does not exploit any parallelism [4, 6].

Architecture SIMD

Architecture composed of many processing units that are overseen by a single control unit [4, 6].

- Each processing unit performs the same instruction (SI) on distinct data. All processing units simultaneously execute identical instructions, consequently the synchronization of the processors.
- The memory shared between the different processors can be subdivided into several modules.
- The processors access the memory via an interconnection network.

This type of architecture is suitable for regular processing, such as matrix calculation on dense matrices or image processing (Fig. 3).

Architecture MISD

A single data stream (SD) feeds multiple processing units (MI). The same data is processed by several processors in parallel. This type of architecture is rarely implemented.

Each processing unit works on the data in its own way, using its own instruction stream, it is primarily of theoretical importance [7] (Fig. 4).

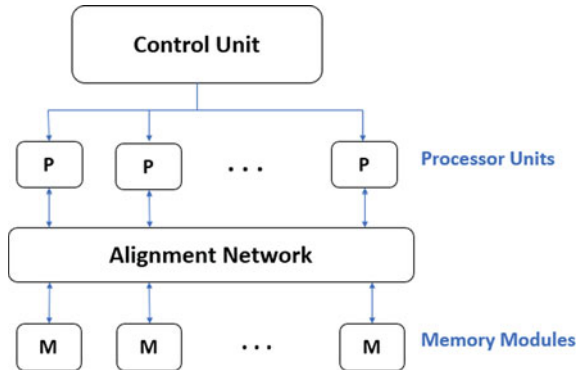


Fig. 3 SIMD architecture

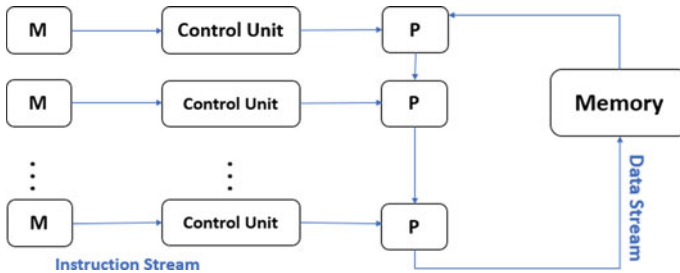


Fig. 4 MISD architecture

6.2 Architecture MIMD

Multiple Instruction (MI) Multiple Data Stream (MD), it is composed of several processors that operate asynchronously:

- Each processing unit can handle a different instruction flow.
- Each unit can operate on a different data stream.

The execution can be synchronous or asynchronous.

Two variants of the MIMD model:

- Shared Memory MIMD
- Distributed Memory MIMD.

We can combine the two variants to have a MIMD architecture hybrid shared-distributed memory architecture [4, 6].

Shared Memory MIMD

The general design of shared memory architectures is based on the following principle [4, 6]:

- All processors share a global memory space (a global memory) visible to all processors.
- Each processor works independently of the others. The exchanges between processors are done via the global memory by read/write operations.
- The modifications made by a processor in global memory are visible to all the others (Fig. 5).

Distributed Memory MIMD

The general design of distributed memory architectures is based on the following principles [4, 6]:

- For each processor, there is a local memory, and there is no global memory accessible to all processors.
- The processors work independently. Any modification of a local memory has no effect on the memories of the other processors.
- The processors are linked between them by a network to communicate between the memories of the different processors. These links are diverse with very variable levels of performance levels (Fig. 6).

Hybrid Memory Architecture

- The most powerful computers in the world today are a mix of shared and distributed memory.
- The basic brick (node) is a multiprocessor with shared memory.
- These bricks are interconnected by a network (ethernet type, myrinet, Infiniband, ...)

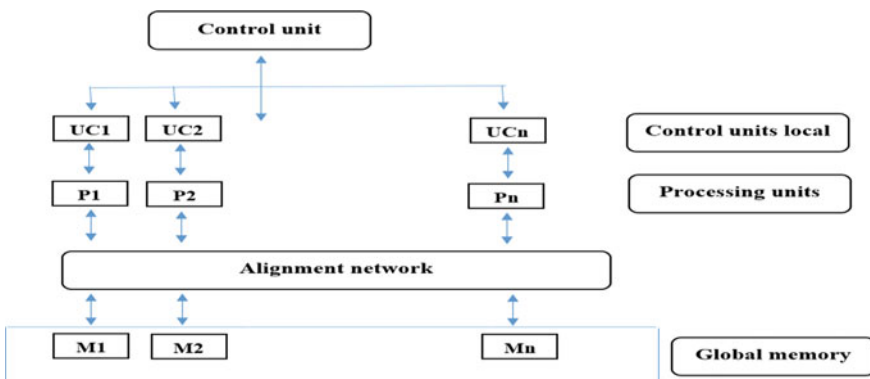


Fig. 5 Shared memory MIMD architecture

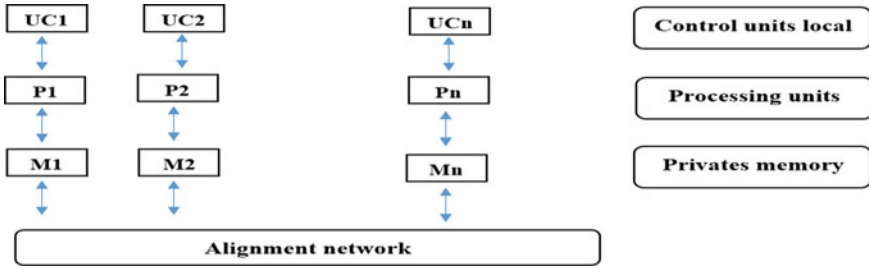


Fig. 6 Distributed memory MIMD architecture

7 Performance Analysis

For the performance analysis of parallel algorithms it is necessary to have a model and a framework to present and analyze the algorithms.

There are several models and the situation is more complex. The performance depends on factors such as:

- concurrency
- processor allocation and scheduling
- communication
- synchronization.

7.1 The Acceleration Factor (Speedup)

Allows to measure the gain brought by using p processors.

- Let Q be a problem and n the size of the input data
- $T^*(n)$ is the time taken by the most efficient sequential algorithm to solve the given problem Q [2].
- $T_p(n)$ is the parallel algorithm's time taken for solving problem Q with p processors.
- The speedup is defined by:

$$S_p(n) = \frac{T^*(n)}{T_p(n)}$$

- In the ideal case we have:

$$S_p(n) = p$$

i.e. The Algorithm is Purely Parallel.

= The parallel algorithm runs p times faster than the sequential algorithm:

$$T^*(n) = pT_p(n)$$

- In reality we have ($1 \leq S_p(n) \leq p$). This is due to several factors:
- Not enough parallelism (parts of the problem are executed sequentially).
- Overhead (extra time) communication time between processors.
- Synchronization and control problems.
- Problems that the processors do not execute the same workload.

Note:

- The calculation of the speedup involves using the time taken by the best sequential algorithm ($T^*(n)$) rather than the time taken by the parallel algorithm when running on a single processor ($T_1(n)$).

7.2 Efficiency

Measures the effective utilization rate of the p processors (measures the percentage of the p processors utilization).

$$E_p(n) = \frac{T^*(n)}{pT_p(n)} = \frac{S_p(n)}{p}$$

- In the ideal case, we have: $E_p(n) = 1$, indicates that $T_p(n)$ runs p times faster than the sequential algorithm, which means that all p processors have been used during the whole calculation.
- In reality, we have: $E_p(n) \leq 1$.

8 Experimental Study of Parallel Algorithms

8.1 Paralleled K-Nearest Neighbors Algorithm

In a first study of KNN algorithms, a comparison was made between the sequential version and the parallel version using the parallel random-access machine (PRAM) with CUDA to obtain their performance in terms of acceleration [8] (Fig. 7).

Execution time versus sample size: Based on the experiment, the execution time for both the sequential and CUDA versions were nearly identical for small sample sizes. But with the increase in the number of execution samples, the difference between sequential and CPU execution became more relevant (Fig. 8).

This demonstrates the effectiveness of using a parallel approach to the KNN algorithm with large dataset.

Fig. 7 Execution time as a function of sample size [8]

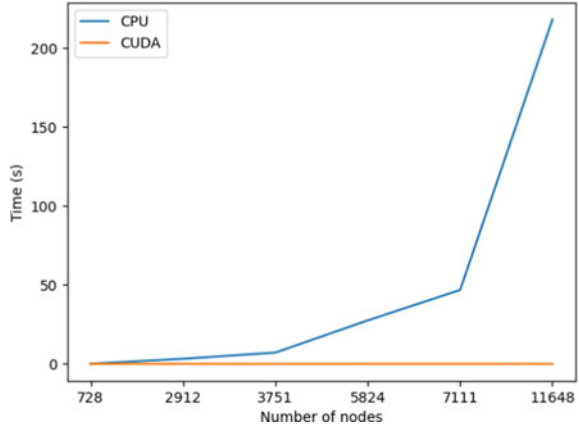
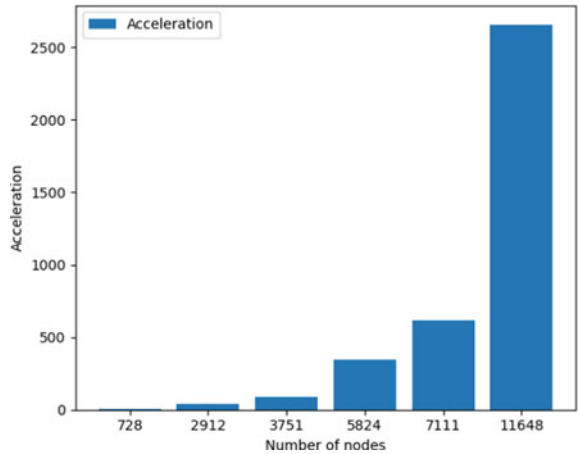


Fig. 8 Acceleration of CUDA over CPU as a function of the sample size [8]



8.2 Paralleled Top_{kws} Algorithm

In this study of Top_{kws} Algorithm, a comparison was made between the sequential version of Threshold algorithm (TA), No random-access algorithm (NRA), Top_{kws} Algorithm and its parallel version using the open-source Hadoop and Spark framework ($SPTop_{kws}$) [9] (Fig. 9).

As approved in the previous study, this work generalizes the effectiveness of parallelism over sequential techniques by extending the study and examining more algorithms.

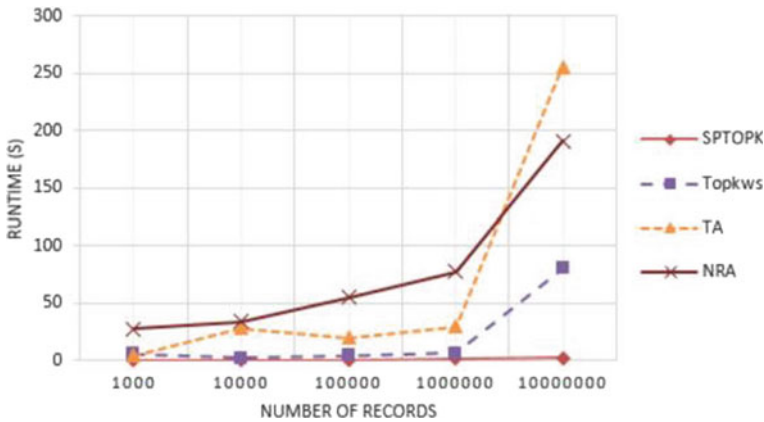


Fig. 9 The variation's effect of the number of records on the algorithm's execution time [9]

9 Conclusion

In this chapter, we have dealt with parallelism by exploring its principle, its many architectures and its advantages over sequential architecture. In future research, we will apply parallel computing to sequential algorithms to study the effectiveness of Artificial Intelligence and Data Processing when working with Parallelism. Concretely, we plan to proceed with the parallelization of the algorithms published in [10–20] to apply them to certain distributed domains.

References

1. D. Etiemble, F. Cappello, Introduction au parallélisme et aux architectures parallèles. Techniques de l'Ingénieur (2017).
2. Daoudi EL Mostafa, T. Gautier, A. Kerfali, R. Revire, J.L. Roch. Algorithmes parallèles à grain adaptatif et applications. TSI. Technique et science Informatiques **24**(5), 505–524 (2005).
3. T.-D. Diep, P.H. Ha, K. Furlinger, A general approach for supporting nonblocking data structures on distributed memory systems. J. Parallel Dist. Comput. **173** (2023).
4. l'informatique parallèle, <https://le-site-d-isn.webnode.fr/>
5. glossaire-definition/Parallélisme-informatique.html, <https://www.techno-science.net/>
6. flynns-classificationnotes classification parallel, <https://byjus.com/gate/>
7. misd-notes, <https://byjus.com/gate/>
8. O. El Atyqy, H. Filali, H. Ba-mohammed, Paralleled K-nearest neighbors algorithm
9. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. IEEE Syst. J. **15**(4), 4876–4886 (2020). <https://doi.org/10.1109/JSYST.2020.3019368>
10. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. Int. J. Internet Technol. Secured Trans. **10**, (2020)
11. A. Idrissi, K Elhandri, H Rehioui and M Abourezq, Top-k and Skyline for cloud services research and selection system. Proc. Int. Conf. Big Data Adv. Wirel. Technol. (2016)

12. A. Idrissi, C.M. Li, J.F. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks. in 18th IEEE International Conference on Tools with Artificial Intelligence, Washington, USA (2006)
13. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. arXiv preprint arXiv: 1709.10372 (2015)
14. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data Min.* (2017)
15. M. Abourezq, A. Idrissi, Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comput. Sci. Appl.* **6**(6), (2015)
16. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Architect.* **9**(2–3), 136–148 (2020)
17. M. Essadqi, A. Idrissi, A. Amarir, An effective oriented genetic algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**, 170–179 (2018)
18. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Elect. Eng.* **73**, 289–303 (2018)
19. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Robot. Intell. Syst.* **14**(3), 65–70 (2020)
20. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment. *Proc. Int. Conf. Big Data Adv. Wirel. Technol.* (2016)

A Bi-LSTM Neural Network to Forecast Stock Market Index



Zakaria Al Bakkari , Ikram El Azami , and Adil El Makrani 

Abstract Rapid advances in machine learning and artificial intelligence technologies, the power of GPUs and TPUs, the availability of rich data, and increased computing power in machines have opened the door to the development of sophisticated methods for predicting stock prices, stock price forecasting has been the subject of much research, but so far, scientists have not found an acceptable solution. Recent deep learning research has prompted researchers to practice neural networks to predict stock market indices. In this study, we investigate a bidirectional long-short-term memory (bi-LSTM) neural network model for forecasting the future value of the S&P 500 Index. The suggested model can predict the next day's index based on the precedent value of the index. Results show that our model exceeds several benchmarks with above-average accuracy compared to other statistical and machine learning models.

Keywords Deep learning · Bi-LSTM · Recurrent neural network for finance · S&P 500 index

1 Introduction

Stock market forecasting is a classic problem that interests financiers and computer scientists. In response to this problem, Fama et al. suggested the famous Efficient Market Hypothesis (EMH). Fama [1], where the prices of financial stocks always

Z. Al Bakkari (✉)

Laboratory of Research in Informatics FS, Ibn Tofail University, Kenitra, Morocco
e-mail: zakaria.albakkari@uit.ac.ma

I. El Azami · A. El Makrani

Department of Informatics, Laboratory of Research in Informatics FS, Ibn Tofail University, Kenitra, Morocco
e-mail: akram_elazami@yahoo.fr

A. El Makrani

e-mail: adil.elmakrani@uit.ac.ma

reflect all relevant available information. Supported by network big data collection and graphics parallel processing. New advances in processing units (GPUs) and computing technologies have made the task of stock market prediction a task for artificial intelligence researchers, especially deep learning models, which will generate great interest in this topic in the future.

Striving to start with the story of fundamental analysis and technical analysis: first, the technical analysis relies only based on charts and trends. Second, Fundamental analysis evaluates stock prices based on their prices. After introducing linear models as stock market forecasting solutions, including (ARIMA) autoregressive integral moving average (Hyndman and Athanasopoulos 2018), machine learning Development of models such as B. Logistic Regression and [2] Support Vector Regression SVR (Cao and Tay 2003), Adaboost (Yoav Freundt Robert Schapire 1997) and Extreme Learning Machines [3]. Today, with a large amount of historical financial data and model nonlinearity and extremely complex functions. Deep learning is represented by different forms of deep neural networks that come as an emerging solution to detect the various nonlinear and complex relationships in the historical stock price and also the short-term and long-term dependencies on future stock prices. financial series have a nonlinear relationship that deep learning can entirely extract the complex features of influencing factors. The progress of technology permits Researchers to access a greater amount of information promptly and perform very complex calculations thanks to the parallel processing capability of graphics processing units (GPUs).

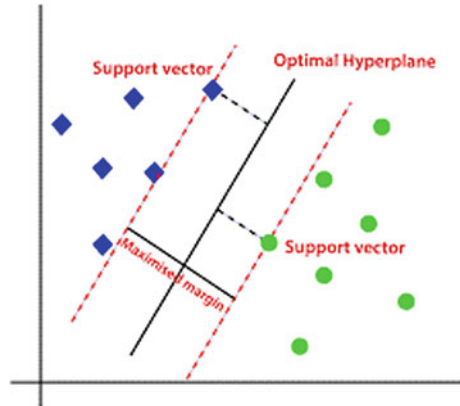
2 Related Work

This section define the recent models to predict the stock market index, we can divide these techniques into three categories: statistical techniques, machine learning techniques, and deep learning techniques.

2.1 *Statistical Techniques*

Conventional statistical methods, comprising linear regression, exponential smoothing, and moving average, are used in stock price forecasting. But they weren't effective because these models principally suppose a linear relationship in the data. So, the main flaw of statistical techniques is to assume the linearity of data and ignore stock market stochastics. The most popular stock market predicting method is the Autoregressive Integrated Moving Average (ARIMA): The model can identify various standard temporal structures from time series data. As a financial forecasting and analysis tool, ARIMA has been widely used in financial analysis.

Fig. 1 A basic structure of SVM



2.2 Machine Learning Techniques

There are a lot of methods of ML such as an Extreme learning machine, Random Forest, and Hybrid model. However, the most popular machines learning technique used in finance is the support vector machine:

Support Vector Machine:

Support Vector Machine (SVM) [4] is a supervised machine learning technique for filtering hyperplanes from a amount of data points, and its main property is that the definition of model parameters corresponds to a convex optimization problem. The global optimal solution is therefore any local solution (Fig. 1) by A. Saini (<https://www.analyticsvidhya.com/blog/2021/10/support-vector-machin-essvm-a-complete-guide-for-beginners>).

Its equation is given by:

$$f(a) = \psi^T g(a) + \beta \tag{1}$$

While:

ψ^T : means a T matrix rotation parameters

$g(a)$: means a space transformation for a fixed feature

β : denote a bias

when the variable has continuous values, the name becomes Support Vector Regression (SVR), as the name proposed the SVR is a regression algorithm.

2.3 Deep Learning Techniques

We have a deep neural network as a deep learning model to forecast the stock market:

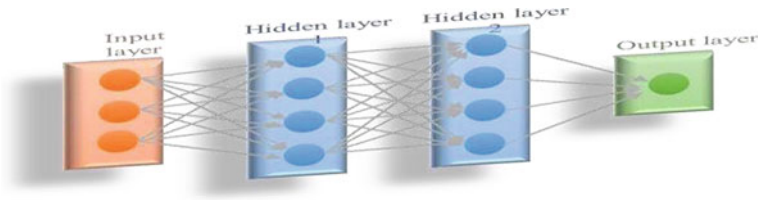


Fig. 2 A basic composition of the deep neural network

Deep Neural Network:

A deep neural network (DNN) is a type of network that consists of a complex composition of two or more hidden layers that help it process complex views. The deep neural network composition in (Fig. 2).

3 Methodology

3.1 Dataset

No external data download is needed, we use the yfinance library with the ticker “^GSPC” of the S&P 500 index, The baseline model only uses OHLCV data. The length of the data of the series studied is from 1st January 2012 to 2 June 2022. The daily data are used to forecast the price of the stock market index.

We generally use Min Max scaling in stock prediction but the problem is that prices are always increasing, in the other words there is no maximum value, we can’t use the maximum value in the train set because this is not the true maximum value. Standardization of stock returns in data preprocessing is more suitable for the prediction stock market [5]. The distribution of my data changes from train to test, the solution to this problem is also turning the price of the stock market index into a stationary series for future predictive modeling [6]. We split my sample dataset into a training dataset which represents 80% of the sample and a test data set which represents 20% of the sample (Table 1). Google Colab was used as a computational instrument for finance. The python programming language was used in the DL algorithms in this study.

Table 1 Division of data

Trading period	Days	Beginning	Ending
All dataset	2621	03/01/2012	01/06/2022
Training dataset	1965	03/01/2012	24/10/2019
Test dataset	656	26/10/2019	01/06/2022

Table 2 Explanatory variables for bi-LSTM model

Number	Variables
1	Opening
2	Closing
3	High
4	Low
5	Volume

We choose the Optimizer Adam boost and the batch size is 64, four hundred as the number of EPOCHS, the number of neurons is equal to 256, the dropout rate is equal to 0.4.

Multiple variables can be in used as time series of SP&500 [7], here is the definition of these variables (Table 2):

The **opening** price is the price at which the SP&500 index trades for the first time at the opening of the day.

The **closing** price generally refers to the last price at which a stock market index was traded during normal trading hours.

A **high** price is the highest price of SP&500 index trades during a trading day, usually greater than or equal to the closing price.

The lowest price is the lowest price at which the SP&500 index trades during the trading day, usually lower than the closing price or equal to the opening price.

Trading **volume** is simply the number of shares traded for a given stock, index or other assets during a given period time.

In this study, we focus on the closing price of the SP&500 stock market index as the time series variable.

3.2 Our Model

In this paper, we suggest a bi-LSTM model forecast the stock market index.

Before treating the model architecture, let’s show the development of these kinds of models:

Recurrent Neural Network (RNN) Model

Conventional artificial neural networks cannot achieve the function of inference using information about subsequent previous events using previous inference information about subsequent events. The RNN algorithm is very useful in dealing with time series data. RNNs have a recurrent network of closed circles that permit information to be maintained [8]. These closed circles permit us to pass information from the current node to the next node, however, the forward and backward transmission of this information back and forth has long-term dependencies. Iterative RNNs use previous information. In a conventional feedback neural network, the information is propagated from the input layer to the output layer passing through to the hidden

layer, where they are fully connected, however, the nodes are connected many times. Such As networks that are inconceivable with time series. But in an RNN, nodes of the hidden layer are connected at different times and the previous state is a function of the current output. The network remembers the precedent information and transforms it into the result of the current output. RNNs are neural networks divided by weights in the time dimension. The output of the model is a function of the state of the hidden layer ht at time t . it depends on the input xt at time t and the state of the hidden layer $ht - 1$ at time $t - 1$. Until the hidden state time is a function of the RNN output model depends on the information of entry at the precedent instant.

Model's output:

$$y_t = \sigma_0(P_t y_t + b_0) \quad (2)$$

Model's hidden state:

$$h_t = \sigma_0(P_x X_t + P_h h_{t-1} + b_h) \quad (3)$$

it analyzes the relationship between the normal time series and the RNN model. Considering univariate time series, let $h = c$, $P_x = a$, $y_t = ht$, and σh an identity transformation, then the RNN model becomes as:

$$y_t = ax_t + cy_{t-1} + b \quad (4)$$

This brings exogenous variables in the model as a first-order autoregression model. So, RNN treats a nonlinear complex time series. For the optimization of the below RNN model parameters, we employ the backpropagation through time algorithm to calculate the gradient and to expand it in time series. Until the hidden state ht of RNN is bound by the hidden state $ht-1$ of the previous moment, the RNN gradient is a function of time.

The current moment is t , the hidden state vector at the precedent moment is $ht-1$, the parameters are P_x and P_h , and a bias item is b . Consider only one memory location, which is a number in the meantime 0 and 1. If it = 1, all information is entered into the cell and the gate is open; if it = 0, no information is entered into the cell and the gate is closed. LSTM's forget gate:

$$f_t = \sigma(P_{xf} x_t + P_{hf} h_{t-1} + b_f) \quad (5)$$

The RNN gradient is shown in Fig. 3.

However, with the methods described above, parameter learning suffers from the problem of vanishing and exploding gradients. This makes it difficult for RNNs to model long-term temporal dependencies.

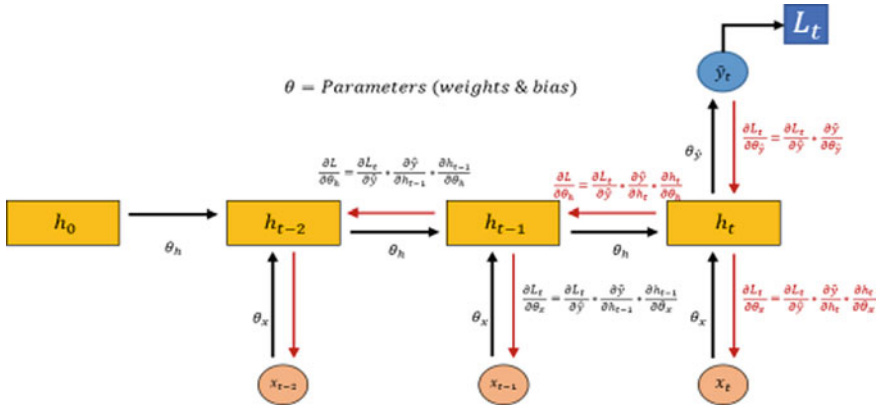


Fig. 3 The RNN gradient parameters

Long Short-Term Memory (LSTM)

The LSTM finds a solution to the vanishing and exploding gradient problems of simple RNNs and has achieved success in many domains of machine learning [9]. The starting point to the model LSTM is the entry of a gated cell system that retains past information about the cell state as shown in Fig. 4. It learns dynamically by using different gates when it needs to be updated with new information and when the past information is forgotten by the network. LSTMs use gates to sort some information. At time t.

The storage unit stores the past information up to the present time, which is processed by principal gates: the forget gate defines that the internal storage unit must store the previous time information, and the input gate defines the new information input to the internal storage unit; the output gate definite the ²output information of the internal storage unit.

The entry gate of LSTM:

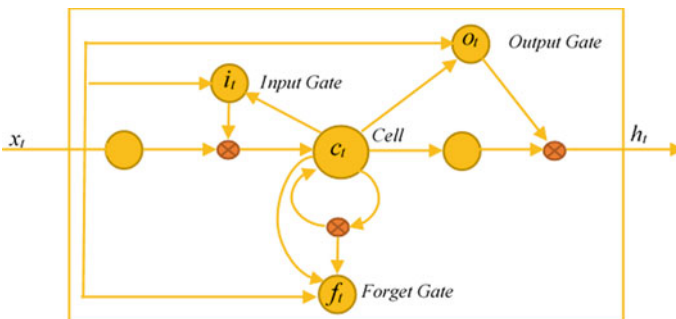


Fig. 4 The architecture of LSTM

$$i_t = \sigma(P_{xi}X_t + P_{hi}h_{t-1} + b_i) \quad (6)$$

The activation function is σ and the input X_t .

The type of activation function is sigmoid.

f_t is a number in the meantime 0 and 1. If $f_t = 1$, the state of the cell is completely carried over to the precedent cell, and the gate is opened; if $f_t = 0$, the precedent cell state is discarded and the gate is closed.

LSTM's output gate:

$$O_t = \sigma(P_{xo}X_t + P_{ho}h_{t-1} + b_o) \quad (7)$$

o_t is a number in the meantime 0 and 1. When $o_t = 1$, the state of the cell can be output and the door is open; when $o_t = 0$, the state of the cell cannot be output and the door is closed.

The update of the state of the internal memory unit is given by the formula:

$$C_t = f_t c_{t-1} + i_t \tanh(P_{xc}X_t + P_{hc}h_{t-1} + b_c) \quad (8)$$

First, the information after the last cell state is monitored by the forget gate, and finally, the information after the input information is controlled by the input gate.

LSTM's output is:

$$h_t = O_t \tanh(c_t) \quad (9)$$

Bidirectional RNNs: biRNN

A bidirectional RNN (biRNN) concatenates two hidden layers in opposite senses with the same output. In this form of deep learning, the output layer has information about future (forward) and past (backward) states. Invented by Schuster and Paliwal in 1997, [11] BRNN is used to rise the size of input information available on the network.

The basis of the BRNN is to divide the neurons of an RNN into two opposite senses, first, for the first-time sense (forward state). Second, the other is for the opposite time sense (backward state). These two-state outputs Are connected to the opposite sense state inputs. The general structure of BRNN is represented in Fig. 5. we use input information from the future and past of the current time setting can be used, such as standard RNN which requires delays for comprising future information and two-time directions [11].

BiRNNs are trained using an algorithm similar to RNNs in that the two-direction neurons do not interact. However, when practicing time-based backpropagation, an additional process is essential because the updates of the input and output layers are not performed simultaneously. For the reverse information flow, the general process of training is as follows, the information of the output neuron flows first, then the forward state and the reverse state flow, and then the output neuron transmits. After performing the forward and backward passes, update the weights [11].

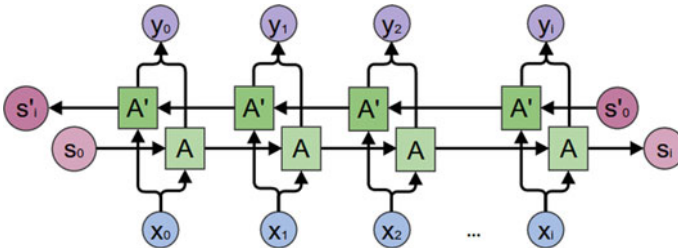


Fig. 5 A basic structure of bidirectional RNNs [10]

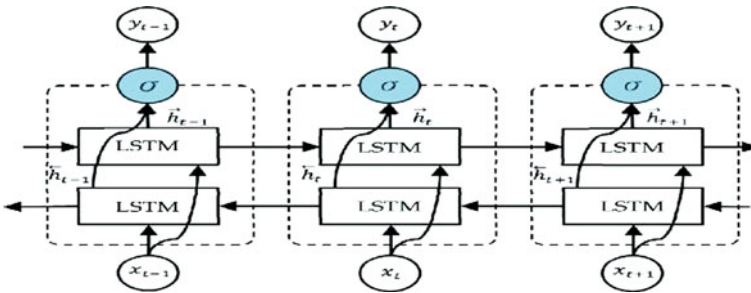


Fig. 6 A basic structure of Bidirectional LSTM [13]

Bidirectional LSTM: biLSTM

A deep bidirectional LSTM [12] is a type of LSTM model described where two LSTMs are applied to the input data (Fig. 6). The LSTM is first applied to the input sequence (i.e. the feed-forward layer). In the second round, the inverse of the input sequence is included in the LSTM model (i.e. the reverse layer). Practicing LSTM twice improves long-term dependency learning and thus improves model accuracy [13].

We choose variable x_t as the historical sp500 closing price and the model y_t as the predicted closing price.

Some studies focus only on price movements [14], in this study, we focus on predicting the price of the SP&500 index from historical prices of the SP&500 index.

4 Experimental Results

4.1 Model Performance

We choose as a rule of evaluation three performance metrics: mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE).

The formula is:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_{actual} - y_{predicted})^2}{n}} \quad (10)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{actual} - y_{predicted}| \quad (11)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_{actual} - y_{predicted}|}{y_{actual}} \quad (12)$$

RMSE is used to reveal forecast error.

MAE is used to measure bias.

MAPE is used to judge the accuracy of predictions.

We have mentioned that the smaller value in the above metric the better performance reflects it.

4.2 Model Evaluation

The model has a smaller performance metric therefore the model has better accuracy compared to other traditional models.

The result of the bi-LSTM model is presented in (Fig. 7).

The bi-LSTM is the best model which has a smaller performance metric compared to other models as shown in Table 3.

4.3 Model Comparison

For example, in Table 3, SVR reached $1.74e + 02$, $4.27e + 02$ and $7.37e + 00$ in MAE, RMSE, and MAPE, respectively. these values are the greatest among all the methods. Second, we explain the conventional Deep Learning models including DNN Compared with Machine Learning, DL models reached better performance (in terms of smaller values of MAE, RMSE and MAPE). the learning massive SP&500

Fig. 7 Bi-LSTM model result from SP&500

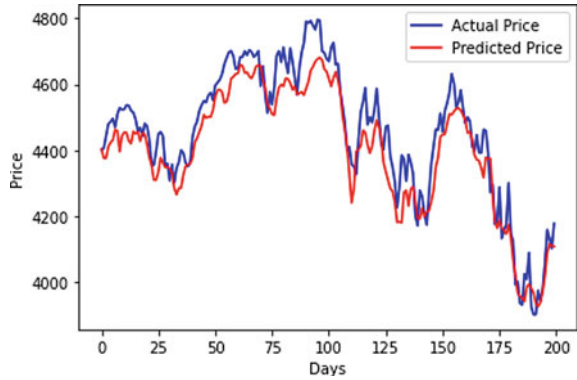


Table 3 Performance metrics comparison with basic models

Model	RMSE	MAE	MAPE
SVM	4.273739e + 02	1.749045e + 01	7.370630e + 00
ARIMA	1.777185e + 02	1.210661e + 01	3.401039e + 00
DNN	1.207928e + 02	9.664789e + 00	2.192550e + 00
Bi-LSTM	1.142861e + 02	9.653662e + 00	2.153341e + 00

closing price data show the power of DL models in generalization. Finally, bi-LSTM reached the smaller performance metric with an RMSE equal to $1.14e + 02$, MAE equal $9.65e + 00$ and MAPE equal $2.15e + 00$. Figures 8, 9 and 10 shows the result of other models.

Fig. 8 ARIMA model

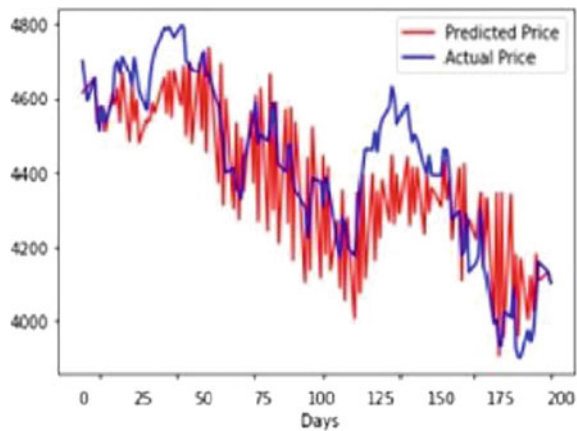


Fig. 9 DNN model

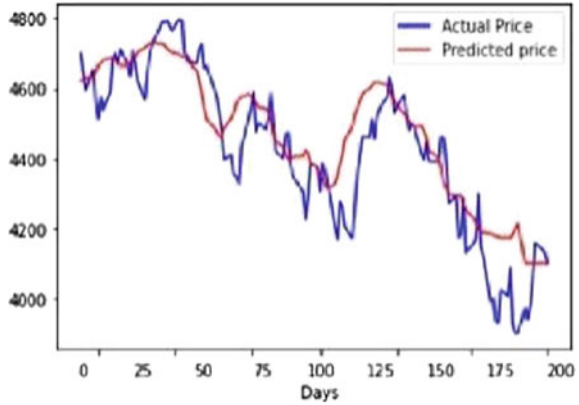
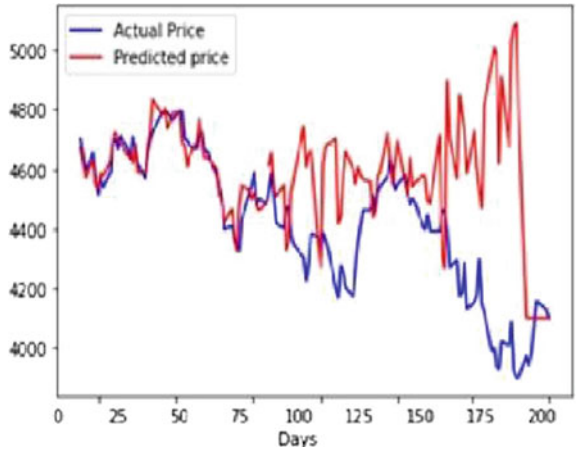


Fig. 10 SVR model



4.4 Model Parameters

The hyperparameters will impact the accuracy of the predictor and if we only have a small dataset, it's possible that playing around with different hyperparameters will give us an "Overpowered neural network" will perform just fine in the set available but after.

We have more hyperparameters in this study: nb layers and neurons. We can modify them until a network has better accuracy, without falling into overfitting, that's why we need more and more data, we need different sets of tests for each architecture to avoid overfitted models.

$N_STEPS = 70$: Window size or the sequence length, $LOOKUP_STEP = 1$ is the next day.

We get the last sequence by appending the last nstep sequence with lookupstep sequence, for instance, if $nsteps = 60$ and $lookupstep = 20$, lastsequence should be

80 lengths and this last sequence will be used to predict future stock market index prices not available in the dataset.

lookup step is by default = 1, the next day.

TEST_SIZE = 0.2: that means the test ratio is 20%.

N_LAYERS = 3: numbers of layers.

UNITS = 256: numbers of neurons of the Bi- LSTM neural network.

The activation function: linear.

Dropout rate = 0.4: Dropout rate is used with most types of neural networks. This is a great tool to minimize overfitting in a model. This is much better than the available regularization methods and can also be combined with max- normalization which provides a significant boost over just using dropout.

In the future, we will improve our results by combining Convolutional neural network (CNN) and LSTM for more accuracy [15, 16] and we may also choose another training ratio and an approach for dynamic optimization of parameters during the backtesting process by using training–testing rolling window [17]. we can also mix these models in a hybrid model to improve accuracy [18].

5 Conclusions

The domain of SP&500 index prediction using bi-LSTM interests many researchers and traders and improving prediction accuracy will give more profits. This paper will treat a stock market index SP&500 prediction structure to help traders in the trading stock market. The suggested bi-LSTM achieves a better accuracy which exceeds the performance of statistical and machine learning models such as SVM, and DNN. Finally, we will examine to improve our suggested model by adjusting different numbers of bi-LSTM Layers and increasing data size and time horizon.

References

1. E.F. Fama, Random walks in stock market prices. *Financ. Anal. J.* **51**(1995), 75–80 (1995)
2. L. Cao, F. Tay, Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Netw.* (2003)
3. S. Kumar Chandar, M. Sumathi, S.N. Sivanadam, Forecasting gold prices based on extreme learning machine. *Int. J. Comput. Commun. Control (IJCCC)*, 372–380 (2016)
4. Y. Chen, Y. Hao, A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Syst. Appl.* **80**, 340–355 (2017)
5. C. Montenegro, R. Armas, Augmented data deep learning model to prediction of S&P500 index: a case study including data of COVID-19 period. *Inf. Technol. Syst.* **414**,175–184 (2022)
6. W. Ahmed, M. Bahador, The accuracy of the LSTM model for predicting the S&P 500 index and the difference between prediction and backtesting (2018)
7. F. Wang, Predicting S&P 500 market price by deep neural network and ensemble model. *E3S Web of Conferences* 214, 02040 (2020)

8. Gu, N. Lu, L. Liu, A novel recurrent neural network algorithm with long short-term memory model for futures trading, *J. Intell. Fuzzy Syst.* **37**(4), 4477–4484 (2019)
9. S. Hochreiter, J. Schmidhuber, Long short-term memory. *2020 Neural Comput* **9**(8), 1735–1780 (1997)
10. G. Miao, G. Shi, et S. Li, « Online Prediction of Ship Behavior with Automatic Identification System Sensor Data Using Bidirectional Long Short-Term Memory Recurrent Neural Network ». *Sensors* **18**, n° 12 (30 November 2018): 4211
11. M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997)
12. S. Siami-Namini, N. Tavakoli, A. Siami Namin, The performance of LSTM and BiLSTM in forecasting time series. *IEEE International Conference on Big Data: Proceedings*, Dec 9–Dec 12, (Los Angeles, CA, USA, 2019)
13. Y-H. Li, L.N. Harfiya, K. Purwandari, Y-D. Lin, Real-time cuffless continuous blood pressure estimation using deep learning model. *Sensors*, **20**(19) (2020)
14. D.M.Q. Nelson, A.C.M. Pereira, R.A. de Oliveira, Stock market's price movement prediction with LSTM neural networks, *International Joint Conference on Neural Networks (IJCNN)* (2017)
15. S. Selvin, R. Vijayakumar, E.A. Gopalakrishnan, V.K. Menon, K.P. Soman, Stock price prediction using LSTM, RNN, and CNN- sliding window model. In: *Proceedings IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 1643–1647
16. I.E. Livieris, E. Pintelas, P. Pintelas, A CNN– LSTM model for gold price time-series forecasting. In *Proceedings of the International Symposium on Emerging Technologies for Education* (Springer, Cham, Switzerland, 2017), pp. 548–556
17. R. Mateuszkijewski, Predicting the price of S&P500 index using classical methods AND recurrent neural network'. *Working papers.* 27/2020 (333)
18. M. Asiful Hossain, K. Rezaul, R. Thulasiram, N.D.B. Bruce, Y. Wang', Hybrid deep learning model for stock price prediction, in 2018, *IEEE Symposium Series on Computational Intelligence (SSCI)*

Artificial Intelligence and Advanced Technologies

Fast Yolo V7 Based CNN for Video Streaming Sea Ship Recognition and Sea Surveillance



Abdelilah Haijoub, Anas Hatim, Mounir Arioua, Slama Hammia, Ahmed Eloualkadi, and Antonio Guerrero-González

Abstract The object detection in the maritime field is becoming more important thanks to the different offers it provides, such as: monitoring, traffic management, coastal control, etc. However, the maritime environment is known for its complexity, which poses difficulties in detecting targets, especially with the classical method. Hence the importance of introducing an intelligence layer to the acquisition data that will support the decision support part and facilitate the detection of an undesirable event. This paper aims to make video surveillance in the maritime domain intelligent through the use of the advantages of machine learning, and more specifically the implementation of the YOLOv7 model that will allow us to provide real-time detection, rigorous precision of the different types of vessels, plus the speed of processing of frames in the configuration used. The experimental results prove that the model YOLOv7 as the latest version of yolo is the model that gives the most efficient results when compared with the other existing models.

A. Haijoub (✉) · M. Arioua · A. Eloualkadi

Laboratory of Information and Communication Technologies (LabTIC), National School of Applied Sciences of Tangier, Abdelmalek Essaadi University, Tangier, Morocco
e-mail: Haijoub.Abdel@gmail.com

M. Arioua

e-mail: M.Arioua@uae.ac.ma

A. Eloualkadi

e-mail: a.Eloualkadi@uae.ac.ma

A. Hatim · S. Hammia

Laboratory of Information Technology and Modeling (TIM Team), National School of Applied Sciences of Marrakech, Cadi Ayyad University, Marrakech, Morocco
e-mail: A.Hatim@uca.ma

S. Hammia

e-mail: Hammia.Slama@gmail.com

A. Guerrero-González

Department of Automation, Electrical Engineering and Electronic Technology, Universidad Politécnica de Cartagena, Cartagena, Spain
e-mail: Antonio.Guerrero@upct.es

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,

Studies in Computational Intelligence 1102,

https://doi.org/10.1007/978-3-031-33309-5_8

Keywords YOLOV7 · Ship recognition · Convolutional neural network · Object detection · Machine learning · Video surveillance

1 Introduction

Over the years, the importance of the control of the maritime domain, especially the coasts and the sea, is becoming more and more important, especially in civil and military applications. And given the difficulty and complexity of manual management, which requires a lot of personnel and a great potential on their part, the introduction of artificial intelligence in this field has become a fundamental element. Several applications of detection and recognition in the maritime domain exist, namely: the control of pollutant spills, the monitoring of maritime traffic, the management of fishing and illegal smuggling, having the identity of ships and navigation information, etc. However, and as the sea is characterized by its various constraints (waves, clouds, rain, fog and reflections) that present real challenges to the detection and recognition in the maritime domain, hence the need for the intelligence layer that will improve security, give support to personnel and also evolve the management and monitoring.

Three sources of images that come into play are: spaceborne optical images, synthetic aperture radar and video surveillance. synthetic aperture radar and spaceborne optical images, are known for their large images that require a lot of processing and acquisition, Moreover, they do not support real-time data acquisition. And since the application requires fast and accurate processing, we opted for a solution based on video surveillance.

Vessel detection is influenced by several factors, such as land constructions, complex waves, light on the water surface, fog, etc. In order to filter the interference and noise, several approaches are challenged, namely: the image processing method, combining machine learning algorithms with image processing, etc. However, the combination method generates results that are often not practical on all environments where the processing latency is high.

To eliminate these interferences, several researchers have worked on this topic. Shao et al. [1] proposed a YOLOv2 and saliency detection-based technique; it predicts the ships via a CNN, then it uses detection of the global contrast region to adjust the localization, to determine the sea level, they further used color segmentation and edge detection. with regards to the ship to eliminate the interference related to the constructions. Zou and Shi [2] are working on an SVDnet algorithm that is robust to the severe climate of the sea area. Shi et al. [3] use coarse-to-fine method which is an edge detection. Tang et al. [4] propose the DNN and ELM method which is Deep Neural Network-based and Extreme Learning Machine for fast ship detection. Kim et al. [5] exploits the faster R-CNN using SeaShip dataset, it increased the detection rate by recovering the missed ships using the IoU of the bounding box between consecutive images. Wawrzyniak et al. [6] proposed the Moving Vessel Detection Algorithm (MVDA) which detects all types of moving vessels, it also assigns a

unique identifier to each vessel passing the camera. Li et al. [7] add a CBAM module to Yolov3 (convolutional block attention module) to minimize external interference. Fu et al. [8] improved the performance of Yolov4 by adding the CBAM layer.

This paper's contribution may be summarized as follows: the detection of ships with Yolo version 7, and the comparison of Yolov7 [9] with older versions of Yolo [10] and the state of the art results on the same application.

The remaining of the paper is structured as follows: Sect. 2 details the architecture of the Yolov7 model. Section 3 describes the dataset used for training the models, then it presents the hardware configuration used, and finally it presents the results of the experiments we performed. Finally, will close with a conclusion.

2 Yolov7 Based Ship Recognition

Object detection algorithms are classified into two main families: single stage algorithms and multistage algorithms. The first type, single stage, also called regression-based, is characterized by its speed and enough precision (YOLO [9], SSD [11], SSD MobileNet [12]). On the other hand, the second type, also called region-based, is known for its high precision and speed (Faster RCNN [13], Mask RCNN [14]) which is less than that of the single stage algorithms.

The speed of detection in video surveillance applications and especially in the maritime domain is a priority. therefore, we paid attention on the first family of single stage algorithms, then we made comparisons between the different versions of the yolo and the researches mentioned in the state of the art as explained in Sect. 3, where we deduced that the YOLOV7 [9] model is more efficient thanks to its speed and accuracy, while taking into account the conditions of the maritime domain.

YOLOV7 is the most recent version of the YOLO architecture, it is characterized by its high speed and high accuracy improved compared to other versions. It contains E-ELAN [9] (Extended Efficient Layer Aggregation Network) and the Model Scaling, which allow it to increase the learning ability and raise their performance. The E-ELAN module uses merge cardinality, expand cardinality and shuffle cardinality to increase the learning ability. Figure 2 illustrates the difference between the proposed E-ELAN and other ELANs.

In general, the different versions of YOLO contain a neck, a head where the outputs obtained and backbone. However, the YOLOV7 model contains two heads instead of one [9], namely lead head and auxiliary head; the lead head is responsible for the classification and the auxiliary head helps in the training of middle layer.

We have implemented the YOLOV7 model on a hardware configuration described in Sect. 3. Figure 1 shows the architecture of the suggested solution; let a graphic card linked to a surveillance camera for the frame acquisition, these frames pass through the YOLOV7 model's different layers in classifying the video surveillance images.

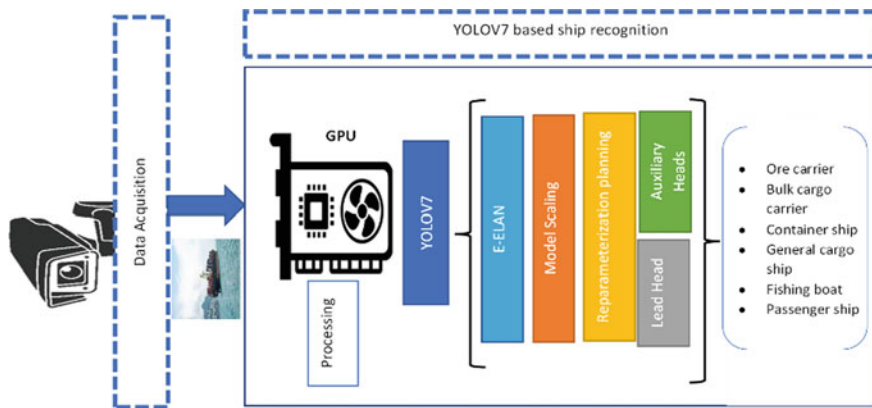


Fig. 1 Flow architecture for vessel detection via video-surveillance

3 Experimental Results

3.1 Experimental Environment

To train the models and test their performance in a hardware configuration, we used a high-performance computer equipped with a Quadro RTX 5000 Nvidia graphics card, 32 GB memory, and an Intel Xeon processor CPU 3.60 GHz \times 8.

This experiment was implemented with ubuntu 18.04 LTS, CUDA 11.7.0 for training acceleration and OpenCV for image processing. We trained YOLOV3 [15], YOLOV4 [16], YOLOV5 [17] and YOLOV7 [9] in two frameworks namely PyTorch and Darknet.

3.2 Dataset

In order to achieve a robust classifier, we need to build a dataset that contains training images of various conditions such as images with random objects, images with changing lighting conditions, images with various backgrounds, images where the region of interest is partially hidden and images with overlap...

These conditions are verified in the Seaship dataset [18], It was created using images taken by cameras of a real deployed video surveillance system. A total of 156 cameras are installed at 50 different points along Hengqin Island's northwestern border, covering coastal regions totaling 53 km².

There are generally three cameras installed, one low-light HD camera and two bolt-type HD (high definition) cameras. The suggested dataset images are processed from the high-quality surveillance video that these cameras provide.

Table 1 Number of each class in SEASHIP dataset used

Classes	Number of images
Ore carrier	2200
Bulk cargo carrier	1953
Container ship	1505
General cargo ship	901
Fishing boat	2190
Passenger ship	474
Total	9223

We used 7000 images extracted from the seaship dataset, this dataset is divided into several classes. The following Table 1 presents the different classes with the number of images of each.

In order to test the models, we used test videos bases on images that do not appear in the dataset with the real conditions of a usual sea ship detection environment (waves, fog and reflections). The tests are done using the described configurations in Sect. 3 (Fig. 2, 3).

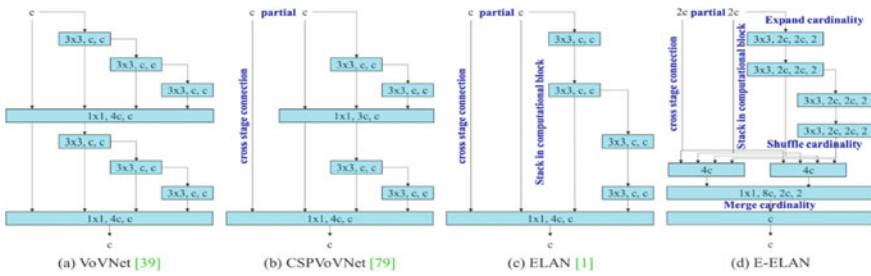


Fig. 2 Extended efficient layer aggregation networks and the proposed extended ELAN

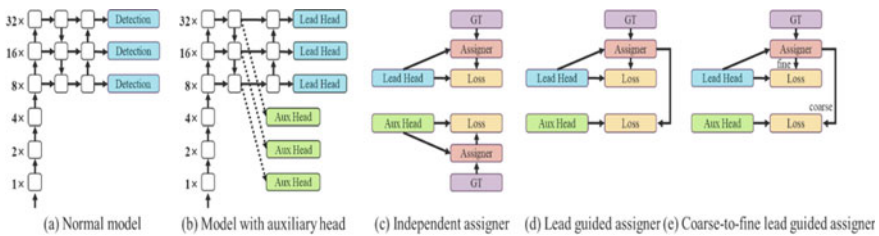


Fig. 3 Model scaling for concatenation-based models

3.3 Ship Detection Performance Indicators

To evaluate the performances of the trained models in a hardware environment, we used the following scores:

Precision: this indicator allows to calculate the number of correctly predicted; it is the ratio of predicted positives (True Positive + False Positive) to the number of correctly predicted positives (True Positive).

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Recall: this indicator allows to calculate the percentage of results that our model correctly predicted; the percentage of correctly predicted results to total predictions (True Positive) divided by the total number of positives (True Positive + False Negative).

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

F1 score: this indicator allows to have a good assessment of our model's performance, this indicator also called the harmonic mean, it calculates the average of the recall and precision percentages.

$$F1 = \frac{2 * recall * precision}{recall + precision} \quad (3)$$

Iou: It calculates the amount that two borders overlap. The amount of overlap between the predicted boundary and the actual boundary of the item is measured using this formula. To identify whether a prediction is a true positive or a false positive, a threshold IoU (such as 0.5) is established in certain works.

$$Iou = \frac{area\ of\ overlap}{area\ of\ union} \quad (4)$$

mAP: the average accuracy scores across all classes are summed, and the result is divided by the total number of classes, to provide this indicator.

$$map = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \quad (5)$$

n = number of classes. AP_k: Average Precision of the class k.

FPS: used to get an idea of the speed of an algorithm according to the test hardware configuration, it shows the number of frames processed in a second.

Table 2 Hyperparameters training

Parameters	Value
Batch size	16
Momentum	0.949
Learning rate	0.0013
Number of iterations	600 epoch /12000 for yolov4
Fixed image size	640 × 640
Weight decay	0.0005

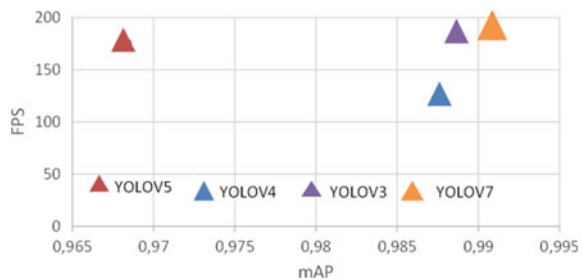
3.4 Results Analysis

In this section, we will discuss the results from training and testing for the various versions of the Yolo, namely: yolov3, yolov4, yolov5 and yolov7. Then compare the different performance indicators to show the efficiency of each one of them. The models were trained under the same conditions using the introduced database. The dimensions of the training images size are 640 *640. In order to evaluate the FPS metric, we used real time videos which contains images that do not appear in the dataset and present the real conditions of our application (waves, fog and reflections). The configurations used for training are presented in the following table (Table 2).

After comparing the different results, we can conclude that all these models succeed to achieve a good detection but with different performances. Based on the indicators we can conclude about the efficiency of each of them. Indeed, the model YOLOV5 achieves the recognition with high precision compared to the others, and YOLOV4 achieves the best Recall score. The YOLOV7 possess the higher FPS and mAP indicators. The Fig. 4 presents the relation between these models in FPS/mAP and shows that YOLOV7 detects objects more accurately and quickly than the other algorithms, which is suitable for real time surveillance. The Figs. 5, 6, 7 and 8 present a comparison of the indicators for each model.

Indeed, Fig. 5 illustrates the comparison of the accuracy indicator, Fig. 6 shows the results of the recall indicator which measures the model’s ability to identify positive samples, Figs. 7 and 8 gives the evolution of the map indicator which measures average accuracy ratings according to the epochs between the versions of yolo 4,5 and 7.

Fig. 4 mAP by FPS ratio between YOLO models



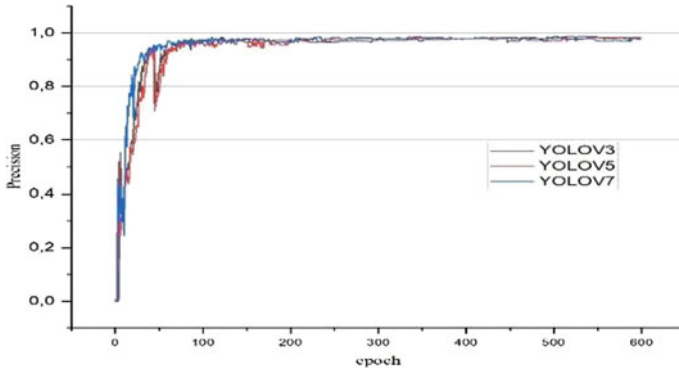


Fig. 5 Comparison of precision between YOLOV5, YOLOV3 and YOLOV7

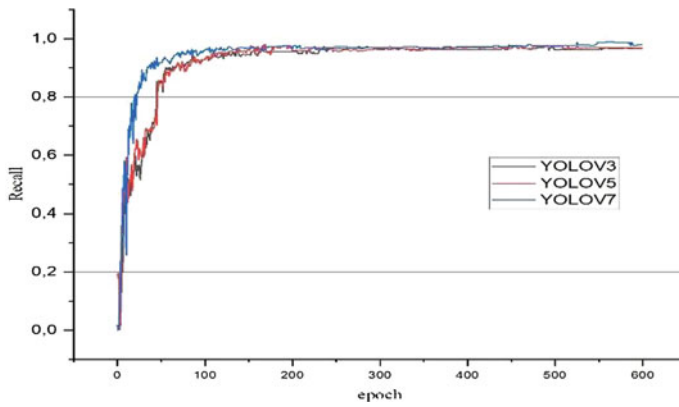


Fig. 6 Comparison of recall between YOLOV5, YOLOV3 and YOLOV7

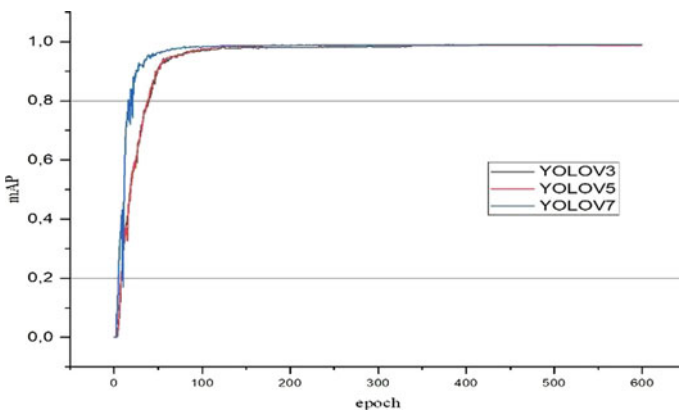
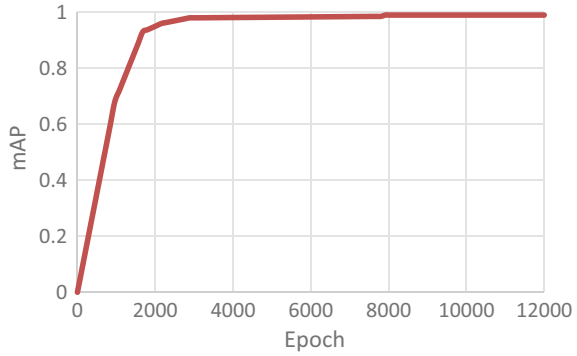


Fig. 7 Comparison of mAP between YOLOV5, YOLOV3 and YOLOV7

Fig. 8 YOLOV4 mAP during training



The Table 3 shows the comparison between the models trained on the seaship dataset based on the performance metrics illustrated in Sect. 3. Figure 4 presents the FPS versus mAP indicators for each model.

The Fig. 9 presents a detection of fishing boat using YOLOV4, YOLOV3, YOLOV5 and YOLOV7.

LI et al. [7] trained the YOLOV3 tiny model with improvements in the backbone layer, and added a CBAM layer to reduce external interference. The model was trained on the same dataset, it is small and dedicated to restricted environments, and it performs less than our model in precision, map and recall but has a high FPS, which shows that tiny models are fast but not very precise. Fu et al. [8] also worked on the same concept, he combined CBAM with the Yolov4 model to improve the YOLOV4 model in ship detection, this combination improved the mAP50 from 82.04 to 84.06 and mAP75 from 66 to 67.85 but it reduces the FPS from 53.6 to 50.4. We can conclude that our YOLOv7 model performs better than Huixuan Fu’s proposed model.

Shao et al. [1] suggested a saliency sensitive CNN based on Yolov2 to increase detection, this model is inferior to ours in terms of FPS and map (Table 4).

Table 3 Comparison between models according to performance indicators

YOLO	V3	V4	V5	V7
Precision	0.982	0.96	0.984	0.966
Recall	0.966	0.98	0.968	0.977
F1	0.97	0.97	0.98	0.98
mAP@0.5 IoU	0.989	0.988	0.968	0.990
FPS	181	122	175	188
Training time (h)	17.04	20	16.79	16.15

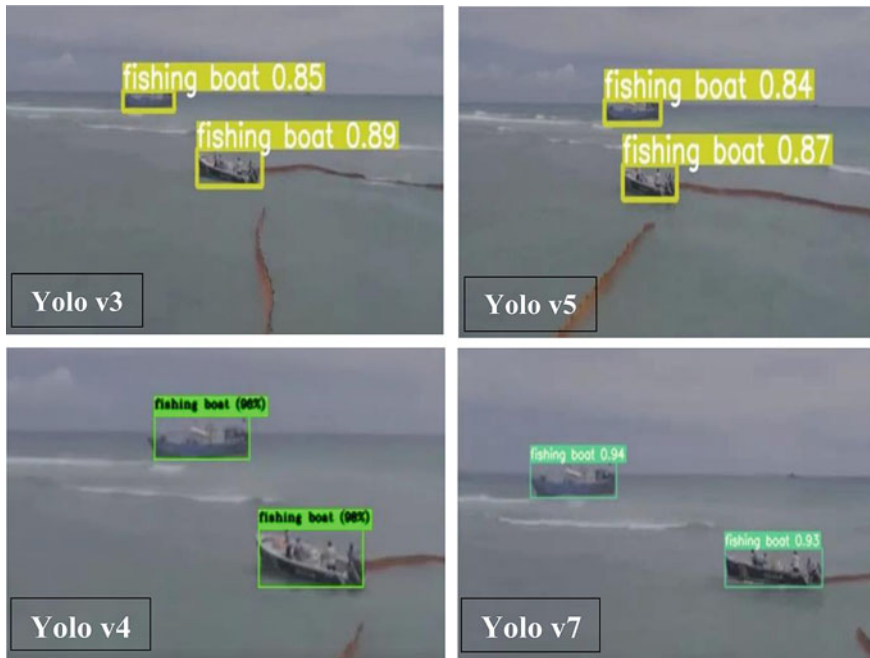


Fig. 9 Ship detection test results of each version of YOLO

Table 4 State of the art results versus Yolov7 results comparison

	Method	mAP	FPS
[1] [2018]	YOLOV2	0.874	49
[7] [2021]	Enhanced YOLOV3 tiny	0.741	135
[8] [2021]	YOLOV4	0.8406	50.4
Ours	YOLOV7	0.990	188

4 Conclusion

The present paper is a communication of the researches and experiments realized in the first phase of the study of the recognition and detection of ships in the maritime domain. Indeed, this work consists in first studying the performances of the different models that exist in the application of detection of the ships in order to produce a precise detection and in real time. To do this, we chose seaship dataset as a comparison base and then we pre-parsed it to be compatible with yolo style dataset. This dataset contains 6 types of ships and it is prepared under different conditions (illumination, backgrounds, various occlusion conditions, etc.). Then, we trained the different versions of the YOLO family in a high performance computer, and after analyzing and comparing the results of the trainings performed with the searches

performed in the same application, we deduced that yolov3, Yolov4 and yolov5 are performing well, except that the model YOLOV7 gives results with much higher performance in the detection of vessels.

In conclusion, we deduce that the new YOLOV7 with seaship dataset have shown their efficiencies thanks to their high performances, their rigorous precisions and their speed.

References

1. Z. Shao, L. Wang, Z. Wang, W. Du, W. Wu, Saliency-aware convolution neural network for ship detection in surveillance video. *IEEE Trans. Circuits Syst. Video Technol.* **30**(13). 781–794 (2019)
2. Z. Zou, Z. Shi, Ship detection in spaceborne optical image with SVD networks. *IEEE Trans. Geosci. Remote Sens.* **10**(154), 5832–5845(2016)
3. Z. Shi, X. Yu, Z. Jiang, B. Li, Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature. *IEEE Trans. Geosci. Remote Sens.* **8**(152):4511–4523 (2013)
4. J. Tang, C. Deng, G. Huang, B. Zhao, Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Trans. Geosci. Remote Sens.* **3**(153):1174–1185 (2014)
5. K. Kim, S. Hong, B. Choi, E. Kim, Probabilistic ship detection and classification using deep learning. *Appl. Sci.* **6**(18), 936 (2018)
6. N. Wawrzyniak, T. Hyla, A. Popik, Vessel detection and tracking method based on video surveillance. *Sensors* **23**(119), 5230 (2019)
7. H. Li, L. Deng, G. Yang, J. Liu, Z. Gu, Enhanced YOLO v3 tiny network for real-time ship detection from visual image. *IEEE Access* **9**, 16692–16706 (2021)
8. H. Fu, G. Song, Y. Wang, Improved YOLOv4 marine target detection combined with CBAM. *Symmetry* **13**(14), 623 (2021)
9. C.Y. Wang, A. Bochkovskiy, H.Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv. 2207.02696* (2022)
10. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 7263–7271 (2017)
11. W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C.-Y. Fu, C.B. Alexander, *SSD: Single Shot MultiBox Detector* (Springer, Cham, 2016), pp. 21–37
12. P.D. Hung, N. Kien, *Ngoc: SSD-MobileNet Implementation for Classifying Fish Species* (Springer, Cham, 2019), pp. 399–408
13. R. Girshick, Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, 1440–1448 (2015)
14. A.O. Vuola, S.U. Akram, J. Kannala, *Mask-RCNN and U-net Ensembled for Nuclei Segmentation* (IEEE, 2019), pp. 208–212
15. J. Redmon, A. Farhadi, YOLOv3: an incremental improvement. *arXiv. 1804.02767* (2018)
16. A. Bochkovskiy, C.Y. Wang, H.Y.M. Liao, YOLOv4: Optimal speed and accuracy of object detection. *arXiv. 2004.10934* (2020)
17. Y. Zhao, Y. Shi, Z. Wang, *The Improved YOLOV5 Algorithm and Its Application in Small Target Detection* (Springer, Cham, 2022), pp. 679–688
18. Z. Shao, W. Wu, Z. Wang, S. Zhenfeng Shao, W. Du, C. Li, SeaShips: a large-scale precisely annotated dataset for ship detection. *IEEE Trans Multimed* **20**(110), 2593–2604 (2018)

Graph Convolutional Network for Multilingual Sentiment Analysis



El Mahdi Mercha, Houda Benbrahim, and Mohammed Erradi

Abstract Sentiment Analysis, known also as opinion mining, discover and extract subjective information from source data allowing businesses to better understand the social sentiment around their products or services. The multilingual characteristics of these data require efficient multilingual sentiment analysis tools. In this work, we propose a graph-based approach for multilingual sentiment analysis. We construct a single heterogeneous text graph based on semantic, sequential, and statistical information to represent the entire multilingual corpus. Then, a graph convolution network learns predictive representation for nodes in a semi-supervised manner. Extensive experiments in real-world multilingual sentiment analysis dataset, demonstrate the effectiveness of the proposed approach. Also, it significantly outperforms the baseline models.

1 Introduction

The accessibility of the Web, and social media to a continuously increasing worldwide audience resulted in the rapid generation of opinionated textual data. This textual data carries valuable insights that can be very useful in a variety of fields, such as finance, politics, security, and marketing. These insights provide an opportunity for companies and individuals to discover a user base's thoughts and opinions. Thus, they can make informed decisions to improve their brands and services.

Sentiment analysis is considered as a computational study of sentiments, opinions, emotions, and appraisals to better understand a person's reactions and attitudes, towards several entities [1]. Because of the importance and impact of social interactions and opinions, sentiment analysis has received a great deal of attention, and several methods have been improved and developed for it [2].

E. M. Mercha (✉) · H. Benbrahim · M. Erradi
ENSIAS, Mohammed V University, Rabat, Morocco
e-mail: elmahdi.mercha@um5s.net.ma

E. M. Mercha
HENCEFORTH, Rabat, Morocco

Despite years of sentiment analysis research, the majority of the existing approaches in the field are language-dependent. In addition, these approaches are mainly focused on rich-resource languages like English [2], with the exception of a few studies focusing on languages with limited resources, such as annotated datasets [3]. However, users express themselves in morphologically rich languages that have been explored quite sparsely. Therefore, developing effective multilingual approaches is becoming increasingly critical.

In this work, we propose a graph-based approach for multilingual sentiment analysis. We construct a single heterogeneous text graph to represent the entire multilingual corpus. The whole documents of the corpus and the word vocabulary contribute as nodes in the graph that has been created. The relation between these nodes is created based on semantic, sequential, and statistical information.

Then, a vanilla Graph Convolution Network (GCN) [4] is used to learn efficient representations for both words and documents. Therefore, the task of multilingual sentiment analysis is considered as node classification problem. The experiments in real-world multilingual sentiment analysis dataset, demonstrate the effectiveness of the proposed approach. In addition, it can achieve strong classification performance compared with baseline models without using external sentiment information.

To the best of our knowledge, this is the first study which explore graph neural networks for multilingual sentiment analysis.

The main contribution is the representation of the entire multilingual corpus with a single heterogeneous text graph. We have conducted extensive experiments on real-world multilingual sentiment analysis dataset to demonstrate the effectiveness of the proposed approach.

2 Related Works

Initial studies in the literature on sentiment analysis almost exclusively focused on building monolingual systems. Recently, attention has been shifted to creating multilingual systems [5–7]. Multilingual sentiment analysis methods tend to perform language-agnostic and translation-free systems. These systems learn insightful information directly from multilingual unpaired data and provide predictions. The studies presented in the literature for multilingual sentiment analysis can be divided based on the methodology used for text representation into two categories: word-based [8] and character-based [6].

Word-based methods are extensively used in the development of language independent systems for multilingual sentiment analysis [8, 9]. These methods represent textual data based on either basic statistics of some ordered word combinations or randomly initialized word embedding. References [8, 9] developed architectures based on convolutional neural network and randomly initialized word embeddings that can learn directly from multilingual data. Reference [5] suggested a hybrid architecture based on the convolutional neural network and long short-term memory models to learn both n-gram features and long-term dependencies.

Although word-based are the most commonly used methods for dealing with multilingual sentiment analysis, several studies have proposed character-based methods. Almost character-based methods rely on convolutional neural network to design effective architectures which learn directly from character features [6, 10].

3 The Proposed Method

3.1 Heterogeneous Text Graph Construction

We construct a heterogeneous graph containing both word nodes and document nodes as in TextGCN [11] to model the entire multilingual corpus. We define word document edges based on the term frequency-inverse document frequency (TF-IDF). The weight based on TF-IDF is calculated as follows:

$$\text{tfidf}(w, d, C) = \text{tf}(w, d) \times \text{idf}(w, C)$$

where $\text{tf}(w, d)$ and $\text{idf}(w, C)$ denote respectively the relative frequency of word w within document d and the logarithmically scaled inverse fraction of documents of the corpus C containing the word w . However, the word-word edges are constructed based on two different types of language properties of information: sequential and semantic information. To extract the sequential information, we employ the sliding window strategy to gather statistics on word co-occurrence over the whole corpus. Then, the weight between two word nodes is calculated based on *positive pointwise mutual information* (PPMI), a prominent metric for word associations. Formally, the weight of a word pair w_i, w_j based on the sequential information is calculated as follows:

$$PPMI(w_i, w_j) = \max\left(\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, 0\right)$$

$$p(w_i, w_j) = \frac{\#W(w_i, w_j)}{\#W}$$

$$p(w_i) = \frac{\#W(w_i)}{\#W}$$

where $\#W(w_i, w_j)$ is the number of sliding windows in which both words w_i and w_j co-occurred, $\#W(w_i)$ is the number of sliding windows containing the word w_i , and $\#W$ is the total number of the sliding windows in the whole corpus.

Due to the lack of bridges across languages, building a single heterogeneous text graph to model the whole contents of a multilingual corpus based only on statistical and sequential information may result in a disconnected graph. Therefore, we utilize the semantic information to bridge the gap between languages and to construct a

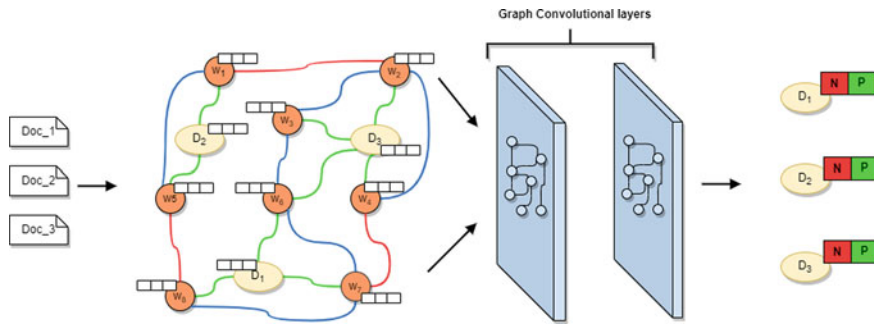


Fig. 1 The proposed graph-based approach for multilingual sentiment analysis

single connected graph. The weights of the semantic edges are calculated based on the similarity between multilingual word embeddings which are aligned in a single vector space. In this study, we utilize the fastText embeddings that have been aligned in a single space [12]. The weight of semantic edge of a word pair w_i, w_j is computed as follows:

$$\cos(w_i, w_j) = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|}$$

where x_i, x_j denote respectively the aligned word embeddings of the words w_i and w_j . The overall approach is illustrated in Fig. 1.

3.2 Graph Convolutional Network

The vanilla GCN is a variant of graph neural networks that extends convolution operations into non-Euclidean data structures, borrowing its concept from the standard convolutional neural network. It is a multilayer neural network that operates directly on a graph and learns representations of nodes depending on the surrounding nodes and the structure of the graph. Specifically, consider a graph $G = (V, E, A)$ where $V(|V| = n)$ is the set of nodes constituting the graph, E is the set of edges, and $A \in \mathbb{R}^{n \times n}$ is the adjacency matrix of the graph. If a pair of words w_i, w_j are linked, the $A(i, j)$ indicate the weight of the edge; otherwise, $A(i, j) = 0$. In addition, for all $w_i \in V$; $(w_i, w_i) \in E$ and $A(i, i) = 1$. We denote the normalized symmetric adjacency matrix of A as $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D denotes the degree matrix of A , and $D_{ii} = \sum_j A_{ij}$. We introduce the feature matrix $X \in \mathbb{R}^{n \times d}$ which hold the embedding vectors of all graph nodes, where d is the dimension of the embedding vectors and each row x_i of the matrix is embedding vector corresponding to the node i . The vanilla GCN model learns the nodes representations based on two-layers as follows:

$$H^{(1)} = \text{ReLU}(\tilde{A}XW^{(0)})$$

$$H^{(2)} = \text{softmax}(\tilde{A}H^{(1)}W^{(1)})$$

here $w^{(0)} \in \mathbb{R}^{d \times h}$ and $w^{(1)} \in \mathbb{R}^{h \times c}$ denote the weights of the first and second layer, respectively, with h is the dimension of the hidden representation, and c is the number of classes. The activation functions are defined as follows:

$$\text{ReLU}(x) = \max(x, 0)$$

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

4 Experiments

4.1 Datasets

To evaluate the proposed approach, we adopt three movie reviews datasets, namely Internet Movie Database (IMDB) [13], Allociné [14], and Muchocine [15], to construct a single multilingual sentiment analysis dataset.

- **IMDB** is an English-language dataset for binary sentiment classification. It consists of 25 k highly polar movie reviews for training, and 25 k for testing.
- **Allociné** is a French-language dataset for sentiment analysis. It contains 100 k positive and 100 k negative movie reviews.
- **Muchocine** is Spanish-language dataset for sentiment analysis. It consists of 3872 longform movie reviews, each review with a 1–5 scale rating.

In order to reduce the problem to binary classification, we mapped the classes with 1 and 2 stars into negative class, and classes with 4 and 5 stars into positive class, and we discard the class with 3 stars which is considered as the neutral class. While the size of the memory required by our approach scales with the size of the graph, we use a sample of data to reduce the memory required as well as the time needed for the training. Table 1 shows the statistics of the constructed multilingual sentiment analysis dataset.

Table 1 Statistics of the constructed multilingual sentiment analysis dataset

Training set	Validation set	Test set
10,646	2138	2138

4.2 Baselines

We compare the proposed approach against several strong baselines and state-of-the-art methods. The baselines include:

- **Bi-LSTM Emb-Non-Static**: a bi-directional long short-term memory [16] with trainable word embedding.
- **CNN Emb-Non-Static**: a convolutional neural network [17] with trainable word embeddings.
- **Bi-LSTM Emb-Static**: a bi-directional long short-term memory with randomly initialized word embedding. The word embeddings are not updated throughout the training process.
- **CNN Emb-Static**: a convolutional neural network which uses randomly initialized word embeddings. The word embeddings are not optimized during training.
- **Char-CNN**: a character-Level convolutional networks for text classification [18].
- **fastText**: a simple baseline method for text classification [19].
- **Text level GCN (TL-GCN)**: text level graph neural network for text classification [20].

4.3 Experiment Settings

In our experiments, we started by applying some standard language-agnostic preprocessing methods, especially, lowering, removing special characters, removing URLs, and removing HTML tags. Then, we tokenized text based on a single space, and we removed the low frequency words which appeared less than 5 times. In the construction of a single heterogeneous text graph, we set the size of the sliding window into 25. The initial word embeddings are initialized with the identity matrix, which means that each node is represented with one-hot vector. The dimension of the node embedding in the first layer of GCN is 200. The GCN is trained for a maximum of 200 epochs, with a learning rate of 0.002. We stop the training if the validation loss does not decrease for five successive epochs.

For the baseline models, we adopt the default settings for Char-CNN, fastText and TL-GCN as mentioned in their original papers. However, for Bi-LSTM Emb-Non-Static and CNN Emb-Non-Static, we performed hyper-parameter optimization to select the optimal parameters. For Bi-LSTM Emb-Non-Static, we set the embedding dimension as well as the number of units as 100, dropout as 0.5, layer weight regularizer as $L2$, with a weight of 0.01. For CNN Emb-Non-Static, the embedding dimension and the number of filters is set to 100, the kernel size as 3. We adopt for Bi-LSTM Emb-Static, CNN Emb-Static the same settings as Bi-LSTM Emb-Non-Static and CNN Emb-Non-Static respectively. All the four models with a learning rate of 0.001. For all baseline models we use randomly initialized word embedding instead of pre-trained word embedding.

4.4 Results Analysis

A comprehensive experiment is conducted on the constructed multilingual sentiment analysis dataset. Table 2 shows the results of the proposed approach against the baseline models. The experimental results reveals that the proposed approach can learn predictive representations and achieves high classification performance. For more detailed performance analysis, we observe that the worst results are achieved by the Bi-LSTM Emb-Static, Char-CNN, and CNN Emb-Static. This can be explained with the fact that representing multilingual texts with only feature characters cannot learn insightful information, due to the syntactic inconsistency between languages. In addition, the results achieved by Bi-LSTM Emb-Static and CNN Emb-Static is due to the high dependence between the representation learned for the whole text and the word embedding. However, random word embeddings do not consider the relation among words and do not capture any semantic or syntactic information. Bi-LSTM Emb-Non-Static and CNN Emb-Non-Static clearly outperform Bi-LSTM Emb-Static and CNN Emb-Static, respectively, as they can learn effective words representation with sentiment information in a supervised manner. The graph-based method TL-GCN performs quite well, and show competitive performance, which indicate the effectiveness of the graph-based approaches. We note that the proposed approach outperforms TL-GCN. This is due to the way proposed for constructing the graph from the entire corpus, which efficiently model the global information between words and documents and bridge the gap between languages. Furthermore, GCN can learn node representations based on semantic, sequential, and statistical information by combining information from graph topologies and attributes into low-dimension embeddings.

The experimental results shows that the proposed approach can learn predictive word and document representations, as well as achieve high multilingual sentiment analysis performance. The major factor which explains the supremacy of our approach is the proposed method for modeling the entire corpus with a single heterogeneous graph based on the semantic, sequential, and statistical information. The constructed graph can capture the global information about words relations in the whole corpus, as well as allows the propagation of the sentimental information between documents based on the word's connections. Also, it bridges the gap between languages through the semantic linkages.

Even though the proposed approach obtained interesting results and overcome the state-of-the-art model, it comes with its own limitations. The proposed approach builds a single graph for the whole corpus, implying that the size of graph scaled with the size of the corpus, resulting in a high memory consumption in large corporuses.

Table 2 Test accuracy on the constructed multilingual sentiment analysis dataset. We evaluate each model 30 times and report the mean \pm standard deviation

Model	Accuracy
Bi-LSTM Emb-Non-Static	78.29 \pm 1.57
CNN Emb-Non-Static	79.78 \pm 0.92
Bi-LSTM Emb-Static	53.24 \pm 0.05
CNN Emb-Static	63.99 \pm 1.40
Char-CNN	54.02 \pm 0.57
fastText	67.68 \pm 3.66
TL-GCN	82.86 \pm 1.08
Our approach	87.55 \pm 0.26

5 Conclusion and Future Works

We have introduced a graph-based approach for multilingual sentiment analysis. The proposed approach constructs a single heterogeneous text graph based on semantic, sequential, and statistical information to represent the entire multilingual corpus. The graph the vanilla GCN is used to learn representations for both nodes. The results achieved in real-world multilingual sentiment analysis dataset reveals the efficiency of the proposed approach against strong baseline models.

Future work concerns enhancing the performance of the proposed approach and evaluate its robustness against language variation.

Acknowledgements The authors would like to thank HENCEFORTH for its financial support for this research project.

References

1. B. Liu, Sentiment analysis: mining opinions, sentiments, and emotions (2020)
2. L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey. *Wiley Interdiscip. Rev. Data Mining Knowledge Discovery* **8**(4), e1253 (2018)
3. A. Dahou, S. Xiong, J. Zhou, M.H. Haddoud, P. Duan, Word embeddings and convolutional neural network for arabic sentiment classification. In: *Proceedings of coling 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (2016), pp. 2418–2427
4. T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks. [arXiv:1609.02907](https://arxiv.org/abs/1609.02907) (2016)
5. M.H. Shakeel, S. Faizullah, T. Alghamidi, I. Khan, Language independent sentiment analysis. In: *AECT (IEEE, 2020)*, pp. 1–5
6. J. Wehrmann, W. Becker, H.E. Cagnini, R.C. Barros, A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In: *IJCNN (IEEE, 2017)*, pp. 2384–2391
7. E.M. Mercha, H. Benbrahim, Machine learning and deep learning for sentiment analysis across languages: a survey. *Neurocomputing* (2023)

8. L. Medrouk, A. Pappa, Deep learning model for sentiment analysis in multi-lingual corpus. In: International Conference on Neural Information Processing Springer (2017), pp. 205–212
9. M. Attia, Y. Samih, A. Elkahky, L. Kallmeyer, Multilingual multi-class sentiment classification using convolutional neural networks. In: LREC. (Miyazaki, Japan, 2018)
10. J. Wehrmann, W.E. Becker, R.C. Barros, A multi-task neural network for multilingual sentiment classification and language detection on twitter. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing (2018), pp. 1805–1812
11. L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 (2019), pp. 7370–7377
12. A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou, Word translation without parallel data. [arXiv:1710.04087](https://arxiv.org/abs/1710.04087) (2017)
13. A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011), pp. 142–150
14. T. Blard, French sentiment analysis with bert. GitHub repository. <https://github.com/TheophileBlard/french-sentiment-analysis-with-bert>
15. Spanish movie reviews. <http://www.lsi.us.es/fermin/index.php/Datasets>
16. P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (2016), pp. 2873–2879
17. Y. Kim, Convolutional neural networks for sentence classification. In: EMNLP, pp. 1746–1751 (2014)
18. X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 1 (2015), pp. 649–657
19. A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2. (Short Papers, 2017), pp. 427–431
20. L. Huang, D. Ma, S. Li, X. Zhang, H. Wang, Text level graph neural network for text classification. In: EMNLP-IJCNLP, pp. 3444–3450 (2019)

Predicting Blood Glucose Levels in Type 1 Diabetes Using LSTM



Dounia Nasir, Mohamed Elmehdi Ait Bourkha, Anas Hatim, Said Elbeid, Siham Ez-ziymy, and Khalid Zahid

Abstract The prediction of blood glucose level (BGL) appears necessary for patients with type 1 diabetes (T1D). In this work, we develop an automatic predictor using a clinical data from 12 patients with T1D. This system accurately informs the patients about their BGL in the future. As it is important, early detection of hypoglycemia or hyperglycemia is crucial for T1D patients. In this study, we proposed a method for predicting blood glucose levels (BGL) using a long short-term memory (LSTM) neural network with a single input (past samples of BG). Our results are based on LSTM neural network. To evaluate our model, we use the root mean square error (RMSE) and we find that the RMSE is 2.23 mg/dL for a prediction horizon of 15 min and 7.38 mg/dL for a prediction horizon of 30 min.

Keywords Type 1 diabetes (T1D) · Blood glucose (BG) · Continuous glucose monitoring (CGM) · Hypoglycemia · Hyperglycemia · Long-short term memory (LSTM) neural network

1 Introduction

Diabetes is one of the most prevalent endocrine diseases affecting many people worldwide. It's classified into two types depending on the patient's conditions: non-insulin dependent and insulin-dependent [1]. Patients with insulin-dependent

D. Nasir (✉) · M. E. A. Bourkha · A. Hatim
TIM Team, ENSA Marrakech, Cadi Ayyad University, Marrakech, Morocco
e-mail: dounia.nasir@ced.uca.ma

S. Elbeid
CISIEV Team, ENSA Marrakech, Cadi Ayyad University, Marrakech, Morocco

S. Ez-ziymy
LGEMS, ENSA Agadir, Ibn Zohr University, Agadir, Morocco

K. Zahid
Ibn Zohr Hospital, Ministry of Health, Marrakech, Morocco

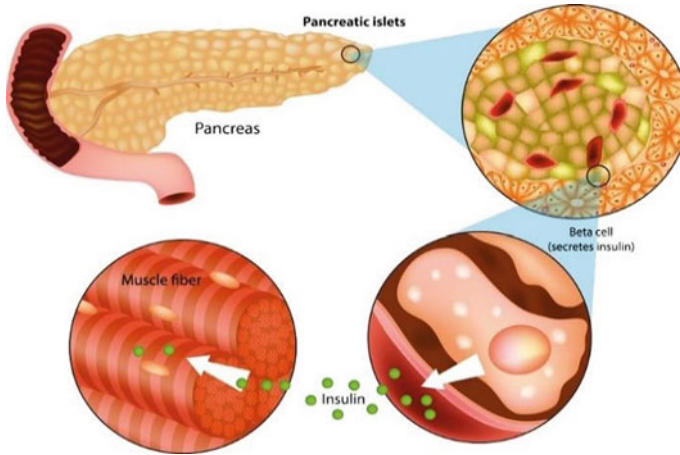


Fig. 1 Secretion of insulin in blood through β cells

diabetes require insulin treatment. Although the blood glucose level of non-insulin dependent patients can be controlled through diet, a significant proportion of them may still require insulin therapy. In this paper, we will focus on insulin-dependent diabetes, also known as Type 1 Diabetes (T1D).

T1D can be very harmful for a person's health. It is progressively at the rate of 3% per year [2, 3]. In the human body, two hormones are responsible for the blood glucose control (insulin and glucagon). Blood glucose level for patients with T1D is higher due to the loss or destruction of beta cells (β) in the pancreas "Fig. 1", which lead to a hyperglycemia (when the blood glucose is greater than 125 mg/dL while fasting, and greater than 180 mg/dl 1 to 2 h after eating).

Hyperglycemia recognized as the major factor that promotes the atherosclerosis "Fig. 2" (the main reason for diabetic nephropathy and retinopathy) [4]. Moreover, it associates generally with some complication such as: Dry mouth, Shortness of breath, Loss of consciousness...

We proposed LSTM neural network for blood glucose prediction in patients with T1D. We validate our model using the hold out validation method, where 80% of our data is used for training and the remaining 20% is used for the testing.

2 Related Works

Many works focused on predicting blood glucose level and its changes in patients with T1D. This works can be divided into two parts (i.e., directions). The first part is based on mathematical methods that are simple and easy to be implemented in electronics devices. However, these mathematical models often show inaccurate forecasting of

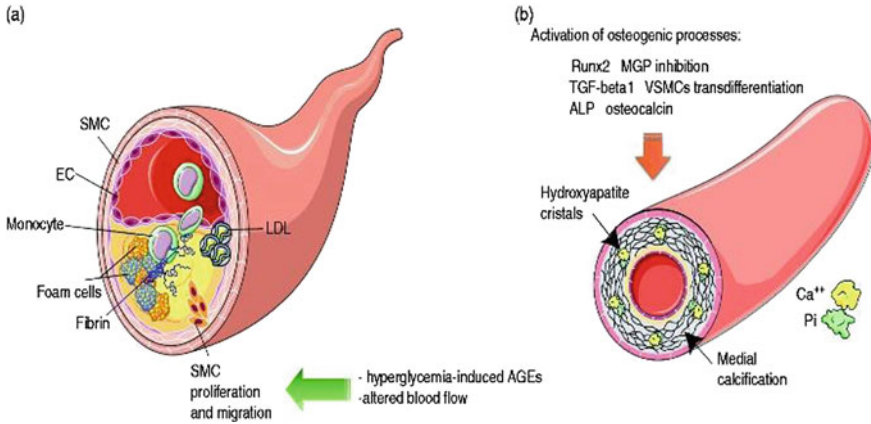


Fig. 2 Changes from hyperglycemia to atherosclerosis

blood glucose levels due to their dependence on the patient’s activities, which can be difficult to measure accurately.

The second part is based on artificial intelligence and the advanced signal processing techniques, which demonstrates a high performance of predicting BGL in patients with T1D. However, the implementation of artificial intelligence (AI) for predicting BGL can be challenging due to its complexity.

Pérez-Gandia et al. [5] proposed a method based on neural network with a single previous inputs (CGM), and used 6 real patients with T1D.

Hamdi et al. [6] proposed a method based on support vector regression (SVR) utilizing differential evolution (DE) algorithm with a single previous inputs (CGM), and used 12 real patients with T1D.

Daskalaki et al. [7] proposed a method based on artificial neural network utilizing two inputs features: (CGM) and insulin, and used 30 T1D’s database from silico.

Georga et al. [8] proposed a method based on (SVR) and Random-forest with multiple inputs, and used 15 real patients with T1D.

Ben Ali et al. [9] proposed a method based on ANN utilizing optimized input for each patient with a single previous input (CGM), and used 12 real patients with T1D.

Martinsson et al. [10] proposed a method based on recurrent neural network (RNN) with a single previous inputs (CGM), and used 6 real patients with T1D.

Wang et al. [11] proposed a method based on adaptive-weighted-average Framework based on ARMAX (AR), ELM and SVR with a single previous inputs (CGM), and used 10 real patients with T1D.

Turksoy et al. [12] proposed a method based on recursive ARMAX model with multiple inputs, and used 14 patients with T1D.

Zecchin et al. [13] proposed a method based on Feed-forward NN and first-order polynomial model with two inputs features: (CGM) and meal, and used 15 real patients with T1D.

Table 1 Related work's performance

Study	PH (min)	BGL predicted (mg/dl)
Daskalaki et al. [7]	30	2.8–4.5
	45	4.0–6.3
Georga et al. [8]	30	5.7
Turksoy et al. [12]	30	11.7
Zecchin et al. [13]	30	16.6
Midroni et al. [14]	30	16.21
Zarkogianni et al. [15]	30	11.42
	60	19.58
	120	31.00
Pappada et al. [16]	75	43.9

Midroni et al. [14] proposed a method based on XGBoost's feature importance to select optimal features with multiple inputs, and used 6 real patients with T1D.

Zarkogianni et al. [15] proposed a method based on Self—Organizing Map (SOM) with two inputs features: (CGM) and physical activities, and used 10 real patients with T1D.

Pappada et al. [16] proposed a method based on ANN with multiple inputs, and used 27 real patients with T1D.

Alfian et al. [17] proposed a method based on ANN with additional time-domain features with a single previous inputs (CGM), and used 12 real patients with T1D.

In Table 1, we present a summary of the performance of the previous studies mentioned before.

3 Methodology

“Figure 3” shows a summary of different steps that we follow to predict BGL values in the future.

A. Data acquisition

The data used in this paper is from DirectNet platform, its protocol is to give an evaluation of counter-regulatory hormone responses during hypoglycemia and the accuracy of continuous glucose monitors in children with T1D. This clinical data obtained from 12 children aged 3 to 7 years and 12 to 17 years. Patients record their blood sugar every 5 min for about 7 days using a CGM device called Guardian-RT. The protocol used to collect the CGM dataset was approved by DirecNet, Jaeb Center for Health Research.

This data is publicly available online at the following link: (<https://public.jaeb.org/direcnet/stdy/167>). In Table 2, we present the periods of CGM's recording for each 12 patients.

Fig. 3 Different steps of prediction

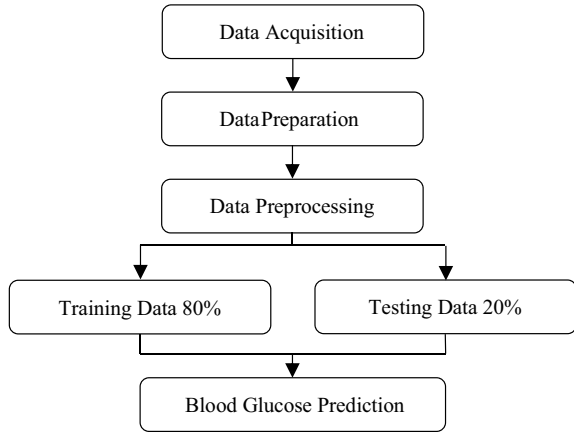


Table 2 Periods of CGM’s recording

Patient ID	Period (Days)
19	9
11	8
13	7
32	7
4	7
22	7
16	7
15	7
18	7
8	7
6	7
21	7

B. Data preparation

The data includes record ID, patient ID, date of blood glucose reading, time of blood glucose reading, blood glucose level, calibration, and file type (Guardian 1 or Guardian 2). During the preparation of our training data, we noticed that some missing BGL points in the CGM had a value of 0.00 mg/dl. Training our model with missing BGL values can lead to low performance of BGL prediction. Hence, we use linear interpolation to fill in missing BGL samples.

We only applied a linear interpolation for missing BGL data, this data did not exceed a duration of 3 h. This is because coding these points with long duration interpolation can lead to inaccurate numerical values. However, interpolation can also lead to bias problems. In Table 3, we present the data points before and after interpolation.

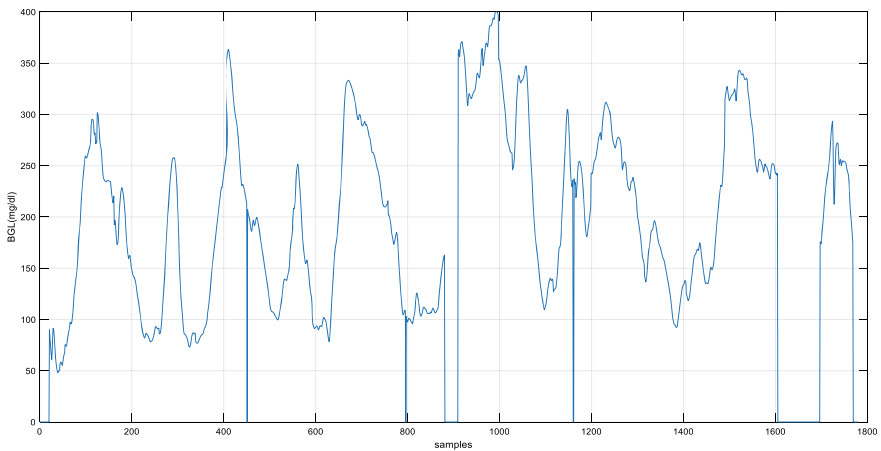
Table 3 Data points before and after interpolation

Patient ID	Data points	
	Before interpolation	After interpolation
19	2720	2742
11	2393	2439
13	2298	2340
32	2231	2234
4	2212	2294
22	2173	2181
16	2159	2172
15	2145	2171
18	2132	2042
8	2097	2154
6	2087	2149
21	2076	2185

“Figure 4” presents the plot of BGL’s data points before the interpolation, while “Fig. 5” presents the plot of BGL’s data points after the interpolation.

C. Data preprocessing

In this study, the CGM is the recording of blood glucose as a time series, where the sensor (Guardian RT 1 and 2) measures the BGL every 5 min. The BGL measured by the sensors may be subject to noise -that come- from certain physiological responses and sensor delays. Noisy BGL data can reduce the ability of our model to predict future values of BG. So, to improve the signal-to-noise ratio and the accuracy of our

**Fig. 4** CGM before interpolation

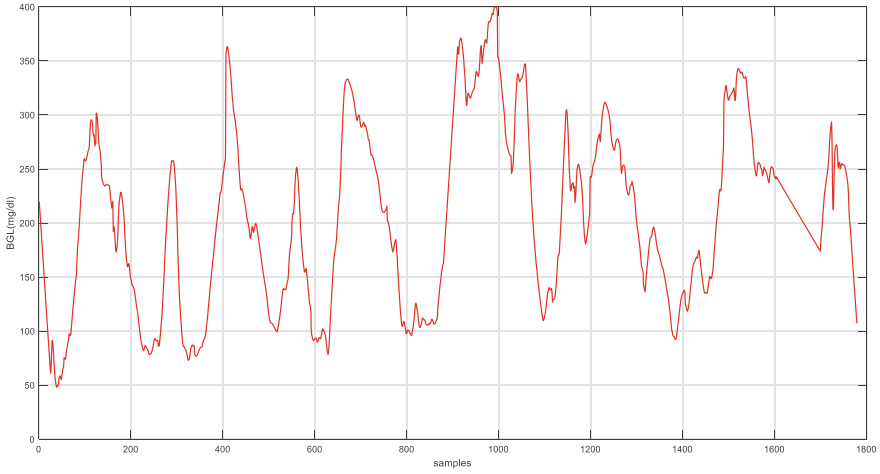


Fig. 5 CGM after interpolation

model, we applied the Moving Average filter [18]. The noise can be deleted using the next formula shown in “(1)”. “Figure 6” presents the BGL after filtering.

We consider the time series of BGL $\{G(1), G(2), \dots, G(n)\}$

$$G(i) = \frac{1}{l} \sum_{q=0}^{l-1} G(i - q) \tag{1}$$

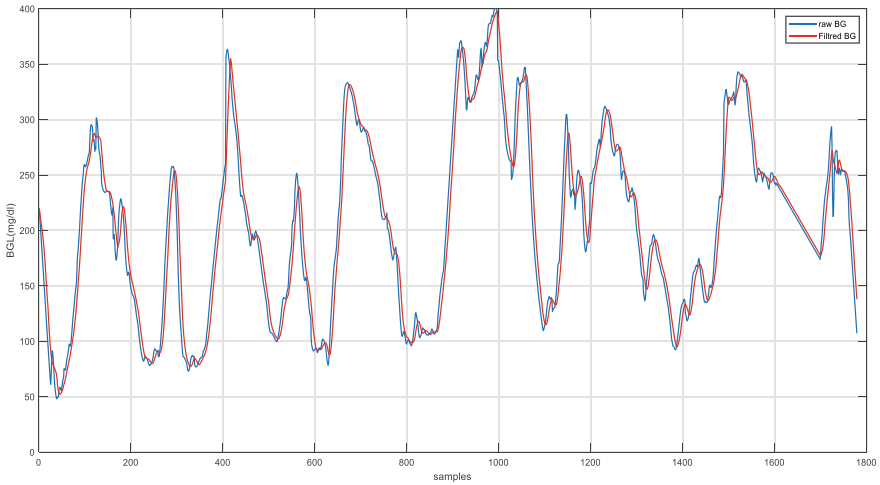


Fig. 6 BGL after filtering

With: $i \in [1, n]$, and l is the width of the sliding window filter (i.e., we used the moving average filter with a window size $l = 10$).

The BGL for each patient after filtering is represented as a vector with a length of n . Our model predicts future BGL values based on the last 30 min of CGM data.

The BGL is collected every 5 min. And, the input vector contains $h = 6$ BGL values, which are measured from the last 30 min of continuous glucose monitoring (CGM) data.

Assuming a prediction horizon $PH = 15$ min, our model uses the last 30 min of historical data as features to predict the next 15 min of blood glucose (BG) values. The prediction output is represented by $l = 3$ values, which correspond to the predicted values for each 5 min interval within the $PH = 15$ min.

The input matrix A , and the output vector B are obtained by using the sliding window's techniques applied on our time series of BGL $\{G(1), G(2), \dots, G(n)\}$.

The input matrix A in "(2)" is formulated by A_i vectors with $i \in [1, n]$, each vector is expressed as follow:

$$A_i = [G_i G_{i+1} \dots G_{i+h-1}] \quad (2)$$

The output vector B "(3)" is formulated by B_i values as:

$$B_i = G_{i+h+l} \quad (3)$$

$$A = \begin{bmatrix} A_i \\ A_{i+1} \\ \dots \\ A_{n-h-l} \end{bmatrix} \quad (4)$$

$$B = \begin{bmatrix} B_i \\ B_{i+1} \\ \dots \\ G_n \end{bmatrix} \quad (5)$$

$$B_{predicted} = f_i(A) \quad (6)$$

With f_i is our prediction model.

To evaluate the performance of our predictive model, we used the holdout validation techniques. Specifically, 80% of each patient's data was randomly selected for training, and the remaining 20% was held out for testing.

D. Development of a BGL predictor using LSTM-Based model

The LSTM (Long short-term memory) is a type of Recurrent Neural Networks (RNN) architecture as shown in "Fig. 7". The LSTM was created to give an accurate model of a temporal and large sequences.

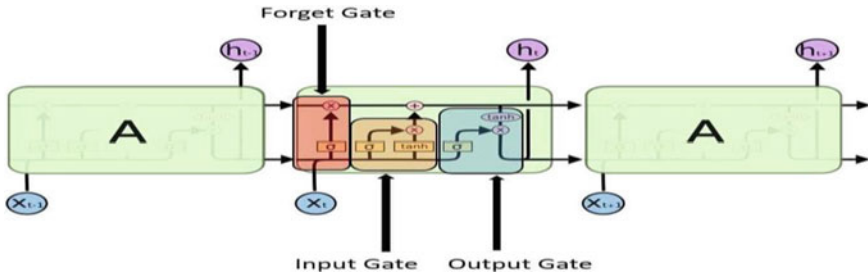


Fig. 7 Architecture of LSTM

In our model, we used a sequence-to-sequence LSTM, it is also called encoder decoder LSTM. The idea is to use one LSTM to read the input sequence, one time-step at a time, to obtain large fixed-dimensional vector representation, and then to use another LSTM to extract the output sequence from that vector (Sutskever et al. [19]).

The LSTM neural Networks can be more accurate and effective than the conventional (RNN) as the LSTM can overcome technical problems such as vanishing gradient and exploding gradient (Graves et al. [20]).

The activation function used in the hidden layers is the Rectified Linear Unit function (ReLU), which is defined as $\max(0, x)$. It has several advantages, such as: its derivative only taking two values 0 or 1, which makes the training faster. Additionally, the gradient values are 1 when the input is positive, and 0 when the input is negative or equal to 0, guaranteeing no vanishing gradient.

In Table 4, we present various pieces of information about the LSTM neural network utilized in this study.

Table 4 Information about LSTM neural network

Name	Type	Activation	Learnables
Sequence Sequence inputs with 1-dimensions	Sequence input	1	–
LSTM LSTM with 100 hidden units	LSTM	100	Inputs weights 400*1 Recurrent weights 400*100 Bias 400*1
ReLU-1 ReLU	ReLU	100	–
FC 1 fully connected layer	Fully connected	1	Weights 1*100 Bias 1*1
ReLU-2 ReLU	ReLU	1	–
Regressionoutput Mean-squared-error	Regression output	1	–

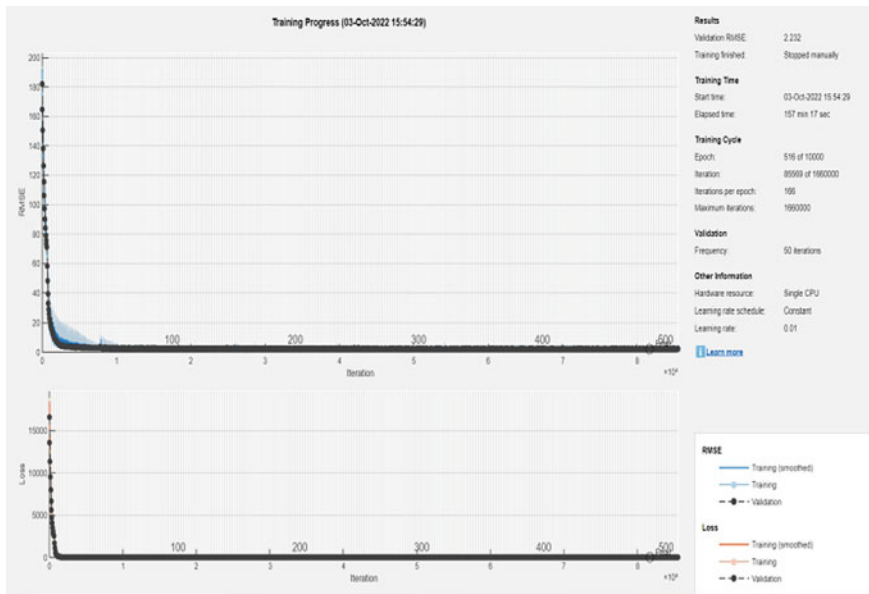


Fig. 8 Training progress with LSTM for PH = 15 min

“Figure 8” presents the training progress of the LSTM neural network described earlier.

The data was normalized using the Z-score method in order to expedite the training process.

E. Evaluating performance

- RMSE (Root Mean Square Error) represents how far the prediction’s errors are from the regression line of the data points. It measures the prediction’s accuracy. It can be calculated using the next formula, with $G_{predicted}$ is the predicted CGM level and G_{real} is the actual measured CGM.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (G_{predicted_i} - G_{real_i})^2} \quad (7)$$

- MAPE (Mean Absolute Percentage Error) measures the accuracy of a forecasting problems. It can be expressed as a percentage using the following formula.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{G_{real_i} - G_{predicted_i}}{G_{real_i}} \right| \quad (8)$$

- R^2 is a correlation coefficient that examines the quality of linear regression. It provides information about the goodness of fit of a model. It can be calculated using the next formula (i.e., in percentage), with $\overline{G_{real}}$ is the mean value of G_{real} .

$$R^2 = 1 - \frac{\sum_{i=1}^n (G_{real_i} - G_{predicted_i})^2}{\sum_{i=1}^n (G_{real_i} - \overline{G_{real}})^2} \tag{9}$$

The proposed model based on sequence-to-one LSTM neural network gives the best RMSE at PH = 15 min, which is 2.23 mg/dl. While in PH = 30 min, the RMSE was 7.38 mg/dl.

We also tried other machine learning models to predict BGL, such as: SVM, decision tree, Medium neural network and Wide neural network. However, LSTM neural network give better results in BGL prediction compared to other machine learning models.

“Figures 9” and “10” present the predicted CGM using our LSTM neural network -our model- and the real CGM data for patient ID 19 in PH = 15 min and 30 min, respectively.

The results show that our model was able to successfully predict the BGL of the patients with T1D.

In Table 5, we present the evaluation indicators RMSE, MAPE and R^2 on the testing data for each patient for PH = 15 min and 30 min, respectively.

In Table 6, we present the LSTM neural network tested on the testing data of all 12 real patients with T1D.

The performance of our BGL predictor is very promising.

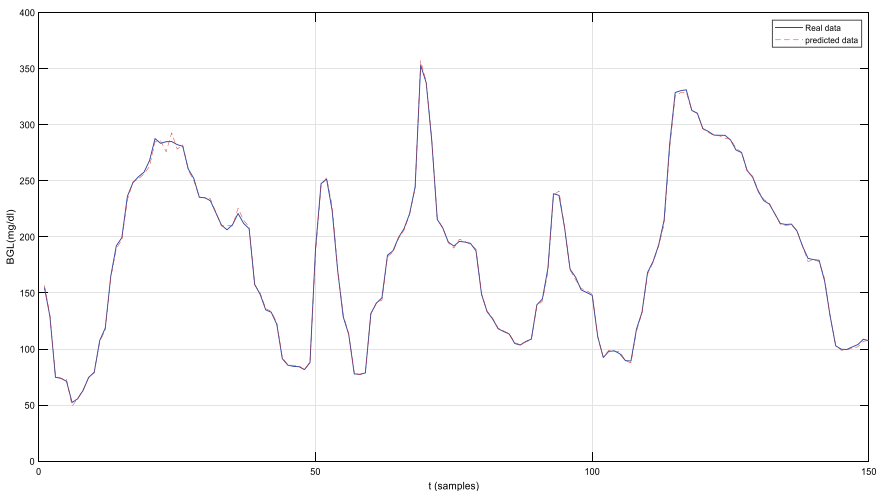


Fig. 9 Prediction of BGL in PH = 15 min for patient ID 19

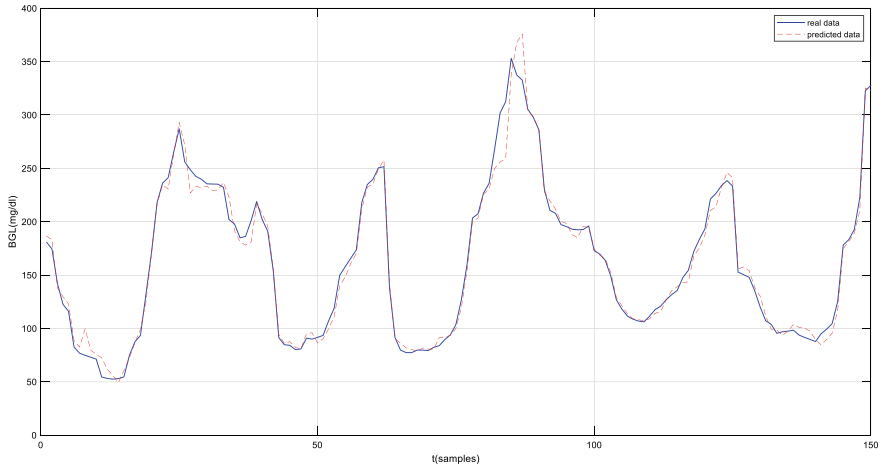


Fig. 10 Prediction of BGL in PH = 30 min for patient ID 19

Table 5 RMSE, MAPE and R2 for PH = 15, 30 min

PH	15 min			30 min		
	Patient ID	RMSE (mg/dl)	MAPE (%)	R2	RMSE (mg/dl)	MAPE (%)
19	2.00	0.70	0.99	7.55	2.95	0.99
11	2.56	1.07	0.99	9.00	3.87	0.97
13	2.21	0.78	0.99	6.59	2.61	0.99
32	2.68	1.11	0.99	8.95	4.17	0.98
4	1.83	0.93	0.99	6.08	3.33	0.98
22	1.29	0.83	0.99	4.69	3.06	0.98
16	2.25	1.17	0.99	8.20	4.35	0.98
15	1.83	0.69	0.99	6.04	2.37	0.98
18	2.23	0.84	0.99	6.45	2.91	0.99
8	2.51	1.17	0.99	6.97	3.93	0.98
6	2.98	1.07	0.99	9.65	3.74	0.98
21	1.85	0.87	0.99	6.48	3.35	0.99

Table 6 RMSE, MAPE and R2 on testing data for all patients

PH (min)	RMSE (mg/dl)	MAPE (%)	R2
15	2.23	0.93	0.99
30	7.38	3.38	0.98

Table 7 General comparison with some previous studies

Study	PH (min)	RMSE (mg/dl)
Pérez-Gandia et al. [5]	15	9.70
	30	17.50
Wang et al. [11]	15	10
	30	19
Hamdi et al. [6]	15	9.44
	30	10.78
Ben Ali et al. [9]	15	6.43
	30	7.45
Martinsson et al. [10]	30	18.87
Alfian et al. [17]	15	2.82
	30	6.31
Proposed method	15	2.23
	30	7.38

The mean RMSE for PH = 15 min was 2.23 mg/dl and the mean RMSE for PH = 30 min was 7.38 mg/dl.

The average R2 index was greater than 0.98 for both prediction horizons: PH = 15 min and 30 min.

4 Comparison with Some Other Previous Works

Comparing with some other previous works, we can notice that our model is optimal at PH = 15 min with RMSE of 2.23 mg/dl, and 7.38 mg/dl at PH = 30 min.

The results show that our proposed method exceeds the state-of-the-art in various previous works. However, this general comparison cannot be used as the main proof of the performance, because performance depends on many other factors, such as: the nature of the data (the number of patients used in the study) and the preprocessing techniques (method used to fill missing data, the type of the filter to denoise the CGM, and data splits).

Furthermore, the previous works mentioned in Table 7, below used a different database, except for Alfian et al. [17] who used the same database with the same 12 real patients with T1D.

5 Conclusion

This paper presents an efficient method based on LSTM neural network to predict BGL and provide earlier information on the changes in blood glucose.

In this work, we focused on the prediction horizons of 15 and 30 min. The sliding window was used -in this study- to create a matrix of input ‘Blood Glucose Level’, and a vector of output ‘Blood Glucose’ that represents the future values for each of the prediction horizons used.

To improve the performance of prediction, we used the moving average window to denoise CGM data. The features used are based solely on the last 30 min of blood glucose values, which makes our approach simpler and less complex.

In this study, we used clinical data from 12 real patients with T1D (12 children), this makes the general evaluation of the model’s performance difficult.

Our LSTM-based model shows a good performance by generating low RMSE, low MAPE and high R2. The RMSE of the prediction from LSTM neural network is: 2.23 mg/dl for PH = 15 min, and 7.38 mg/dl for PH = 30 min.

References

1. R.F. Hamman, Genetic and environmental determinants of non-insulin-dependent diabetes mellitus (NIDDM). *Diabetes. Metab. Rev.* **8**(4), 287–338 (1992). <https://doi.org/10.1002/dmr.5610080402>
2. D. Aronson et, E. J. Rayfield, How hyperglycemia promotes atherosclerosis: molecular mechanisms. *Cardiovasc. Diabetol.* **1**:1 (2002). <https://doi.org/10.1186/1475-2840-1-1>
3. dc_40_s1_final.pdf. Consulté le: 27 octobre 2022. [En ligne]. Disponible = sur: https://professional.diabetes.org/files/media/dc_40_s1_final.pdf
4. Atherosclerosis, Wikipedia. 24 octobre 2022. Consulté le: 27 octobre 2022. [En ligne]. Disponible sur: <https://en.wikipedia.org/w/index.php?title=Atherosclerosis&oldid=1117983239>
5. C. Pérez-Gandía et al., Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technol. Ther.* **12**(1), 81–88 (2010). <https://doi.org/10.1089/dia.2009.0076>
6. T. Hamdi, J. Ben Ali, V. Di Costanzo, F. Fnaiech, E. Moreau, et J.-M. Ginoux, Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. *Biocybern. Biomed. Eng.* **38**(2), 362–372 (2018). <https://doi.org/10.1016/j.bbe.2018.02.005>
7. E. Daskalaki, A. Prountzou, P. Diem, et S.G. Mougiakakou, Real-time adaptive models for the personalized prediction of glycemic profile in type 1 diabetes patients. *Diabetes Technol. Ther.* **14**(2), 168–174, (2012). <https://doi.org/10.1089/dia.2011.0093>
8. E.I. Georga, V.C. Protopappas, D. Polyzos, et D.I. Fotiadis, Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models. *Med. Biol. Eng. Comput.* **53**(12), 1305–1318 (2015). <https://doi.org/10.1007/s11517-015-1263-1>
9. J. Ben Ali, T. Hamdi, N. Fnaiech, V. Di Costanzo, F. Fnaiech, et J.-M. Ginoux, Continuous blood glucose level prediction of type 1 diabetes based on artificial neural network. *Biocybern. Biomed. Eng.* **38**(4), 828–840 (2018). <https://doi.org/10.1016/j.bbe.2018.06.005>
10. J. Martinsson, A. Schliep, B. Eliasson, et O. Mogren, Blood glucose prediction with variance estimation using recurrent neural networks. *J. Healthc. Inform. Res.* **4**(1), 1–18 (2020). <https://doi.org/10.1007/s41666-019-00059-y>
11. Y. Wang, X. Wu, et X. Mo, A novel adaptive-weighted-average framework for blood glucose prediction. *Diabetes Technol. Ther.* **15**(10), 792–801 (2013). <https://doi.org/10.1089/dia.2013.0104>

12. K. Turksoy, E.S. Bayrak, L. Quinn, E. Littlejohn, D. Rollins, et A. Cinar, Hypoglycemia early alarm systems based on multivariable models. *Ind. Eng. Chem. Res.* **52**(35), 12329–12336 (2013). <https://doi.org/10.1021/ie3034015>
13. C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, et C. Cobelli, Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. *IEEE Trans. Biomed. Eng.* **59**(6), 1550–1560 (2012). <https://doi.org/10.1109/TBME.2012.2188893>
14. C. Midroni, P.J. Leimbigler, G. Baruah, M. Kolla, A.J. Whitehead, et Y. Fossat, Predicting glycemia in type 1 diabetes patients: experiments with XGBoost, p. 6.
15. K. Zarkogianni et al., Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. *Med. Biol. Eng. Comput.* **53**(12), 1333–1343 (2015). <https://doi.org/10.1007/s11517-015-1320-9>
16. S.M. Pappada et al., Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes Technol. Ther.* **13**(2), 135–141 (2011). <https://doi.org/10.1089/dia.2010.0104>
17. G. Alfian et al., Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features. *Biocybern. Biomed. Eng.* **40**(4), 1586–1599 (2020). <https://doi.org/10.1016/j.bbe.2020.10.004>
18. S.K. Salih, S.A. Aljunid, S.M. Aljunid, et O. Maskon, Adaptive filtering approach for denoising electrocardiogram signal using moving average filter. *J. Med. Imaging Health Inform.* **5**(5):1065–1069 (2015). <https://doi.org/10.1166/jmih.2015.1495>
19. I. Sutskever, O. Vinyals, et Q. V. Le, Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems*, vol. 27 (2014). Consulté le: 28 octobre 2022. [En ligne]. Disponible sur: <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>
20. A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, et J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(5), 855–868 (2009). <https://doi.org/10.1109/TPAMI.2008.137>

Sarcasm Detection on Social Media using Machine Learning Approach



Chahrazad Lagrini and Abdellah Idrissi 

Abstract With sarcastic language, the speaker can say the exact opposite of what they mean, usually in a provocative, arrogant, and spiteful tone. Sarcasm can also be used to express annoyance or disappointment. Sarcasm is difficult to detect on social media due to the absence of several clues, such as tone and gestures. This paper seeks to find a powerful implementation, in order to detect sarcasm on social media. In this context, we propose a Machine Learning approach to predict if a statement is sarcastic or not. The sarcasm detection has garnered interest from researchers in several fields, including Artificial Intelligence and especially in Natural Language Processing (NLP). However, this subject is little addressed by researchers. In this paper, we will study the theoretical part of sarcasm detection, followed by the methodology adapted to the dataset, to finally present the results. The results obtained are promising, we were able to achieve an accuracy of 84% using RNN and 81% using SVM.

Keywords Sarcasm · Sentiment analysis · Machine learning · Social networks

1 Introduction

Over the years, as the number of users has grown, social media platforms have incorporated new features. Text interactions in social media are becoming more multimodal through the integration of images, videos, and more. Due to people's varied cultural backgrounds, interactions on these platforms expanded as the number of users grew. While most users were relatively young in the early days of social networks, people of all ages are now participating in these platforms. Big Data

C. Lagrini (✉) · A. Idrissi
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: chahrazad.lagrini@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,
Studies in Computational Intelligence 1102,
https://doi.org/10.1007/978-3-031-33309-5_11

137

researchers were drawn to social media platforms because of the variety of user interaction data that can be found there. These researchers used this data to retrieve information, identify interaction patterns, and assess sentiment. Sentiment analysis is a fairly mature field in computer science, but most often treat it as a binary classification problem. Clearly, human emotions are so complex that it is useful to consider categories other than “positive” and “negative.” Recently, researchers have been trying to classify positive emotions into finer categories, such as happiness and surprise. As well as negative emotions into finer categories such as sadness and anger. In contrast, sarcasm has received less attention as a form of emotion. The recognition of sarcasm is acknowledged in the published works to be a more intricate process as compared to identifying emotions. Requires contextual information, making detection difficult. Sarcasm can deceive users about what other users are thinking if it is not seen as a normal aspect of user engagement on social media, leading to misunderstandings and “internet debates.”

2 Previous Work

Text mining is the main source of the current research in computer science for sarcasm detection. Sarcasm studies are typically derived from the prevailing research in linguistics or the perspectives of scholars regarding sarcasm usage. Sarcasm can be better understood through the lens of linguistics when explored from three different angles:

– First of all, for sentiment analysis, there are two primary emotions that might be associated with sarcasm: a visible emotion and an intentional emotion.

Visible emotion in this context alludes to the sentiment expressed by a statement’s literal interpretation. However, what a person attempts to imply and expects the audience to understand is known as their desired or intentional emotion.

According to some studies, the visible emotion of the declaration is positive while the intentional emotion is negative. Some research, however, claim against such a broad statement. They argue that when expressing sarcasm, the surface emotion and intentional emotion of the declaration will be different, which is coherent with the previous group of researchers’ perspectives.

– Secondly, using sarcasm in communication goes against Grice’s rules for constructive conversation, specifically the quality maxim calls for one to be honest and fair in a conversation by not providing any false information, sarcasm is seen as a false statement about thoughts and feelings. Then, the maxim of manners calls for being succinct, organized, and unambiguous in one’s speech; sarcasm, by its very nature, raises questions about what is being said.

– And at last, linguists propose that certain cues in common patterns frequently accompany sarcasm. For instance, they contend that changes in speech amplitude and rate, as well as nonverbal cues like quotes in the air, are reliable signs of sarcasm.

Many researches in psychology used Grice Maxims to detect sarcasm and/or applied these linguistics theories to understand context,, analyse data like uppercase text, discern conflicting attitudes in various parts of larger speeches... Yet, as these language researches focus on interaction in real life, they may be inaccurate in social media. Therefore, AI researchers do not employ the paraverbal characteristics or indicators of sarcasm and instead focus on the verbal features of sarcasm presented through text.

Until now, according to the article “The Sarcasm Detection in News Headlines Based on Machine Learning Technology”, published in 2021, “the best accuracy was shown by the neural network model with scales created using the Glove method: 80.5%. The following model of the Bayesian classifier using TF-IDF vectors: 78.9%. Next is a model of a neural network with scales created using the Word2Vec method: 77.2%. Following neural network without weights: 76.6%. The lowest logical regression model in the table using TF-IDF vectors: 74%” [1]. The authors of this article worked on the same dataset we used, available on Kaggle and GitHub websites, where two American websites’ news headlines are compiled.

3 Methodology

The realization of this project requires a programming language. We therefore chose the most used language in the NLP: Python.

Python is one of the most popular programming languages, created by Guido van Rossum. It provides a wide spectrum of libraries that support many programming tasks. These libraries are easy to install, open source and well documented.

3.1 Data Preprocessing

Data Preprocessing is a Data Science technique that transforms raw data into data more suitable for study. This is a preliminary step that requires all available information to organize, sort and merge.

We applied several preprocessing techniques according to the following steps:

- Removing links: There is no extra information on these Links. During the preprocessing, they are eliminated.
- Removing punctuations: Punctuations are removed before extracting features from the text.
- Lowering the case: We will convert all uppercase letters to lowercase. This operation allows us to obtain uniform data.
- Removing stop words: This manipulation is often performed in text processing. Stop words are very common words in the language studied (“a”, “in”, “the”... in English)

which do not provide any informative value for understanding the “meaning” of a document. They are often words which are therefore very frequent and slow down the work.

– Stemming and Lemmatization: Lemmatization looks for the base form of any supplied word, for example, it transforms a conjugated verb to its infinitive or to transform a noun to the masculine singular. The basic form of a word will help to identify user’s feelings.

Stemming or rooting is the reduction of the word to its root. We therefore remove the prefixes, suffixes and others to keep in the end only the original word. For example, in our dataset based on tweets in English, if we have a word like “historical”, stemming will change it to “history, or stemming will change it to “histori”. Now, you may have the impression that the story is the best representation, but the computer doesn’t understand either word.

Lemmatization works best if you want higher accuracy, because there are words that are completely different from each other but have a similar root. On the other hand, they will not have the same lemmatized form. So, it’s a compromise between speed and performance. You want a fast program; stemming is your option. You want accuracy, opt for lemmatization.

– Bag of Words: A bag of words or Bow model is a method for collecting features from text so that text input can be used with ML algorithms. Every document is transformed into a vector image.

– Tf-IDF: Term Frequency-Inverse Document Frequency (Tf-IDF) measures the relevance of a term for a document from a corpus and is based on the idea that words that appear several times in a document provide a lot of information, while words that appear in all documents provide little or no information about a document. Tf-Idf increases the importance of words with many occurrences in a document and decreases the importance of words that appear in many different documents.

– Tokenization: Tokenization is the act of dividing a text into a list of words or tokens. Tokenization helps to create a list of words from a tweet. These tokens aid in context comprehension or the creation of an NLP model. By examining the order of the words, tokenization aids in interpreting the meaning of text. For example, the text “Twitter is down” can be tokenized into “Twitter”, “is”, “down”.

3.2 *Sarcasm Detection Classifiers*

In what follows we will use supervised classifiers to detect sarcasm.

– Naïve Bayes:

The naive bayes classification method is a supervised language algorithm that makes the “naive” assumption that all features are independent.

The Naïve Bayes Classifier is the simplest and fastest classification algorithm. This learning model requires only a small number of training samples compared to other models.

The Naïve Bayes Classifier assumes that the variables are independent, which is still not true in real cases. Indeed, if the correlation between the characteristics is high, this learning model will give a poor performance.

The Naïve Bayes formula is:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

– Support Vector Machine:

SVM stands for Support Vector Machine.

“SVMs are a set of unsupervised ML learning methods that can be used for classification, anomaly detection, and regression. One of the strengths of SVMs is that they are very efficient when your data is high dimensional and also efficient in cases where the number of dimensions is greater than the number of samples. SVMs are particularly good at solving text and hypertext categorization since the application decreases the use of label formation opportunities” (Witten and Frank, 2005).

The SVM classifies categories by dividing them by a hyperplane. A hyperplane is a subspace one dimension smaller than the surrounding space. In n-dimensional space, where n refers to the number of features, each piece of data is represented as a point, with each feature’s value corresponding to a separate coordinate. The hyper-plane is then used to carry out the classification, thereby separating the two classes.

– Random Forest:

As its name suggests, several distinct decision trees constitute Random Forest, which functions as a whole. The class with the most votes determines the predictions for the model among each tree in the Random Forest.

The accuracy and prevention of overfitting issues improve with increasing forest density.

– Decision Tree:

A decision tree is a flowchart that is created around a single central concept and further divided based on the outcomes of your choices. The name of this tool is explained by its strong resemblance to a tree and its many branches.

Such trees allow you to visually represent and analyse the results, expenses, and alternative outcomes of a tough decision. This tool is ideal for estimating the value of each expected result based on the decision made, but also for clarifying the consequences of your choice. Then compare the different results to quickly determine the most sensible and effective course of action. A decision tree can be used to manage

expenses, find opportunities, and solve problems. The decision tree is also called classification tree, decision tree, or decision tree.

– Recurrent Neural Network:

The Recurrent Neural Network (RNN) is a set of nodes that share the same weight within each network layer. Also, to lessen the loss, the weights and basis are separately modified throughout gradient descent.

In order to forecast the output of a layer, the foundation of RNN is the idea that the output of one layer should be kept and brought back to the input.

RNNs are made to recognise the sequential traits and data usage patterns necessary to forecast the following most likely scenario.

RNNs are used in the development of models that simulate the activity of the human brain system. They are particularly powerful in scenarios involving context in predicting an outcome.

4 Results

The evaluation of the models will allow us to compare them and choose the best model for the final code.

We used four performance analysis metrics to evaluate the results as below:

TP, TN, FP and FN stand for:

TP = True Positive.

TN = True Negative.

FP = False Positive.

FN = False Negative.

– Accuracy: It displays the proportion between all the accurate predictions and all the potential predictions in the test data. The accuracy is determined by dividing the total of the true positives and true negatives divided by the total of the entire positive and negative classes.

$$\text{Accuracy} = (\text{TN} + \text{TP}) / \text{Total observations}$$

– Precision: The precision metric displays the proportion of real positives to all real positives in the sample. Also, the precision shows how many forecasts for the Positive class are positive. This metric shows the proportion of true positives (TP) to the total of true positives and false positives (FP).

$$\text{Accuracy} = \text{TP} / (\text{TP} + \text{FP})$$

Fig. 1 Evaluation of Models using Stemming

	Model	Accuracy	F1 Score	Recall	Precision
0	Naive Bayes	0.78	0.76	0.80	0.72
1	SVM	0.79	0.78	0.79	0.76
2	Decision Tree	0.70	0.69	0.68	0.71
3	Random Forest	0.76	0.73	0.79	0.69
4	Recurrent Neural Network	0.80	0.78	0.80	0.76

Fig. 2 Evaluation of Models using Lemmatization

	Model	Accuracy	F1 Score	Recall	Precision
0	Naive Bayes	0.80	0.77	0.83	0.72
1	SVM	0.81	0.80	0.82	0.77
2	Decision Tree	0.71	0.70	0.70	0.69
3	Random Forest	0.77	0.74	0.81	0.68
4	Recurrent Neural Network	0.84	0.83	0.82	0.84

– Recall: Recall is the proportion of the True Positive class that was correctly identified. It displays TP (true positive) divided by the product of TP and FN (false negative).

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

– F1 Score: The weighted or harmonic average of recall and precision is shown by the F1 score, often well-known as the F-Score. Thus, both false positives and false negatives are considered.

$$\text{F - Score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

By applying these evaluation modes to the models used, we notice that the performance was better using lemmatization. The most efficient model is RNN with an accuracy of 84% followed by SVM with an accuracy of 81% (Figs. 1 and 2).

5 Conclusion and Future Work

In this work, we tried to detect sarcasm in tweets using specific models and our results were promising.

However, this is still insufficient since the detection of sarcasm is not an obvious task. Sometimes, we find ourselves faced with several criteria to analyze to simplify detection, such as: the nature of the user’s profile (if it is often sarcastic or never), the user’s beliefs, ideology and many more...

In the future work, it will be better to, at least, adopt the addition of emojis, memes, as well as the introduction of new dictionaries for the Slang used in social media will bring a plus. Moreover, the introduction of the methods exposed in [17–28] would be interesting to improve our investigations.

References

1. M. Zanchak, V. Vysotka, S. Albota, The Sarcasm detection in news headlines based on machine learning technologie, in 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), September 2021
2. J. Tepperman, D. Traum, S. Narayanan, “yeah right”: Sarcasm recognition for spoken dialogue systems, in *Ninth International Conference on Spoken Language Processing* (2006)
3. A. Rathee, Sentiment analysis: what is it and why does it Matter? (2018)
4. P. Haiyun, E. Cambria, A. Hussain, A review of sentiment analysis research in Chinese language
5. C. Nav Chandra, S. Gupta, R. Pahade, Sentiment analysis and its challenges. *IJERT* **4**(3) (2015)
6. M. Ebrahimi, A.H. Yazdavar, A. Sheth, Challenges of sentiment analysis for dynamic events, in *IEEE Intell. Syst.* **32**(5) (2017)
7. L. Chen, et al., Extracting diverse sentiment expressions with target-dependent polarity from Twitter, in *Sixth International AAI Conference on Weblogs and Social Media* (2012)
8. I. Roldós, Sentiment analysis applications and examples (2020)
9. T.M. Mitchell, *The Discipline of Machine Learning* (Carnegie Mellon University, Pittsburg, PA, 2006)
10. P. Domingos, *A Few Useful Things to Know about Machine Learning* (University of Washington, Seattle, WA, 2012)
11. R. Swanson, S. Lukin, L. Eisenberg, T.C. Corcoran, M.A. Walker, Getting reliable annotations for sarcasm in online dialogues, arXiv preprint [arXiv:1709.01042](https://arxiv.org/abs/1709.01042) (2017)
12. R. Schifanella, P. de Juan, J. Tetreault, L. Cao, Detecting sarcasm in multimodal social platforms, in *Proceedings of the 2016 ACM on Multimedia Conference* (ACM, 2016), pp. 1136–1145.
13. P.M. Pexman, K.R. Rostad, C.A. McMorris, E.A. Climie, J. Stowkowy, M.R. Glenwright, processing of ironic language in children with high functioning autism spectrum disorder. *J. Autism Dev. Disord.* **41**, 1097–1112 (2011). <https://doi.org/10.1007/s10803-010-1131-7>
14. J. Li, H.R. Rao, Twitter as a rapid response news service: an exploration in the context of the 2008 China Earthquake. *Electron. J. Inf. Syst. Dev. Ctries.* **42** (2010)
15. A multimodal approach to sarcasm detection on social media, by Dito Das (2019)
16. Detection on sarcasm using machine learning classifiers and rule based approach by K. Sentamilselvan, P. Suresh, G.K. Kamalam, S. Mahendran, D. Aneri, published in 2021
17. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data. Min.* (2017)
18. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High. Perform. Syst. Archit.* **9**(2-3):136–148 (2020).
19. M. Abourezq, A. Idrissi, Integration of QoS Aspects in the Cloud Service Research and Selection System. *Int. J. Adv. Comput. Sci. Appl.* **6**(6) (2015)
20. A. Idrissi, CM. Li, JF. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks, in 18th IEEE International Conference on Tools with Artificial Intelligence, (2006)
21. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. arXiv preprint [arXiv: 1709.10372](https://arxiv.org/abs/1709.10372). (2015)
22. A. Idrissi, K. Elhandri, H. Rehioui, M. Abourezq, Top-k and Skyline for Cloud Services Research and Selection System. *Int. conf. on Big Data Adv. Wireless technol* (2016)

23. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10** (2020)
24. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework *IEEE Syst. J.* **15**(4):4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
25. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Rob. Intel. Syst.* **14**(3):65–70, (2020)
26. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment, in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (2016)
27. M. Essadqi, A. Idrissi, A. Amarir, An Effective Oriented Genetic Algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput Sci.* **127**, 170–179 (2018)
28. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electr. Eng.* **73**, 289–303 (2018)

School Dropout Prediction using Machine Learning Algorithms



Said Ouabou, Abdellah Idrissi , Abdeslam Daoudi, and Moulay Ahmed Bekri

Abstract The objective of this article is to develop a method that integrates Machine Learning models to predict whether a student is at risk of dropping out or not, based on a set of data. First, we proceeded to collect, analyze and prepare a set of data, to make them usable by machine learning algorithms. Second, we tested this data on several algorithms such as Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, Neural Networks, and K-Nearest Neighbours. Then, we exposed the evaluation and the deployment of these models. Finally, we have developed a web application that integrates these models, makes predictions, visualizes this data and models its performance.

Keywords Machine learning · Artificial Intelligence · Dropping out of school

1 Introduction

The education system of the national education system has recognized a great evolution thanks to structural reforms that aim at performance and manage to give it the quality expected from its teaching.

Admittedly, a lot of effort has been made, but it still faces various challenges such as school dropout, which is one of the major challenges facing the Ministry of National Education “because of its negative impact on dropouts and its economic and social cost...” [1]. According to statistics, 304,558 students dropped out of school in the 2021–2022 school year, an increase of 0.8% in the primary cycle compared to the previous year.

S. Ouabou (✉) · A. Idrissi · A. Daoudi · M. A. Bekri
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: said-ouabou@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

Indeed, to help students whether their social, family or economic situations prove that they may be dropouts and to avoid the multiple consequences of this phenomenon either cognitive, social or economic, and by integrating new technologies, including machine learning techniques, we worked on the development of a web application that uses machine learning algorithms to predict the risk of dropping out of school, which will contribute at the right time and find suitable solutions.

The essential objective of this paper is to create a model of prediction of school dropout and to create a web application for this problem.

2 Proposed Solution and Approach

A. Type of problem to be solved

Machine learning is a computer programming technique that uses statistical probabilities to give computers the ability to learn for themselves without explicit programming.

There are four types of machine learning: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Since we have a problem that we want to predict whether a student is at risk of dropping out or not, and by supervising and training the algorithm from a database, then the problem to be solved is of a supervised learning type, of which there are two categories of problem: regression, classification.

The problem mentioned above is considered a classification problem since this type of algorithm is used to classify data in two groups (binary classification) and not continuous values.

B. The procedure to be followed

See Fig. 1.

3 Data Collection and Analysis

A. Data collection

Data collection is a crucial step in such a project, because without relevant data, one cannot have satisfactory results, even with the best algorithms of classification. This phase is therefore crucial and we must devote time to it.

B. Data presentation and analysis

After cleaning and deleting the invalid data, the database is ready for the analysis phase.

Based on the analysis of this database, the following information can be extracted: We have two types of variables (Tables 1 and 2):

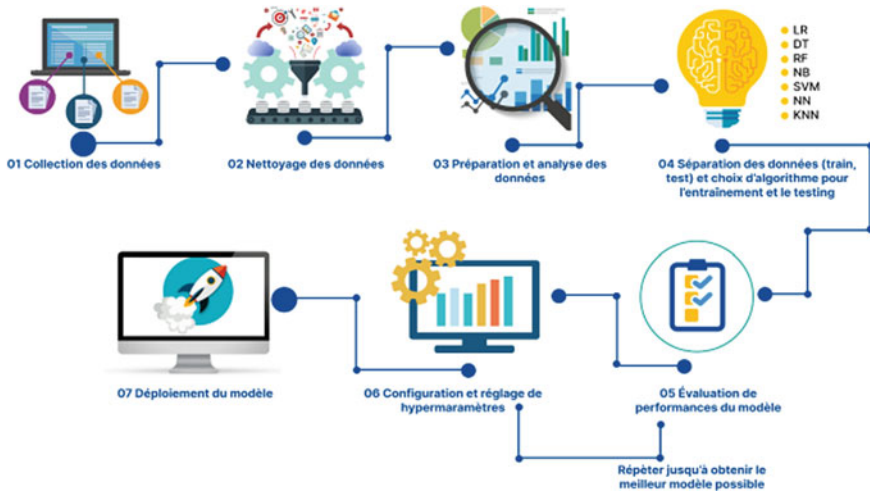


Fig. 1. Flowchart showing the steps of building the ML model [20].

Table 1 Continuous variables and their maximum and minimum values

Continuous variables	Max value	Min value
'pct_days_absent'	3153.0	0.0
gpa	9.96	0.0

Table 2 Number of dropout observations

Dropout	Total	(%)
No	42,227	82.42
Yes	9005	17.57

Continuous: ['pct_days_absent', 'gpa'].

Categorical: ['dropout', 'gender', 'gifted', 'rural/urban', 'still_enrolled'].

4 Results and Interpretation

A. Measurement Metric

(a) Cross-validation in ML

We use repeated stratified k-fold cross validation which allows us to repeat the k-fold cross validation procedure several times and report the average result on all folds (fold) of all executions.

Here we have defined the scoring ‘accuracy’ we can choose precision, recall or f1-score (Table 3).

(b) ROC curve

Figures 2, 3, 4, 5, 6, and 7.

From these results, it can be inferred that these models are acceptable (0.71–0.78), LR has the high value 0.788.

In this section, we discussed the model performance evaluation phase, a very important step in a ML project. Cross-validation, confusion matrix, and ROC curve were used for this phase. The results were good (between 85% and 86.25%).

Table 3 Repeated stratified K fold on different ML models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR	85.9	76.8	29.1	42.2
DT	85.5	78.9	25.6	38.4
RF	85.1	61.7	42.1	49.8
NB	85.5	63.1	46.7	53.6
SVM	85.4	84.4	23.1	36.2
NN	86.25	73.75	36.23	48.59
KNN	85.0	64.1	37.7	47.5

Fig. 2 ROC LR

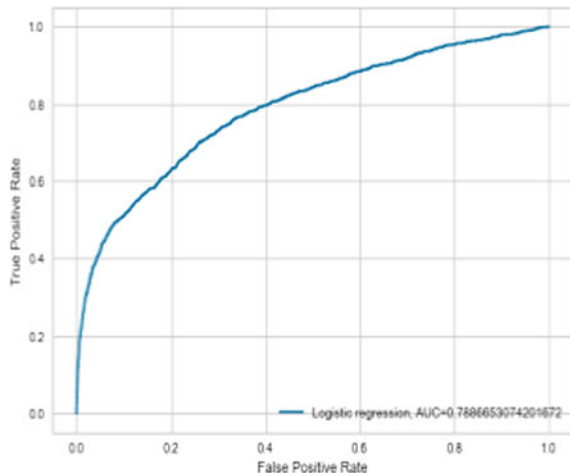


Fig. 3 ROC DT

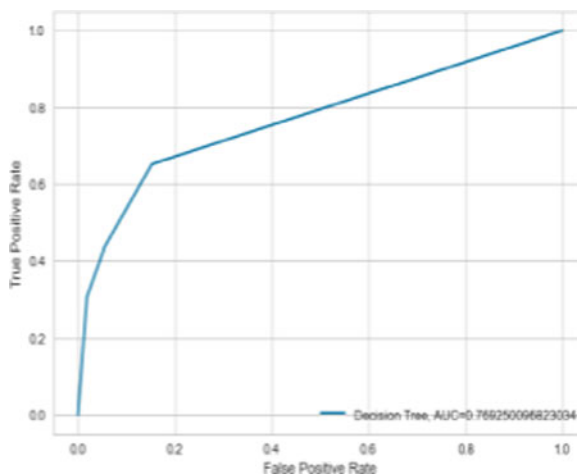


Fig. 4 ROC RF

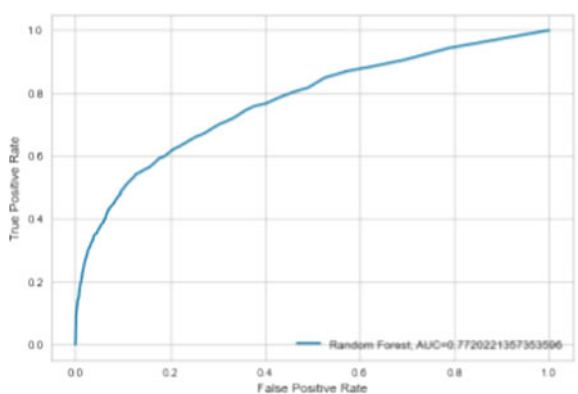


Fig. 5 ROC NB

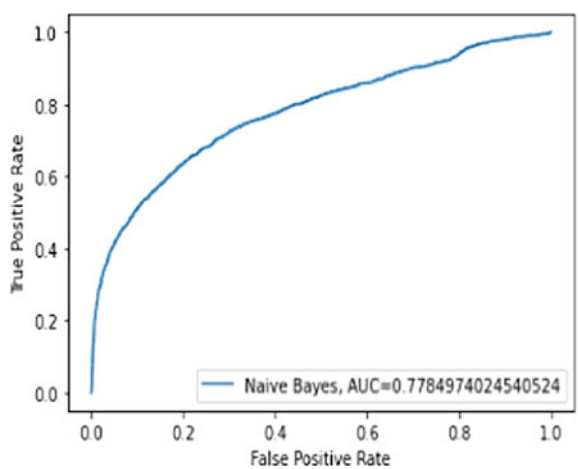


Fig. 6 ROC SVM

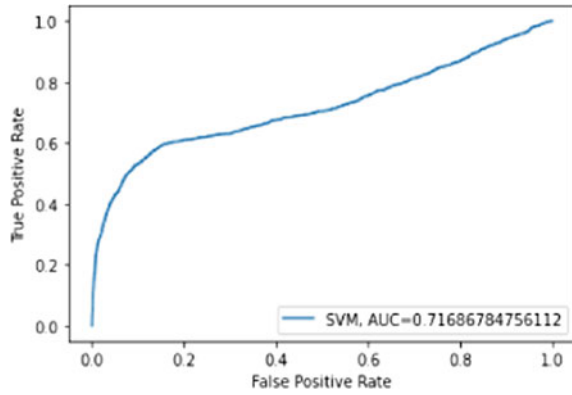
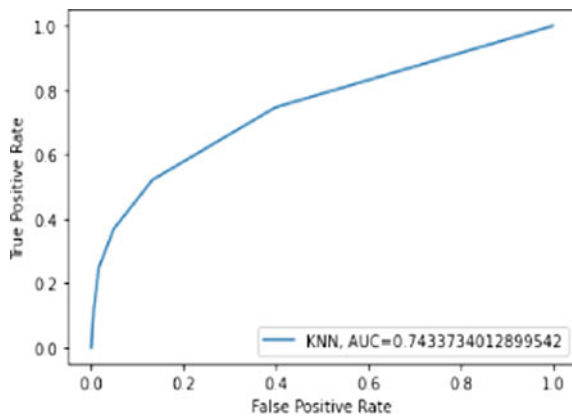


Fig. 7 ROC KNN



5 Implementation

Validation of a ML model is not usually the last step in a ML project. Indeed, the deployment of the model remains a crucial stage in the life cycle of ML. The model must be deployed in order to be available in an environment where it can make predictions.

In this chapter we focus on the tools and technologies used for the realization of this paper, then the interfaces of the application.

A. Tools

This section focuses on the development tools used to carry out the paper.

(a) *Anaconda*

Anaconda is a free and open source distribution tool for Python and R programming.

This software is essential for all developers in the field of data science. It allows to collect and transform data on a large scale thanks to the tools it offers [2].

(b) Jupyter Notebook

The use of Jupyter Notebook allows a better understanding of the results of a statistical analysis. It is common to share such results in the form of a static graph, however, this practice has its limitations. With the Jupyter Notebook, users can deepen the analysis by playing with data or graphics in a totally interactive way [3].

(c) Visual studio code

Visual Studio Code is a free open source code editor developed by Microsoft. It runs on Windows, Mac OS and Linux. It provides developers with both an integrated development environment with tools to advance technical projects, from editing to construction to debugging. It supports several dozen programming languages such as C, C++, C#, Python, HTML, CSS, PHP, Javascript, Markdown, Java, etc.

Visual Studio Code also allows developers to create and use extensions through its API, to customize their use of the tool [4].

(d) Streamlit

Streamlit is a Python open source framework designed for machine learning engineers and data scientists. It enables the creation of web applications that can easily integrate machine learning models and data visualization tools.

It offers a very interesting alternative for building and sharing web applications, and allows to create front-end in an innovative way.

The Streamlit framework can connect to multiple software platforms. It is compatible with most data visualization frameworks (matplotlib, plotly, seaborn, ...) and machine learning (pandas, pytorch, etc.) [5].

(e) SQLite3

SQLite is a lightweight relational DBMS designed specifically for local data storage. It is the most used DBMS in the world, It requires no configuration, nor the installation of a database server. It is based on writing in C, an imperative programming language, and on accessibility via SQL (Structured Query Language).

SQLite works in different environments including Linux, Windows, macOS, Android and iOS [6].

B. Deployment

There are several approaches to deploy a model, it is possible to embed it directly in a web application (streamlit). To do this, the model must be saved for reuse, using the Pickle module and the dump() method (Fig. 8).

Pickle for the serialization of python objects.

The pickle module implements binary protocols for serializing and de-serializing Python objects. Serialization is the process by which a hierarchy of Python objects is converted into byte streams. De-serialization is the reverse operation [7].

The dump() method saves the object to the file specified in the arguments (Fig. 9).

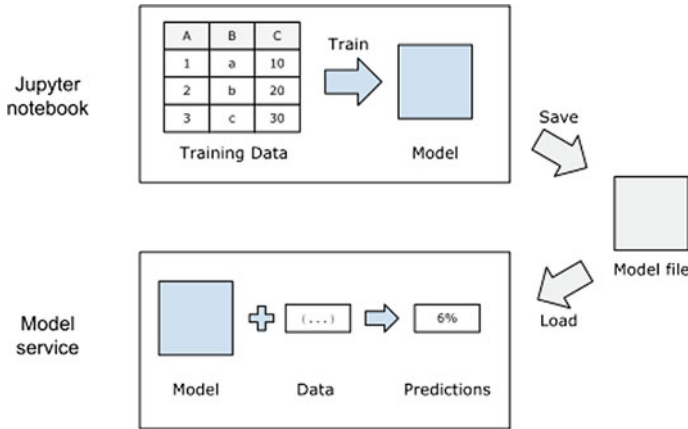


Fig. 8 ML model registration process [21].

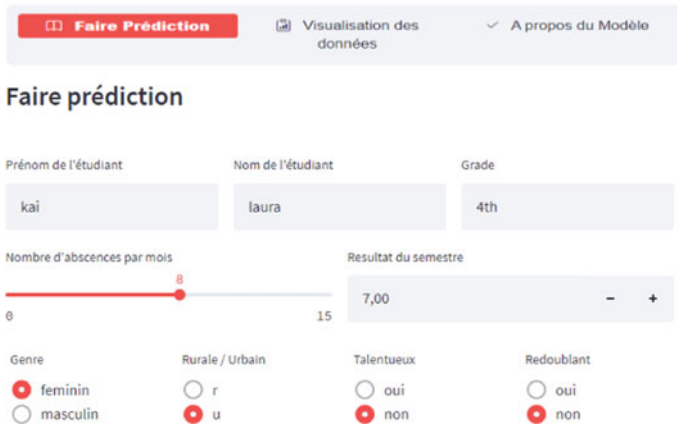


Fig. 9 Interface prediction 1

(a) Interfaces Graphiques

The user enters the student's information and makes a prediction according to the chosen model, he can see both the results of all models (button All models) (Fig. 10).

Based on this result, it can be said that NN is the only one who predicted that the student would drop out of school, unlike other models.

Save saves this prediction, view data displays a table containing student info and the result of each model.

Fig. 10 Interface prediction
2



(b) Data visualization interface

In this last part, we described the different development tools used, so we deployed the model in Streamlit, and we presented the prototype realized. The various graphical interfaces of the web application summarizing the operation of the application (Fig. 11).

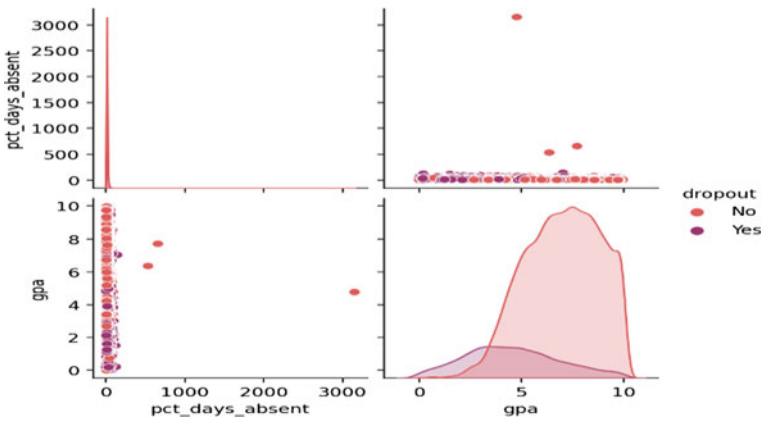


Fig. 11 Pairplot

6 Conclusion

In order to create a web application to predict school dropout using the various machine learning techniques, in terms of results, the evaluation of model performance revealed that accuracy (accuracy) These models are good enough to predict whether a student is likely to drop out or not.

Indeed, during this paper, we studied and implemented the different algorithms of machine learning, more precisely classification algorithms, such as logistic regression, decision tree, decision tree forest, Naïve Bayesian, support vector machine, neural networks, K closer neighbors, not to mention the data analysis part and the deployment of models.

Moreover, the paper has put into practice various technological tools such as Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, Keras, Tensorflow and Streamlit offered by the programming language Python.

Based on the evaluation of the models and the comparative study conducted, it was concluded that:

Logistic regression and neural networks gave better results compared to other algorithms.

We also see that data quality plays an important role in achieving a good result.

Finally, according to the work done, it was found that the subject dealt with is very interesting and important, school dropout is not a fatality, if the actors are aware of all the facets of the phenomenon in order to detect a school dropout situation at its first signs. This leads us to think about expanding the work on model improvement from the analysis of data resulting from prediction and comparing it with real results to improve the model. In this context, we claim to rely on the results provided in [8–19] to improve our approach and thus be able to make it more efficient.

References

1. B. Hassan, Education: benmoussa face au défi de l'abandon scolaire. le360.ma. [En ligne] 03 Septembre 2022. [Citation : 12 Juin 2022.] <https://fr.le360.ma/politique/education-benmoussa-face-au-defi-de-labandon-scolaire-256317>
2. Anaconda python. data-transitionnumerique.com. [En ligne] [Citation : 05 Juin 2022.] <https://www.data-transitionnumerique.com/anaconda-python/>
3. Jupyter Notebook. lebigdata.fr. [En ligne] [Citation : 05 Juin 2022.] <https://www.lebigdata.fr/jupyter-notebook>
4. Visual studio code. blogdumoderateur.com. [En ligne] [Citation : 05 Juin 2022.] <https://www.blogdumoderateur.com/tools/visual-studio-code/>
5. Streamlit ou l'outil pour présenter votre travail de Machine Learning. datascientest.com. [En ligne] 05 Avril 2022. [Citation : 05 Juin 2022.] <https://datascientest.com/streamlit-ou-loutil-pour-presenter-votre-travail-de-machine-learning#:~:text=Une%20bibrairie%20Python%20open%20source,sur%20les%20besoins%20de%20chacun>.
6. sqlite définition. journaldunet.fr. [En ligne] 08 Janvier 2019. [Citation : 05 Juin 2022.] <https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203607-sqlite-definition/>.

7. Pickle.docs.python.org. [En ligne] [Citation : 11 Juin 2022.] <https://docs.python.org/fr/3/library/pickle.html?highlight=getattr#:-:text=Le%20module%20pickle%20impl%C3%A9ment%20des,convertie%20en%20flux%20d'octets>
8. A. Idrissi, K. Elhandri, H. Rehioui, M. Abourezq, Top-k and Skyline for Cloud Services Research and Selection System. *Int. Conf. Big Data and Adv. Wireless technol* (2016)
9. A. Idrissi, CM. Li, JF. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks, in *18th IEEE International Conference on Tools with Artificial Intelligence* (2006)
10. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. *arXiv preprint arXiv: 1709.10372*. (2015)
11. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data. Min* (2017)
12. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet. Technol. Secured Trans.* **10**, (2020)
13. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4):4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
14. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment, in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (2016)
15. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Rob. Intell. Syst.* **14**(3):65–70 (2020)
16. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electr. Eng.* **73**:289–303 (2018)
17. M. Abourezq, A. Idrissi, Integration of QoS Aspects in the Cloud Service Research and Selection System. *Int. J. Adv. Comput. Sci. Appl.* **6**(6) (2015)
18. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Architect* **9**(2–3):136–148, (2020)
19. M. Essadqi, A. Idrissi, A. Amarir, An Effective Oriented Genetic Algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**:170–179 (2018)
20. Etapes Machine Learning: comprendre le processus, Talend - A Leader in Data Integration & Data Integrity. <https://www.talend.com/fr/resources/etapes-machine-learning/> [Citation : 11 juin 2022]
21. A. Grigorev, Introduction to machine learning. *Machine Learning Bookcamp: Build a portfolio of real-life projects.* <https://livebook.manning.com/book/machine-learning-bookcamp/chapter-1/> [Citation : June 11, 2022]

Survival Prediction in Patients with Nasopharyngeal Cancer Using some Machine Learning Methods



Abdellah Idrissi, Hasna Lakrim, and Mehdi Bouskri

Abstract The staging system of tumor lymph node metastases currently ranks first in predicting the prognosis of nasopharyngeal cancer. However, this system alone is not sufficient for this type of prediction and patients at the same stage may show significant clinical heterogeneity and distinct oncological findings. The plasma Epstein-Barr virus (EBV) DNA titer remains the only clinically useful biomarker in patients with nasopharyngeal cancer. The factors affecting the prognosis of nasopharyngeal cancer and the total effect of these factors on the prognosis of this cancer remain an open subject for further investigation. To provide answers to related questions, this study examined prognostic factors for survival using machine-learning techniques. In this project, we have the following result: An accuracy of 97%, a precision of 100%, a recall of 90% and 95% for the F1 score, using Random Forest algorithm as well as the Voting Classifier algorithm.

Keywords Nasopharyngeal cancer · Radiotherapy · Prediction · Machine learning · Random forest · Voting classifier

1 Introduction

Nasopharyngeal cancer is difficult to detect early, and this may be due to the fact that nasopharyngeal cancer is not easily diagnosed, given the likely symptoms of other, more common conditions. Treatment for nasopharyngeal cancer usually includes radiation therapy, chemotherapy, or both. In general, patients with this cancer have a poor prognosis and low survival rate.

A. Idrissi · H. Lakrim · M. Bouskri (✉)
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: mehdi.bouskri@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

This reaffirms the importance of survival prediction or survival analysis. With this analysis, medical professionals can give patients hope, enable better communication, and better prepare for impending death, thereby avoiding futile treatments and helping patients, families, and caregivers. We can provide the best quality palliative care.

2 Nasopharyngeal Cancer

(1) Definition and Risk Factors

Nasopharyngeal cancer (nasopharynx or cavity) is a rare cancer that forms in the part of the throat that connects the back of the nose to the mouth (nasopharynx). This cancer is more common in southern China and Southeast Asia.

(2) Symptoms

Most people with nasopharyngeal cancer notice a lump or mass in the back of the neck and go to see a doctor. There may be bumps on both sides of the neck towards the back. The bump is usually not tender or painful. They are caused by cancer that has spread to lymph nodes in the neck, causing them to swell. Many symptoms and signs of nasopharyngeal cancer are more commonly caused by other less serious conditions. However, if you have these problems, it is important to see your doctor right away to determine the cause and treat it if necessary.

(3) Diagnostic

Tests to diagnose nasopharyngeal carcinoma

Tests and procedures used to diagnose nasopharyngeal carcinoma include:

- Physical examination. Diagnosis of nasopharyngeal carcinoma usually begins with a general examination.
- Examination using a camera to see inside the nasopharynx, and nasal endoscopy may require local anesthesia.
- Test to remove a sample of suspicious cells.

Tests to determine the stage of cancer

Once the diagnosis is confirmed, the doctor will order other tests to determine the extent (stage) of the cancer, such as imaging tests.

Imaging tests may include:

- Computed tomography (CT).
- Magnetic Resonance Imaging (MRI).
- Positron Emission Tomography (PET).
- X-ray.

The stage is used along with several other factors to determine treatment planning and prognosis. A lower number means the cancer is smaller and confined to the nasopharynx. A higher number means that cancer has spread beyond the nasopharynx to lymph nodes in the neck or other parts of the body.

(4) Treatment

The patient and doctor plan the treatment together based on several factors, including: Cancer stage, treatment goals, general health status, and acceptable side effects.

Treatment of nasopharyngeal cancer usually begins with radiation therapy or a combination of radiation therapy and chemotherapy.

- **Chemotherapy**

Chemotherapy is a drug treatment that uses chemicals to kill cancer cells. Chemotherapy drugs can be given as tablets, intravenously, or both. Chemotherapy given at the same time as radiation therapy. When the two treatments are combined, chemotherapy improves the effectiveness of radiation therapy. This combination therapy is called combination therapy or chemoradiation therapy.

- **Surgery**

Surgery is rarely used to treat nasopharyngeal cancer. It can be used to remove cancerous lymph nodes in the neck. In some cases, it can be used to remove tumors from the nasopharynx. This usually requires the surgeon to make an incision in the palate to access an area where cancerous tissue can be removed.

(5) Can nasopharyngeal cancer be prevented?

Although many cases of nasopharyngeal cancer cannot be prevented, the following measures can help reduce the risk of nasopharyngeal cancer.

- Avoid fish and salty meat.
- Avoid Smoking.
- Avoid consuming large amounts of alcohol.

3 Technical Study

A. Dataset

The trial will include 143 patients between the ages of 12 and 80. Project dataset typically include:

- Clinical prognostic markers:
 - o Age, alcohol consumption, lymph node involvement, metastatic status, disease stage.

- Metabolic and anatomic prognostic markers:
 - o N-SUV max, NTR and MTV derived from PET/CT [18F]FDG.
 - o Bone invasion of skull base as defined by HN MRI.
- Biological prognostic markers
 - o The dynamic change of viral load in EBV DNA in pre-, end- and post-treatment.

Since this dataset contains missing values, we searched for those values and found some columns with multiple missing values.

These columns have since been removed.

B. Classifiers Used

1. Logistic Regression

Logistic regression is the primary classification method. It belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. Logistic regression is fast and relatively easy. It's basically a binary classification method, but it can also be applied to multiclass problems. The sigmoid function and the natural logarithm are used to describe logistic regression.

The sigmoid function has values very close to 0 or 1 over most of its domain. This fact makes it more suitable for use in classification methods.

2. K-Nearest Neighbor (KNN)

The KNN (K Nearest Neighbor) algorithm is a supervised machine learning algorithm that can be used to solve both classification and regression problems.

The K Nearest Neighbor algorithm estimates the probability that a data point belongs to one of two groups, based on the data point's closest neighbors.

KNN algorithms can be used for both classification and regression problems. This is classified as a lazy learner. In other words, instead of running training steps, it only stores a set of training data.

Furthermore, this means that all computations are performed when classification or prediction is done. It is also called memory-based learning because it stores all training data in memory.

KNNs have two main characteristics. First, KNN is a nonparametric algorithm. This means that no assumptions are made about the data set when using the model. Instead, the model is built entirely from the data provided.

Second, when using KNN, the data set is not split into training and test sets. This is because KNN does not distinguish between training and testing sets. All data is used when asking the model to make predictions.

To determine the class of unobserved observations, KNNs essentially use a voting mechanism and indicate that the class with the most votes is the class of relevant data points. If K is equal to 1, we only consider the nearest neighbors of the data points when determining class. If K equals 10, the 10 nearest neighbors are used.

KNN makes very accurate predictions. It can compete with the most accurate SOTA model (the state-of-the-art model). Therefore, KNN algorithms can be used for applications that require high accuracy, but at the same time do not require human-readable models. The accuracy of the prediction depends on the measured distance. This makes the KNN algorithm suitable for applications with sufficient domain knowledge. This understanding will help you choose an appropriate scale.

For best results, we recommend normalizing the data to the same scale. In general, normalization ranges from 0 to 3. Apart from that, the *hyper parameter* tuning of K and the distance metric are also important. You can test the KNN algorithm with different values of K using cross-validation techniques. The model with the highest accuracy can be considered the best option.

3. Support Vector Machine (SVM)

Support vector machine is one of the supervised machine learning technique that can be used for both classification and regression problems. However, it is often applied to classification problems. The SVM approach plots each data item as a point in n-dimensional space. Each feature value is the value of a separate coordinate, and then, the classification is performed by getting the hyperplane that separates the two classes very well.

Hyper-planes are decision limits that help classify data points. Data points on either side of the hyperplane can be viewed as separate classes. The dimension of the hyperplane also affects the number of lines. If the number of input features is 2, the hyperplane is just a line. If the number of input features is 3, the hyperplane will be a 2D plane. It becomes hard to imagine when the number of features exceeds three.

Support vectors are the closest points to the hyperplane and affect the hyperplane's position and orientation. We optimize the classifier margins by applying these support vectors. Deleting support vectors changes the position of the hyperplane. These are the points that help create the SVM.

4. Decision Trees

A simple and widely used classification approach is the decision tree (DT) classifier. Decision trees are a supervised machine learning approach used as a classification technique. A decision tree represents a classification model as a tree structure. The deeper the tree, the more complex the decision rules and the better the model. Divide the data set into smaller subsets, creating incrementally related decision trees. The final result is a tree with decision nodes and leaf nodes. Furthermore, the main purpose of this method is to develop a model that predicts the values of predictors where leaf nodes indicate class labels and decision trees use tree representations to solve the problem where features are represented internally.

- Decision tree assumptions:

- Initially, we consider the entire training as a root.
- Feature values should be categorical. If the values are persistent, they are converted to discrete values before building the model.

- Data records are distributed recursively according to characteristic values.
- Order features as root or interior nodes using mathematical techniques such as entropy, information gain, and gain.

5. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for ML classification and regression problems. It is based on the concept of ensemble learning, which combines multiple classifiers to solve complex problems and improve model performance.

A random forest is a classifier that takes a set of decision trees over different subsets of a given dataset and averages them to improve the prediction accuracy of that dataset. Instead of relying on decision trees, random forests get predictions from each tree. Predict the final output based on the majority vote of the predictions.

Here are some points that explain why the random forest algorithm should be used:

- Short training time compared to other algorithms.
- Predict outputs with high accuracy, even when processing large datasets efficiently.
- It can also maintain accuracy even when large amounts of data are missing.
- Advantages of Random Forest:
 - Random Forest can perform both classification and regression tasks.
 - Can handle large datasets with high dimensions.
 - This improves model accuracy and avoids overfitting problems.

- Disadvantages of Random Forest:

Random forests can be used for both classification and regression tasks, but are less suitable for regression tasks.

6. Ensemble learning:

Ensemble learning is the process of learning multiple models that: Classifiers or experts strategically generated and combined to solve specific problems in computer intelligence.

The goal of ensemble methods is to combine the predictions of multiple baseline estimators generated by a particular learning algorithm to improve the performance of a single estimator.

7. Voting Classifier:

A voting classifier is a machine learning estimator that trains different models or base estimators and makes predictions based on the aggregate results of each base

estimator. Aggregation criteria can be combined voting decisions for each estimator output. It has two types of voting criteria:

- Hard Voting: The vote is calculated on the planned exit class.
- Soft Voting: The vote is calculated on the predicted probability of the exit class.

Voting Classifier is a machine learning algorithm commonly used by Kagglers to improve model performance and rank.

Model interpretability is reduced because the Shap or Lime packages cannot interpret the model.

Unlike other models, scikit-learn does not provide an implementation for computing the best-performing functions of voting classifiers.

8. Stacking classifier:

The simplest form of stacking can be described as an ensemble learning technique that uses predictions from multiple classifiers as new features to train a Meta classifier. A Meta classifier can be any classifier.

4 Results

After building the correlation matrix, we found that there was no high correlation between features!

We then used five different machine learning classifiers as base models.

The most commonly used classifiers are:

Logistic regression; KNN classifier; Supports Vector Machine, Decision tree classifier; random forest classifier. Each model is tuned for optimal hyperparameters.

Random forest classifier proved to be the best model for this dataset and this classification problem.

We are interested in a good recall score because we don't want to reduce false negatives. Next, we added ensemble models.

For ensemble models, two classifiers are used:

- Voting Classifier.
- Stacking Classifier.

Note: Random forests are also ensemble classifiers (Adaboost, bagging classifiers, gradient boosting classifiers), but they are ensembles of the same base model, not a mixture of different types of classifiers, so they do not get in the way.

In this experiment, the voting classifier turned out to be the best model (Table 1).

Table 1 Final result

Model	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.86	0.80	0.80	0.80
KNN Classifier	0.86	0.75	0.60	1.00
Support Vector Machine	0.90	0.86	0.90	0.82
Decision Tree Classifier	0.69	0.31	0.20	0.67
Random Forest Classifier	0.97	0.95	0.90	1.00
Voting Classifier	0.97	0.95	0.90	1.00
Stacking Classifier	0.93	0.90	0.90	0.90

5 Conclusion

This work has added value as the results obtained were compared with the paper by Melek Akcay, Durmus Etiz and Alaattin Ozen, “**Using Machine Learning to Assess Prognosis of Nasopharyngeal Carcinoma**” published on 6 March 2020. In this article, the author applied the Gaussian Naive Bayes algorithm to achieve 88% accuracy, but by applying the random forest algorithm and voting classifier, our project achieved 97% accuracy, 100% precision and 90% Recall, and an F1 score of 95%. In addition, several algorithms can be used in this context. We can mention, among others, the methods presented in [1–12]. Our future work will focus on these different approaches.

References

1. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Rob. Intel. Syst.* **14**(3), 65–70 (2020)
2. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment, in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (2016)
3. M. Essadqi, A. Idrissi, A. Amarir, An effective oriented genetic algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia. Comput. Sci.* **127**, 170–179 (2018)
4. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electr. Eng.* **73**, 289–303 (2018)
5. M. Abourezq, A. Idrissi, Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comput. Sci. Appl.* **6**(6) (2015)
6. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Archit.* **9**(2–3), 136–148 (2020)
7. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data Min.* (2017)
8. A. Idrissi, C.M. Li, J.F. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks, in *18th IEEE International Conference on Tools with Artificial Intelligence* (2006)

9. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. arXiv preprint arXiv:1709.10372 (2015)
10. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
11. A. Idrissi, K. Elhandri, H. Rehioui, M. Abourezq, Top-k and skyline for cloud services research and selection system, in *International Conference on Big Data and Advanced Wireless Technologies* (2016)
12. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10** (2020)
13. <https://journals.sagepub.com/doi/full/10.1177/1533033820909829#tab-contributors>
14. <https://cancer.ca/fr/cancer-information/cancer-types/nasopharyngeal/prognosis-and-survival>
15. <https://www.mdpi.com/2075-4426/11/8/787>
16. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8398698/>
17. <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002730&ref=https://githubhelp.com>
18. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5402773/>
19. <https://tel.archives-ouvertes.fr/tel-03188077/document>
20. <https://www.sciencedirect.com/science/article/pii/S0007455121006822>
21. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7438151/>
22. 'Long-term cancer survival prediction using multimodal deep learning', Luis.A Vale-Silva et Karl Rohr, 29 juin 2021
23. 'Survival prediction models: an introduction to discrete-time modeling', Krithika Suresh, Cameron Severn & Debashis Ghosh, 26 juillet 2022
24. <https://jamanetwork.com/journals/jama/fullarticle/193279>
25. <https://www.webmd.com/cancer/nasopharyngeal-cancer>

Comparative Analysis of Skyline Algorithms used to Select Cloud Services Based on QoS



El Khammar Imane, Abdellah Idrissi, Mohamed El Ghmary,
and Kaoutar El Handri

Abstract The Cloud computing has grown in popularity. Consumers now have a wide variety of cloud services to choose from. The task of choosing the perfect cloud service from a variety of available options has become complex for non-IT users. The service selection must optimize the overall quality of service. A perfect cloud service must meet each customer's specific quality of service requirements. This research will focus on Skyline algorithms, We present a comparison study of three Skyline algorithm Block Nested Loop, Sort First Skyline and branch and bound Skyline algorithms for quality of service sensitive cloud service selection, we will end with some experiments which are presented to show the potential of these three algorithms to help developers of cloud service selection tools choose the right skyline algorithm for their work.

Keywords Cloud computing · Cloud service selection · Skyline algorithm · Block nested loop algorithm · Sort-filter skyline algorithm · Branch-and-bound skyline algorithm · Quality of service (QoS)

1 Introduction

As internet technology evolves cloud computing is fast growing in popularity [1]. Cloud computing lets users store, manage and even build models and structure. Cloud computing also provides computing services that permit them to access the

El K. Imane · A. Idrissi (✉)

IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: idrissi@um5r.ac.ma

El K. Imane

e-mail: imane.elkhammar@um5r.ac.ma

M. El Ghmary · Kaoutar El Handri

Department of Computer Science, FSDM, Sidi Mohamed Ben Abdellah University Fez, Fez,
Morocco

Internet's resources. It aims to make large data accessible, secure and reliable, as well as scalable and easily shared through web technology [2]. Cloud computing allows computing devices as well as platforms to work together and reduces the requirement for platform adequacy. Clients often employ a third-party cloud platform for their storage and computing requirements, paying exclusively for the services they actually use. Cloud computing offers a myriad of advantages over traditional computing, in terms of price, efficiency and reliability [3].

Cloud services are service that is hosted on a remote cloud infrastructure and accessible to users through the Internet by using a variety of methods of provisioning, like software (SaaS), (PaaS), (IaaS), and (DaaS) [4].

Community, private, public and hybrid deployment techniques are the most frequently used for cloud-based services [5].

Cloud computing has revolutionized the way that small, medium and large companies use IT services [6]. Cloud computing has many advantages that include lower IT costs, enhanced IT elasticity, massive resources, scalability, and flexibility of services. Cloud computing is growing rapidly and has witnessed an explosive growth in its use [7]. Cloud computing is becoming in demand. A growing number of cloud service providers are now providing various cloud based services. Cloud computing is becoming more popular, and more companies seeking outsourcing their IT needs to cloud [8].

One of the most challenging aspect of choosing a cloud service is the diversity of cloud computing services, since several cloud providers offer similar services, making it hard to compare one cloud service to another [9]. Other obstacles to selecting a cloud service are not having a thorough understanding of the non-functional characteristics of services cloud provider, and they offer services that are based on various qualities of (QoS) attributes, like security, performance, reliability and prices of their systems. We also have to consider that the structure and the representations of services differ between cloud service providers, and each service provider can provide an alternative interpretations (QoS) attributes that are related to cloud services. When you consider (QoS) features the selection of a service becomes more complex. This makes it essential that cloud users select the most suitable service [10].

The remainder of the document is laid out as follows: Sect. 2 will focus on similar research initiatives. Section 3 shows how we approached the QoS question. A description and analysis of the three skyline algorithms BNL, SFS and BBS are presented in Sect. 4. Experiment in Sect. 5 shows the efficacy and efficiency of the three algorithms. A conclusion can be found in Sect. 6.

2 Related Work

Numerous kinds and variants have been discussed in the literature on databases. Skyline methods were primarily developed to accelerate Skyline query processing, however they aren't able to handle massive databases or situations in which cloud

services are vulnerable to (QoS). We'll be discussing the similar methods for choosing and processing cloud services in the following section [11].

The authors of [12] have suggested an algorithm for simulation to aid in cloud service selection to be reliable. It employs three-level representation of QoS attributes. They also created A (MCDM), Analytical Hierarchy Process (AHP) and Simple Additive weighting (SAW) to rank services based on various QoS attributes. However, the methods for selecting aren't as effective like the skyline.

They [13] developed ELECTREIsSkyline an improvement to their prior work (CSRSS) an agent to choose Cloud services, by considering The question of QoS and how it can be adapted to incorporate (QoS). This gives users the ability to define the attributes of (QoS) they require.

They employs [14] developed a Web Agent using the Skyline approach To identify the Cloud services which are most suitable for users' needs. This algorithm is based upon the BNL Skyline algorithm.

They specifically [15] addressed the input/output procedures used to calculate the skyline. To solve the problem (MinMax) Distance and Euclidean Distance were utilized to create and sort the input lists to calculate the skyline.

They [16] handled QoS-based selection of web services. They compared Skyline's SFS algorithm with BBS algorithm. This algorithm allows for more efficient selection and runs faster.

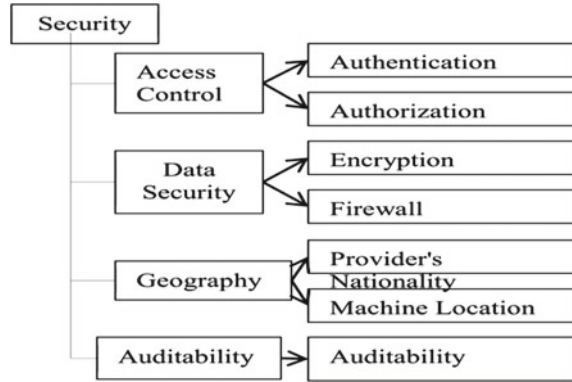
3 QoS Modelling

QoS refers to all the features that a service needs to offer to satisfy the implicit and stated needs of the user. When evaluating Cloud services There are a variety of QoS-related aspects to be considered such as security, cost and time.

Reference [17] to depict the QoS attributes of cloud services, we have created a hierarchical model. This helps cloud service customers comprehend and define their needs. It also provides QoS attributes so that cloud service users can meet their needs using cloud-based services.

The proposed model is divided into two parts that organize and define the attributes of cloud services that are QoS. The first step is categorization of QoS attributes. Based on our research on cloud QoS attributes, We found that different software and providers employ different ways to categorize or organize the attributes QoS and, in addition, to address the technical and commercial aspects of cloud services, Our model outlines the QoS characteristics into three areas: economic, strategic, and organizational. Cloud service selection serves the main goal of helping organizations and businesses select cloud services. It is essential to consider business requirements when choosing cloud services. Structure and representation QoS Attributes is the second phase. This is the phase that represents QoS attributes in three levels in a hierarchical manner, with seven top-level attributes: business performance pricing, security usability, compliance, usability, and security. Figure 1 illustrates an example

Fig. 1 Three different hierarchical levels of security attributes



of the 3-level structure for the Security characteristics in this model. This hierarchical structure allows consumers to be flexible.

This is in particular with the purpose of the model which includes knowledge of consumer regarding cloud computing. It is much easier to put the more general traits at the top and the more specific ones lower by employing a hierarchical framework. By doing this the proposed model is able to provide the capability to provide general high level QoS attributes to novice users, and more precise lower level attributes to experts. This type of flexibility isn't available in existing selection tools. Our model is inspired by research published in [13–15, 18].

4 Skyline Algorithms for Service Selection

Skyline was launched recently to address the issue of picking the best cloud service. Skyline selects services to be the most suitable service. We have two types of Skyline computation techniques that are based on whether or not they depend upon recalculated indexes of data. Although index based methods are more efficient since they don't access all the data available however, they are limited in their use because they require an indexed data. Furthermore, multidimensional indexes such as R-trees have their own drawbacks due to the well-known problem of dimensionality. They don't require any specific access structures to calculate the skyline, index-based methods can be more general. We select algorithms from two categories, from not index based techniques we have Block Nested Loop algorithm and the Sort Filter Skyline and in index-based techniques we have Branch-and Bound Skyline algorithm, which is derived.

4.1 The BNL Algorithm

For the calculation of the skylines, a naive approach would be to compare all the points of the data set two by two and to return as skyline all the points which are not dominated by any other. The BNL (Block Nested Loop) algorithm builds on this approach by fully analyzing the entire data set and storing a list of possible skyline points in central memory. If central memory is saturated, the algorithm manages a temporary file stored in secondary memory in which all the candidates not considered due to lack of space are stored. They will be addressed in a next iteration.

This algorithm works well in the context of small or medium-sized databases, because the number of candidate points for the skyline is reduced. The temporary file is then little used since there is practically always room in the main memory [19]. The pseudo code that is used to implement BBS algorithm is then shown:

```

L : input list of tuples for which the Skyline is to be computed
D: input list of dimensions
a, b: tuples
S: output list of the tuples forming the Skyline
Function CalculateSkyline
  Foreach a in L do
    If S =  $\emptyset$  Then
      S = {a}
    Else
      Foreach b in S - {a} do
        result = Compare (a, b, D)
        If result = count (D) then
          S = S + {a} - {b}
        Elseif result  $\neq$  0 and b is the last
          tuple in S then
            S = S + {a}
        Else
          Goto (*)
        End IF
      End Foreach
    (*) End If
  End Foreach
Return S
End Function

```

4.2 The SFS Algorithm

It sort the input data by ascending order based on the monotone preference formula. This function is the total of all coordinates for an individual point in every dimension

or to be an algorithm for entropy. Pre-sorting is the process of ensuring that a particular point a is visited first, before another point b . This permits the gradual behaviour of SFS and an increase in the amount of pairwise comparisons between the points. The algorithm analyzes data points in accordance with the order in which they are scored. Additionally, it put an in-memory memory buffer that has identified skyline-related points similarly to the one found on BNL. The buffer is empty at the start. Its state at the beginning is that of being empty. A point is retrieved from the data that has been sorted and when it's not dominated by a Skyline point within it, it's added to it. SFS conducts dominance tests by looking for all skyline points that are currently in existence. SFS is not without a major disadvantage. It is unable to adapt to the preferences of users. It instead has to look through all the data in order to give an entire skyline. Similar to BNL. However, certain skyline points are able to be restored earlier when it is shut down. SFS has the benefit of needing fewer comparisons than BNL [16]. The pseudo code that is used to implement BBS algorithm is then shown:

```

Input: A Sorted Dataset D (Heap).
Output: The Set of skyline points of dataset D.
nonfinished = True
While (nonfinished) do
    H=open_cursor(Heap)
    nonfinished = False
    While (next_tuple(H,h)) do
        if ("H is not dominated") then
            if ("window is full") then
                nonfinished = True
                Break
            Else
                "Add t to window."
    if (nonfinished) then
        R=open_new_file_(second_pass)
        Write(R,h)
        While (next_tuple(H,h)) do
            if ("t is not dominated") then
                Write (R , h)
free(Heap)
close(R)
Heap=SecondPass
"Write window tuples to output."
"Clear window."

```

4.3 The BBS Algorithms

BBS is only able to traverse R-tree once unlike NN. Data is organized in R-tree. Each R-tree node can hold three entries. Each leaf can host three input items. Similar to NN in that data points are organized in accordance with their minimum distances from the source (mindist) or minimum bounding rectangles (MBR), Calculating the (mindist) of a point is as simple as adding its coordinates, while calculating the mindist of an (MBR) is based on measuring the distance between the (MBR)'s lower-left corner and its point of origin. The algorithm chooses the closest tree-lined points to the source among all the unexplored places at each step. The method also stores these points in memory to validate the dominance phase. The pseudo code that is used to implement BBS algorithm is then shown.

```

D=∅ // D is a set of dominant points
Fill the root E by all entries in the heap
While E not empty yet do
  x is removed // x is the top entry from E
  if x is dominated by other points in D remove x
  else
    if x is an intermediate entry
      for each child xi of x
        if xi is not dominated by some point in D insert xi into heap
        else // x is a data point
          insert xi into D
        end
      end
    end
  end
end while

```

BBS algorithms perform better than other skyline strategies due to their low input/output price as well as a low number of accesses to R-tree nodes and a shorter duration of the process [20].

BBS algorithms perform better than other skyline strategies due to their low input/output price as well as a low number of accesses to R-tree nodes and a shorter duration of the process [20].

5 Experimentation and Results

We used a DELL laptop for the studies. It has an Intel Core i7 3.60 GHz Processor, 8 GB RAM, Windows 10 Professional as our operating system and Oracle Database as our database management system. The algorithm is implemented in Python.

Our experiments are conducted using a set of real Cloud service QoS information collected by us and our own database comprising 150,000 Cloud services. Dataset is extracted like a text-based document. Any line of the document includes a set of 9 parameters that are detached by the use of commas.

We compared the results achieved using the Skyline algorithms BBS, SFS and BNL. For an input list of 150,000 Cloud services, The number of dimensions can be varied between 1 and 7 (Table 1).

It has been found that the BBS algorithm is more effective at determining which cloud services are most suitable for users, including their QoS requirements. As shown in Fig. 2 and table 1, the BBS algorithm takes less time to calculate the 150,000 service than the SFS 35 s and BNL45s. Figure 3 shows how the final solution can be reduced to 7% of the input size using the BBS algorithm. However, it doesn't preclude SFS and BNL from being used in cases where the initial cloud service number is not large. The BBS algorithm, based on experimental results and Skyline's Size, can reliably and efficiently perform service select for service-oriented system and compound cloud applications. It is also the most suitable skyline algorithm when working with large, high dimensionality data bases.

Table 1 Execution time of BBS, SFS and BNL ALGORITHMS

Number of cloud services	Time execution (in s)		
	BBS	SFS	BNL
1000	0	2	3
5000	1	3	4
10000	3	5	7
50000	10	17	25
100000	19	25	37
150000	22	35	45

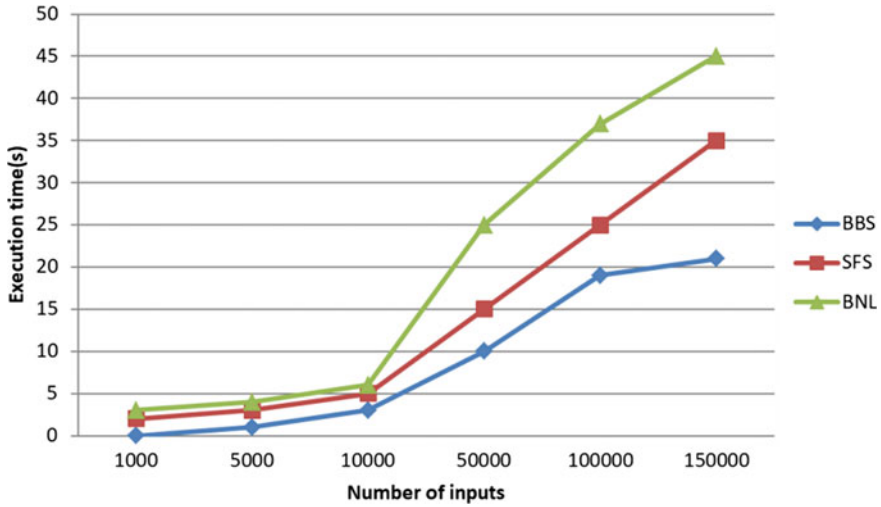


Fig. 2 Comparison of execution times in different input sizes

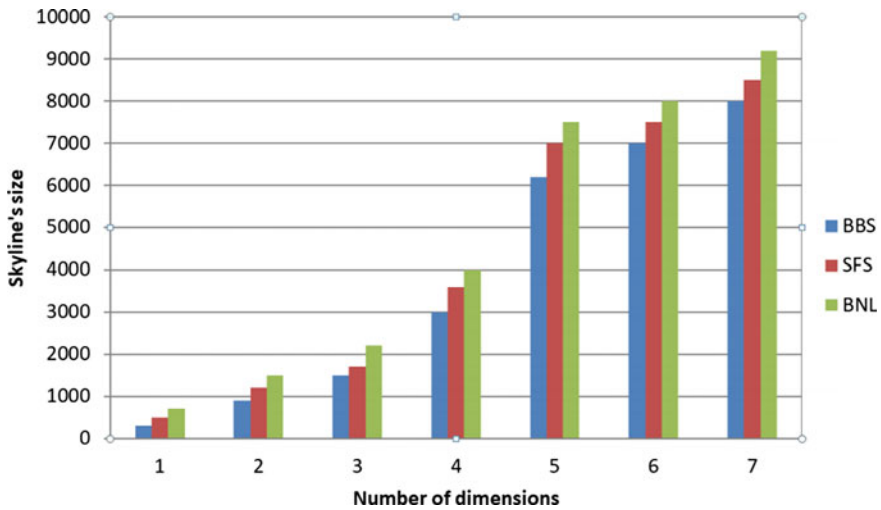


Fig. 3 Skyline's Size/Number of Dimensions for each algorithm

6 Conclusion

The cloud service selection challenge is designed to find quickly the most effective cloud service from an array of data points. In this article, we have addressed the question of QoS based service selection in this article. We have compared three Skyline algorithms: the SFS algorithm, BNL algorithm and the BBS algorithm. The BBS algorithm is more efficient with a shorter execution time and offers better results. As future work, we will rely on the different algorithms presented in [21–31]

to improve our search and service selection approaches. Note that all these techniques could be adapted to other fields such as health, environment, transport, education, etc.

References

1. M. Abourezq et al., An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Archit.* **9**(2–3), 136–148 (2020). <https://doi.org/10.1504/IJHPSA.2020.111557>
2. H. Alabool et al., Cloud service evaluation method-based multi-criteria decision-making: a systematic literature review. *J. Syst. Softw.* **139**, 161–188 (2018). <https://doi.org/10.1016/j.jss.2018.01.038>
3. D. Belkasmı et al., On fuzzy approaches for enlarging skyline query results. *Appl. Soft Comput.* **74**, 51–65 (2019). <https://doi.org/10.1016/j.asoc.2018.10.013>
4. H. Bypour et al., An efficient secret sharing-based storage system for cloud-based internet of things. *Int. J. Eng.* **32**(8), 1117–1125 (2019). <https://doi.org/10.5829/ije.2019.32.08b.07>
5. Y. Cheng, Y. Morimoto, Cheng. *Bull. Netw. Comput. Syst. Softw.* **8**(2), 81–86 (2019)
6. F. De la Prieta et al., Survey of agent-based cloud computing applications. *Future Gener. Comput. Syst.* **100**, 223–236 (2019). <https://doi.org/10.1016/j.future.2019.04.037>
7. H. Du, et al., A two phase method for skyline computation, in ed. by Y. Jia, et al *Proceedings of 2019 Chinese Intelligent Systems Conference* (Springer, Singapore, 2020), pp. 629–637. https://doi.org/10.1007/978-981-32-9682-4_66.
8. M. Fariss et al., Comparative study of skyline algorithms for selecting web Services based on QoS. *Procedia Comput. Sci.* **127**, 408–415 (2018). <https://doi.org/10.1016/j.procs.2018.01.138>
9. M. FARISS, et al., Prefiltering approach for web service selection based on QoS, in *2019 International Conference on Systems of Collaboration Big Data, Internet of Things & Security (SysCoBloTS)*. pp. 1–5 (2019). <https://doi.org/10.1109/SysCoBloTS48768.2019.9028043>.
10. Z. Huang et al., An efficient algorithm for skyline queries in cloud computing environments. *China Commun.* **15**(10), 182–193 (2018). <https://doi.org/10.1109/CC.2018.8485480>
11. S.K. Keshari, V. Kansal, S. Kumar, A systematic review of quality of services (QoS) in software defined networking (SDN). *Wireless Pers Commun.* **116**, 2593–2614 (2021)
12. M. Eisa, M. Younas, K. Basu, I. Awan, Modelling and Simulation of QoS-Aware Service Selection in Cloud Computing. *Simul. Model. Pract. Theory* **103**, 102108 (2020)
13. M. Abourezq, A. Idrissi, Integration of Qos aspects in the cloud service research and selection system. *IJACSA* **6** (2015)
14. M. Abourezq, et A. Idrissi, A cloud services research and selection system, in *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, avr. (2014)
15. M. Abourezq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Archit.* **9**, 136–148 (2020)
16. M. Fariss, H. Asaidi, M. Bellouki, Comparative study of skyline algorithms for selecting Web Services based on QoS. *Procedia Computer Science.* **127**, 408–415 (2018)
17. H. Wang et al., Integrating reinforcement learning and skyline computing for adaptive service composition. *Inf. Sci.* **519**, 141–160 (2020). <https://doi.org/10.1016/j.ins.2020.01.039>
18. J. Araujo, P. Maciel, E. Andrade, G. Callou, V. Alves, P. Cunha, Decision making in cloud environments: an approach based on multiple criteria decision analysis and stochastic models. *J Cloud Comp.* **7**, 7 (2018)
19. Integrating multi-objective genetic algorithm based clustering and data partitioning for skyline computation | SpringerLink, <https://link.springer.com/article/https://doi.org/10.1007/s10489-009-0206-7>, last accessed 2022/07/07.

20. Introducing a New Supply Chain Management Concept by Hybridizing TOPSIS, IoT and Cloud Computing | SpringerLink, <https://link.springer.com/article>, <https://doi.org/10.1007/s40032-020-00619-x>, last accessed 2022/07/07.
21. F. Mourad, N. el Allali, H. Asaidi, M. et Bellouki, An improved approach for QoS based web services selection using clustering. *Adv. Sci. Technol. Eng. Sys. J.* **6**, 616–621 (2021). <https://doi.org/10.25046/aj060270>
22. K. Elhandri, A. Idrissi. Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Tech. Sec. Transac.* **10** (2020)
23. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
24. H. Rehioui, A. Idrissi. A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data Min.* (2017)
25. A. Idrissi, K Elhandri, H Rehioui, M. Abourezq. Top-k and skyline for cloud services research and selection system. *International conference on Big Data and Advanced Wireless technologies.* (2016)
26. A. Idrissi, C.M. Li, J.F. Myoupo. An algorithm for a constraint optimization problem in mobile ad-hoc networks. *18th IEEE International Conference on Tools with Artificial Intelligence.* (2006)
27. A. Idrissi, F. Zegrari. A new approach for a better load balancing and a better distribution of resources in cloud computing. arXiv preprint arXiv: 1709.10372. (2015)
28. F. Zegrari, A. Idrissi, H. Rehioui. Resource allocation with efficient load balancing in cloud environment. *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies.* (2016)
29. F. Zegrari, A. Idrissi. Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Robot. Intell. Syst.* **14**(3), 65–70. (2020)
30. S. Retal, A. Idrissi. A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comp. Elect. Eng.* **73**, 289–303. (2018)
31. M. Essadqi, A. Idrissi, A. Amarir. An effective oriented genetic algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comp. Sci.* **127**, 170–179. (2018)

Artificial Intelligence Applied to Blockchain and IoT

A State-of-the-Art Survey on Ransomware Detection using Machine Learning and Deep Learning



Loubna Moujoud, Meryeme Ayache, and Abdelhamid Belmekki

Abstract The world has noticed an alarming surge in ransomware cyberattacks these last years, causing an important financial losses to various organizations. Ransomware attacks are types of malware that usually lock the users' devices or encrypt their data files and request them to pay money (ransom) to unlock the devices or to recover the encrypted files. Several researches proposed techniques to detect this kind of malwares in their early stages of propagation. However, most of these detection methods followed a signature-based technique, which have difficulties to detect zero-day and unknown ransomware. New techniques that can dynamically identify and stop this type of ransomware are thus desperately needed. In this direction, machine learning and deep learning techniques are recently applied in ransomware detection, spam detection, image recognition, ... etc. In this paper, we provide a survey about the ransomware detection studies using machine learning and deep learning techniques, conducted from 2017 to 2022. This paper also provides an in-depth list of possible directions for future study.

Keywords Machine learning · Deep learning · Ransomware · Ransomware detection

L. Moujoud (✉) · M. Ayache · A. Belmekki
Department of Mathematics, Computer Science and Networks, INPT, Rabat, Morocco
e-mail: moujoud.loubna@inpt.ac.ma

M. Ayache
e-mail: ayache@inpt.ac.ma

A. Belmekki
e-mail: belmekki@inpt.ac.ma

1 Introduction

A rapid innovation of ransomware's infection and propagation techniques may be seen in the exponential growth of new ransomware variants in recent years. The goal of ransomware, an active kind of modern malware, is to obtain money from its victims by denying them access to their systems' data in exchange for a ransom. According to TrendMicro research [1], 2016 was marked by a surprising increase of newly discovered ransomware families. The two most frequent types of ransomware are encryptors and screen lockers.

Encryptors, as the name implies, encrypt data on the victim's system, making the content unusable without the decryption key. Unfortunately, many victims pay the ransom in order to get the key, and restore access to their data.

Screen lockers, on the other hand, restricts the user access by locking his device, asserting that the system is encrypted, while the mouse and the keyboard are accessible. That allows the victim to fulfill the ransom demand.

The difference between these two types is presented in the Fig. 1.

The number of companies affected by ransomware increased from 37% in 2020 to 66% in 2021, according to the Sophos State of Ransomware 2022 Study [2]. Organizations that choose to pay a ransom increased from 26% in 2020 to 46% in 2021. Only 26% of firms were able to restore encrypted data using backups in 2021, according to the same global poll, and ransoms were also paid. "The findings confirm the brutal truth that when it comes to ransomware, it doesn't pay to pay. Despite more organisations opting to pay a ransom, only a tiny minority of those who paid got back all their data," said Sophos principal research scientist Chester Wisniewski.

Below are just a few examples of some infamous ransomware detected over the last few years [3]:

- **Locky** [4] was first used for an attack in 2016 by a group of organized hackers. Locky is a refined malware that infects networks by infected attachments. It is distributed exploiting social engineering techniques to spread the malicious

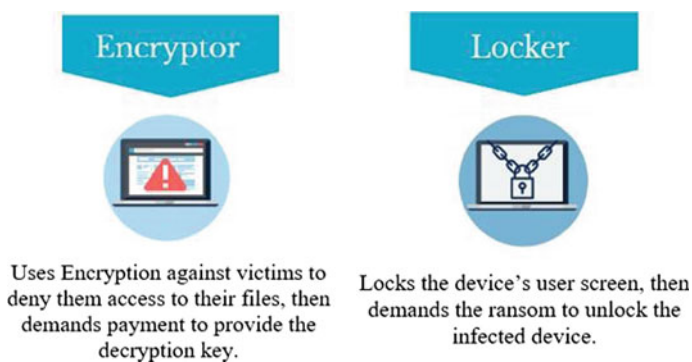


Fig. 1 The difference between Encryptor and Locker Ransomware

code. The large email campaigns were spearheaded by the Necurs Botnet [5], which was considered one of the most largest botnets before it went dormant. Locky ransomware targets completely different file types in order to trigger the infectious process, such as Microsoft Office files, JScript, HTML Application, Windows Executable, etc. The most commonly reported file type used in Locky Ransomware Attacks is Microsoft Word, containing malicious macros

- **WannaCry** WannaCry was a ransomware attack that spread to over one hundred fifty countries in 2017. It propagated through EternalBlue [6], an exploit leaked from the National Security Agency (NSA), and stolen by a group referred to as The Shadow Brokers. Microsoft had released patches previously to close the exploit. Unfortunately, many organizations had not applied these patches, and so were left exposed to the attack. It makes use of a vulnerability in the network resource sharing protocol created by Microsoft. An attacker can send specially created packets to any machine that accepts data from the public internet on port 445 due to the exploit. The WannaCry attackers demanded a ransom of \$600 worth of Bitcoin. The worldwide financial damage caused by WannaCry was approximately \$ 4 billion. Figure 2 presents the Countries initially affected in WannaCry ransomware attack [7].
- **Ryuk** Russian hackers created the encryption Trojan known as Ryuk, which was first discovered in August 2018. It has targeted numerous sizable businesses that use Microsoft Windows and associated platforms. The exploit begins when a victim opens a malicious Microsoft Office document that was sent as an email attachment. When a user opens the document, a malicious macro begins a PowerShell command that tries to download the banking Trojan Emotet [8]. Many of the US organizations that were targeted paid the required ransom money as a result of the substantial impact. The total estimated damage is more than \$640,000.
- **DarkSide** is a ransomware attack that began at the start of August 2020, connected to the DarkSide group [9], and currently functions as ransomware-as-a-service (RaaS). The attack starts with brute force techniques and gains access by taking advantage of well-known weaknesses in the remote desktop protocol. The Colonial Pipeline Firm ransomware outbreak, which was found to be caused by the DarkSide ransomware group in May 2021 [10], led the company to decide to proactively and temporarily shut down the pipeline.

According to the threat intelligence team of Malware-bytes labs, in April 2022, three new ransomware-as-a-service (RaaS) groups were detected: **Onyx**, **Mindware**, and Black **Basta** [11].

- The **Black Basta gang** started in February 2022. However, in just two months, it arrived to add almost 50 victims located primarily in the United States, Canada, Australia, the United Kingdom, and New Zealand in various English-speaking nations. In June, and under the collaboration with QBot malware to spread laterally throughout the network [12]. Furthermore, Black Basta had a Linux variant that targeted the VMWare ESXi servers [13].



Fig. 2 WannaCry ransomware attack

- The *Onyx* is based on the old Chaos builder. This ransomware does not encrypt data, but it destroys any files larger than 2 MB. This behavior is related to a bug in the poorly-written ransomware builder.
- The *Mindware gang* appeared at the first time in mid March using a well-known ransomware strain called SFile2. However, it did not start practicing till April 2022. This ransomware steal the data from the victim before encrypting it. Hence the victim is faced with the twin threats: encrypted data and sensitive information leakage.

Traditional detection systems will have a harder time identifying ransomware since it frequently evolves and new versions often behave differently than their predecessors. According to research that employed deep learning and machine learning techniques to detect ransomware, there is a significant improvement in detection rates and false positive rates at large scale compared with traditional methods. In order to prove this, we presented in this article the ransomware detection studies conducted from 2017 to 2022, dedicated to Windows, Linux or IoT platform. This study provides also the motivation for the use of machine learning and deep learning approaches for analyzing and classifying different types of ransomware.

The rest of the work are structured as follows: in Sect. 2, we present a brief history of the ransomware malware. Section 3 describes the lifecycle of ransomware attack which is divided into three main stages. Section 4 presents the literature review of the existing mechanism and techniques to detect ransomware attack. In Sect. 5, we opened some research directions and open challenges to improve the field of ransomware detection. Section 6 concludes the paper.

2 A Brief History of Ransomware

Ransomware has become more mature, advanced, and destructive in the last years. The 1989 AIDS Trojan [14], created by a medical researcher who tried to blackmail other researchers via malware distributed on floppy disks, was the first known and recorded occurrence of ransomware. The AIDS Trojan laid the groundwork for cyber criminals and they realized that they could monetize ransomware on a far wider scale.

With the Internet making it easier, Ransomware took off in popularity in the mid-2000s. In 2006, attacks began utilizing more sophisticated and stronger encryption algorithms such as RSA encryption. Archives was the first ransomware to use RSA encryption, which encrypted everything in the MyDocuments directory and required victims to purchase items from an online pharmacy to obtain a 30-digit password to unlock their files.

The year 2008 saw the invention of Bitcoin, which made it much easier for ransomware authors to collect money from their victims. By providing an easy and untraceable payment service, this period has known an evolution of ransomware shaped by several variants.

The year 2012 marks ransomware's move into the big time. New ransomware families have been discovered, including Cobralocker, Gimemo, Kovter, Lokibot, Lyposit, Reveton and Urausy. With the emergence of ransomware as a service (Raas), potential criminals with non-technical abilities now have more access to ransomware. As suggested by its name, RaaS is a ransomware-as-a-service model that allows affiliates to perform ransomware attacks using already developed ransomware tools. Without having to know how to code, it enables cybercriminals to buy ransomware software on the dark web and conduct ransomware attacks. Reveton was the first ransomware as a service discovered in 2012, which fraudulently claims to be from a legitimate law enforcement authority and freezes the compromised machine's operating system and demands a ransom to unlock it.

In 2017, ransomware attacks were becoming more largescale. WannaCry ransomware attack was one of these exploits, which became the biggest and most famous in history. Reports from 2017 indicate that total damages from ransomware are reaching 1 billion\$ [15]. The year 2017 saw a sharp increase in new discovered ransomware families. According to Statista [16], 327 new family was detected in 2017, and 32 new families were uncovered in 2021. In Q1 2022, 8 new ransomware families were detected with 3083 new modifications of the existing ransomware malwares [17]. The graphical statistics according to the number of newly discovered ransomware families are given in Fig. 3.

Ransomware is evolving at a rapid pace jeopardizing several industries related to: (a) *healthcare* which lost over \$157 million due to ransomware attacks since 2016 [18]; (b) *Education* especially with the remote mode caused by Covid 19; and (c) 90% of the *financial* institutions had experienced ransomware attacks in 2020 (especially the banking sector). Therefore, it has become mandatory to invest in an organizational culture of cyber resilience to prevent incoming variants of ransomware attacks. In fact, according to *Gartner*, ransomware payments and negotiations will

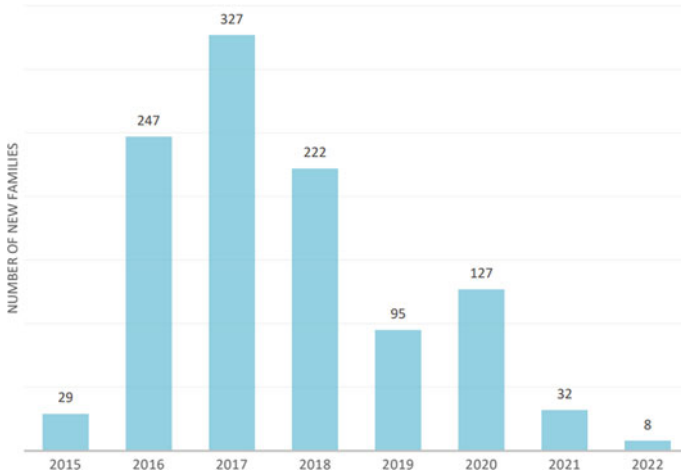


Fig. 3 Number of newly discovered ransomware families from 2015 to 2022

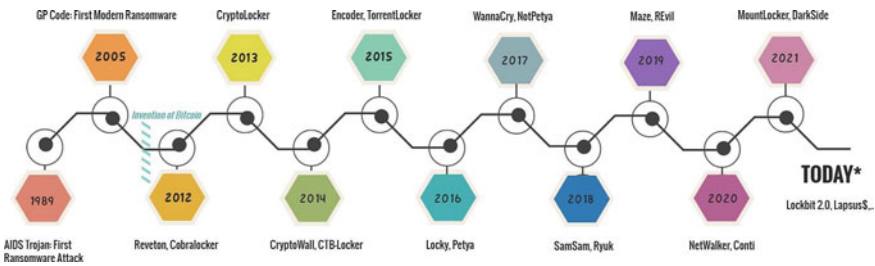


Fig. 4 Ransomware timeline

rise to 30% by the end of 2025 [19]. This shows a shift in the ransomware landscape from previous years to more sophisticated and creative methods of attack.

The timeline presented in Fig. 4 shows the rapid increase of Ransomware as well as its evolution, since its first appearance in 1989.

3 How Ransomware Works?

Although the tactics and techniques used by ransomware operators may evolve over time, but the lifecycle of a ransomware attack remains relatively the same. The ransomware attack lifecycle can be organized into three primary stages, as shown in Fig. 5:

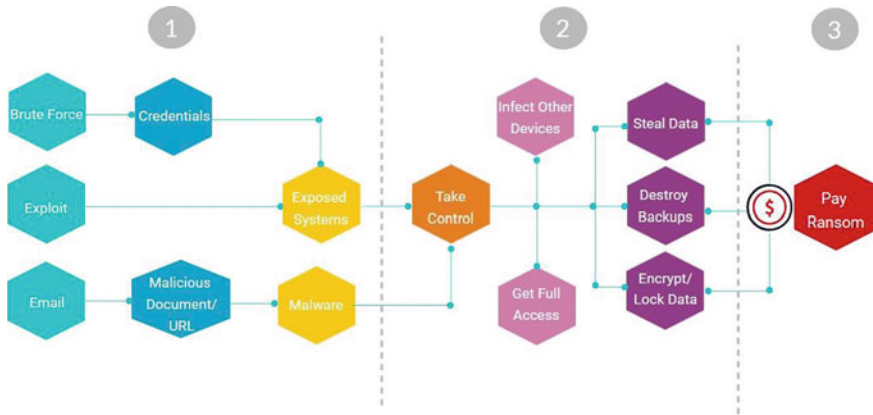


Fig. 5 Lifecycle of a ransomware attack

- Stage 1: Initial Access** This first stage is where the attacker sets up the ransomware to infiltrate the system, and trying to establish a foothold in the target network. This can be done in several ways, such as sending out phishing email attacks, which allows the attacker to send malicious email that contains a link to a website that has been compromised or an attachment that has malware embedded in it, and therefore the access is most often gained. It can be done also by setting up malicious websites, exploiting weaknesses and vulnerabilities discovered in software or hardware with an Internet-facing interface, or attacking software vulnerabilities directly. Attackers also employ brute-force attacks as a third tactic, to target weak and easy usernames and passwords.
- Stage 2: Data Infection** The second stage occurs once the ransomware has infiltrated the system. The malicious code will establish a line of communication, and begins gathering information. The ransomware attacker can download additional malicious software using this communication line. At this point, the ransomware can remain hidden and dormant for days, weeks or months before the attacker decides to launch the attack. The ransomware may attempt to move laterally across other systems in to access as much data as possible. They also target backup systems to weaken the victim’s capacity for recovery, preventing it from being able to repair its system and restore data, and forcing it to pay the ransom. Once an attack has been activated, ransomware begins infecting or locking as many computers as it can.
- Stage 3: Ransom Demands** is the final phase. The ultimate purpose of the ransomware threat actors is to extort money from the victim restricting the victim’s access to its resources. Once systems have been locked up or encrypted, the malicious parties ask the victim for a ransom in exchange for decrypting the data. Typically, ransoms are requested in untraceable digital currencies like Bitcoin or Monero [20].

4 Literature Review

Several works have been done to analyze and detect ransomware. The available literature in the field of ransomware detection is presented in this section.

A. Machine Learning Detection Studies

- (1) *Multi-level Analysis*: Poudyal et al. [21] proposed a framework for detecting ransomware based on machine learning, static analysis, and reverse engineering techniques. The framework is employed to carry out multi-level analysis in order to more thoroughly review malware code functions and segments. Benign and malicious samples are used within the experiments. Different supervised ML algorithms are used to categorize the samples once they have been pre-processed to extract their features. Experimental findings demonstrated the effectiveness of the suggested framework. Depending on the different machine learning (ML) techniques utilized, ransomware samples had detection accuracies ranging from 76 to 97%. Seven of these algorithms performed well, with a detection rate of 90%. In fact, the rate of ransomware detection is minimal for Logistic Regression at 89.18% and highest for Random Forest with 97.95%.
- (2) *Decision Tree*: Wan et al. [22] suggested a framework for detecting ransomware attacks in networks based on Decision Tree (DT) model. To replace the packet-based data, the framework uses a flow-based Biflow. The open malicious traffic datasets are converted into binary data reflecting network flows using Argus. The results showed that when the DT algorithm is paired with six feature selection techniques, it performs better. The ransomware might avoid the classifier, though, if its behaviors change.
- (3) *SVM*: Support vector machine (SVM) has been presented by Takeuchi et al. [23] as a ransomware detection model for Microsoft Windows systems. Using Cuckoo Sandbox, they dynamically retrieved features from ransomware API invocation sequences. Their approach is based on the extraction of features from the API call history generated by the ransomware attacks. According to their experimental results, 276 ransomware samples could be distinguished from benign samples with an accuracy of roughly 97%.
- (4) *PEDA*: For the purpose of identifying crypto-ransomware before any encryption takes place, the pre-encryption detection method (PEDA) has been proposed by Kok et al. [24]. The PEDA has two distinct detection levels. A signature repository (SR) is the first, and it shows any signature matches with well-known ransomware. A learning algorithm (LA) is used in the second detection level to find both known and undiscovered crypto-ransomware. Data from the application program interface (API) is used by LA to train the predictive model using machine learning. Based on the results, it can be concluded that LA was successful in detecting crypto-ransomware before the encryption became functional.

- (5) *CF-NCF*: Bae et al. [25] proposed a ransomware detection approach using CF-NCF (Class Frequency—Non-Class Frequency) and machine learning methods. In a dynamic analysis environment, they extracted Windows API call sequences and tested them using n-grams with various n values. In order to generate feature vectors, CF-NCF was used. Based on the results, RF outperforms LR, SGD, NB, SVM and KNN in ransomware detection. The experimental results demonstrate that the suggested approach has a ransomware detection accuracy of up to 98.65%, able to distinguish between malicious and benign files.
- (6) *PFEs*: A new model was suggested by Cusack et al. [26] using programmable forwarding engines (PFEs), a recent development in networking hardware, that enable rapid per-packet and network monitoring data collecting. The authors used this information to track the network activity between an infected computer and the command and control (C&C) server after the data collection phase. From this traffic, they derived high-level flow features, that are used to categorize ransomware. The classification model reached a false negative rate under 11% while reaching a detection rate of more than 86%. According to the results, a flow-based fingerprinting technique is feasible and precise enough to detect ransomware before encryption.
- (7) *DNAact-Ran*: The DNAact-Ran system was proposed by Khan et al. [27], which combined revolutionary digital DNA sequencing with machine learning for ransomware detection. Features were chosen from a supplied dataset using MOGWO and BCS, and digital DNA was created using k-mer frequency and DNA sequencing. Naïve Bayes, Decision Stump and AdaBoost classification algorithms were compared with the suggested approach, and the results show that Naïve Bayes has a detection accuracy of 78.5%, Decision Stump's is 75.8%, AdaBoost's is 83.2%, and the proposed algorithm with the highest rate 87.9%.
- (8) *Random Forest*: A novel static analysis-based technique to identify ransomware was presented by Khammas [28]. The essential characteristic of the suggested approach is the elimination of the disassembly process in favor of direct feature extraction from the raw byte using frequent pattern mining, which significantly speeds up detection. The Gain Ratio feature selection approach was found to work best with 1000 features for the detection procedure. The findings demonstrated that, in terms of accuracy and time needed, trees with seed numbers of 100 and 100 produced the greatest outcomes. The proposed model achieved a high ransomware detection accuracy of 97.74%.
- (9) *IoT*: Dash et al. [29] presented a method that uses network traffic flow analysis to detect malware on IoT devices. They specifically employed the specific local fingerprint of the network usage behavior of ransomware to distinguish it from non-malicious programs. The results showed that the suggested method had a detection rate of 93.76% and a precision rate of 89.85%.
- (10) *WEKA*: Egunjobi et al. [30] demonstrated a classification method that combines static and dynamic variables to improve the ransomware detection and classification. They used a test set to train supervised machine learning algorithms

and a confusion matrix to track accuracy, allowing for a thorough evaluation of each approach. In their research, supervised algorithms including the Naïve Bayes algorithm had an accuracy of 96%, SVM had an accuracy of 99.5%, random forest had an accuracy of 99.5%, and IB1 had an accuracy of 96%.

- (11) *Markov Model & RF*: In their paper, Hwang et al. [31] suggested a two-stage mixed Markov model and Random Forest model. To control both the false positive (FPR) and false negative (FNR) error rates, they employed the Markov model to capture the properties of ransomware and a Random Forest machine learning model to assess the remaining data. They used the two-stage mixed detection approach to attain an overall accuracy of 97.3% with 4.8% FPR and 1.5% FNR.

N-gram: Zhang et al. [32] suggested a static technique based on text analysis. Five machine learning algorithms—DT, RF, KNN, NB, and GBDT—along with various opcode N-grams (2-g, 3-g, and 4-g) and feature dimensions were employed to generate the classifiers. They chose feature N-grams with excellent precision using TF-IDF. Findings demonstrate that Random Forest outperforms alternative algorithms. Also, the outcomes demonstrate that classifiers with N-grams of different lengths can perform effectively with various feature dimensions.

- (12) *Feature selection-based*: Masum et al. [33] provided in their study a framework based on feature selection and using various machine learning techniques, such as neural network-based architectures, to categorize the security level for ransomware detection and prevention. For this purpose, they used a variety of machine learning methods, including DT, RF, NB, LR, and NN-based classifiers. According to the experimental results, RF classifiers perform better than other approaches in terms of accuracy, F-beta, and precision scores.
- (13) *RansomWall*: RansomWall is a multi-layered security system for avoiding Cryptographic Ransomware which was initially described by Shaukat et Et [34]. It consists of a Strong Trap layer for early detection, a Machine Learning layer for zero-day attacks, and a File Backup layer. They created RansomWall for the Microsoft Windows operating system and tested it against 574 samples from 12 families of cryptographic ransomware under real-world conditions. During the testing of RansomWall utilizing multiple Machine Learning techniques, the GTB Algorithm achieved a detection rate of 98.25% and had nearly no false positives.

B. Deep Learning Detection Studies

- (1) *Deep Convolutional Neural Networks*: Ashraf et al. [35] used both traditional machine learning methods and Deep learning based on Transfer Learning in their research for ransomware detection. Two distinct datasets (static and dynamic) were used to undertake feature engineering and analysis. Through experiments, it has been found that the most crucial features for ransomware detection are Registry modifications, API calls, and DLLs. According to the results, static features outperformed the chosen dynamic features. Additionally, Random

Forest performed well in predicting the benign class, whereas SVM performed well in predicting the positive class.

- (2) LSTM: LSTM networks were suggested as a technique by Maniath et al. [36] to distinguish malicious executables from benign executables. They identified the issue as a binary sequence categorization issue of the executable's API calls. The API calls that the executable made while it was running were extracted using a dynamic analysis. In detecting the ransomware behavior, they achieved an experimental accuracy of 96.67%.
- (3) Autoencoder: A deep learning approach to dealing with ransomware was proposed by AbdulsalamYa'u et al. [37]. Based on behavioral data, a deep learning autoencoder was trained to identify the samples as benign or ransomware. They achieved a true positive rate of 99.7%.

The Table 1 below summarises the models we reviewed. In this table we presented the advantages and the inconvenient of each proposed approach.

C. Discussion

In this study, we have highlighted the machine learning and deep learning algorithms that have been used in each work. These algorithms are perfectly adjusted to the complex nature of ransomware. A summary of different algorithms used in the detection studies are shared in the Fig. 6.

According to the evaluation results in Table 1, RF outperforms other algorithms in terms of precision and accuracy when it comes to ransomware detection. That explains the employment of Random Forest classifier in several research, as seen in Fig. 6.

With the rise in ransomware attacks, datasets are extremely important for malware detection. The Table 1 summarizes the various datasets and their repositories used in the detection studies from 2017 to 2022 from different platforms. According to the data, VirusTotal samples are used in the majority of studies on ransomware detection. Hybrid Analysis, theZoo, and VirusShare are further recognized repositories.

The Table 1 also presents the analyzed ransomware families considered in each study. It indicates that current studies have limited scope as the samples were considered from few and popular ransomware families. We can conclude that none can detect all variants of ransomware, especially unknown and uncategorized ransomware. A powerful and accurate ransomware detection system can be achieved by combining machine learning and deep learning algorithms, to increase the detection rate of ransomware attacks.

5 Research Direction

In this paper, we presented the available literature about machine and deep learning approaches used in ransomware detection. Based on this survey, we highlighted some research directions which might be a subject to improve the field of ransomware

Table 1 Summary of related works on ransomware detection approaches

Study	Year	Ransomware family name	Dataset source	No. of samples	Evaluation results
[21]	2018	Cryptowall, Torrentlocker, Cryp-tolocker, Zerolocker, Cryptorlocker, Ct-blocker, Xorist, and Wannacrypt	Virus Total, Virus Share and the Zoo [38]	302 samples	The RF performs better than the compared algorithms, but new types of ransomware decrease the detection rate of the proposed method
[22]	2018	Locky & Cerber	Network traffic [39]	–	The results showed that DT performs well, but if the ransomware's actions alter, it may be able to avoid the classifier
[23]	2018	WannaCry, Cerber, Petya & CryptoLocker	Hybrid Analysis website [40]	588 samples	The suggested schema increases prediction accuracy and lowers the ransomware missed rate
[24]	2020	–	Old, Virus Total, and the Zoo [41]	1846 samples	Detect crypto-ransomware before encryption, but samples with small size can decrease the detection rate
[25]	2020	–	Windows 7 system directories and VirusTotal [42]	2200 samples	The RF outperforms the LR, NB, SGD, KNN, and SVM at ransomware detection

(continued)

detection. Several open issues and challenges identified in the field of these studies can be summarized as follows:

- *Data Collection:* No known dataset that could be used to train machine and deep learning models fully captures all ransomware attack behaviors.
- *Analysis:* Some of the evaluations are limited to a certain amount of ransomware like signature-based approach.
- *Evasion and obfuscation:* Ransomware authors are increasingly using advanced obfuscation techniques to hide their malicious codes.
- *Hardware complexities:* Most of the developed techniques require advanced hardware.

Table 1 (continued)

Study	Year	Ransomware family name	Dataset source	No. of samples	Evaluation results
[26]	2019	Cerber	Network traffic [43]	100 MB of ransomware traffic & 100 MB of clean traffic	Based on network activity, the solution can detect ransomware before it encrypts files, however it ignores the UDP protocol
[27]	2020	Critroni, CryptoLocker, CryptoWall, Kollah, Koyter, Locker, Matsnu, PGPCoder, Reveton, TeslaCrypt & Trojan-Ransom	Github-PSJoshi [44]	1524 samples	The accuracy of the proposed algorithm is is 87.9%
[28]	2020	Cerber, TeslaCrypt & Locky	VirusTotal [45]	1680 samples	The proposed method achieved a high accuracy of 97.74%, but if the number increases to more than 1000, that led to a degradation in accuracy
[35]	2019	–	VirusTotal, VirusShare and Windows 7 Operating System	45,000 samples	In the proposed method, two datasets have been used: static and dynamic. This method has performed well with static dataset than dynamic dataset
[29]	2018	–	Network data collected from IOT devices in pcap file format	–	KNN has shown to be an accurate algorithm, easy to train and update, and fast. The results shows that the Network Traffic data analysis performed less well on infected data than benign data
[30]	2019	Locker and Crypto Locker	VirusTotal& WEKA [46]	200 samples	The RF gave a relatively high detection accuracy of 99.5%

(continued)

Table 1 (continued)

Study	Year	Ransomware family name	Dataset source	No. of samples	Evaluation results
[31]	2020	–	VirusShare [41]	6393 samples	The two-stage mixed detection model gave 97.28% overall accuracy. Unfortunately, ransomware can avoid this detection system by mimicking malware
[36]	2017	15 families of ransomware	Online sources, honeynets and Windows 7	157 samples	The accuracy of 96.67% is higher than earlier research in the field of using deep learning in ransomware detection
[34]	2018	TeslaCrypt, Cerber, Jigsaw, TorrentLocker, Locky, CryptoLocker, CryptoDefens, Hidden Tear, CryptoFortress, and CrypVault	VirusShare [41]	574 samples	Using GTB, the suggested approach has a detection rate of 98.25%. Two Ransomware variants unexpectedly stopped working after encrypting a small number of files, according to an analysis of False Negatives. Only.vcf files were encrypted by another ransomware sample
[32]	2019	CryptoLocker, CryptoWall, Cryrar, Locky, Petya, Reveton, Teslacrypt & Wannacry	VirusTotal [45]	1787 samples	The proposed method proved that N-gram could identify a known ransomware sample to its family. Unfortunately three families, namely, cryptowall, locky, and reveton, couldn't be identified well

(continued)

Table 1 (continued)

Study	Year	Ransomware family name	Dataset source	No. of samples	Evaluation results
[37]	2019	Critroni, CryptLocker, CryptoWall, KOL-LAH, Kovter, Locker, MATSNU, PGP-CODER, Reveton, TeslaCrypt & Trojan-Ransom	Resilient Information Systems Security (RISS) [47]	1524 samples	The proposed method achieved a high accuracy of 99.7%
[33]	2022	–	Ransom Dataset from Github [48]	138 047 samples	The suggested technique shown that RF classifiers perform better than other techniques in terms of accuracy, F-beta, and precision scores. Yet, when compared to other approaches, LR failed to provide satisfying F-beta scores and recall scores

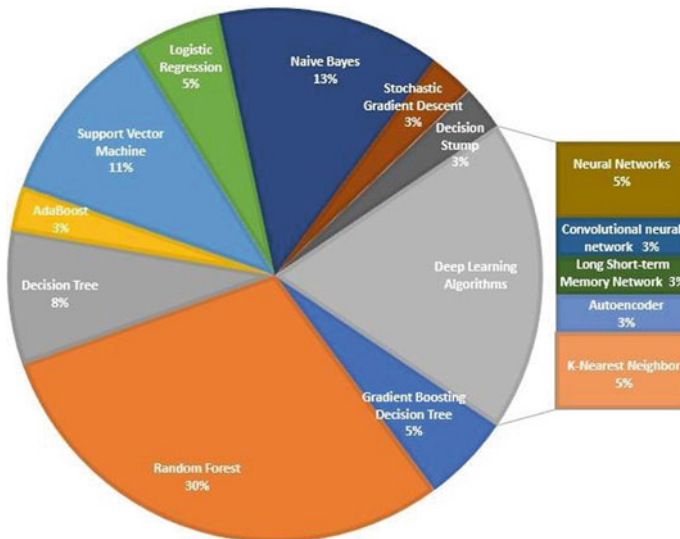


Fig. 6 Machine learning and deep learning algorithms used in ransomware detection studies

- *Unsatisfying experimental results:* It is challenging for current ransomware detection methods to have both a high detection rate and a low rate of false positive detection.
- *Time Limitation:* Some studies were difficult to implement because of the time constraints on sample analysis.

Our work in this paper is motivated by the interest in understanding the limitations of current ransomware detectors, in order to propose an effective ransomware detector prototype in future research.

6 Conclusion

The detection of ransomwares has become an important undertaking that involves various advanced solutions for improving security. In this paper, a survey is provided to present existing methods in the literature, and the advantage/limitation of each study. This survey also aims to provide a manual that the researchers can follow as they explore the technologies and methods now available for ransomware attack detection. Each study in this survey have shown that ML and DL are two effective approaches to detect new variants and families of ransomware. Nevertheless, additional effort is required to keep up with new ransomware attacks to fulfill all security defense requirements.

In our future work, we intend to investigate the possibility of combining machine learning and deep learning approaches, in order to develop a ransomware detector that can deal with sophisticated ransomware, and capable of identifying ransomware from both benign files and other variants of malware.

References

1. Trend Micro Incorporated, The next tier—8 security predictions for 2017—security predictions, 2017. Accessed 25 June 2022
2. Sophos, Sophos state of ransomware 2022 report (2022)
3. Kaspersky, Ransomware attacks and types—how encryption trojans differ (2022). Accessed 4 April 2022
4. L. Constantin, New locky ransomware version can operate in offline mode (2016). Accessed 30 June 2022
5. T. Burt, New action to disrupt world’s largest online criminal network (2020). Accessed 23 April 2022
6. C. Burdova, What is eternalblue and why is the ms17-010 exploit still relevant? (2020). Accessed 23 April 2022
7. Map of how tens of thousands of computers were infected with wan-nacy (2017). Accessed 30 June 2022
8. Malwarebytes Threat Intelligence,. What is emotet malware and how to protect yourself (2021). Accessed 1 April 2022
9. Wikipedia, Darkside hacker group (2021). Accessed 22 May 2022

10. A. Hobbs, *The Colonial Pipeline Hack: Exposing Vulnerabilities in us Cybersecurity* (In SAGE Business Cases. SAGE Publications, SAGE Business Cases Originals, 2021)
11. Threat Intelligence Team, Ransomware: April 2022 review (2022). Accessed 25 May 2022
12. B. Toulas, Qbot now pushes black basta ransomware in bot-powered attacks (2022). Accessed 25 June 2022
13. S. Gatlan, Linux version of black basta ransomware targets vmware esxi servers (2022). Accessed 25 June 2022
14. Dr. J. Popp, Aids trojan horse (2021). Accessed 23 May 2022
15. J. De Groot, A history of ransomware attacks: The biggest and worst ransomware attacks of all time (2022). Accessed 1 June 2022
16. J. Johnson, Number of new ransomware families 2020 (2021). Accessed 12 June 2022
17. Kaspersky, Iformaiton technology threat evolution in q1 2022. non- mobile statistics (2022). Accessed 28 June 2022
18. Ayed Al Qartah, *Evolving Ransomware Attacks on Healthcare Providers*. PhD thesis, Utica College (2020)
19. Kasey Panetta, The top 8 cybersecurity predictions for 2021–2022 (2021). Accessed 30 June 2022
20. Ransomware actors increasingly demand payment in monero. Accessed 1 June 2022
21. S. Poudyal, K.P. Subedi, D. Dasgupta, A framework for analyzing ransomware using machine learning, in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (IEEE, 2018), pp. 1692–1699
22. Y.-L. Wan, J.-C. Chang, R.-J. Chen, S.-J. Wang, Feature-selection-based ransomware detection with machine learning of data analysis, in *2018 3rd International Conference on Computer and Communication Systems (ICCCS)* (IEEE, 2018), pp. 85–88
23. Y. Takeuchi, K. Sakai, S. Fukumoto, Detecting ransomware using support vector machines, in *Proceedings of the 47th International Conference on Parallel Processing Companion* (2018) pp. 1–6
24. S.H. Kok, A. Azween, N.Z. Jhanjhi, Evaluation metric for crypto-ransomware detection using machine learning. *J. Inf. Secur. Appl.* **55**, 102646 (2020)
25. S.I. Bae, G.B. Lee, E.G. Im, Ransomware detection using machine learning algorithms. *Concurr. Comput.: Pract. Exp.* **32**(18), e5422 (2020)
26. G. Cusack, O. Michel, E. Keller, Machine learning-based detection of ransomware using sdn, in *Proceedings of the 2018 ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization* (2018), pp. 1–6
27. F. Khan, C. Neube, L. Kumar Ramasamy, S. Kadry, Y. Nam, A digital dna sequencing engine for ransomware detection using machine learning. *IEEE Access* **8**, 119710–119719 (2020)
28. B. Mohammed Khammas, Ransomware detection using random forest technique. *ICT Express* **6**(4), 325–331 (2020)
29. A. Dash, S. Pal, C. Hegde, Ransomware auto-detection in iot devices using machine learning. *no. December* (2018). pp. 0–10
30. S. Egunjobi, S. Parkinson, A. Crampton, Classifying ransomware using machine learning algorithms, in *International Conference on Intelligent Data Engineering and Automated Learning* (Springer, 2019. pp. 45–52
31. J. Hwang, J. Kim, S. Lee, K. Kim, Two-stage ransomware detection using dynamic analysis and machine learning techniques. *Wireless Pers. Commun.* **112**(4), 2597–2609 (2020)
32. H. Zhang, X. Xiao, F. Mercaldo, S. Ni, F. Martinelli, A.K. Sangaiah, Classification of ransomware families with machine learning based onn-gram of opcodes. *Futur. Gener. Comput. Syst.* **90**, 211–221 (2019)
33. M. Masum, Md J. Hossain Faruk, H. Shahriar, K. Qian, D. Lo, M. Islam Adnan. Ransomware classification and detection with machine learning algorithms, in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (IEEE, 2022). pp. 0316–0322
34. S. Kashif Shaukat, V.J. Ribeiro, Ransomwall: A layered defense system against cryptographic ransomware attacks using machine learning, in *2018 10th International Conference on Communication Systems & Networks (COMSNETS)* (IEEE, 2018). pp. 356–363

35. A. Ashraf, A. Aziz, U. Zahoor, M. Rajarajan, A. Khan, Ransomware analysis using feature engineering and deep neural networks. *arXiv preprint arXiv:1910.00286* (2019)
36. S. Maniath, A. Ashok, P. Poornachandran, VG Su- jadevi, Prem Sankar AU, and Srinath Jan. Deep learning lstm based ransomware detection, in *2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE)* (IEEE, 2017). pp. 442–446
37. G. AbdulsalamYa'u, G. Kuwunidi Job, S. Mustapha Waziri, B. Jaafar, N. Ado SabonGari, I. Zahraddeen Yakubu, Deep learning for detecting ransomware in edge computing devices based on autoencoder classifier, in *2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECOT)* (IEEE, 2019). pp 240–243
38. Thezoo: Make the possibility of malware analysis open and available to the public. Accessed 1 April 2022
39. Malware-traffic-analysis. a source for pcap files and malware samples. Accessed 1 April 2022
40. Inc. hybrid analysis gmbh. free automated malware analysis service - powered by falcon sandbox. Accessed 1 April 2022
41. Virusshare.com—because sharing is caring. Accessed 1 April 2022
42. Virustotal. api scripts. Accessed 2 April 2022
43. A source for packet capture (pcap) files and malware samples. Accessed 2 April 2022
44. A real-world dataset. Accessed 2 April 2022
45. Virustotal. Accessed 2 April 2022
46. Weka. Accessed 2 April 2022
47. Riss: Resilient information systems security—ransomware dataset. Accessed 4 April 2022
48. Ransomware detection using machine learning—github. Accessed 4 April 2022

Improving the Application of Blockchain Technology for Tracking Processes in the Supply Chain Integrated Business Intelligence



Khadija El Fellah, Adil El Makrani, and Ikram El Azami

Abstract Trust, traceability, and especially real-time tracking and transparency emerge as critical factors in the design of circular blockchain platforms in the management of supply chains consisting of multiple parties, by adopting various tracking and tracing technologies. Therefore, it influences the perspective of the business intelligence process. In this article, we illustrate the needs and requirements for managing supply chain projects in companies by tracking and tracing their products. This type of tracking and tracing is particularly needed in distributed architectures engaged in project-based enterprises where multiple suppliers are involved in a single project. This tracking and tracing data can be widely used for the business intelligence process. In that case, it will affect the performance of the supply chain by using Ethereum smart contracts to help companies manage a crisis in a targeted manner, after the discovery of such a crisis, and provide distributed trust. In addition, tracking and tracing data allows consumers to ensure that this information has not been tampered with.

Keywords Blockchain · Tracking · Supply chain · Business intelligence · Performance

1 Introduction

An increasing number of customers want to know where their items come from and want to be sure that the products are authentic, which is a major concern after many food scandals. Customers no longer trust the food they buy and are looking for more

K. El Fellah (✉)

Laboratory of Research in Informatics, FS, UIT, Kenitra, Morocco

e-mail: khadija.elfellah@uit.ac.ma

A. El Makrani · I. El Azami

Department of Informatics, Laboratory of Research in Informatics, FS, UIT, Kenitra, Morocco

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

201

A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,

Studies in Computational Intelligence 1102,

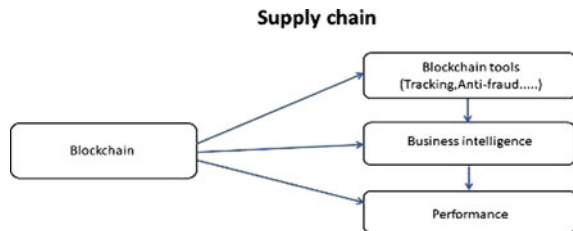
https://doi.org/10.1007/978-3-031-33309-5_16

accurate information about its origins. Product tracking could potentially be revolutionized by a recently created technology called a blockchain [1]. The fact that this technology is decentralized and safe is one of its main advantages, according to many people. Moreover, it lacks a modifiable owner or authority. For this reason, organizations are thinking about integrating blockchain technology into their operations [2]. Digital money, notably the well-known “Bitcoin,” was initially developed by Satoshi Nakamoto in 2008 using this technology [3]. Eventually, a number of other networks were developed, including “Ethereum,” which has increased the possibilities for application in many industries (financial, pharmaceutical, food, etc.) owing to smart contracts. Additionally, since the blockchain system stipulates the collection of raw data and maintains it in an immutable ledger, we advise utilizing such features to conduct business intelligence in supply chain management. Because blockchain is by far the most common use of business intelligence(BI), no other platform is as closely associated with it as BI [4]. Blockchain-based business intelligence solutions can significantly simplify and expand business intelligence capabilities [5]. In the parts that follow, we will go into greater detail about the suggested idea.

2 Methodology

To carry out our work, we adopt the following methodology: Part one will discuss how blockchain can affect supply chain management by using some tracking and tracing tools to collect data from different sources. We also highlight some examples of blockchain application in several sectors so you can see how it impacts supply chain performance. Part two will discuss how business intelligence can benefit from these tracked data to improve their process, and Part three will discuss how business intelligence can affect supply chain management. Finally, concluding by discussing about the results of this work (Fig. 1).

Fig. 1 Conceptual research model. *Source* prepared by the authors



3 Literature Review

A. Blockchain in the Supply Chain

(1) Blockchain

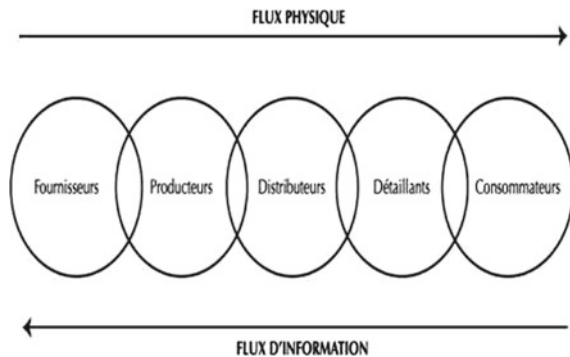
Blockchain is a type of distributed ledger technology that stores records of digital data or “exchanges” in a manner that makes them tamper-resistant. Therefore, when a transaction is requested through the system, the system broadcasts it to a peer-to-peer network of multiple linked computers called nodes. Each of them calculates equations to verify the transaction’s consistency across the network and check for errors. There are numerous validation procedures to demonstrate an honest validation of a block, The proof of work (PoW) and proof of stake (PoS) mechanisms are the two that are most frequently utilized [6]. Therefore, after being verified, the transaction is collected with other transactions to form a block of information for the ledger.

The blockchain allows for sophisticated interactions Rather than just storing or transferring currency thanks to the smart contracts concept, a smart contract is a piece of software that executes predetermined conditions on each node of a blockchain network, automatically [3] verifying that the contract can be done without the need for a third party. If all conditions are met, the contract will be executed.

(2) Supply Chain and Logistic.

The supply chain is composed of three flows: the **information flow**, which involves partner coordination to manage the daily movement of goods and materials up and down the supply chain; it also involves long-term planning; this flow includes, among other things, product characteristics, supplier information, supply strategies such as lead times and prices, information related to customer service providers (customs), sales history, etc. The **physical flow** comes next, and it addresses the processing, transportation, and storage of products and materials. Finally, there is the **financial flow**, which deals with the transfer of money between suppliers and within the business. If these three flows are properly coordinated, client needs will be addressed (Fig. 2).

Fig. 2 Supply chain diagram. *Source* Supply Chain Management, 2019, p. 33 [7]



Logistics is a component of the supply chain that involves managing the flow of materials and information within an organization using available resources, strategies, and IT procedures. The primary objective of logistics is to get finished products to their end user at the lowest possible cost without sacrificing quality. In logistics activities, we find the receiving, storage, preparation, and transportation of products. Logistics is “the technology used to control the physical flow of materials and goods transported in a supply chain between multiple industrial, commercial and service companies”.

(3) *Blockchain in the Supply Chain*

The companies may track any transaction using blockchain technology, which also enables the sharing of documents, personal information, and digital currencies. It is extremely difficult to manipulate the ledger because it is widely dispersed over the network. A modification to the ledger requires that it be simultaneously posted to all network nodes. In the absence of this, the network detects a discrepancy between a record and the others and reports the transaction as counterfeit.

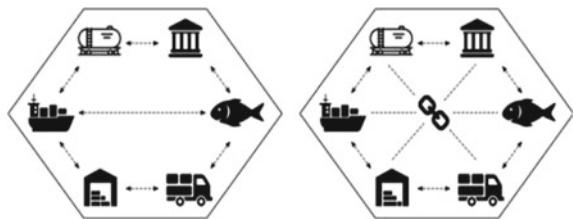
Therefore, every business that manufactures or sells products that rely on external resources or labor may profit from blockchain systems’ support and encouragement of transparent oversight. Supply chain management and logistics are made possible by data tracking with thorough oversight in an immutable record.

The supply chain can benefit greatly from blockchain. In fact, this technology can be used in a variety of supply chain activities, but the applications that generate the most attention would be the traceability of a product since its manufacture [8].

Supply chains will benefit from the immutability and transparency of blockchain. In a blockchain-based supply chain, both suppliers and traders are miners. Miners send new transactions with other miners when the state basket product is updated. Validation of transactions must be checked by everyone involved in the product lifecycle [9].

As illustrated in the image above (Fig. 3), which compares the representation of interactions in a traditional supply chain on the left with a blockchain implementation on the right, all information communicated between participants is recorded on a distributed ledger. Only network users, such as suppliers, manufacturers, assembly facilities, and distribution centers, have access to transactions.

Fig. 3 Traditional supply chain versus supply chain using blockchain. (Source chowdhury mohammad et colman alan, 2018) [10]



(4) *Supply Chain Efficiency Tools Based on Blockchain*

Blockchain has undoubtedly transformed Bitcoin from a little-known virtual currency with little value to a currency that can be used for transactions anywhere in the world, and the supply chain can adopt these features to enhance its performance thanks to its benefits such as security, traceability, and anti-counterfeiting.

(a) *Traceability*

According to Rouse, traceability is defined as the “Possibility of following a product at its production, transformation, and marketing, especially in the food industry” (Rouse definition).

The traceability component is particularly crucial in the supply chain because it enables customers to be certain of the product’s origin and the path taken before arriving at his basket [11]. On the other hand, traceability is a guarantee of product quality for the supplier in the view of the customer and control organizations. The supply chain must then accomplish this goal using a massive array of data and traceability-related information to function in good faith. These data must be dependable, accurate, sincere, correct, believable, precise, and of high quality.

In this sense, the supply chain benefits from blockchain technology, which creates a system of data collection and security throughout a product’s life cycle. The supply chain’s participants enter information into the blockchain that is relevant to them. As a result, everyone involved in the supply chain can know where this product has been and how each of their partners has contributed. The consumer can access all the information at the conclusion of the cycle.

The most popular traceability tools include bar codes, labels with radio frequency identification (RFID), cards, and others.

(b) *Fight Against Fraud and Counterfeiting*

The biggest obstacles to international trade are fraud and counterfeiting, which are adversely affected by the introduction of technology 4.0 [12]. It has an impact on all industries, including those producing food, automobile parts, pharmaceuticals, children’s games, and other goods [13].

Major consequences include fabricated goods that do not conform to sanitary production requirements and diseases brought on by the use of illegal substances in the manufacturing process. Every year, counterfeit medications result in the deaths of 700,000 individuals globally [14]. Blockchain can be the solution to this problem thanks to a system of product information control and verification to prevent the circulation of falsified products.

(5) *Blockchain Implementation Cases in Industries*

The following are some examples of sectors that have integrated blockchain into their supply chains:

(a) *Food Industry*

It should be noted that some companies in the industry have implemented blockchain technology with the aim of traceability of food products, and this is to make the brands more transparent, in response to food frauds that have caused scandals such as that

of 2013 related to a retailer who sold beef containing horsemeat. Additionally, the adoption of the blockchain prevents any modification or deletion of data referring to the stages of manufacture from production to sale. As a result, monitoring of data in the blockchain is done in an encrypted manner, which encourages transparency between the actors in the chain and enhances the security of the data [7].

Take the example of TE FOOD, IBM Food Trust, Provenance and Ambrosus.

(b) *Pharmaceutical Industry*

The goal of the pharmaceutical supply chain is to combat fake medications. Ten percent of medicines in use in low- and middle-income countries are inferior or fake, according to a study performed by the OMS in 2017.

Blockchain technology is being used in clinical trials with the aim of promoting transparency and traceability while preventing data manipulation. Moreover, another technology integrated with blockchain is the Internet of Things (IoT), which aims to gather patient data.

The pharmaceutical sector would be able to keep a digital record of every action taken on pharmaceuticals from manufacturing through distribution, and the actors in the pharmaceutical supply chain may use this record to confirm the legitimacy of the items.

Thus, Blockpharma tracks each box of medication manufactured by a laboratory throughout the supply chain cycle by registering it in a chain of private blocks. As a result, this company provides its customers with an application that enables them to scan a QR code of a box of medicine from their smartphone and quickly see all the information about the product, enabling them to confirm the legitimacy of the drug [7].

B. *Blockchain in Business Intelligence*

(1) *Business Intelligence*

In various fields, the term “business intelligence” (BI) has different meanings. BI is the technical term for the extraction, transformation, management, and analysis of business data for decision-making. This process is primarily based on large data sets, especially data warehouses, which distribute information or knowledge throughout the company at all levels—strategic, tactical, and operational.

(2) *Blockchain in Business Intelligence*

Historical data are essential for making long-term decisions due to their vital role in strategic analysis, an operational organization can perform accurate analyses, and get more valuable business knowledge by gathering historical data. Since businesses need real-time data to run their operations, blockchain technology offers immediate access to enormous volumes of information. This makes it an ideal tool for seeking out business intelligence. Some of the advantages of using blockchain for this are low cost, perfect security, and automation of collecting data [6].

Users can perform quantitative analysis on supply chain data by using smart contracts to filter out all irrelevant data from blocks and store information about the

progress or current state of the supply chain in their smart contracts. Subsequently reports are created based on analytics and business intelligence using these data. In this scenario, the blockchain system defines the data model, collects raw data, stores it in an immutable block, and then assists with performing BI by executing one or more well designed smart contracts, finally converting these raw data into analytical reports and profitable BI.

In this case, using Ethereum's public blockchain enables seamless authorization of property from multiple sources. It ensures authorization data is immutable and helps build trust between everyone involved in the product lifecycle [9].

(3) *Blockchain Implementation Cases in Business Intelligence*

The possibility of using blockchain to improve the process of business intelligence emerged with the aid of the delivchain Framework, which is suggested to address the drawbacks of the traditional OTIF (on time in full) model, which is the most widely used metric for delivery performance in supply chain management and is additionally used to solve the problem of traditional SCM's lack of trust and transparency [6].

The DelivChain framework's capacity is used to collect all the data about the latest status or progress throughout the supply chain, and then to transform that raw data into insightful analytical reports and useful business intelligence [6].

Business intelligence makes real-time data analytics accessible, which enhances supply chain management. The extracted business data must be affordable and secure. Blockchain technology has many advantages, including low cost, perfect security, and high levels of automation, and it is currently what is driving the growth of business intelligence [6].

C. Business Intelligence in the Supply Chain

It is worth mentioning that many companies that intend to improve their supply management and make more efficient decisions are opting for business intelligence (BI) [5]. Nevertheless, logistics professionals find it difficult to exploit the large volumes of data that are widely dispersed. This requires making the best decisions to optimize the company's supply chain, based on harmonized and reliable KPIs.

(1) *The Real Benefits of Business intelligence for the Supply Chain*

- **Real time savings:** BI allow us to predict the demand with more accuracy and thus to ship the products as soon as possible. It also facilitates the choice of a suitable carrier according to the identified needs and constraints. Thus, business intelligence allows us to immediately transform raw and unstructured data into usable information.
- **Cost and risk reduction:** Thanks to BI, decision-makers have a reliable view of the company's results, but also of its costs and expenses. This requires improvements to reduce costs and increase margins. Inventory management can be optimized to have the right goods at the right time, in the right quantity. The exploitation of data also allows us to reduce the costs related to suppliers.

- **More effective collaboration:** Thanks to the use of business intelligence dashboards, the company's collaborators can share their results quickly and safely. Therefore, adopting a business intelligence solution is an excellent way to improve collaboration at all levels of the company. Finally, bi allows the company to optimize its supply chain at different levels. It is also supposed to help managers make the right strategic choices.
- **The pressure of real time:** The deployment of business intelligence (bi) in supply chain management has become commonplace. The main goal is to facilitate decision making for better supply chain planning, i.e., the ability to render, visualize, analyze and share structured data through dashboards. The bi tools do not allow us to respect the cut-off times of the carriers, but they allow us to receive information in real time and to be notified by email or sms at each important step.
- **Limited business intelligence and analysis in the supply chain:** The bi has advantages in the long term and is not instantaneous but for the supply chain, the information is brought back late, which prevents the piloting of operations. However, thanks to the emergence of new technologies to improve real-time visibility, companies can now ensure the availability of the product ordered and the smooth running of operations related to its transport.

4 Discussion and Results

A complicated structure known as the supply chain regulates the transportation of goods. Currently, transparency, relatively low transaction costs, and immediate applications of technology might help in product security. Blockchain's robust and decentralized capabilities are often applied to financial systems, but they might also be quickly adapted to other operations such as supply chain tracking and contracts.

Different organizations can profit from using business intelligence software in a variety of ways. Research demonstrates that integrating BI systems can reduce expenses and provide organizations with a competitive advantage in addition to enhancing performance, increasing efficiency, planning resources, and encouraging collaboration with suppliers and buyers. This is because they may gather data from multiple sources using tracking and tracing tools rather than several spreadsheets. These tools include tracking KPIs and goals connected to the supply chain. Teams may visualize their supply chains with the use of reports and dashboards provided by BI software. Through the use of this software, teams may track goals such as KPIs utilizing a blockchain-based tracking system. Using an immutable ledger of transactions can ensure data security and trust.

The drawbacks of blockchain technology are still worth discussing despite the variety of potential it brings to the sector. We highlight three issues that need to be

solved in order to enable widespread adoption of the technology in regard integrating blockchain into the current SCM system.

- **Scalability:** The total number of transactions will significantly increase once a blockchain system is implemented. Due to the distributed ledger's distributed nature, each network member must maintain a separate copy of the ledger to verify transactions and mine new blocks. This practice inevitably leads to data redundancy and database overload [6].
- **Performance:** Blockchain uses the Merkle tree to more efficiently and safely encrypt blockchain data. If many transactions are entered into each block, the calculation of the Merkle tree hash could cause the processing to take more time overall. Therefore, we may conclude that comparably poor performance is one of the main challenges to blockchain integration [6].
- **Privacy:** Participants in a blockchain system are identified by their key pairs. Other users cannot directly discover the true identity by reading the ledger on the distributed network. However, being anonymous does not mean you cannot be found. Classical blockchains cannot perfectly protect user privacy, as there is still an opportunity to reveal identity by observing one or more fixed transaction patterns in the ledger [6].

Finally, industrial organizations need a powerful application and systematic system capable of operating in real-time, providing insightful tracking of supply chains, logistics, and operations that are closely linked to sales tracking applications, hourly, daily, and monthly production tracking, financial data tracking and many other business data sources for business performance management purposes. Combining technologies such as RFID and IOT can achieve that kind of real-time tracking data.

5 Conclusion

According to several academics, blockchain technology is being positioned as a future technology in multiple industries as it can bring benefits to users while striving to advance their business operations regardless of the industry. Thus far, if blockchain can successfully integrate many stakeholders (participants or users), this will be their biggest challenge going forward, not to mention security concerns and the ability of blockchains to handle ever-increasing data volumes.

Supply chain companies can use blockchain to record production updates on a single shared ledger, providing complete data transparency and a single source of truth. Organizations can access the status and location of products at any time, as transactions are always time-stamped and up-to-date. This reduces problems with counterfeit goods, substandard goods, delays and waste. Additionally, the ledger audit trail ensures compliance with laws and regulations and enables rapid reaction to situations such as product recalls. In addition, the supply chain can realize automated monitoring of production, transportation, and quality control by combining

blockchain with smart technologies such as the Internet of Things. Companies can also choose to share track and trace data with customers as a tool to confirm product legality and ethical supply chain practices.

Developing a blockchain-based tracking system with built-in powerful technology to track data from various sources such as IoT or RFID can help curb the spread of false or modified data. Promote trust, transparency and traceability, and optimize communication between network participants.

We propose that a blockchain-based tracking system can use these tracking data in the business intelligence process. By executing one or more well-designed smart contracts and then finally converting the raw data into analytical reports, supply chain performance is impacted, as supply chain management dashboards enable organizations to track KPIs through rigorous tracking (e.g., cash-to-cycle-time, Perfect Order Rate, Customer Order Cycle Time, and Inventory Turnover), investigate key metrics with detailed analysis in each widget, and identify risks in their processes and mitigate them through action plans.

References

1. J. Biggs, S.R. Hinish, M.A. Natale, M. Patronick et al., Blockchain: revolutionizing the global supply chain by building trust and transparency
2. Y. Zhang, C. Zhang, Improving the application of blockchain technology for financial security in supply chain integrated business intelligence. *Secur. Commun. Netw.* **2022**, 1-8 (2022). <https://doi.org/10.1155/2022/4980893>
3. S. Nakamoto, Bitcoin: a peer-to-peer electronic cash system
4. X. Tan, K. Mojtahed et al., The safe financial processing method for realizing supply chain integrated business intelligence using blockchain application scenarios. *Mob. Inf. Syst.* **2022**, 1-15 (2022). <https://doi.org/10.1155/2022/1454855>
5. F. Ji, A. Tia, The effect of blockchain on business intelligence efficiency of banks. *Kybernetes* **51**(8), 2652–2668 (2022). <https://doi.org/10.1108/K-10-2020-0668>
6. M.H. Meng, Y. Qian et al., The blockchain application in supply chain management: opportunities, challenges and outlook. *EasyChair* (2018). <https://doi.org/10.29007/cv1j>
7. C. Juré, What is the role of blockchain technology in logistics and supply chain? the case of a logistics company in Geneva, p. 63. Juré, Quel est le rôle de la technologie Blockchain dans la Logistique et la Supply Chain ? : le cas d'une entreprise de logistique à Genève, p. 63
8. M.P. Caro, M.S. Ali, M. Vecchio, R. Giuffreda et al. Blockchain-based traceability in Agri-Food supply chain management: A practical implementation, in *IoT Vertical and Topical Summit on Agriculture-Tuscany (IOT Tuscany)*. Tuscany, Mai 2018, pp. 1–4 (2018). <https://doi.org/10.1109/IOT-TUSCANY.2018.8373021>
9. S. Su, K. Wang, H.S. Kim et al., Smartsupply: smart contract based validation for supply chain blockchain, in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Halifax, NS, Canada, Juill 2018, pp 988–993 (2018). https://doi.org/10.1109/Cybermatics_2018.2018.00186
10. M.J.M. Chowdhury, A. Colman, M.A. Kabir, J. Han, P. Sarda, Blockchain versus database: a critical analysis, in *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. New York, NY, USA, Août 2018, pp 1348–1353 (2018). <https://doi.org/10.1109/TrustCom/BigDataSE.2018.00186>

11. A. Kumar, S.K. Srivastava, S. Singh et al., How blockchain technology can be a sustainable infrastructure for the agrifood supply chain in developing countries. *J Glob Oper Strateg Sourc* **15**(3), 380-405 (2022). <https://doi.org/10.1108/JGOSS-08-2021-0058>
12. K. Rauniyar, X. Wu, S. Gupta, S. Modgil, A.B. Lopes de Sousa Jabbour, et al. Risk management of supply chains in the digital transformation era: contribution and challenges of blockchain technology. *Ind. Manag. Data Syst.* **123**(1), 253–277 (2023). <https://doi.org/10.1108/IMDS-04-2021-0235>
13. P. Dutta, T.M. Choi, S. Somani, R. Butala et al., Blockchain technology in supply chain operations: applications, challenges and research opportunities. *Transp. Res. Part E Logist. Transp. Rev.* **142**, 102067 (2020). <https://doi.org/10.1016/j.tre.2020.102067>
14. D. Peltier-Rivest, C. Pacini et al., Detecting counterfeit pharmaceutical drugs: a multi-stakeholder forensic accounting strategy. *J. Financ. Crime* **26**(4), 1027–1047. <https://doi.org/10.1108/JFC-06-2018-0057>

Studying Consensus Mechanisms for Blockchain



Hamza El Mezouari and Fouzia Omary

Abstract In the world of computing, Peer-to-Peer and Blockchain systems are gaining ground as they operate without central parties like companies, organizations, or individuals. Despite their popularity, the performance of distributed systems is limited by the consensus mechanism used. There are various algorithms in use, and there is always room for improvement. Some of these protocols are Proof of Work (PoW), Proof of Stake (PoS), and Proof of Reputation (PoR). Hence, it's crucial to compare these consensus algorithms for a better understanding of blockchain execution. Our paper examines various factors that impact blockchain performance, such as algorithms, security, scalability, and energy efficiency. We explore ways for developers to enhance the consensus mechanism by changing the way nodes participate in the blockchain based on factors like work, stake, and reputation. We compare a few selected consensus algorithms based on Energy usage, Scalability, Immutability, Tolerance of adversaries and Throughput.

Keywords Consensus · Peer to Peer · Blockchain · Proof · Reputation · Distributed systems

1 Introduction

Blockchain has undergone four transition steps so far and is in the preparation phase for the fifth. The first transition, known as Blockchain 1.0, saw the emergence of cryptocurrencies, such as bitcoin, for gambling, small value payments, and foreign exchange. The second transition, Blockchain 2.0, saw the introduction of smart contracts through platforms like Ethereum, Hyperledger, and Codius.

H. El Mezouari (✉) · F. Omary

“Intelligent Processing and Security of Systems”, Computer Science Department, Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco
e-mail: hamza_elmezouari@um5.ac.ma

F. Omary

e-mail: f.omary@um5r.ac.ma

The third transition, Blockchain 3.0, has led to the application of blockchain in various domains such as healthcare, access control systems, identity management, and electronic voting. The fourth transition, Blockchain 4.0, has focused on exploring the application of the technology in smart cities, the Internet of Things (IoT), manufacturing, and agriculture. The fifth transition, Blockchain 5.0, is currently underway, with the primary goal of merging blockchain with artificial intelligence and the semantic Web 3.0.

In Refs. [1, 2], Blockchain operates on a peer-to-peer architecture system, where nodes are interconnected without a central server to coordinate them. The network is comprised of simple nodes and miner nodes. Miners communicate with each other to validate transactions, form new blocks, and add them to the chain. In a blockchain system, there is no central party that decides whether a transaction is accepted or a block is mined. Rather, these systems rely on consensus among peers in the network to make changes to the chain. New blocks are linked using hash functions and stored on each node in the network.

For new blocks to be added to the chain, a consensus mechanism must be executed. The miner must receive validation from other nodes before deciding to add or reject the block. Thus, the consensus process is a crucial step in a blockchain application. There are numerous algorithms currently in use, with room for improvement and innovation. The most commonly used protocols are Proof of Work, Proof of Stake, and Delegated Proof of Stake.

Outline:

The rest of this paper is organized as follows: In Sect. 2, we briefly cover popular consensus protocols. In Sect. 3, we examine alternative consensus methods based on innovative aspects. In Sect. 4, we explore the potential for developers to create new or enhance existing consensus methods. In Sect. 5, we compare PoW with other consensus algorithms (Fig. 1).

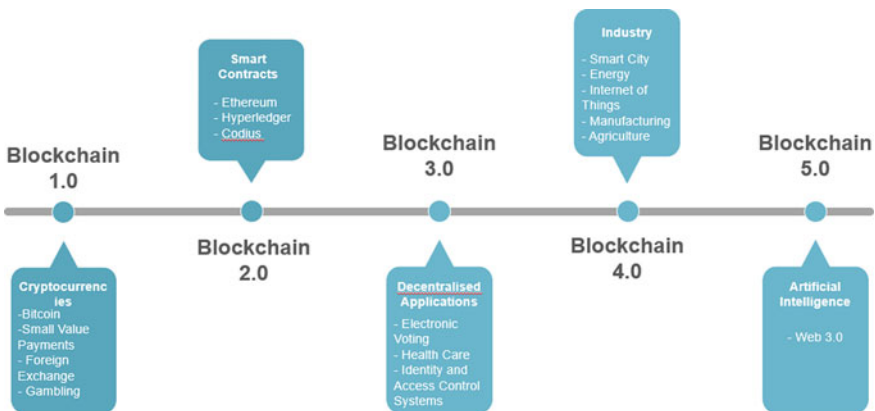


Fig. 1 Overview of blockchain transitions

2 Popular Consensus Protocols

In this paragraph we present the most used consensus in blockchains such as Proof of Work, Proof of Stake, Delegate Proof of stake.

a. Proof of Work:

Proof of Work (PoW) was the first consensus mechanism used by Bitcoin. PoW and mining are closely related. Calculating PoW requires a vast amount of computing power, which is why PoW blockchains are secured and validated by nodes called miners located all over the world who compete to solve a mathematical problem and add new transactions to the chain. The miner who succeeds is rewarded by the network with a predefined amount of cryptocurrency (Fig. 2).

PoW has several significant advantages, making it a reliable method for establishing a secure P2P network. As the value of cryptocurrency increases, more miners join, adding to the security and power of the network. The vast amount of processing power required makes it infeasible for a malicious node to hack the system.

However, PoW is an energy-intensive process that can struggle to process and validate a large number of transactions simultaneously. This is why other algorithms, such as Proof of Stake, have been developed and are becoming more popular.

b. Proof of Stake (PoS):

Many individuals have found that PoW has several drawbacks, with scalability being the primary issue that must be addressed.

Proof of Stake (PoS) aims to find a new way of achieving consensus in a peer-to-peer system. Instead of solving a puzzle, miners who wish to create a block must demonstrate that they own a certain number of coins in order to be selected for the next consensus process and validate new transactions.

The network selects a winner based on the amount of cryptocurrency each validator holds and the length of time they have held it. Once the last block is

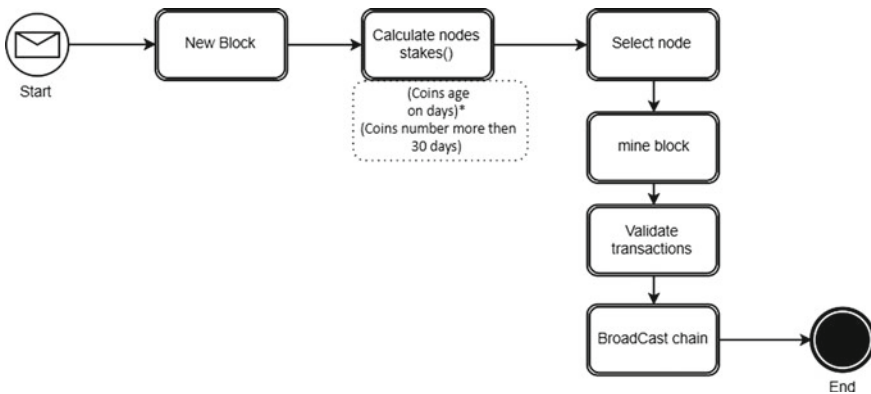


Fig. 2 PoS workflow

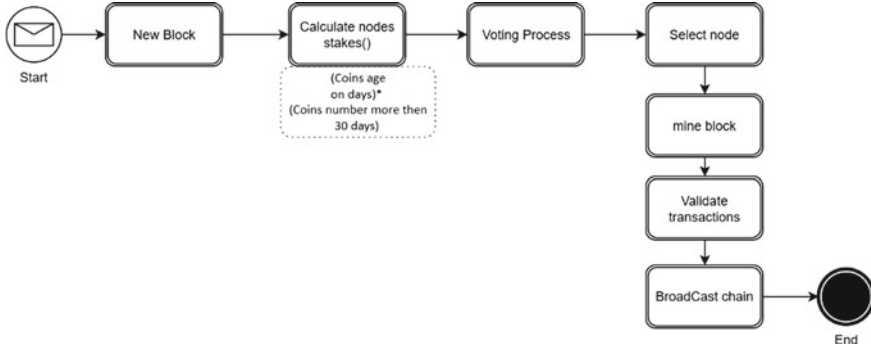


Fig. 3 DPoS workflow

validated by the highest-ranked participant, other validators must confirm its accuracy. The blockchain is updated only if a sufficient number of confirmations are received (Fig. 2).

c. **Delegated proof of stake (DPoS):**

DPoS, similar to PoS, allows miners to create blocks based on their accumulated stake. However, PoS operates through direct democracy, while DPoS uses a representative democracy where nodes vote for delegates to create and validate blocks (Fig. 3).

The main duty of elected delegates in DPoS consensus is block production. They are referred to as witnesses in this consensus mechanism. The witnesses are responsible for generating blocks and validating transactions. They receive a reward for each successful block publication, which is then shared with users who participated in the elections.

3 Alternative Protocols

a. **Proof of reputation:**

In Refs. [3, 4], PoR, participants are judged based on their confidence and reputation. Participants who publish new blocks receive trust, which increases their reputation. Like other protocols that award participants and miners with network factors for creating blocks or good behavior, the PoR rewards participants with trust, which isn't transmittable. Participants whose trust value is constantly high can give better and more stable services than those who do not.

Unlike PoW, which selects the node that solves the hash problem fastest to publish a block, PoR selects the node with the highest reputation in the group. When the validators validate and publish the block, the node's reputation increases (Fig. 4).

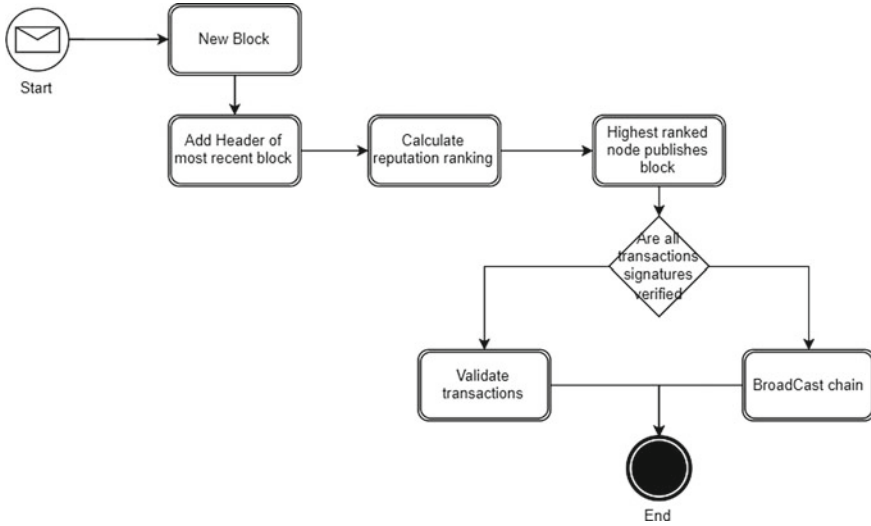


Fig. 4 PoR workflow

At the start of each consensus cycle, a group of validators must be defined, consisting of nodes with reputation values higher than a predefined threshold and a combined reputation value greater than 50% of all node reputations. One member of the group is selected randomly to play the leader role and carry out the mining task.

The leader performs the following functions:

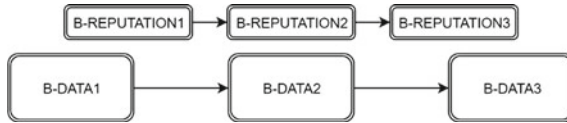
- Filling a new block with all valid transactions from the pending transaction list.
- Calculating the new reputation values of nodes in the network.
- Broadcasting the commit message to the group of validators.
- Receiving validation messages from the group of validators.

The blockchain storage is divided into two chains, one for reputation storage and one for data storage, represented by the Reputation Block and Data Block respectively:

$$\text{Reputation Block}_i = (\text{ReputationBlockHeader}_i, \text{ReputationList}_i).$$

- ReputationBlockHeader_i: header that contains meta information about a specific block.
- ReputationList_i: list that contains all nodes in the network, with each node linked to its reputation value from the previous round.

When a new data block is added to the transaction chain, a corresponding new reputation block (ReputationBlock) is also included in the reputation chain.



b. Proof of Experience:

In PoEx, a novel consensus mechanism based on previous proof of work, the aim is to reduce energy consumption and increase throughput. Miners who validate new blocks will see their mining difficulty decrease (Fig. 5).

At the start of the blockchain, all nodes have the same value for the local target (local difficulty level), which is equivalent to the global target (global difficulty level). The more experienced node is, the more its local target is reduced from the global target. The miner’s local target is recalculated and verified by other nodes after each new block publication.

Compared to Proof of Work, to increase the number of possibilities and prevent miner nodes from being limited by the number of possible nonces, PoEx increases the nonce size to 64 bits.

c. Novel Consensus Mechanism Based on Quantum Teleportation:

In the near future, the potential of quantum computing will be realized. There is already a number of research being done in Refs. [5–8], on how to utilize the qubit state to create an innovative and promising consensus mechanism. The consensus proposed is based on quantum teleportation, which is fundamentally different from classical consensus algorithms as it only executes quantum operators and does not consume computing power (Figs. 5 and 6).

In a quantum blockchain, there are two types of nodes: miners and simple nodes. Miners compete to find solutions and add blocks to the chain, while simple nodes use quantum currency for transactions. To validate transactions, users have to create

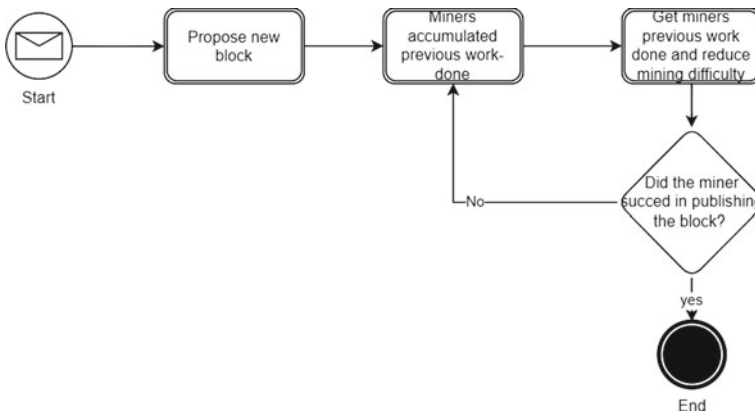


Fig. 5 PoE workFlow

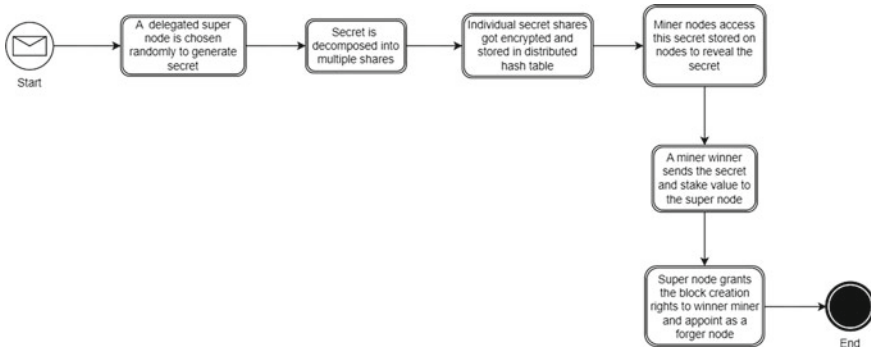


Fig. 6 DPoA workflow

a quantum state ψ comprised of N photon sequences. Each block consists of a block header and a block body. The header contains the sequence number of photons from the previous block header and a random value, while the body contains a set of transactions.

For each voting process, quantum nodes must generate a vote v_i in the form of a quantum password. The leader node then uses the AND operator to calculate the result $R = \text{AND}(v_1 \otimes v_2 \otimes v_3 \otimes \dots \otimes v_n)$. The result is represented by quantum states $\{|0,0\rangle$ and $\{|1,1\rangle$, indicating a “yes” or “no” response, respectively.

d. Delegated Proof of Accessibility (DPoA):

DPoA was designed in 2022 by researchers at an Indian university in Ref. [9]. It is a new consensus mechanism that combines proof of stake and a secret sharing technique. There are five types of participants in the P2P network: (Fig. 6).

- **Delegated Super Nodes:** Delegated nodes must randomly choose one of them to launch the block creation process.
- **P2P Nodes:** Simple nodes connected to the network, each with a secret and public key.
- **Miner Nodes:** Miners are capable of finding the shared secret key and must communicate with other nodes to intercept transactions from simple nodes, forming new blocks and adding them to the chain.
- **Secret Shareholders:** Nodes that hold a part of the shared secret key.
- **Forger Node:** The node that successfully brings the distributed shared keys together, responsible for creating the next block.

Consensus Algorithm Steps:

- A randomly chosen delegated super node generates a secret.
- The secret is divided into multiple shares.
- Individual secret shares are encrypted and stored in a distributed hash table.

- Miner nodes access the secret stored on nodes to reveal the secret.
- The winning miner sends the secret and stake value to the super node.
- The super node grants the block creation rights to the winning miner and appoints them as the forger node.

4 Discussion

In this paragraph, we discuss the possibilities for developers to create new or improve existing consensus mechanisms. Most new consensus mechanisms are based on Proof of Work (PoW), with slight modifications to the implementation algorithms, such as changing the way the proof is calculated or selecting miners. PoEx is an example of a PoW-inspired mechanism where miners benefit from their work experience to mine new blocks.

Another approach is to improve the hash function, by implementing a novel function that is personalized and optimized based on specific characteristics, instead of using predefined functions like MD5, SHA-1, SHA-2, etc.

There is also potential to change the communication protocol between nodes, by using web service APIs instead of sockets. As quantum computers become more advanced, some new consensus mechanisms are exploring the use of quantum teleportation for sending messages in the form of a series of qubits.

Finally, hybridizing two or three consensus mechanisms to take advantage of their strengths and overcome their limitations can also be a promising avenue for improvement. This approach could result in enhanced security, energy efficiency, throughput, and other parameters.

5 Consensus Comparison

Some characteristics of a consensus blockchain include:

1. Scalability: As the number of users and transactions increases, the blockchain can be scaled to accommodate that growth.
2. Immutability: Transactions on the blockchain are recorded in a tamper-proof manner, making them unchangeable once added to the chain.
3. Interoperability: Blockchain networks are designed to communicate with other networks.
4. Throughput(TPS): Transactions are processed in near real-time, making it a more efficient way to transfer value than traditional systems (Fig. 7).
 - a. **Throughput(TPS)**

As we can see in the chart novel consensus tend to have higher throughput than PoW-based networks because validators are chosen based on their stake, experience

Protocols	Energy consumption	Scalability	Immutability	Interoperability	Throughput(TPS)
PoW					
PoS	Less	High	Less	Less	High
DPoS	Less	High	Less	Less	High
PoR	Less	High	Less	Less	High
PoE	Less	High	Less	Less	High
CQB	Less	High	Less	Less	High
DPoA	Less	High	Less	Less	High

Fig. 7 Consensus comparison chart, in Ref. [10]

or reputation etc. and do not need to solve complex mathematical problems, which leads to faster confirmation times and a higher number of transactions that can be processed per second. However, the exact throughput may vary depending on the specific implementation and the number of authorities used.

PoW-based networks tend to have lower throughput than the theoretical CQB, as it uses quantum computing which is much more powerful than classical computing. However, CQB is still in research phase and there is no actual implementation of CQB yet, therefore, it’s hard to predict the exact throughput of CQB based networks.

b. Immutability:

PoW based blockchains have a higher level of immutability compared to PoS, as it is more computationally expensive to alter previous blocks in the chain. So novel consensus may be less immutable than PoW, as it could be theoretically possible for a validator with a significant stake, experience, reputation etc. in the network to alter previous blocks if it is compromised or malicious.

Quantum computing has the potential to solve complex mathematical problems much faster and more efficiently than traditional computers, which could greatly increase the security and immutability of a blockchain network. However, the technology is still in its early stages, and there are some concerns about the potential for quantum-based attacks on CQB systems, which could make them less immutable than PoW-based systems.

c. Interoperability:

PoW is a widely used consensus mechanism that is implemented in many different blockchain networks, such as Bitcoin and Ethereum. Because of its widespread use, it is relatively easy for different PoW-based networks to interoperate with each other through the use of atomic swaps or other cross-chain communication protocols. However, it’s still possible for other blockchain networks to interoperate with other networks, but it could be more difficult to achieve.

In summary, PoW-based networks tend to be more interoperable than CQB-based networks, due to the widespread use of PoW and the existence of standard protocols for cross-chain communication. However, CQB being a theoretical concept and not

an actual implementation yet, it's hard to predict how interoperable it would be with other networks.

d. Energy consumption:

Overall, alternative consensus systems are generally considered to be more energy-efficient than PoW systems, as they do not require as much computational power and therefore consume less energy. However, the energy consumption of PoS systems can vary depending on the specific implementation and the number of validators participating in the network.

It is hard to compare the energy consumption of PoW and Quantum blockchain as the latter is still in research phase and hasn't been implemented yet, but in theory it is expected that quantum blockchain would consume less energy as it uses less computational power.

e. Scalability:

Alternative consensus systems are generally considered to be more scalable than PoW systems, as they do not require as much computational power to validate transactions and create new blocks. In PoS for example, validators are chosen based on the amount of cryptocurrency they hold and are willing to "stake" as collateral, which allows for a larger number of validators to participate in the network and validate transactions. This allows for a higher throughput and faster confirmation times, resulting in better scalability.

It is hard to compare the scalability of PoW and CQB as the latter is still in research phase and hasn't been implemented yet, but in theory it is expected that CQB would provide better scalability than PoW.

6 Conclusions

In this article, we examined and compared the features of commonly used and innovative consensus mechanisms, as well as how existing consensus mechanisms can be improved to better meet the demands of the IT industry. The advent of quantum computing is certainly the next major transformation for blockchain systems, given its tremendous computing power, which can solve scalability and security issues. We also discussed various approaches that researchers take to create new consensus mechanisms by modifying existing ones or developing new technologies that are revolutionary in the IT domain.

References

1. D. Berdik, S. Otoum, N. Schmidt, D. Porter, Y. Jararweh, A survey on blockchain for information systems management and security. *Inf. Process. Manage.* **58**(1), 102397 (2021)
2. S. Nakamoto, Bitcoin: a peer-to-peer electronic cash system, s. d. 9
3. F. Gai, B. Wang, W. Deng, W. Peng, Proof of reputation: a reputation-based consensus protocol for peer-to-peer network, in *Database Systems for Advanced Applications*, éd by J. Pei, Y. Manolopoulos, S. Sadiq, J. Li, vol. 10828. *Lecture Notes in Computer Science.* (Springer International Publishing, Cham, 2018), pp. 666-81
4. O. Aluko, A. Kolonin, Studying the applicability of proof of reputation(PoR) as an alternative consensus mechanism for distributed ledger systems, in *Computer Science & Information Technology (CS & IT)* (AIRCC Publishing Corporation, 2021) pp 41–53
5. A.R. Faridi, F. Masood, A.H.T. Shamsan, M. Luqman, M.Y. Salmony, Blockchain in the quantum world. *Int. J. Adv. Comput. Sci. Appl.* **13**(1) (2022)
6. J. Seet, P. Griffin, Quantum consensus, s.d. 9
7. H. Wang, J. Yu, A blockchain consensus protocol based on quantum attack algorithm, in *Computational Intelligence and Neuroscience*, ed by D. Zhang (2022), pp. 1–6
8. X. Wen, Y. Chen, W. Zhang, Z.L. Jiang, J. Fang, et al. Blockchain consensus mechanism based on quantum teleportation. *Mathematics* **10**(14), 2385 (2022)
9. M. Kaur, S. Gupta, D. Kumar, C. Verma, B.C. Neagu, M.S. Raboaca, Delegated proof of accessibility (DPoAC): a novel consensus protocol for blockchain systems. *Mathematics* **10**(13), 233 (2022)
10. D.P. Oyinloye, D. Peter, J.S. Teh, N. Jamil, M. Alawida et al., Blockchain consensus: an overview of alternative protocols. *Symmetry* **13**(8), 1363 (2021)

Design and Construction of a Smart Agricultural Greenhouse



Moulay Ahmed Bekri, Abdellah Idrissi , Said Ouabou, and Abdeslam Daoudi

Abstract A smart greenhouse is a greenhouse that integrates Internet of Things technology to improve the productivity of vegetables, fruit and plants, rationalize water consumption and automatically monitor the greenhouse. In this way, Internet of Things technology is used to collect and analyze bioclimatic indicators of the greenhouses in real time, so that the necessary measures and actions (automatic, semi-automatic or manual) can be taken. Various sensors (with or without internet connection) are used to monitor the greenhouses and measure environmental standards according to the needs of each crop. This eliminates the need for static monitoring in the greenhouses. These sensors provide information on water level, pressure, humidity and temperature and automatically control the triggers to turn on the irrigation pumps, turn on the lights, control the heaters and turn on the fans. This paper presents an integrated system used to measure temperature, humidity, light, and soil moisture in greenhouses and control water levels in irrigation ponds. The measurement data is shared and managed using IoT. The data collected is recorded in a database in order to make the necessary and optimal decisions for the greenhouse (like FIRBASE). The system allows farmers to monitor their greenhouses from their mobile phones or computers connected to the Internet.

Keywords IOT · Smart green house · Sensor network · Temperature · Humidity

M. A. Bekri (✉) · A. Idrissi · S. Ouabou · A. Daoudi
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco
e-mail: moulayahmed.bekri@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,
Studies in Computational Intelligence 1102,
https://doi.org/10.1007/978-3-031-33309-5_18

225

1 Introduction

By 2050, the WHO predicts that we will need to produce 70% more food due to the exponential rise in global population. The need to increase agricultural productivity has become urgent due to the loss of agricultural land and the depletion of natural resources.

Growing market competition and high quality standards have driven greenhouse farming's development in recent years. Systems for greenhouse production are becoming more complex and hence significantly more expensive.

Therefore, in order to maximize their investments, greenhouse producers must better control their production conditions if they hope to remain competitive. Automatic system adjustments are necessary to fully benefit from expanded plant opportunities.

Today, improved climate control in greenhouses is a crucial global issue. In the world of agriculture, this is especially accurate. Modern greenhouse systems must be improved to better accommodate the usage of novel cultivation methods and associated equipment [1].

Crops must be grown in ideal circumstances in order to increase profitability. Therefore, it is crucial to keep an eye on the temperature, humidity, light, soil moisture, etc.

The autonomous control of climatic processes has substantially benefited from advancements in computer technology. Climate research in greenhouses now requires a vital tool: computerized management. enhances the climate and environment for greenhouse cultivation. However, it is the sole rational method for predicting climate change in greenhouses.

Currently, the technology trend is to use wireless technologies such as the Internet of Things. It is integrated in several areas such as: Examples: military, industry, agriculture, etc. Internet of Things is a very broad and rich term. One can imagine a whole world interconnected and able to communicate through the exchange of information between objects.

In agriculture, agricultural greenhouses can be transformed into smart greenhouses using the Internet of Things.

Indeed, intelligent agricultural greenhouses are able to adapt to a certain evolution of the environment using inputs and outputs but also means of communication and even electronic processing interfaces.

As a result, the agriculture of tomorrow will be automated and agricultural production will be based on the concept of the intelligent agricultural greenhouse which will manage not only the criteria of humidity, light and watering but also the ease of control and of access by the farmer which will certainly allow him a more optimal production on all the plans [2].

The Internet of Things allows farmers to remotely control their agricultural fields, especially with regard to parameters such as:

- Humidity.
- The light necessary for agricultural production.

- Watering.
- The temperature.
- Water level.
- Etc.

There are two key sections to our work. The initial step entails creating an automated, self-regulating agricultural greenhouse with a collection of sensors, actuators, and microcontrollers.

The creation of an Android app for remote control of the greenhouse using IOT is the second step. Therefore, the objective is to showcase every piece of hardware and software that was needed to complete the assignment.

2 Hardware Tools Used in the Greenhouse





A. *Sensors:*

A crucial part of greenhouse monitoring systems are sensors. Each sensor records its continuous readings of a particular condition, such as temperature or humidity in a particular area, and sends them to the system. One of the input terminal strips on the base unit is linked to each sensor. You must match your needs with the number of inputs available because each condition requires a different input.

We utilized the sensors listed in the following Fig. 1 in our smart greenhouse:

To measure soil humidity, temperature, and humidity, we install a soil moisture sensor, a temperature and humidity sensor, and both inside the greenhouse. The microcontroller authorizes the pump to start watering the smart greenhouse when the temperature, humidity, or soil moisture surpasses a particular threshold. Once

Fig. 1 Sensors used

Sensors	Image
Soil Moisture Sensor	
Water level sensor	
Humidity and temperature sensor	
Light Sensor	

the greenhouse become watering, the pump automatically stops thanks to the action of the soil moisture sensor [3].

2 *Actuators:*

The microcontroller continuously analyzes the digitalized data parameters of different sensors, compares them to predefined threshold values, and determines whether corrective action is necessary for the condition at this moment. If so, it triggers the actuators to carry out a controlled operation. Relays, contactors, switches, and other actuators can be employed in the system. They are used to turn on or off the pumps that are used to water plants or fill irrigation basins, as well as to turn on fans and greenhouse lights.

In our smart greenhouse, we used the actuators described in the following Fig. 2.

3 *Communication Module:*

We need to send readings of temperature and other quantities (voltage, luminosity, etc.). This information can be entered into a database, used in calculations or otherwise, it all depends on what is fact, for this we need a Transmitter, a Receiver and a transmission medium, in our case we have worked with the HC05 Bluetooth module shown in the following Fig. 3.

4 *Arduino UNO:*

Arduino Uno is a microcontroller board based on the ATmega328P. It has 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analog inputs, a 16 MHz ceramic resonator (CSTCE16M0V53-R0), a USB connection, a power jack, an ICSP header and a reset button. It contains everything needed to support the microcontroller; simply connect it to a computer with a USB cable or power it with a AC-to-DC adapter or battery to get started.. You can tinker with your Uno without worrying too much about doing something wrong, worst case scenario you can replace the chip for a few dollars and start over again (Fig. 4) [4].

Fig. 2 Actuators used




Actionneur	Image	Rôle
Pump		The water pump's duties include supplying water to the water basin and irrigating the greenhouse.
Fan		Ensures ventilation of the greenhouse
Lamp		The lamp is used to illuminate the greenhouse when needed

Fig. 3 HC-05 bluetooth module

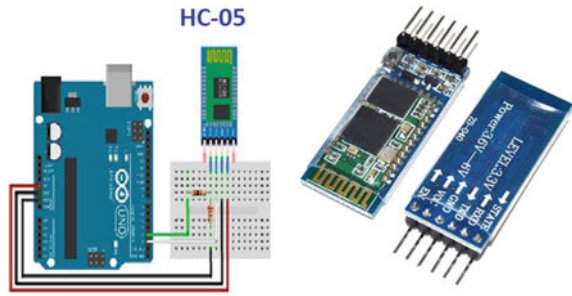
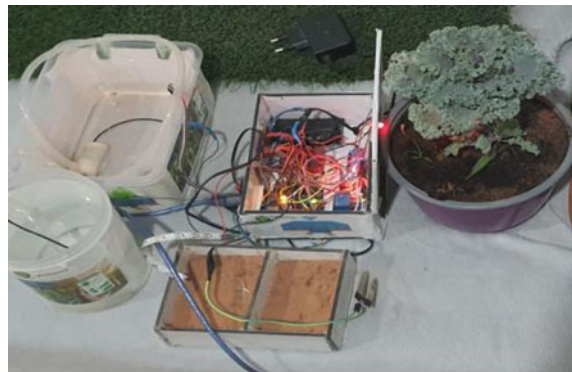


Fig. 4 Arduino UNO card



Fig. 5 Greenhouse implementation



3 Realization and Implementation of Our Greenhouse

We used the ARDUINO IDE, APPInventor for the Android application to remotely manage and monitor our greenhouse, and Google sheet to gather and record data for the need of analysis and data processing (Fig. 5).

The developed application contains various interfaces that have been successfully implemented to facilitate the use of applications and provide a good user experience, namely the remote control of the greenhouse as needed.

Fig. 6 Database interface



The Fig. 6 is the Database interface where the storage and recording of all greenhouse data in order to analyze them to make the right decisions.

4 Conclusion and Perspectives

As part of our work, we built an automated “smart” greenhouse for agriculture. An Android application developed was used to remotely control the greenhouse using Internet of Things technology. Farmers can use it to control parameters such as temperature, humidity and water level via Smartphones.

This system that can help improve agriculture by facilitating the management of information carried out by the application developed.

We enjoyed working on this subject, because this work is an extraordinary idea that will help us in our daily lives. This system thus developed will constitute a step that can contribute to improving the economy of the country thanks to a series of agricultural compensations that avoid dependency, food security, etc.

This work is lively, stimulating and motivates new research. We think we have seen some of our future professional lives.

Despite the efforts we have made to accomplish this humble task, despite the concepts we have appropriated, we see that our contribution is only the beginning of a long journey.

The works expressed in [5–16] deserve to be applied in this environment. Concretely, the work we have done can be improved, supplemented and continued in a number of ways, including:

- Machine learning and use of Big Data.

- Use of the Raspberry card to replace the control computer and the wifi module to connect the capture interface with this card.
- Adapt our platform for the management of large volume greenhouses.
- Possibility of controlling several greenhouses in a network.
- Control more parameters.
- Use other types of advanced commands such as neural networks.

References

1. K. Mesmoudi, Etude Expérimentale et Numérique de la Température et de l'Humidité de l'Air d'un Abri Serre Installé dans les Haut Plateaux d'Algérie, Région des Aurès. Thèse de Doctorat Physique Energétique, option énergétique Université de Batna, 2010
2. Y. Elafou, Contribution au contrôle des paramètres climatiques sous serre. Thèse de Doctorat Université Lille 1, 2014
3. Z. Ala-Eddine, Une approche IoT pour la mise en œuvre des serres intelligentes connectées. Mémoire de fin d'étude Master, Université de biskra, 2018
4. Kaoutar Hafdi. Proposition et validation formelle d'une architecture Reidy fiable et dynamique destinée aux systèmes IoT - Application aux Smarts Grid. Thèse De Doctorat, Novembre 2020.
5. M. Essadqi, A. Idrissi, A. Amarir, An Effective Oriented Genetic Algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**:170–179, 2018
6. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electr. Eng.* **73**:289–303, 2018
7. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment, in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, 2016
8. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Rob. Intell. Syst.* **14**(3):65–70, 2020
9. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. arXiv preprint arXiv: 1709.10372. 2015
10. A. Idrissi, CM. Li, JF. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks, in *18th IEEE International Conference on Tools with Artificial Intelligence*. Washington, USA, 2006
11. A. Idrissi, K. Elhandri, H. Rehioui, M. Abourezq, Top-k and Skyline for Cloud Services Research and Selection System. *Int. conf. Big. Data. Adv. Wireless technol.* 2016
12. M. Abourezq, A. Idrissi, Integration of QoS Aspects in the Cloud Service Research and Selection System. *Int. J. Adv. Comput. Sci. Appl.* **6**(6), 2015
13. M Abourezq, A Idrissi and H Rehioui. An amelioration of the skyline algorithm used in the cloud service research and selection system. *International Journal of High Performance Systems Architecture* 9 (2-3), 136-148. 2020.
14. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data. Min.*, 2017
15. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10**, 2020
16. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>

Implementation and Management of a Home Automation Control System (Smart Home)



Abdeslam Daoudi, Abdellah Idrissi , Moulay Ahmed Bekri, and Said Ouabou

Abstract Recently, the internet has grown in incredible ways. Moreover, the use of the Internet is not limited to network management, but also has extended to the management of objects, this is called Internet of things, among the most in view of the use of this new technology we find the field of home automation, currently called the smart home. Indeed, the smart home market is expected to experience increasing demand, due to the availability of comfort, protection and security equipment, as well as the reduction in energy costs. In this paper, we made a model on which we integrated a system to have a real preview. This system contains the main standards that must contain a smart home and all this will be controlled thanks to an application installed on your smartphone. We created a database for the storage and analysis of data in order to understand the behaviour of the inhabitants to make the house more automatic, that is to say to have a self-management of the house following the usual behaviours residents already registered in our database without using the application itself.

Keywords Internet of Things · Home automation · Smart home

A. Daoudi · A. Idrissi · M. A. Bekri (✉) · S. Ouabou
IPSS Team, Artificial Intelligence and Data Science Group, Computer Science Department,
Faculty of Science, Mohammed V University in Rabat, Rabat, Morocco

A. Daoudi
e-mail: abdeslem.daoudi@um5r.ac.ma

A. Idrissi
e-mail: idrissi@um5r.ac.ma

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,
Studies in Computational Intelligence 1102,
https://doi.org/10.1007/978-3-031-33309-5_19

1 Introduction

Currently, electronics are increasingly being replaced by programmed electronics. We also speak of embedded systems (on-board computing). The objective is to simplify electronic diagrams, reduce the use of electronic components and the cost of manufacturing a product [1].

The Internet of Things (IoT) promises to be an unprecedented evolution. Objects are able to communicate with each other, to exchange, to react and to adapt to their environment on a large scale. The latter marks a new stage in the evolution of cyberspace. This revolution facilitates the creation of intelligent objects allowing advances in multiple fields. Indeed, one of the areas most affected by the IoT is home automation [2], since the home is a place of great importance. All individuals, and especially the elderly, spend the majority of their time at home, hence the considerable influence of the home on the quality and the way of their life. Improving feeling, comfort and security in their own habitat has a deep social impact.

In recent years, technology has been applied to the creation of the Smart House. Smart House is defined as a residence equipped with technologies of electronics, automation, computing and ambient telecommunications which aim to assist the inhabitant in various situations of his domestic life.

2 Problematic

As the contemporary human life has become more complicated and the person has a lot of worries, he needs a revolution both in his daily life and at home, or rather an assistant who completes the routine tasks at its place. In other words, it is the use of an application allowing the execution of the orders given by the inhabitant. Nevertheless, it is sometimes possible that the latter forgets to give this order, in particular at the level of the security of the house and the person. This can cause serious accidents.

3 Approach and Methodology

In our solution, we are interested in the realization of a small model on which we have integrated a system to have a real preview. This system contains the main standards that a smart home must contain (Arduino board, sensors, actuators, etc.). All of this will be controlled through an application created with the online mobile application development tool (Mit App Inventor) installed on the Smartphone. We then created a database under the platform (Firebase) for the storage and analysis of data in order to understand the behavior of the inhabitants and to make the house more automatic, namely a self-management of the latter following the usual behaviors of

the inhabitants already registered in our database without resorting to the application itself.

4 Functional and Technical Study

A. *Study of needs*

The specification of needs is the starting phase of any mobile application development. Through this part we will identify the needs of our application, focusing on functional and non-functional needs.

B. *Functional needs*

The field of artificial intelligence applied to the home is increasingly broad, the applications are multiplying at the rate of technological evolution. Thus, it becomes difficult to identify all the intelligence needs in a habitat, it all depends on our imagination and our creativity.

Each stage of our life has specific needs. These needs of everyday life revolve around the following axes:

- **Comfort:** Living well at home, as long as possible.
- **Security:** The home automation network watches over the occupants.
- **Communication:** guarantee permanent and close contact with loved ones.
- **Health:** ensure follow-up adapted to specific needs.
- **Energy:** this is the optimization of our energy expenditure.

C. *Non-functional needs*

These are the needs that characterize the system. These are performance, hardware type, or design type needs.

Respect for privacy: the presence of sensors to collect relevant information for monitoring activities or adapting intelligent behavior must in no way disturb an individual's privacy.

- **Availability:** in close connection with safety, dependability is an essential characteristic.
- **Interoperability:** limit the heterogeneity of communication technologies and protocols in a habitat.
- **User-friendliness:** our system must be easy to use. Indeed, user interfaces must be user-friendly, ergonomic and adapted to all inhabitants.
- **The speed of processing:** it is essential that the processing time be as short as possible, especially in emergency situations.
- **Confidentiality:** insofar as the data handled by our system makes it possible to know everything that takes place in a habitat, we must guarantee great responsibility and optimal security. Thus, the access rights to the system must be properly assigned.

5 Realization and Tests

D. *Presentation of the Home Automation Functions Offered*

For this home automation system, we have chosen as functions to establish:

a. Lighting management function:

This function allows the management of the lighting of three rooms with the aim of saving electrical energy. Two Sensors perform this function: the PIR motion detector for Passive Infrared Sensor and the LDR photoresistor, light of pendent resistor which is a light sensor. We save energy by switching lights on when someone is around and the light level is very low in the room [3].

b. Main door and garage opening function:

This function ensures the opening of the main door of the habitat and the garage in a more secure way by adopting a keypad access system.

c. Window shutter opening management function:

The user can more easily control the opening of the windows by pressing a simple button on the control application with your smartphone or tablet.

d. Temperature and ventilation acquisition function:

The acquisition of the temperature is done via a DHT11 temperature sensor to control the climate inside the habitat with ventilation [3].

E. *Conception*

a. UML language

UML is a modeling language that makes it possible to express and develop object models, independently of any programming language. The UML is under the full responsibility of the OMG (Object Management Group). It was designed to serve as a support for an analysis based on object concepts. It is defined as a graphical and textual modeling language intended to understand and describe needs, specify and document systems, sketch software architectures, design solutions and communicate points of view [4].

F. *Diagrams*

To give an overall view of the functional behavior of our system, we used the use case diagram. This diagram presents the interactions that will allow actors to achieve their objectives using our system (Fig. 1).

If the actor is the idealization of a role played by an external person, process or thing that interacts with a system, then ours will mostly have only one actor, namely its end user and lesser extent the administrator.

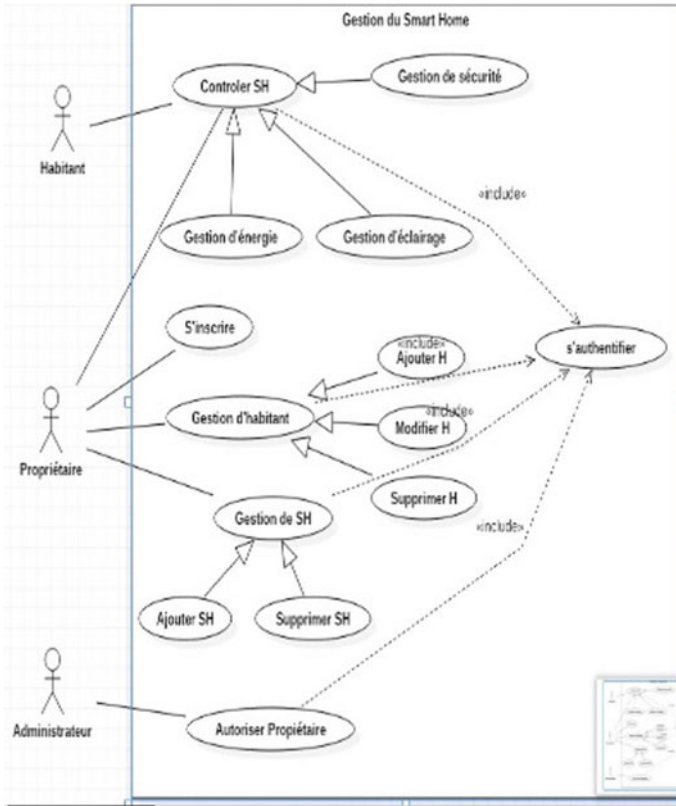


Fig. 1 Usecase diagrams

The user: this is the holder of a smartphone or tablet and a valid account and who requests a service concerning the house in question while respecting his rights of use. For security purposes we define two categories of users:

- *The owner:* this is the main and official user of the house. So he is an administrator at the smart home level.
- *The inhabitant:* This is the secondary user, he can be a family member, friend or other.

The administrator: It is he who ensures the proper functioning of the said system and the management of user accounts in terms of validity and access rights. The role of the administrator is not limited to solving problems, but he must also propose solutions in line with the needs of the user.

G. Application level design

In order to better detail and describe the behavioral aspect of our system, we use sequence diagrams as a popular dynamic modeling solution. This type of modeling is

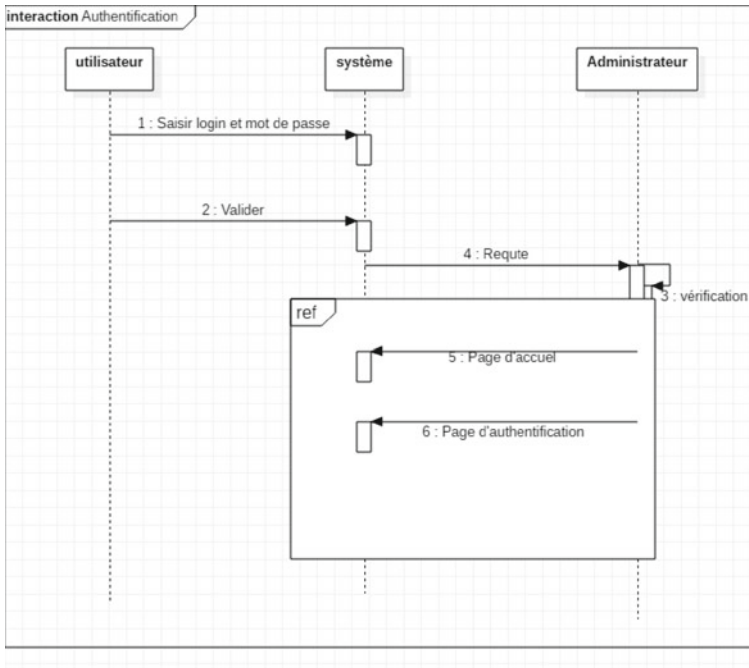


Fig. 2 Sequence diagram

interested in the interactions occurring inside our system according to a chronological order.

H. “Authentication” sequence diagram

Through this diagram, we will describe the scenario of the “Authenticate” use case. First, the user asks the system to allow him to authenticate. Once the requested interface is displayed, it introduces the necessary fields and validates the operation. In turn, the system first ensures the uniqueness and validity of the data before saving it to the database (Fig. 2).

6 Building the House

The first step is to make a scale model of a house. We chose a three-room house with a window and a door. This model would make it possible to present certain functionalities of home automation through seven scenarios: motion detection, secure access to the habitat, opening of the shutters, the window, remote lighting, acquisition of the temperature, detection of gas/smoke and ventilation. These scenarios will be automated through Arduino boards running computer programs (Fig. 3).



Fig. 3 Photo of our smart home

A. Breadboard mounting

To test the proper functioning of our system it was necessary to mount it on a test plate to see if there is no problem before realizing the printed circuit (Fig. 4).

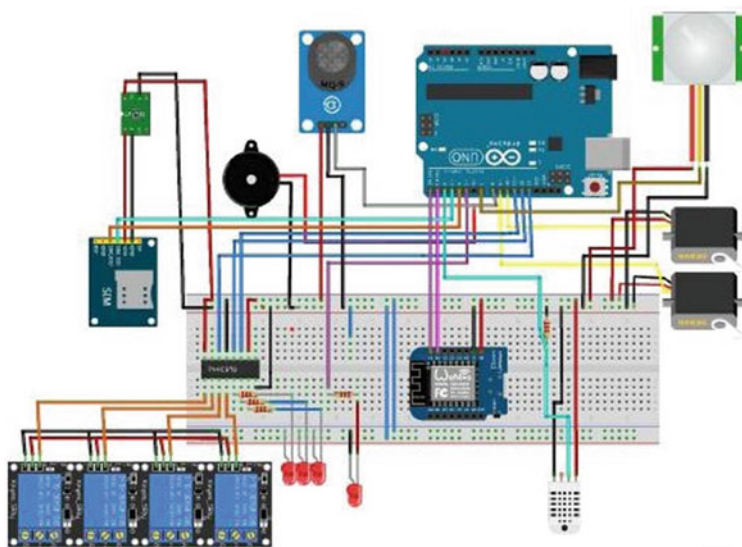


Fig. 4 The test assembly of our smart home with Tinkercad

7 Application Interfaces

A. Home

Opening the application gives access to the home interface.

Besides, the home page is one of the most important pages of the application. It is on it that most visitors to the application make their first impression and their feelings about your content,

It must be clear, readable...

The home page of our application contains access to the main parts of the latter which can be divided into four parts approached in the form of buttons:

- Lighting.
- Home appliance.
- Security.
- Door/curtains (Fig. 5).

B. Website GUI

In this part, we present the visual graphic charter of the website which aims to market the application that we have developed, also to provide details and information such

Fig. 5 Home page





Fig. 6 Website

as download links and to show its main functionalities to everyone in the world (Fig. 6).

C. *Firebase Database:*

Firebase is considered a platform for web applications. It helps developers create quality apps. Data is stored in JSON (JavaScript Object Notation) format, insert, update, delete, or add data without using queries. It is the backend of the system used as a database to store data.

Available services are:

- Firebase Analytics
- Firebase Cloud Messaging (FCM)
- Firebase Auth
- Real-time Database
- Firebase Storage
- Firebase Test Lab for Android
- Firebase Crash Reporting
- Firebase Notifications
- Using firebase in our android application:

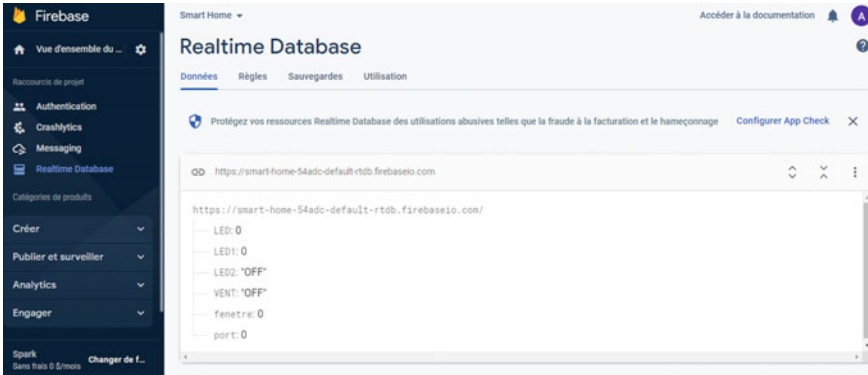


Fig. 7 Database

Firestore Realtime Database can provide data value and updates to a simple API. Thanks to real-time synchronization, users of our application can consult their data from any device (on the web or from their mobile). Note that this database comes with mobile and web SDKs and allows the creation of amplifications without using servers.

For this reason, a database called BD_SH was created at the level of our mobile platform of google FIREBASE (Fig. 7).

8 Conclusion

The purpose of the end-of-studies project is to develop the autonomy and responsibility of the students, to create a group dynamic as well as a spirit of collective work, it also allows the students to put into practice the lessons received, the know-how and skills acquired.

In this project, we focused on the design and realization of a remote control system to control electrical installations for home automation, using a wireless communication protocol (Bluetooth) with a Smartphone.

We have created a real-time measurement system for all physical phenomena based on an Arduino UNO board as a control unit, the role of the Arduino UNO board is to process the data delivered by the sensors and to control different actuators used.

At the beginning, we tried to connect the system to the display medium (PC) by a USB cable to ensure the correct functioning of the sensors. The program written on Arduino IDE allows to display the results on the serial monitor.

Secondly, we developed a command interface under Android (a smartphone application) with the MIT App Inventor environment.

Then, we created a Firestore database, which allows the storage and analysis of data in order to understand the behavior of the inhabitants and to make the house

more automatic, that is to say a self-management of the house following the usual behaviors of the inhabitants already registered in our database without using the application itself.

Such an achievement is not without difficulties. We can say that despite these difficulties, the results obtained through this study, whether practical or theoretical, have proven to be very satisfactory.

Smart home systems are revolutionary systems destined to evolve even more in the future. It must use more effective techniques, like those mentioned in [5–16], to be able to move forward in this project.

References

1. A. Krama, A. Gougui, *Etude et réalisation d'une carte de contrôle par Arduino via le système Androïde* (Université KasdiMerbah Ouargla, Algérie, Mémoire Master Académique, 2015)
2. J. Krumm, *Ubiquitous Computing Fundamentals* (2006)
3. H. Hamouche, Conception et réalisation d'une centrale embarquée de la domotique « Smart Home ». Mémoire Master en Génie électrique, Université Mohammed V École Normale Supérieure d'Enseignement Technique–Rabat (2015)
4. C. Solnon, Modélisation UML, INSA de Lyon-3IF (2013–2014)
5. S. Retal, A. Idrissi, A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks. *Comput. Electr. Eng.* **73**, 289–303 (2018)
6. F. Zegrari, A. Idrissi, H. Rehioui, Resource allocation with efficient load balancing in cloud environment, in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (2016)
7. F. Zegrari, A. Idrissi, Modeling of a dynamic and intelligent simulator at the infrastructure level of cloud services. *J. Autom. Mob. Rob. Intell. Syst.* **14**(3):65–70 (2020)
8. M. Essadqi, A. Idrissi, A. Amarir, An Effective Oriented Genetic Algorithm for solving redundancy allocation problem in multi-state power systems. *Procedia Comput. Sci.* **127**, 170–179 (2018)
9. A. Idrissi, C.M. Li, J.F. Myoupo, An algorithm for a constraint optimization problem in mobile ad-hoc networks, in *18th IEEE International Conference on Tools with Artificial Intelligence*, Washington, USA (2006)
10. A. Idrissi, F. Zegrari, A new approach for a better load balancing and a better distribution of resources in cloud computing. *arXiv preprint arXiv:1709.10372* (2015)
11. A. Idrissi, K. Elhandri, H. Rehioui, M. Abouzeq, Top-k and skyline for cloud services research and selection system, in *International Conference on Big Data and Advanced Wireless Technologies* (2016)
12. H. Rehioui, A. Idrissi, A fast clustering approach for large multidimensional data. *Int. J. Bus. Intell. Data. Min.* (2017)
13. K. Elhandri, A. Idrissi, Comparative study of Top-k based on Fagin's algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10** (2020)
14. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
15. M. Abouzeq, A. Idrissi, H. Rehioui, An amelioration of the skyline algorithm used in the cloud service research and selection system. *Int. J. High Perform. Syst. Archit.* **9**(2–3), 136–148 (2020)
16. M. Abouzeq, A. Idrissi, Integration of QoS aspects in the cloud service research and selection system. *Int. J. Adv. Comput. Sci. Appl.* **6**(6) (2015)

A Comparative Study of Consensus Algorithms Used in Blockchain and Their Adaptation to the IoT Networks



Mohamed Aghroud, Mohamed Oualla, Abdeslam Jakimi,
and Lahcen Elbermi

Abstract IoT devices are typically designed to operate in low-power environments, which means they have limited processing power, memory, and energy resources. Despite these limitations, they are still capable of communicating with other devices without the need for human intervention. There are several challenges for IoT networks such as, heterogeneity of IoT systems, complexity of networks, and resource constraints of IoT devices. Blockchain technology brings opportunities to overcome these challenges. It relies on consensus algorithms; these replace centralized validation. Although blockchain technology has successes in the field of cryptocurrencies such as bitcoin, it can also be applied in other areas such as IoT networks. For this purpose, this paper presents a comparative study on blockchain consensus algorithms for validation and authentication, as well as their adaptations with resource-constrained IoT networks. The comparison between these algorithms is based on the following criteria: computational power, latency, and storage capacity.

Keywords Blockchain · IoT · Consensus algorithm · Consensus protocol · IoT devices

M. Aghroud (✉) · M. Oualla · A. Jakimi · L. Elbermi
Department Computer Science, GLISI Teams, FST Errachidia, My Ismail University, Errachidia,
Morocco
e-mail: mo.aghroud@edu.umi.ma

M. Oualla
e-mail: m.oualla@umi.ac.ma

A. Jakimi
e-mail: a.jakimi@umi.ac.ma

L. Elbermi
e-mail: l.elbermi@umi.ac.ma

1 Introduction

IoT and blockchain are recent technologies, they are arousing enormous interest in the field of scientific research. On the one hand IoT allows to seamlessly interconnect objects so as to create a physical network whose process of data collection, and communication are automatically controlled and managed without human intervention [1]. With the advent of smart homes and smart cities, IoT has become an important development area. On the other hand, blockchain technology with its mechanisms such as security, and decentralization; overcomes the challenges faced in the Internet of Things. To realize an autonomous Internet of Things network, different hardware devices need to be communicated in a distributed manner, then for the communicated data to be valid there is a need for communication protocols, these protocols are called consensus algorithms. Many of the consensus algorithms used in blockchain networks, such as Proof-of-Work (PoW) and Proof-of-Stake (PoS), require significant computational power and data storage capacity. This can make these algorithms unsuitable for resource constrained IoT devices, which may have limited processing power and storage capacity., therefore in this paper we will discuss the different consensus algorithms and their adaptation with IoT networks [2].

This work is structured by parts: in part two we will discuss basics concepts of a blockchain technology. Part three is an overview on IoT. In part four we will examine the integration of blockchain technology into IoT networks. In part five will discuss different consensus algorithms and their adaptation with IoT networks. In part six we will discuss some study's problems. In last part we will give the conclusions.

2 An Overview on Blockchain Technology

A. *Understanding Blockchain Technology*

The foundations of the blockchain were proposed in 2008 by Satoshi Nakamoto [3]. It is a decentralized system for storing and transmitting information in a secure and transparent manner, without the need for a central authority or controlling body [4]. So the Blockchain technology consists of two components: the network and the register, the network can be public or private, and the register as without authorization or with authorization [5], that's why there are three types of blockchains, the first is called public blockchain which anyone can join the network, read and write [6]. The second type named private blockchain which gives access to a certain number of participants [5]. The third type called permissioned blockchain which is a hybrid whose validation process is restricted to a certain number of well-defined nodes [6].

Literally, blockchain refers to digital containers on which information of all kinds are stored. Together, these blocks form a database similar to the pages of a decentralized ledger. Figure 1 shows an example of the blockchain [7].

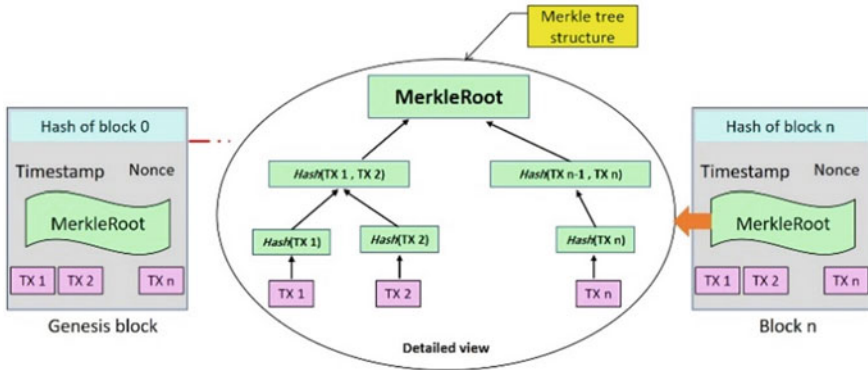


Fig. 1 Block constituents in Blockchain technology

From his definitions we can conclude that:

- Blockchain is a technology for storing and transmitting information.
- Blockchain is a digital database that cannot be falsified.
- Blockchain is a decentralized and open ledger.
- The blockchain allows the creation of trust machines.

Blockchain technology uses consensus algorithms to verify transactions made and update the ledger set, there are several consensus algorithms that we will discuss in Sect. 5.

B. Characteristics of Blockchain

Blockchain brings together the following characteristics:

- Immutability: data cannot be modified or deleted once recorded, it leaves a permanent record of all transactions made on the network since its creation [8].
- Data transparency: is an important aspect of blockchain technology. In a blockchain network, all transactions are recorded on a distributed ledger that is visible to all participants in the network. This means that the transaction is timestamped and cannot be modified, deleted, or reversed without the consensus of the network [9].
- Disintermediation: is one of the key features of blockchain technology. It refers to the removal of intermediaries or middlemen from transactions, with the aim of reducing costs, increasing efficiency, and enhancing security [9].
- Security: the decentralized architecture and the block code guarantee the inviolability of the information, because all the data are copied in the different servers [9].
- Autonomy: In a blockchain network, the computing power and hosting space required to maintain the network are provided by the network nodes, which are typically the users themselves. Therefore, there is no need for central infrastructure [9].

3 An Overview on IoT

A. Understanding the IoT

The Internet of Things (IoT) refers to a network of physical devices and objects that are connected to the internet and can communicate with each other. These devices can range from simple sensors and smart home devices to complex industrial machinery and medical equipment. These devices can include everything from sensors and actuators to household appliances and vehicles [10, 11]. Each object is identified by a unique address in the internet, and is accessible, and programmable [12].

In an IoT network, individual devices are typically identified by their own unique ID, which allows them to communicate and exchange data with other devices in the network [13].

B. IoT Architecture

A typical IoT system consists of layered subsystems that work together to provide a range of industrial services. These subsystems are often organized into different layers, each of which performs a specific set of functions as shown in Fig. 2, starting from bottom to top there is a perception layer which has a wide variety of IoT devices that can sense and collect data from the physical environment [13]. The second communication layer is responsible for network connection which is enabled by various communication protocols such as Bluetooth [14]. The third layer includes industrial applications such as manufacturing, supply chain, healthcare, and vehicle internet [15].

C. Challenges of IoT

The IoT provides the connection of mounted objects with various sensors, actuators, and software systems capable of sensing and collecting information from the physical

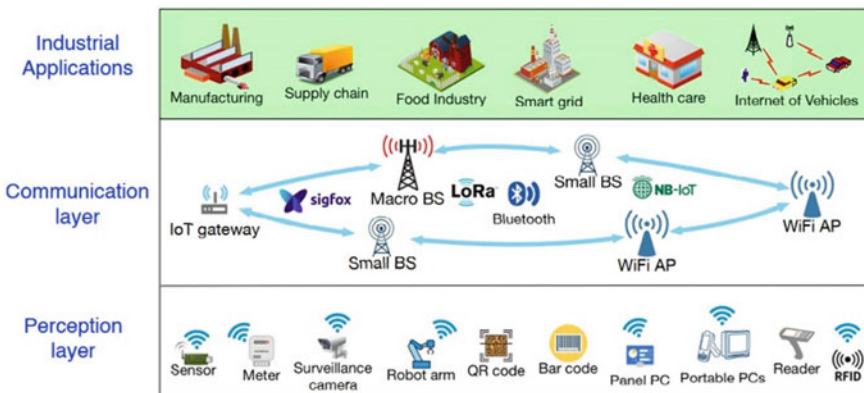


Fig. 2 IoT architecture [15]

environment. The unique characteristics of IoT pose a number of the following research challenges [15]:

- Heterogeneity of IoT systems: refers to the diversity of devices, protocols, and standards used in IoT networks [15].
- Network complexity: there are various communication/network protocols that coexist in IoT such as Bluetooth, 6LoWPAN, Wireless HART, Sigfox, LoRa and NB-IoT [15].
- Low interoperability: the ability of IoT systems (hardware and software) to exchange information. Due to decentralization and heterogeneity of IoT systems IoT interoperability is difficult to achieve [15].
- IoT device resource constraints: IoT devices are limited in storage resource and battery power [16].
- Security vulnerability: The decentralized and heterogeneous nature of Internet of Things (IoT) systems can create significant challenges for ensuring their security [15].

4 Blockchain for IoT Networks

Blockchain technology is based on the P2P model, this type of network allows participants (called network node) to validate the transactions performed according to the consensus algorithm applied, each network node is considered as client and server [17].

Blockchain establishes a P2P network that has several advantages such as reducing installation and maintenance costs, as well as distributing computing and storage requirements among all network devices [17].

There are several reasons why implementing blockchain-based systems in resource-constrained IoT networks can be challenging, despite their built-in mechanisms for ensuring data integrity [17]. However, there are also some challenges associated with using blockchain for IoT networks. One of the main challenges is the scalability of the network. As the number of devices in the network grows, so does the amount of data that needs to be processed by the blockchain. This can lead to slower transaction times and higher fees. Additionally, the security of the network is only as strong as the security of the devices that are connected to it, so it is important to ensure that devices are properly secured and updated to prevent vulnerabilities.

5 Consensus Algorithms

The blockchain can use several algorithms by which different nodes can reach consensus on a new block, Table 1 shows the classification of consensus algorithms. In this section we will present two categories of consensus algorithms namely validation and authentication algorithms. We are going to start with the validation

algorithms. Afterwards, we will compare the first type variants of consensus algorithms (validation algorithms), and we study the second type which is the proof of authentication (PoAh) presented in Table 2 [18].

A. Validation consensus algorithms.

- Proof of Work (PoW)

Is a consensus algorithm used by many blockchain networks to validate transactions and add new blocks to the blockchain. In PoW, nodes on the network compete to solve a complex mathematical puzzle, with the first node to solve the puzzle earning the right to add a new block to the blockchain [19], its most known variants are: Proof of Capacity (PoC) [19], and Proof of Elapsed Time (PoET) [20]. PoC is a consensus algorithm used in some blockchain networks that relies on the storage capacity of a node’s hard disk rather than its computational power., it relies on the storage capacity of their hard disk, indeed this algorithm is not applicable for resource-limited IoT [19].

Miners in the PoET algorithm must solve a hash problem, with the validated miner being randomly selected based on a waiting time. This algorithm is an ideal tool for IoT, since its latency is low, and its throughput is high, But the problem with PoET is that the dependence on Intel, which is at odds with the basics of the blockchain, which is fully decentralized [20].

- Proof of Stake (PoS)

PoS used by some blockchain networks to validate transactions and add new blocks to the blockchain. In PoS, nodes on the network are chosen to validate transactions and add new blocks based on the amount of cryptocurrency they hold or “stake” in the network [19]. This algorithm is not yet an adaptable choice for the resource-constrained IoT because it applies primarily to cryptocurrency which is not applicable for the IoT. Among its variants there are: Delegated Proof of Stake (DPoS) [20], Leased Proof of Stake (LPoS) [21], Proof of Importance (PoI) [22], Proof of Activity (PoA) [19], Casper [23], and Proof of Burn (PoB) [19]. Although its variants bring

Table 1 Classification of consensus algorithms [19]

Consensus algorithms	
Validation algorithms	Authentication algorithms
PoW [19], PoC [19], PoET [20], PoS [19], DPoS [20], LPoS [21], PoI [22], PoA [19], Casper [23], PoB [19], PBFT [19], dPBFT [19], Stellar [24], Ripple [25], Tendermint [26], ByzCoin [19], Raft [19], Tangle [27], Definity [28], Algor and [29], RSCoin [30], Elastico [30], OmniLidger [31], RapidChain [32]	PoAh [33]

Table 2 Comparisons of different consensus validation algorithm [19–34]

Consensus algorithm		Latency ¹	Computing overhead	Storage overhead	Adaptation in IoT
Proof of work	PoW	Haut	Haut	High	No
	PoC	High	Low	Very High	No
	PoET	Low	Low	High	Yes
Proof of stake	PoS	Medium	Medium	High	Partially
	DPoS	Medium	Medium	High	Partially
	LPoS	Medium	Medium	High	No
	PoI	Medium	Low	High	Partially
	PoA	Medium	Haut	High	No
	Casper	Medium	Medium	High	No
	PoB	High	Medium	High	No
Byzantine tuning method	PBFT	Low	Low	High	Yes
	dPBFT	Medium	Low	High	Partially
	Stellar	Medium	Low	High	Partially
	Ripple	Medium	Low	High	Partially
	Tendermint	Low	Low	High	Partially
	ByzCoiin	Medium	Haut	High	No
Raft		Low	Low	High	Partially
Tangle		Low	Low	Low	Yes
Methods based on the VRF	Definity	Medium	Low	N/A	No
	Algorand	Medium	Low	High	No
Sharding-based methods	RSCoin	Low	Low	High	No
	Elastico	High	Medium	High	No
	OmniLedger	Medium	Medium	Low	Partially
	RapidChain	Medium	Medium	Low	Partially

improvements for PoS, but the main drawback in IoT networks is their dependencies on monetary concepts.

- Byzantine agreement algorithms

Byzantine agreement algorithms are a family of consensus algorithms designed to make sure that a distributed network of nodes can agree on a single decision or outcome, even in the presence of faulty or malicious nodes that may attempt to sabotage the consensus process [19].

Practical Byzantine Fault Tolerance (PBFT) is a consensus algorithm designed for use in distributed systems, such as blockchain networks, where a group of nodes must agree on a single decision or outcome, even in the presence of faulty or malicious

¹ High is in the order of minutes, medium is in the order of seconds, low is in the order of milliseconds.

nodes that may attempt to sabotage the consensus process [19]. As long as other variants like Stellar Consensus Protocol (SCP) [24], Delegated Practical Byzantine Fault Tolerance (dBFT) [19], Tendermint [26], Ripple [25], and ByzCoin [19], they rely on monetary concepts, so the low throughput and high latency, therefore cannot be adapted to IoT networks.

- Verifiable Random Function (VRF)

Are algorithms based on random functions whose committee members are randomly selected to participate in the consensus protocol, Definity and Algorand are two variants of this algorithm that are characterized by their higher latency, which is unacceptable for IoT systems [28, 29].

- Sharding algorithms

These algorithms are a type of consensus algorithm used in blockchain networks to increase transaction throughput and scalability. Sharding involves partitioning the blockchain network into smaller, more manageable subsets, called shards, each of which can process transactions in parallel. This allows the network to handle more transactions per second than traditional blockchain networks, which process transactions in a linear fashion [30]. The most recent sharding algorithms are: RSCoin [30], Elastico [30], OmniLedger [31], RapidChain [32] which are based on a centralized monetary supply, so because of their highest latency storage requirements, and security problem, for that they can not be adapted to IoT networks.

- Raft

It is a consensus algorithm for managing a replicated log in a distributed system, such as a distributed database or a blockchain network. It was designed as an alternative to the traditional Paxos algorithm, which can be difficult to understand and implement. However, its performance depends on the dominant node having an absolutely dominant position in the network, which is not very suitable for IoT systems [19].

- Tangle

It is a new technology for distributed ledgers (Heperledger) proposed by the cryptocurrency Iota [20], In the Tangle, each transaction is linked to two previous transactions, forming a web of interlinked transactions. When a new transaction is added to the Tangle, it must confirm two previous transactions, chosen at random from the pool of unconfirmed transactions. This process of confirmation is called “tip selection”. No rules applied by this algorithm to select nodes, which is adaptable for resource-constrained IoT systems [27].

Table 2 illustrates a comparison between previously discussed validation consensus algorithms and their adaptations with IoT systems, Adaptable consensus

algorithms for IoT systems are identified by YES, partially adaptable algorithms are identified by PARTIALLY, and non-adaptable algorithms are identified by NO.

B. Authentication consensus algorithms

In this part, we will discuss the PoAh proof of authenticity, it is proposed to build an adaptable blockchain network for resource constrained IoT systems. This algorithm is based on the authentication mechanism during the validation of blocks [33] (Fig. 3).

Figures 4 and 5 give the comparison of the PoAh and PoW algorithms.

Network participants transact (Trx) from the collected data in order to build a block; in Fig. 4, each block is denoted by the expression $B = \{Trx1, Trx2, \dots, Trxn\}$, nodes disseminate blocks for validation or evaluation, Fig. 5 illustrates the steps for selecting a trusted node [35].

The PoAh algorithm in validating the blocks follows some steps. In the first step, nodes combine multiple transactions to build blocks. In the second step, the nodes sign the blocks with the private key and broadcast them to the network. In the third

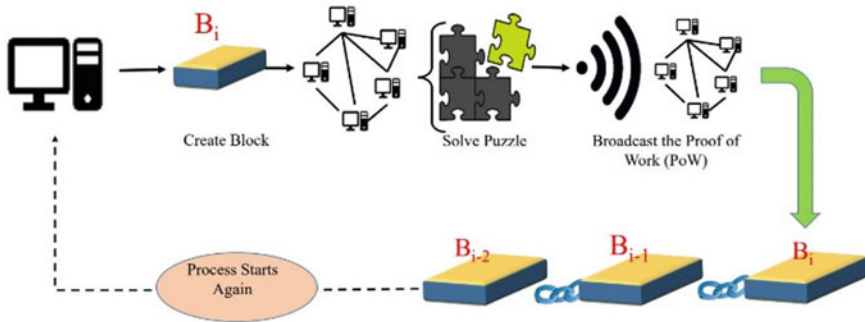


Fig. 3 Proof-of-Work (PoW) consensus algorithm [33]

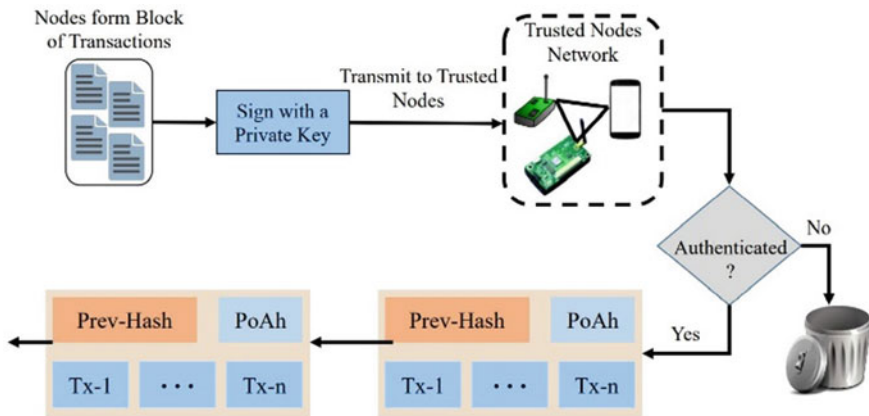


Fig. 4 The Proof-of-Authentication (PoAh) consensus algorithm [33]

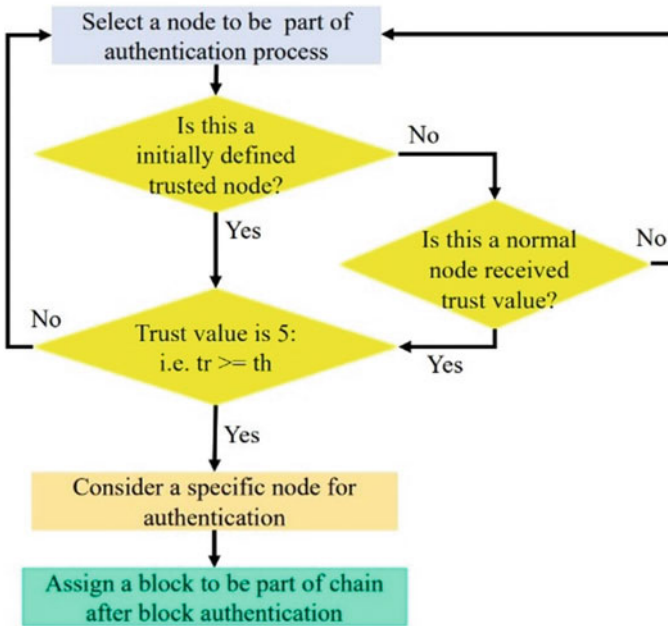


Fig. 5 How to choose an authenticated node to use in PoAh

step the trusted nodes verify the signature of the blocks from the previous step with the source public key, in fact if the block is authenticated then Confidence nodes disseminates authenticated block in the network to add it to the blockchain. And if the block is not authenticated then this block will be deleted, then move to the next block [33].

This algorithm is used blockchain networks to validate transactions and maintain the integrity of the blockchain. The authentication principle uses less resources and energy compared to other mechanisms, which can be very adaptable to IoT systems [33, 35].

C. Discussion

The choice of such an algorithm does not depend on their classification but rather on specific well-defined criteria such as the required computing power, scalability, latency, throughput, security, storage capacity. So from the comparison made previously of different consensus algorithms, we can say that the consensus algorithms of PoET, PBFT, Tangle, and PoAh validation category of authentication algorithms are adaptable to IoT networks, but with some improvements that we will mention in the following section that deals with the research challenges.

6 Open Research Challenges

The major challenge is to apply blockchain technology to meet the specific needs of the desired application, as each application has different specifications, a blockchain implementation is needed for IoT systems, Another research challenge is the improvement of scalability, ensuring security, protecting data privacy, increasing throughput, reducing computational requirements, latency, and limited storage capacity.

In the case of the PoAh consensus algorithm, the criteria for selecting the transactions to form the block is also the selection of the trusted nodes.

Successfull applying blockchain to IoT networks requires solutions that are feasible in practice, scalability to large networks with low latency.

Implementations must be secured against potential attacks, and also must be compatible with resource constrained IoT systems with restricted computing and storage capacity. Each consensus algorithm discussed addresses several of the above issues. However, applications that address all of the mentioned challenges have yet to be implemented.

7 Conclusion

In this paper, we focused on the consensus algorithms used and their practical applicability for resource-constrained IoT devices.

Each of these implementations has addressed some of these limitations, including throughput, latency, computational power, network overhead, scalability, and privacy. However, none of them have been able to address all limitations to an acceptable degree [6].

to realize a large-scale blockchain-based IoT network with low latency would require a hybrid framework or an existent framework with a changed consensus algorithm [6].

Blockchain technology is replacing a centralized authority with consensus algorithms between all the participants to ensure the security of the system in a decentralized system manner [34].

Consensus algorithms that are suitable for large-scale (resource-constrained) IoT networks must satisfy the following constraints:

- Security
- Latency that must be low (in milliseconds order)
- Data storage
- The computing power.

References

1. Md. A. Uddine, A. Stranierie, I. Gondal, V. Balasubramanian, A survey on the adoption of blockchain in IoT: challenges and solutions (2021)
2. M. Salimitari, M. Chatterjee, A survey on consensus protocols in blockchain for IoT networks, p. 1 (2019)
3. S. Nakamoto, Bitcoin : a peer-to-peer electronic cash system (2008) <https://bitcoin.org/en/bitcoin-paper>
4. Blockchain France Associés, La Blockchain décryptée, Observatoire Netexplo (2016)
5. Certified Blockchain Associate, <https://elearning.kba.ai/>, elearning.kba@iitmk.ac.in (2020)
6. R. Romain, G. Ferréol, Principes clés d'une application blockchain. EM Lyon Business School (2016)
7. M. Pignel, la technologie blockchain une opportunité pour l'économie social? p. 6 (2019)
8. Comprendre la Blockchain, Livre blanc sous licence Creative Commons, uchange.co (2016)
9. Pignel, la technologie blockchain une opportunité pour l'économie social? p. 16 (2019)
10. W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge computing: vision and challenges. *IEEE Internet Things J.* **3**(5), 637–646 (2016)
11. A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of things for smart cities. *IEEE Internet Things J.* **1**(1), 22–32 (2014)
12. Introduction to internet of things and cloud, in platform Udemy <https://www.udemy.com/course/a4iot-intro-iot-cloud/learn/lecture/8251204#overview>. Accessed 4 Jan 2022
13. D. Minoli, K. Sohrawy, J. Kouns, IoT security (IoTSec) considerations, requirements, and architectures, dans les Actes de la 14e conférence annuelle de l'IEEE sur les communications et les réseaux grand public (CCNC), pp. 1006–1007 (2017)
14. J. H. Lee, H. Kim, Security and privacy challenges in the internet of things [Security and privacy matters]. *IEEE Consum. Electron. Mag.* **6**(3), 134–136 (2017)
15. H.N. Dai, Z. Zheng, Y. Zhang, Blockchain for Internet of Things: a survey, p. 3. Preprint at [arXiv:1906.00245v3](https://arxiv.org/abs/1906.00245v3) (2019)
16. X. Lu, D. Niyato, H. Jang, D.I. Kim, Y. Xiao, Z. Han, Ambient backscatter assisted wireless powered communications. *IEEE Wirel. Commun.* **25**(2), 170–177 (2018)
17. M. Salimitari, M. Chatterjee, Y.P. Fallah, A survey on consensus methods in blockchain for resource-constrained IoT networks. Department of Computer Science, University of Central Florida, Orlando, FL 32825, United States (2020)
18. M. Salimitari, M. Chatterjee, A survey on consensus protocols in blockchain for IoT networks, p. 2. Preprint at [arXiv:1809.05613v4](https://arxiv.org/abs/1809.05613v4) (2019)
19. J. Debus, Consensus methods in blockchain systems. Frankfurt School of Finance & Management, Blockchain Center, Technical Report (2017)
20. Hyperledger, www.hyperledger.org/blog/2018/11/09/hyperledger-sawtooth-blockchain-security-part-one. Accessed 04 Jan 2022
21. Leasing Proof of Stake, <https://docs.waves.tech/en/blockchain/leasing>. Accessed 04 Jan 2022
22. Proof of Importance, [https://www.techopedia.com/definition/33599/proof-of-importance-poi#:~:text=Proof%20of%20importance%20\(PoI\)%20is,that%20they%20can%20create%20blocks](https://www.techopedia.com/definition/33599/proof-of-importance-poi#:~:text=Proof%20of%20importance%20(PoI)%20is,that%20they%20can%20create%20blocks). Accessed 04 Jan 2022
23. V. Buterin, V. Griffith, Casper the friendly finality gadget. Preprint at [arXiv:1710.09437](https://arxiv.org/abs/1710.09437) (2017)
24. D. Mazieres, The stellar consensus protocol: a federated model for internet-level consensus. Stellar Development Foundation (2015)
25. D. Schwartz, N. Youngs, A. Britto, The ripple protocol consensus algorithm. Ripple Labs Inc White Paper 5 (2014)
26. Tendermint, <https://tendermint.com>. Accessed 04 Jan 2022
27. S. Popov, The tangle, p. 131 (2016)
28. T. Hanke, M. Movahedi, D. Williams, Dfinity technology overview series, consensus system. Preprint at [arXiv:1805.04548](https://arxiv.org/abs/1805.04548) (2018)

29. Y. Gilad, R. Hemo, S. Micali, G. Vlachos, N. Zeldovich, Algorand: scaling byzantine agreements for cryptocurrencies, in *Proceedings of the 26th Symposium on Operating Systems Principles*, pp. 51–68 (ACM, 2017)
30. L. Luu, V. Narayanan, C. Zheng, K. Baweja, S. Gilbert, P. Saxena, A secure sharding protocol for open blockchains, in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 17–30 (ACM, 2016)
31. E. Kokoris-Kogias, P. Jovanovic, L. Gasser, N. Gailly, E. Syta, B. Ford, Omniledger: a secure, scaleout, decentralized ledger via sharding, in *2018 IEEE Symposium on Security and Privacy (SP)*
32. M. Zamani, M. Movahedi, M. Raykova, RapidChain: scaling blockchain via full sharding, in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 931–948 (ACM, 2018)
33. D. Puthal, S.P. Mohanty, V.P. Yanambaka, E. Kougianos, PoAh a novel consensus algorithm for fast scalable private blockchain for large-scale IoT frameworks, p. 9. Preprint at [arXiv:2001.07297v1](https://arxiv.org/abs/2001.07297v1) (2020)
34. M. Salimitari, M. Chatterjez, A survey on consensus protocols in blockchain for IoT networks, p 11. Preprint at [arXiv:1809.05613v4](https://arxiv.org/abs/1809.05613v4) (2019)
35. D. Puthal, S.P. Mohanty, V.P. Yanambaka, E. Kougianos, PoAh a novel consensus algorithm for fast scalable private blockchain for large-scale IoT frameworks, p. 11. Preprint at [arXiv:2001.07297](https://arxiv.org/abs/2001.07297) (2020)

Artificial Intelligence Applied to Some Academic and Real Problems

Dynamics Behavior of Vehicular Traffic Flow in a Scale-Free Complex Network



Siham Lamzabi, Kaoutar El Handri, Marwa Benyoussef,
Hamid Ez-Zahraouy, and Abdelilah Benyoussef

Abstract It is shown in the literature that all the topologies of urban street networks demonstrate a scale-free property. The primary objectives of this paper are to propose a new topology perspective for modeling traffic flow in a scale-free complex network city, understand the dynamics behavior of traffic flow in the proposed model, and identify different parameters that influence the fluidity of the network. The fundamental diagram of the proposed model has been studied. While Analyzing different phases, it is found that throughput is not only a function of the density but depends on several parameters.

Keywords Traffic flow on cities · Scale-free network

1 Introduction

Traffic flow and congestion have become a predominant topic in complex systems due to the rapid increase in vehicles number and transportation demand. Especially during the last decade, the global energy crisis and environmental concerns were the main reasons for improving congestion and optimizing flow. Road traffic is a complex phenomenon, first because of the many factors involved, then because the

S. Lamzabi (✉)

Laboratory of Innovation in Management and Engineering for Enterprise (LIMIE), ISGA Rabat,
Rabat, Morocco

e-mail: siham.lamzabi@isga.ma

K. El Handri

S LIMIE Laboratory, Higher Institute of Engineering and Business, 393 Rte d'ElJadida,
Casablanca, Morocco

IPSS Team, Computer Science Laboratory, Faculty of Science, Mohammed V University in
Rabat, Rabat, Morocco

M. Benyoussef · H. Ez-Zahraouy · A. Benyoussef

Laboratoire de Magnétisme Et Physique Des Hautes Energies, LMPHE (URAC 12) Département
de Physique Faculté Des Sciences, Université Mohammed V-Agdal, Rabat, Morocco

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

261

A. Idrissi (ed.), *Modern Artificial Intelligence and Data Science*,

Studies in Computational Intelligence 1102,

https://doi.org/10.1007/978-3-031-33309-5_21

nature of the network has meshed. Several models have been proposed to describe traffic behavior [1–4]. Using simulations based on special conditioners gives a better understanding. A probabilistic model based on one-dimensional Cellular Automata (CA) was proposed by Nagel and Schreckenberg (NaSch) [5], modeling traffic flow for two and three lanes through cellular automata [6]. This model has been used in several types of research to study vehicular traffic flow [7–9]. In addition, several models describe traffic flow in more than one dimension to model and optimize the flow in cities. The author in [10–14] has studied the traffic on the city network with multiple sources. The first simple two-dimensional cellular automaton model was proposed by Biham, Middleton, and Levine (BML) [15]. This model has been used in several studies [16–18].

Recently, this model has been used to study the behavior of human disease [19] and identify its drug-repurposing [20]. This model has been used in another area to collect high-dimensional data [21].

However, real traffic in a network city is much more complicated. The road network in a big city has a complex topology; it contains several roads connected at intersections that can represent a simple crossroad that combines two, three, or four roads or a big roundabout that can associate several roads. This network can be represented by a graph formed by links and nodes with different degrees. In the last decade, the topological analysis of urban network cities has been adopted in the literature. Based on real data of many network cities, it is found that all topologies demonstrate a small world structure and scale-free property. For any network city, 80% of nodes have a degree less than the average degree of its network, and 20% have a degree more significant than the average. The scale-free property has been studied and analyzed by several researchers [22–24] who have studied some real network cities and demonstrated power law distributions with an exponent around 2. This property has been examined by Jiang [25], and a power law distribution has been confirmed. It is shown in [26] that the evolution of city network topology follows modular growth and preferential attachment mechanisms. The impact of preferential attachment on the development of a network has been investigated. The different properties of the topology of roadway network cities are analyzed, and it is demonstrated that these properties (scale-free property, small world property, and structure–property) remain unchanged. Still, it is noticed that the strength of these properties can be modified by varying the values of specific parameters such as average path length and clustering coefficient.

In order to be closer to the real traffic network, especially during the studies of the big cities, it is interesting to study the traffic according to the BA model [27]. The vehicular network city can be represented following the BA model, where links represent roads and nodes represent intersections which mean crossroads or roundabouts. To understand traffic management in a network city, it is necessary to know local aspects such as priority regulation and how the vehicles are distributed at crossroads. The aim of this paper is to model and study the dynamic behavior of traffic on BA networks using cellular automata, which is based on the NaSch model. While studying the behavior of the fundamental diagram (throughput versus flow density), the different transition phases will be identified. By analyzing those

phases, the conditions and parameters that influence the fluidity of the network will be determined. To study the influence of priority at intersections on traffic flow, the focus is given according to probability P . Subsequently, the effect of various parameters, such as speed and network topology, will be treated.

2 Model and Method

2.1 Barabasi-Albert Model

The majority of networks in real life have a similar characteristic which is their complex topology, World Wide Web, Genetic networks, social sciences, vehicular networks, and so on, are examples of that kinds of networks. These complex networks have a common property which is the behavior of degrees distribution. BA proposed a new model which allows obtaining this propriety. The BA model is based on two concepts observed in real networks (i) the network grows continuously with time. Indeed, real networks are formed by adding new links connected to the already existing links in the system; (ii) preferential attachment. The behavior of degree distribution has been studied in two interdependent BA sub-networks [28]. It is shown how the two subnets must be connected to have the final BA network.

In order to build the BA network, the next step must be followed:

Initially, the network contains m_0 nodes; at each time “ t ,” a new node is added to the network with “ k ” degree, which has a new node as the source node, and their destination is randomly chosen.

The probability “ P_i ” with which a new link is associated to networks is proportional to the degree of each node “ i ” existing in the network. In other words, there is a higher probability that new links will be connected to nodes having a higher degree of connectivity. This probability is formally given as:

$$P_i = \frac{k_i}{\sum_j k_j} \quad (1)$$

where “ k_i ” is the degree of node “ i ” and the denominator is the number of links in all pre-existing nodes “ j .”

The BA networks verify three properties:

- The average degree is $2k$, where k is the number of links associated with each node during the construction of the network.
- The degree distribution follows a power law:

$$P(k) = a \cdot x^{-\alpha} \text{ Where } -3 < \alpha < -2.$$

- The third propriety is to identify a specific node of the network, and at each time step follow up the number of links associated with this node:

$$k_i(t) = m * \sqrt{\frac{t}{t_i}}$$

In this paper, we associate each node $k = 2$, which gives the average degree equal to 4.

2.2 Vehicular Traffic Model

The vehicular networks in the big cities can be considered BA networks. That vehicular network contains several inputs (road entrance), outputs (exit road), and intersections (crossroads). In that case, in the built BA network, roads and crossroads are, respectively, links and nodes of that network. The nodes sending without receiving links are the entrance of networks, and the nodes receiving without sending links are the networks' output. In BA networks, one node ($i = 0$) doesn't send links, so other creations are chosen randomly. When the vehicle arrives at one of the outputs, it can exit the network with the probability $Pe = 0.5$.

Each node "i" has $R_j(i)$ input links $j:1,2,\dots$, $link_{in}(i)$, and $R_k(i)$ output links $k:1,2,\dots$, $link_{out}(i)$.

$link_{in}(i)$: Number of input links of node "i".

$link_{out}(i)$: Number of output links of node "i."

P_m : Probability of movement for the vehicle at each intersection.

P : Probability of priority.

Pe : Probability to exit the network at the output.

The vehicles arriving at the network's output will randomly be injected again into the system through one of the inputs, Fig. 1.

We consider all roads having the same length, "L." Each site can be empty or occupied by one vehicle. At each time, each vehicle can move to another place with velocities 0, 1,..., and V_{max} following the Nash model. Those rules are mentioned below:

R1 acceleration:

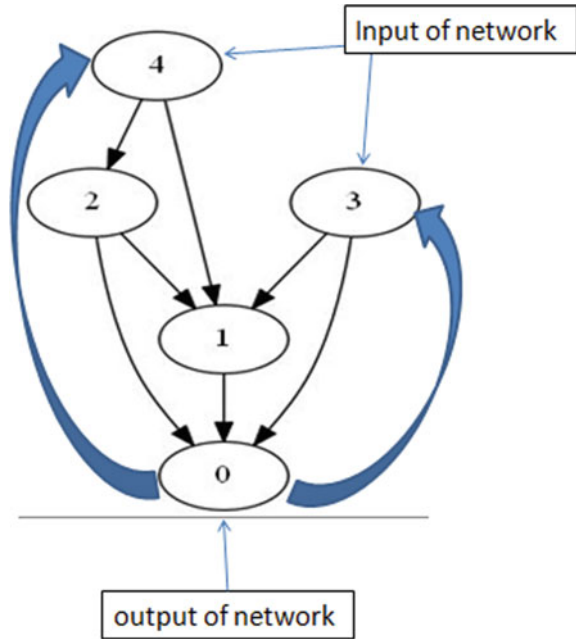
$$V'_i = \min(V_j(t) + 1, V_{max})$$

In other words:

If $V_j(t) < V_{max}$ so the speed of the vehicle "j" grows by one:

$$V_j(t + 1) - > V_j(t) + 1$$

Fig. 1 BA network constituted of $N = 5$ nodes, with periodic boundary conditions



If $V_j(t) = V_{max}$ so the speed remains unchanged.

R2 deceleration:

$V_j(t) > d_j$ the speed decrease to d_j so $V_j'' = \min(d_j, V_j')$.

R3 randomization:

Each vehicle that has $V_j' > 0$ can stop with probability P_b , so:

$$V_j(t + 1) = V_j'' - 1 \text{ with probability } P_b$$

$$V_j(t + 1) = V_j'' \text{ with probability } 1 - P_b$$

R4 movement:

At each time, each vehicle can move to another site with velocities $0, 1, \dots, V_{max}$

$$x_j(t + 1) = x_j(t) + \min(V_{max}, d_j).$$

$x_j(t)$: the position of the vehicle "j" at the time "t"

V_{max} : the maximum vehicle velocity.

d_j : headway of the jth vehicle at time t, $d_j = (x_{j+1}(t) - x_j(t)) - 1$.

For every node N_i , the distance of the much closer vehicle to that node for every road coming to N_i is calculated. In the case of having a similar distance on several roads, the priority is given according to the probability “P” that takes different values to study its influence on the traffic.

The probability “P_m” is the probability of movement in its formula. The mentioned probability “P” is used. The formula below is applied to the vehicle at the same distance:

$$\text{If } R_j(i) < \text{link}_{in} \text{ Then } P_m = \frac{P}{\text{link}_{in}(i) - 1}$$

$$\text{If } R_j(i) = \text{link}_{in}(i) \text{ Then } P_m = 1 - P$$

$R_j(i)$: the road coming from node “j” to node “i.”

$\text{link}_{in}(i)$: Number of input links of node “i”.

At each intersection, the vehicle’s priority destination is chosen randomly.

In the following sections, the behavior of the vehicular traffic in the BA network will be analyzed; the following terms will be used:

Route: the line between one of the inputs of the network to the output, it can contain several nodes; Road: line which connects two nodes; Short route: it is the route that includes the minimum number of nodes; Long route: the route that contains the maximum number of nodes; Fundamental diagram: throughput versus flow density.

3 Results and Discussion

The fundamental diagram of the proposed model has been analyzed. Networks formed by three, four, and N nodes have been studied. Each result is averaged 20 times.

3.1 Network of Three Nodes

This network contains two asymmetric routes Fig. 2a. The behavior of the fundamental diagram of this network is represented in Fig. 3. It is observed that the system exhibits three phases: the free-flow phase (1), the plateau phase (2), and the congested phase (3).

In phase (1), the system is in low-density. All the vehicles can move with maximal velocity for all P values, and the free-flow regime is obtained. In phase (2), for $P \neq 0.5$ in the intermediate density, it can be seen that two platoons are constituted; the first one corresponds to a blockage of some vehicles in the route, which doesn’t have priority. Increasing density, the queue length increases until this route’s congestion

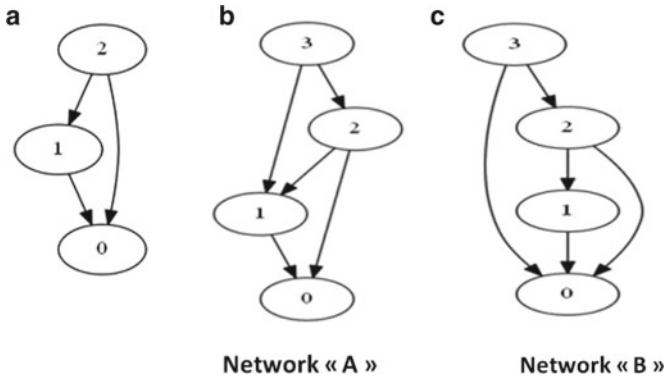


Fig. 2 **a** Representation of network constituted of $N = 3$ nodes; **b** representation of network constituted of $N = 4$ nodes topology “A” **c** representation of network constituted of $N = 4$ nodes topology “B”

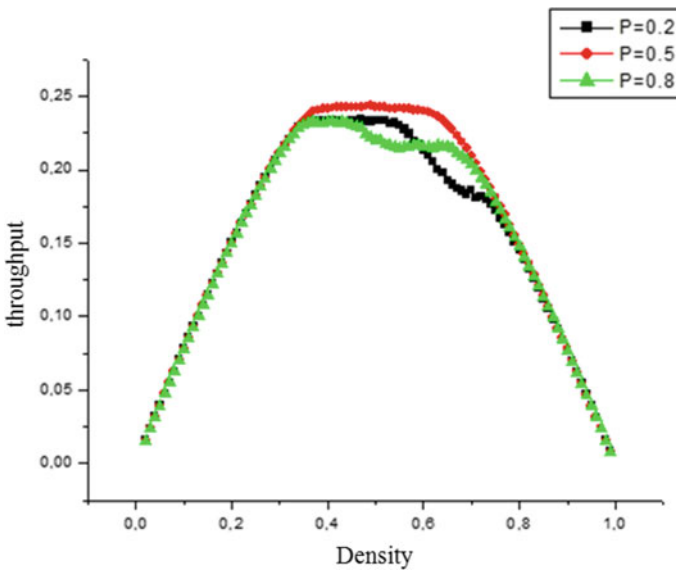


Fig. 3 Fundamental diagram of network constituted of $N = 3$, for different P values, $P_b = 0.2$, $V_{max} = 1$

is reached. Then, a queue in a second route is made up, which gives us a second plateau. At $P = 0.5$ (symmetric case), vehicles are equally distributed on the two routes. Therefore, a queue is created in the two routes simultaneously, which explains that only one platoon is obtained. In phase (3), density is high. Most vehicles do not change their position, so the transition to the congested phase occurs. With changing the maximal velocity, the behavior of the throughput is observed. When $V_{max} =$

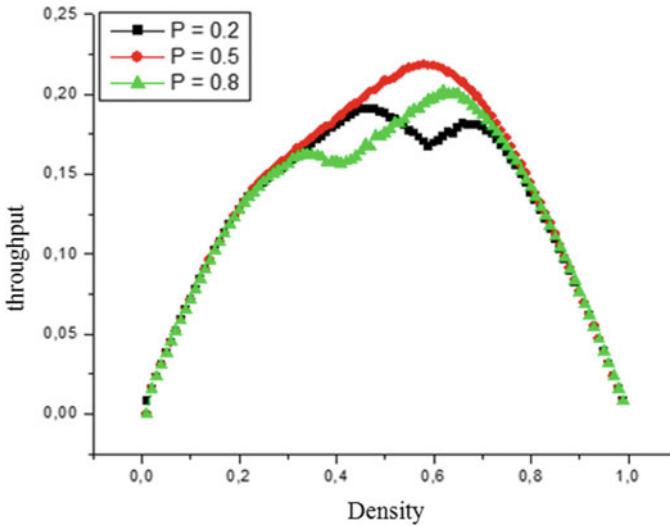


Fig. 4 Fundamental diagram of network constituted of $N = 3$, for different P values, $P_b = 0.2$, $V_{max} = 5$

1, it is observed, for all P values, that in phase (2), the plateaus are created, while when the maximal velocity increases ($V_{max} = 5$), each plateau is replaced by a peak Fig. 4.

A change in the behavior of average velocity characterizes the passage from one phase to another. It is observed in Fig. 5 that in phase 1 (free-flow phase), low-density vehicles can move freely, and the average speed is almost equal to the given maximum speed. The difference observed between the average rate, and the presence of two vehicles causes V_{max} at the same time at intersections. for intermediate density (phase 2), when maximum velocity is high, vehicles in queue move faster, so instead of having a blocking, only a deceleration is observed, which corresponds to decreases in average rate. That causes a change in the slope of the throughput. Finally, for high density (phase 3), the average speed is almost zero, which explains the congested phase.

The effect of braking probability P_b on traffic flow is studied. For all values of P_b , the behavior of throughput is always the same, but when P_b increases, the value of throughput in the intermediate phase decreases.

3.2 Network of Four Nodes

For all values of P , as long as the size of the network increases, the number of plateaus also increases. The simulation shows that even if two networks have the

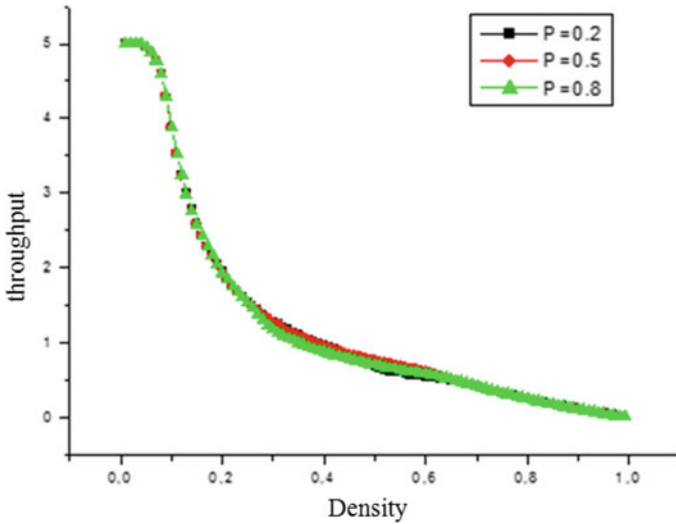


Fig. 5 Representation of the average velocity versus density, $N = 3$, for different P values, $P_b = 0.2$, $V_{max} = 5$

same dimension, the throughput changes behavior; it depends on the structure of the network Fig. 6.

For $V_{max} = 1$, several platoons are obtained in phase (2) (intermediate density). It is observed that the number of the plateau is equal to the number of routes. The value of throughput and the width of the plateau depend on the road length. It is remarked also that the more input links the chain's priority has, the more the value of throughput increases. A peak for a high value of V_{max} replaces each plateau. For more extensive networks, congestion is always made in the same way. The route which has yet to the priority is permanently blocked first; some parameters, such as the length of courses or the number of links associated with each node, don't influence the congestion steps. However, those parameters affect the width of the plateau and the value of throughput.

3.3 Network of n Nodes

For large networks, the throughput changes behavior. For all P values, only two phases are obtained: the free flow phase and the congested phase. The typical fundamental diagram of a large network is represented in Fig. 8a. The network's topology influences the value of throughput, but it is observed that the maximum value of throughput is always obtained at a density of 0.5. The more the number of nodes increases, the more value of throughput decreases; this is caused by blocking at the roundabout. Figure 7 represents the throughput at crossroads and roundabouts; it is

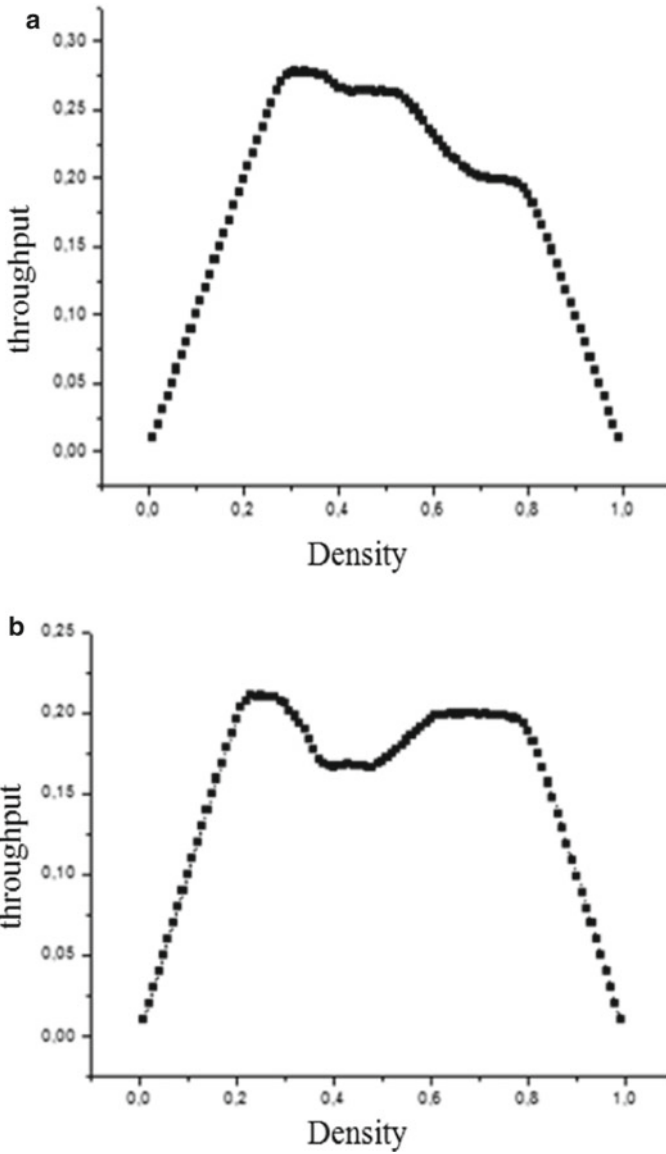


Fig. 6 **a** Fundamental diagram of network constituted of $N = 4$, $V_{max} = 1$, $P = 0.8$, $P_b = 0.2$; **a** Fundamental diagram of network “A”; **b** fundamental diagram of network constituted of $N = 4$, $V_{max} = 1$, $P = 0.8$, $P_b = 0.2$; **b** fundamental diagram of network “B”

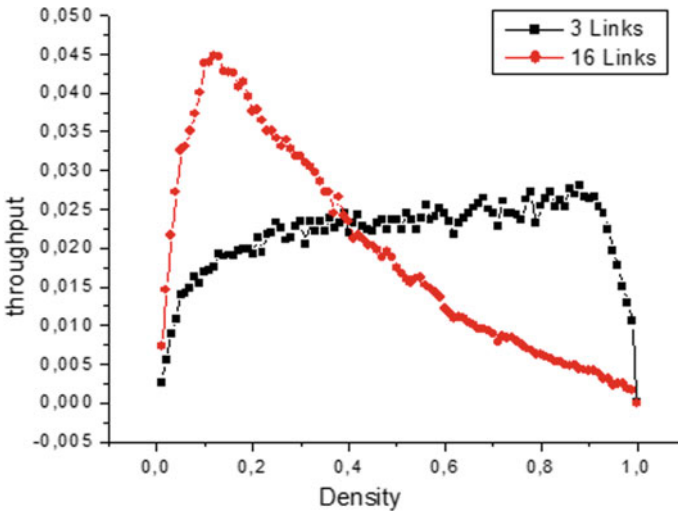


Fig. 7 Throughput at intersections, $N = 50$, $V_{max} = 1$, $P = 0.8$, $P_b = 0.2$, three output

observed that, for high density, the more the number of links associated at each node increases, the more the throughput value decreases. In fact, at intersections, with every time step, one vehicle can move to another road which causes a creation of a queue on other roads. Even by increasing the speed, the exact behavior of throughput is obtained. The average speed value remains very small on networks of big sizes, even for low-density Fig. 8b.

4 Conclusion and Perspectives

In this paper, the behavior of the fundamental diagram of BA networks with periodic boundary conditions has been studied. At each intersection, the vehicles move to another road with probability P , which takes different values. Different transition phases have been determined. For networks of small size, it is shown that the fundamental diagram exhibits three phases. In low-density, the free-flow phase is obtained. In the intermediate density, the platoon phase is observed; different platoons are constituted according to a number of routes from the inputs to the outputs of the network. In high density, a congested phase is reached. It is found that traffic flow depends strongly on the network's topology. For large networks, the same behavior is obtained for every size of network; platoons don't appear anymore, and the maximal throughput is obtained at density 0.5. These results will help understand some physical phenomena happening in many complex systems; for instance, the behavior of traffic flows in cities. It will help in choosing the best way to build the network in order to obtain the maximal throughput.

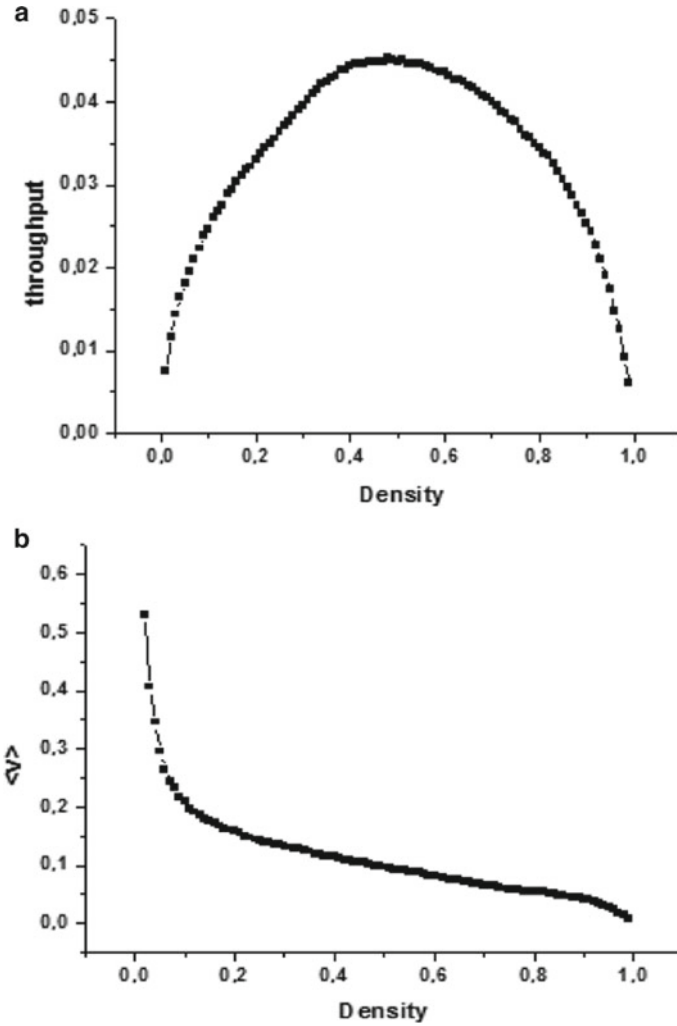


Fig. 8 **a** Fundamental diagram of network constituted of 50 nodes, $P_b = 0.2$, $V_{max} = 1$, $P = 0.5$, three output; **b** average velocity versus density, $N = 50$, $P_b = 0.2$, $V_{max} = 1$, $P = 0.5$, three output

That work has raised several questions, such as:

- How many links should be associated with the roundabout to avoid congestion?
- How can the priority at intersections be managed?
- What is the optimal velocity to get the maximum throughput?

This study proposes a topology for traffic in the city network in order to apply the different algorithms of the Intelligent Transportation System (ITS). These systems allow communication between vehicles and infrastructure to improve road safety and traffic efficiency.

We aim to optimize these models by artificial intelligence techniques such as deep learning while using recommendation for user based on recommendation systems algorithms found in our previous works in [29–34].

References

1. E. Ben-Naim, E.L. Krapivsky, S. Redner, *Phys. Rev. E* **50**, 822 (1994)
2. D. Helbing, *Complex Syst.* **6**, 391 (1992)
3. D. Helbing, *Phys. Rev. E* **51**, 3164 (1995)
4. M. Sasaki, T. Nagatani, Transition and saturation of traffic flow controlled by traffic lights. *Physica A* **325**(3–4), 531–546 (2003)
5. K. Nagel, M. Schreckenberg, A cellular automaton model for freeway traffic. *J. Phys. I* **2**, 2221 (1992)
6. B. Luna-Benoso, J. C. Martinez-Perales, R., Flores-Carapia, *Int. Math. Forum* **8**(22), 1091–1101
7. B. Eisenblatter, L. Santen, A. Schadschneider, M. Schreckenberg, *Phys. Rev. E* **57**, 1309 (1998)
8. A. Schadschneider, *Physica A: Statistical Mechanics and its Applications*, Volume 313, Issue 1, pp. 153–187
9. A. Schadschneider, *Physica A: Statistical Mechanics and its Applications*, Volume 285, Issue 1, pp. 101–120
10. Y. Sheffi, *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods* (Prentice-Hall, New Jersey, 1985)
11. H.S. Mahmassani, R. Jayakrishnan, R. Herman, *Transpn. Res. A* **24**, 149 (1990)
12. M. Hilliges, R. Reiner, W. Weidlich, Modelling and Simulation ESM 93, European Simulation Multiconference, Lyon, A. Pave, ed. (1993) p. 505
13. R. Haberman, *Mathematical Models: Mechanical Vibrations, Population Dynamics and Traffic Flow* (Prentice-Hall, New Jersey, 1977)
14. A.K. Gupta, P. Redhu, The jamming transition of two-dimensional traffic dynamics with consideration of optimal current difference. *Phys. Lett. A*, **377**(34–36), 2027–2033
15. E.M. Shahverdiev, S. Tadaki, Instability control in two-dimensional traffic flow model. *Phys. Lett. A* **256**(1), 55–58
16. O. Biham, A.A. Middleton, D.A. Levine, *Phys. Rev. A* **46**, R6124 (1992)
17. H. Kuang, G.-X. Zhang, X.-L. Li, S.-M. Lo, Effect of slow to start in the extended BML model with four-directional traffic. *Phys. Lett. A* **378**(21), 1455–1460 (2014)
18. X.-M. Zhao, D.-F. Xie, B. Jia, R. Jiang, Z.-Y. Gao, Disorder structure of free-flow and global jams in the extended BML model. *Phys. Lett. A* **375**(7), 1142–1147 (2011)
19. M. Rafiq, A.R. Nizami, N. Ahmad, N. Ahmad, *Alexandria Eng. J.* **62**, 75–83
20. D.M. Gysia, Í. do Vallea, M. Zitnikd, A. Amelib, X. Gana, O. Varola, S.D. Ghiassianf, J.J. Pattenh, R.A. Daveyh, J. Loscalzoi, A.-L. Barabásia, *Biological Sci.*
21. A.D. Monaca, M. Cafaro, M. Pulimeno, I. Epicoco, International Symposium on Distributed Computing and Artificial Intelligence, december 2022, LNNS, volume 583
22. S. Porta, P. Crucitti, V. Latora, The network analysis of urban streets: a dual approach. *Physica A* **369**, 853–866 (2006)
23. J. Buhl, J. Gautrais, N. Reeves, R.V. Sole, S. Valverde, P. Kuntz, G. Theraulaz, Topological patterns in street networks of self-organized urban settlements. *Europ. Phys. J. B* **49**, 513–522 (2006)
24. B. Jiang, A topological pattern of urban street networks: universality and peculiarity. *Physica A* **384**(2), 647–655 (2007)
25. B. Jiang, Small world modelling for complex geographic environments, in *Complex Artificial Environments*. ed. by J. Portugali (Springer, Heidelberg, 2005), pp.259–271

26. Z. Zou, P. Liu, S. Zhou, Y. Xiao, Xu., Xuecai, J. Gao, Analysis on evolving model with modular growth of urban roadway network topology structure. *Kybernetes* **44**(4), 505–517 (2015)
27. A.L. Barabasi, R. Albert, *Science* **286**, 509 (1999)
28. M. Benyoussef, H. Ez-Zahraouy, A. Benyoussef, *Int. J. Modern Phys. C* **25**(9), 1450040
29. K. EL Handri, A. Idrissi, Efficient Top-kws algorithm on synthetics and real datasets. in *International journal of Artificial Intelligent (IJAI)* (2020)
30. K. El Handri, A. Idrissi, Etude comparative de top-k basée sur l’algorithme de fagin en utilisant des métriques de corrélation dans la qualité de service de cloud computing. in *EGC*, pp. 359–360 (2019)
31. K. El Handri, A. Idrissi, système collaboratif d’aide à la décision à base des recommandations multi critères (Sep 03 2020 MA Patent 50776)
32. K. Elhandri, A. Idrissi, Parallelization of Top-k algorithm through a new hybrid recommendation system for big data in spark cloud computing framework. *IEEE Syst. J.* **15**(4), 4876–4886 (2021). <https://doi.org/10.1109/JSYST.2020.3019368>
33. K.E. Handri, A. Idrissi, Comparative study of Top-k based on fagin’s algorithm using correlation metrics in cloud computing QoS. *Int. J. Internet Technol. Secured Trans.* **10**, 143–170 (2020)
34. A. Idrissi, K. Elhandri, H. Rehioui, M. Abourezq, Top-k and skyline for cloud services research and selection system. In: *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, p. 40. ACM (2016)

Customer Journey Map Discovery Approach



Imane El Alama  and Hanae Sbai 

Abstract The customer journey is a marketing concept that describes the path that a customer may take until they purchase a product or service. Therefore, it is pertinent data for organizations, as it allows them to have a better knowledge of customer behavior, this journey can be represented as a map that allows mapping the path of the customer passing through all touchpoints, to present this map, there are several methods, manual by a professional or automatically using algorithms, Another technique for automatically discovering the customer journey map is process mining. This study introduces a new framework based on configurable process mining to find the customer journey map.

Keywords Customer journey map · Configurable process mining · Process model

1 Introduction

In today's digital age, most individuals make purchases through the internet, and e-commerce has become a mainstream sales channel. Companies strive to enhance the customer browsing experience on their website by personalizing their journey. To achieve this, the company must gather information about the customer's preferences to provide the best options.

In marketing, many companies use the customer journey to analyze customer behavior. This journey is a set of interactions between the customer and the company that is mapped on a Customer Journey Map [1]. This mapping aims to improve the customer experience and personalize it. Process Mining is extensively used in this domain, where its techniques are employed to discover the map by transforming web data into an event log using either manual transformation or automatic clustering to apply process mining methods [1–3]. Various studies have utilized process mining

I. El Alama (✉) · H. Sbai
Mathematics, Computer Science, and Application (LMCSA), University Hassan II of Casablanca, Mohammedia, Morocco
e-mail: elalama.imane@gmail.com

discovery algorithms such as Fuzzy Miner, Heuristic Miner, and Alpha algorithm ... to discover the process that groups these interactions. The result of this process is called the Customer Journey Map (CJM) [1–3].

The Customer Journey Map (CJM) contains representative journeys that are common among many customers. For example, a dataset may have 1000 journeys, but the CJM may only contain 10 variant journeys. Therefore, these 10 journeys represent the 1000 journeys. Most studies in this domain use classic process mining algorithms. However, the problem lies in interpreting the output of the process result and identifying the representative journeys from this result.

This paper presents our approach to discovering the customer journey map and identifying the representative customer journeys using configurable process mining discovery.

In the remainder of this paper, we begin by presenting a conceptual overview of the basic concepts of process mining and the customer journey. We then review related work and propose a new framework for analyzing customer journeys using configurable process mining techniques. Finally, we conclude with a summary of our findings and provide suggestions for future research.

2 Basic Concept

2.1 *Customer Journey and Customer Journey Map*

The customer journey is a collection of touchpoints, a touchpoint is an interaction between the consumer and the service provider using a specific channel [4], for example, when a person buys a book from Amazon, there are several steps involved: they search for the book, add it to their cart, make a payment, and then receive confirmation of their order. These steps represent points of contact between the customer and the company, Amazon, through the website channel. The experience that the customer has during this journey is essential for the company to understand their behavior and preferences, as well as how they navigate the website. With the help of information systems and technology, this information can be extracted and analyzed to improve the customer's experience.

The customer journey map is a tool used to visualize and map the customer journey. It helps professionals understand how customers experience a product or service to improve and personalize it. According to [4] the main components of the CJM are as follows:

Customer: a consumer of a service (student, patient, software user...).

Journey: “A CJM contains at least one journey, which is a typical sequence of touchpoints followed by a customer”, there are two kinds of the journey, expected journey and the actual journey, the expected one is designed by professionals of the company to have an ideal journey that corresponds to the company’s expectations, the second contains the experience of the customer that describes his path with the identified problems and needs.

Goal: A customer journey should be mapped with a goal in mind.

Touchpoint: A touchpoint is an interaction between the customer and the service provider, the customer can pass by a Touchpoint more than one time, miss one or several Touchpoints, or abound his journey.

Channel: “The channel is the method chosen by the customer to interact with the touchpoint” such as a “reference desk” or “social media”.

Figure 1 provides an example of a customer journey map. The various touchpoints are represented by nodes labeled T0, T1, and so on. This map represents an expected CJM that can be modified with measurements identified in the actual journey in real-time. In this example, two unknown touchpoints are observed in the actual journey: one inhibits deviation when the customer has doubts about the terms and conditions, and the other activates deviation when the customer interacts with a targeted promotion. After this deviation, the customer follows the rest of the expected journey with the underlying attitude being latent.

Our approach uses the Configurable process mining discovery to obtain the CJM.

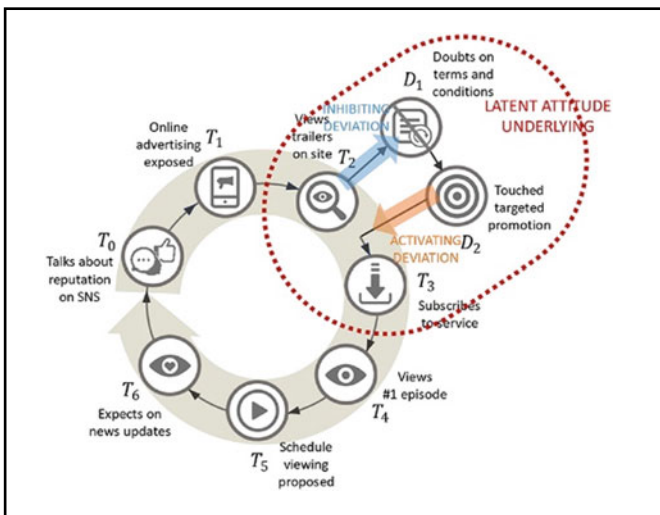


Fig. 1 Example of customer journey map [5]

2.2 *Configurable Process Mining Discovery*

Process mining is a field that combines data mining with the analysis of business processes. The aim of process mining is to discover and potentially redesign a company's business processes through the use of process mining discovery algorithms. This includes controlling the process and comparing it to an existing process. The starting point of process mining is an event log that is stored in the company's databases. This log contains information on the activities, resources, and various attributes such as timestamps, transactional information, resource usage etc.) [6].

In process mining 3 axes are defined [6]:

Process discovery: Automatically discover the process model from the event log.

Conformance checking: Compare the process discovered to its event log to check its conformity.

Process enhancement: improve an existing process model by adding a new perspective from the event log.

A configurable process model is a process that combines many variants of a process into a single model. The model contains non-configurable elements (the common parts between process variants), configurable elements (the parts that allow for several possible choices), and configuration constraints (the rules for deriving configurable process variants). Using configurable process mining discovery, it is possible to discover the process that supports the variability. The configurable process discovery allows for the identification of commonalities and variabilities among different process variants [7].

3 **Related Work**

The goal of utilizing process mining techniques is to identify and map the customer journey from the event log and leverage the information to enhance the expected CJM. Many research studies have employed process mining discovery algorithms such as Fuzzy Miner, Heuristic Miner, and Alpha algorithms using Disco or ProM tool to achieve this. The first step in this process is converting the web dataset into an event log. Most of the studies employ web log data as the input, which contains various data such as IP address, date and time of access, etc. After transforming this file into an event log by using techniques such as transforming URLs into activities using regular expression functions or automatic clustering [1–3].

Reference [8] Presents the CJM-ex using a hierarchical clustering approach, its objective is to allow users to upload and explore their dataset using a customer journey map layout to limit the number of journeys displayed on the same CJM and allow their intuitive exploration.

Reference [9] Uses a genetic approach inspired by Process Mining to discover the CJM, which contains representative journeys.

4 Proposed Discovery Approach Using Configurable Process Mining

Our proposal discovery framework is presented as follows:

The first step of the framework is to automatically transform a set of web logs into a set of event logs. This step involves cleaning the weblogs and then mapping them using associated mapping and transformation rules. These rules include mapping URLs to activity using clustering (whether the clustering method is supervised or unsupervised is yet to be determined); mapping timestamps to timestamps. Each weblog is transformed into an event log. In addition, the framework can use the Customer Journey Map (CJM) model proposed by the author in [4]. In their study, they proposed an XML format for storing the CJM using the components defined in Sect. 2. This format is inspired by XES, which is the IEEE event log standard (Fig. 2) [4].

In the second step, a configurable process mining discovery approach is used to discover a process model from each event log. For example, if there are three event logs, then in this step, three process models will be obtained. The results are merged to obtain a process that supports variability using configurations. This process will contain variant activities, and the path that starts from the first activity passing through a variant one V_x to the end represents the representative customer journey. This approach allows for the identification of the representative journey using configurable process discovery.

After obtaining the output process, the correspondences added by [4] are used to map this process to the CJM or an additional step is added to the discovery step to incorporate correspondences and more information, such as customer feedback data in the last step.

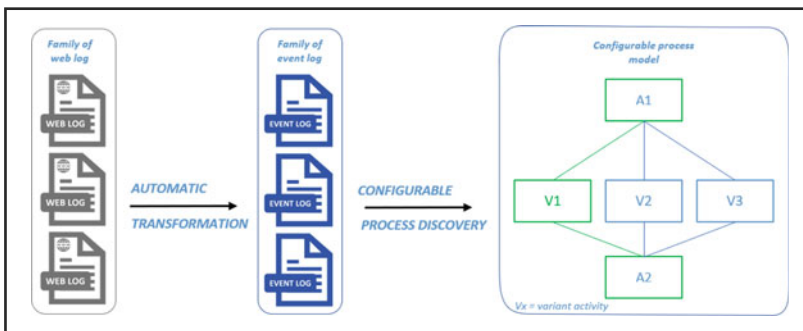


Fig. 2 CJM discovery proposed approach

5 Conclusion

In this paper, we have presented an approach for using configurable process mining techniques to discover the customer journey map. Analyzing the customer journey is a crucial task for companies, particularly for marketing decision-makers. Thus, simplifying this process for them is of utmost importance, as it enables them to save time and analyze the actual customer journey to gain a better understanding of customer behavior and preferences, thereby improving the company's existing customer journey.

The framework presented in this paper provides an opportunity to utilize configurable business process models, which support variability, for identifying representative customer journeys and leveraging all available business process technologies to improve marketing strategies. In the next phase, we will apply our approach to a use case to analyze the data and test its effectiveness.

References

1. J. Goossens, T. Demewez, M. Hassani, Effective steering of customer journey via order-aware recommendation, in 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 828–837 (2018)
2. A. Terragni, M. Hassani, Analyzing customer journey with process mining: From discovery to recommendations, in 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 224–229 (2018)
3. A. Terragni, M. Hassani, Optimizing customer journey using process mining and sequence-aware recommendation, in Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 57–65 (2019)
4. G. Bernard, P. Andritsos, A Process Mining Based Model for Customer Journey Mapping (2017)
5. K. Okazaki, K. Inoue, Explainable model fusion for customer journey mapping. *Front. Artif. Intell.* **5** (2022), <https://doi.org/10.3389/frai.2022.824197>. Accessed 8 Jan, 2023
6. W. Van Der Aalst, Process mining: overview and opportunities. *ACM Trans. Manag. Inf. Syst. (TMIS)* **3**(2), 1–17 (2012)
7. R. Sikal, H. Sbai, L. Kjiri, Configurable process mining: variability discovery approach. in 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), pp. 137–142 (2018). <https://doi.org/10.1109/CIST.2018.8596526>
8. G. Bernard, P. Andritsos, CJM-ex: Goal-oriented Exploration of Customer Journey Maps using Event Logs and Data Analytics. in *Bpm (demos)* (2017)
9. G. Bernard, P. Andritsos, Discovering customer journeys from evidence: a genetic approach inspired by process mining, in International Conference on Advanced Information Systems Engineering, pp. 36–47 (2019)

Unmanned Aerial Vehicle-Assisted Clustered Wireless Sensor Network Data Collection Efficiency Improvement



Mohamed Abid, Said El Kafhali , Abdellah Amzil, and Mohamed Hanini

Abstract Unmanned aerial vehicles, sometimes known as UAVs, have substantial potential applications in a variety of fields, including disaster and emergency scenario management, coverage expansion, and more. UAVs have recently been concentrating their efforts largely on applications in which the presence of humans would be either impossible or harmful. A group of smaller unmanned aerial vehicles, or UAVs, may work together to accomplish missions more quickly and cost-effectively than a single big UAV. However, a large number of problems need to be overcome before it would be possible to build multi-UAV networks that are stable and trustworthy. Then both the number and the trajectories of the UAVs are optimized to reduce data-collecting flight time. We distinguish four trajectory approaches. This includes data collection design recommendations. Then, the K-means clustering of sensor nodes is addressed. Next, we show how well metaheuristic solutions work to plan trajectories.

Keywords Unmanned aerial vehicle · Resource-allocation · Internet of things · Wireless networks · Metaheuristics · Traveling salesman problem

M. Abid · S. El Kafhali (✉) · A. Amzil · M. Hanini
Faculty of Sciences and Techniques, Computer Networks, Mobility and Modeling Laboratory:
IR2M, Hassan First University of Settat, Settat, Morocco
e-mail: said.elkafhali@uhp.ac.ma

M. Abid
e-mail: mo.abid@uhp.ac.ma

A. Amzil
e-mail: a.amzil@uhp.ac.ma

M. Hanini
e-mail: mohamed.hanini@uhp.ac.ma

1 Introduction

In wireless sensor networks (WSNs), several sensor nodes (SNs) are often placed to monitor physical phenomena like pressure, temperature, etc. WSNs are pervasive now because they are an integral aspect of the Internet of Things (IoT) [1, 2], paving the way for innovative uses in sectors as diverse as smart energy networks or smart grids [3], water distribution systems, and intelligent transportation systems (ITS). SNs in these setups are usually low-powered devices with limited ranges.

Traditional terrestrial networks face difficulties in various terrains and the event of equipment failure. UAVs can increase the flexibility and resilience of mobile edge computing network deployment, while also lowering the complexity and expenses of resource management. But when the number of UAVs in use increases, network resource management problems arises, such as how to limit the power and share the spectrum, deal with interference, and divide up tasks [4].

The UAV has evolved in recent years due to the development of sophisticated technology, which has achieved low weight, high adaptability, and extended battery life [5]. UAVs are now useful in a wide range of applications, including the military. Much attention has also been dedicated to the use of UAVs in 5G networks and beyond to improve capacity, as they are simple to install, do not cost much to maintain, and can move about a lot [6].

Most research has focused on a single UAV development, not multi-UAV scenarios. Furthermore, it is impossible to communicate with low-power SNs; thus, cluster heads (CHs) with greater power capabilities are advised for deployment alongside UAVs. Most studies also assumed ideal line of sight (LoS) channels. In cities, however, where non-line-of-sight (NLoS) connections exist, this is not the case. For these reasons, we discuss how to build WSNs with more than one drone to make data collection easier and reduce flight time.

We provide a WSN architecture in which CHs and their positions, as well as UAVs with their paths, gather WSN data as soon as possible. At first, the number and location of CHs are optimized using the clustering of K so that every SN can interact dependably with its associated CHs. Next, we construct UAV trajectories that provide both accurate data capture and the lowest possible overall flight times, using optimal and metaheuristic solutions to the mTSP (Multiple Traveling Salesman Problem) and the TSPN (Traveling Salesman Problem with Neighbors).

The findings provide recommendations regarding the structure of the WSN data collection: Clustering through K -means yields a small number of CHs. As metaheuristic [7, 8], Ant Colony Optimization (ACO), Simulated Annealing (SA), Particle Swarm Optimization (PSO), and genetic algorithm (GA). The Ant Colony Optimization (ACO) was near-optimal for evaluating a UAV trajectory. A UAV may extend its range and duration of flight by hovering close to each CH. The time spent gathering information is affected by factors such as the terrain and the UAV's height. In conclusion, the efficiency of data gathering is enhanced in large UAV WSNs when trajectory fairness is included.

The summary of the contributions compared with the existing work is as follows.

- We proposed a four-phase path planning algorithm, ACO, SA, PSO, GA. However, existing studies only considered the GA algorithm.
- We proposed another strategy to choose the intersection point with the CH perimeter, which is better than the existing work.
- The experimental results demonstrate that ACO improves the capabilities of the system compared to other algorithms.

The rest of this article is structured as follows. We first provide some related work in Sect. 2. Section 3 introduces the system model. In Sect. 4, the formulation of the problem is discussed. In Sect. 5, we describe the proposed approach in detail. In Sect. 6, simulation results are presented and discussed. The paper is concluded in Sect. 7.

2 Related Work

UAV systems have made significant advances in recent decades. Originally developed for use in the military, UAVs are now a common tool in geometrics for data acquisition across a wide range of research and operational fields, including regional security, structure, and infrastructure monitoring, archeological site monitoring, environmental monitoring, agricultural applications, and more [9].

The real-world applications of UAVs in precision agriculture and remote sensing are discussed in [10]. Case studies show that UAVs are more efficient and cost-effective than conventional aircraft or satellite-based platforms. Based on this, several possible local applications of UAVs are indicated, which can help accelerate their usage locally and attract widespread research, investment, and improved policy regulation for their implementation.

Backscatter communication is gaining attention as a potential answer to the IoT battery life issue. With backscatter communication technology, for example, a wireless sensor network can keep an eye on the environment in far-flung places without the need for regular battery maintenance or replacement. Backscatter communication has a restricted range of transmission, which is a major drawback. As a solution to this problem, the authors in [11] offered a multi-UAV-aided data-collection scenario in which the UAV flies near the backscatter sensor node (BSN) to activate it and subsequently collect data. Furthermore, after the collecting mission is complete, the authors reduced the overall flight time of the rechargeable UAVs. Also, some academics have zeroed in on trajectory planning in UAV-assisted MEC systems as a means to fully use a UAV's high mobility. For example, to reduce overall power consumption, [12] improved both the trajectory and the data shared among agents. Under the restrictions of latency and energy budget, the developers of [13] looked at how bits are distributed and where they should be headed.

3 System Model

We will imagine a WSN network in which the locations of the M SNs are completely arbitrary. SNs are capable of sensing a variety of data types, all of which must be sent to the network. As shown in Fig. 1, we will operate under the assumption that a large number of UAVs are sent to gather data from the SNs. Because SNs are often low-power sensors that are unable to directly communicate with UAVs, K CHs were typically installed to gather data and subsequently transport it to UAVs.

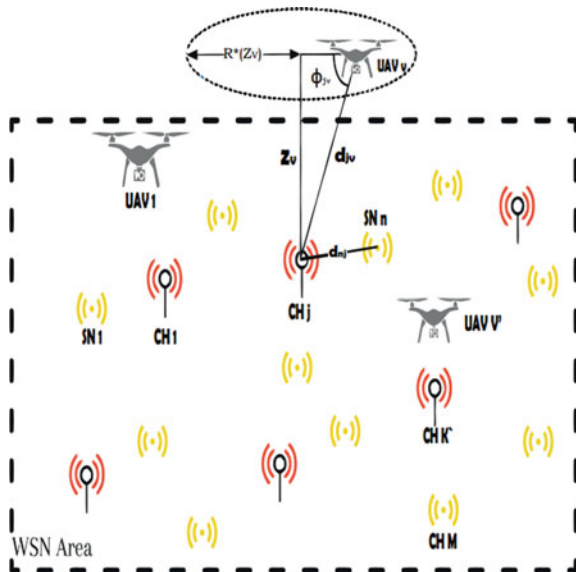
The assumption here is that an SN uses the power P_n to send data to its corresponding CH, and that a CH uses the power $P_h > P_n$ to talk to a UAV. Furthermore, it is believed that there is no interference between the various channels of communication. This is the process that takes place during data collection.

Data sent by SNs after it has been associated with CH_j . At CH_j , μ_j^{ch} , The SNR (Signal-to-Noise Ratio) associated with the received data may be stated as:

$$\mu_j^{ch} = \frac{P_n d_{nj}^{-\alpha}}{\eta_p^2}, \forall j = 1, \dots, K, \forall n = 1, \dots, M \tag{1}$$

where d_{nj} represents the distance between SN n and CH_j , α represents the path loss exponent (PLE), and η_p^2 is the noise power. If the value of $\mu_j^{ch} \geq \mu_{th}$, where μ_{th} there would be a predetermined SNR range. Then this communication connection is judged to have been successful. As a result, the maximum interaction distance for SNs can be computed as follows:

Fig. 1 System model
(Adapted from [14])



$$d_{nj} \leq d_n^{th} = \left(\frac{P_n}{\eta_p^2 \mu_{th}} \right)^{\frac{1}{\alpha}}, \forall n = 1, \dots, M \quad (2)$$

After receiving data, CH_j waits for the UAV to fly over its location before delivering M_j of D bit data packets. M_j is the number of sensors connected to CH_j . In the channel model, we consider LoS and NLoS links to have no interference.

$$\Lambda_{jn} = \mathbb{Q}_{jv}^{LoS} l_{jv}^{LoS} + \mathbb{Q}_{jv}^{NLoS} l_{jv}^{NLoS} \quad (3)$$

where Λ_{jn} represents the mean path-loss across CH_j and UAV ($j = 1, \dots, K, v = 1, \dots, V$), \mathbb{Q}_{jv}^{LoS} the LOS (Line-of-Sight) probability, $\mathbb{Q}_{jv}^{NLoS} = 1 - \mathbb{Q}_{jv}^{LoS}$ is the none-LoS probability, and l_{jv}^{LoS} and l_{jv}^{NLoS} the LoS/none-LoS path losses. It is spelled out as follows:

$$\mathbb{Q}_{jv}^{LoS} = 1 / \left(1 + ae^{-b \left(\frac{180}{\pi} \phi_{jv} - a \right)} \right) \quad (4)$$

$$l_{jv}^m = 20 \log(d_{jv}) + L_{FS}(f_c) + \gamma_m, \forall m \in \{LoS, NLoS\} \quad (5)$$

In which ϕ_{jv} and d_{jv} in Fig. 1 correspond to the angle (in *rad*) and distance (in *m*) between CH_j and UAV, respectively. v , a , and b are environmental constants. $L_{FS}(f_c) = 20 \log(4\pi f_c/c)$, where f_c is the switching frequency, c is the velocity of light and γ_m is the value of the extra path coefficient. The power level at UAV v from CH_j is represented using the Friis formula:

$$P_{jv}^{UAV} = P_h - \Lambda_{jn} \geq P_{th}, \forall j = 1, \dots, K, \forall v = 1, \dots, V \quad (6)$$

If the signal strength of CH_j is greater than the receiver sensitivity of the UAV, denoted by P_{th} , then the communication between CH_j and the UAV v has been successful. Following this, we will construct the combined issue of using some subset of CHs with UAVs to maximize data collection efficiency within strict constraints of time, money, and energy. In doing so, we will be able to maximize the effectiveness of our data collection efforts.

4 Problem Formulation

Let's start by assuming that the actual number of CHs and UAVs being deployed is $K' \leq K$ and $V' \leq V$, respectively. Then we define it by $D_j = \{S_Ns | d_{sj} \leq d_s^{th}\}$ the set of sensors associated with $CH_j, \forall j = 1, \dots, K'$, and $C_v = \{CH_j | P_{jv}^{UAV} \geq P_{th}\}$ the set of CHs associated with UAV $v, \forall v = 1, \dots, V'$.

The problem can be expressed as follows.

$$\begin{aligned}
 & \min_{\mathcal{D}_j, \mathcal{C}_v, \mathcal{W}_v, \mathcal{H}_v} \frac{1}{V'} \sum_{v=1}^{V'} \sum_{j \in \mathcal{C}_v \cup \{0\}} \|\mathbf{w}_{v,j+1} - \mathbf{w}_{v,j}\| / S_v \quad (P) \\
 & j = 1, \dots, K' \\
 & v = 1, \dots, V' \\
 \text{s.t. } & K' \leq K, V' \leq V, \quad (P.a) \\
 & d_{sj} \leq d_s^{\text{th}}, \forall j = 1, \dots, K', s \in \mathcal{D}_j, \quad (P.b) \\
 & P_{jv}^{\text{UAV}} \geq P_{\text{th}}, \forall v = 1, \dots, V', \forall j \in \mathcal{C}_v, \quad (P.c) \\
 & \left| \bigcup_{v=1}^{V'} \mathcal{C}_v \right| = K', \bigcap_{v=1}^{V'} \mathcal{C}_v = \emptyset \quad (P.d) \\
 & \left| \bigcup_{j=1}^{K'} \mathcal{D}_j \right| = N, \bigcap_{j=1}^{K'} \mathcal{D}_j = \emptyset \quad (P.e)
 \end{aligned}$$

where $\mathcal{W}_v = \{\mathbf{w}_v = [x_{v,j}, y_{v,j}, z_{v,j}] | \forall j \in \mathcal{C}_v\}$ is the list of places where the UAV will go to get information from its cluster heads, and \mathbf{w}_v denotes the position in Cartesian coordinates. Furthermore, $\mathcal{H}_v = \{\mathbf{h}_j = [x_j, y_j, z_j] | \forall j \in \mathcal{C}_v\}$ is the list of all CHs the UAV should visit v in the order they were listed. Furthermore, $\mathbf{w}_{v,0} = \mathbf{w}_{v,|\mathcal{C}_v|+1}$ is the dock station for the UAV v , and S_v is the average flying speed for v . Lastly, $|\cdot|$ is the cardinality operator, and $\|\cdot\|$ is the Euclidean norm operator.

Because we are working under the assumption that the amount of time spent communicating is both constant and sufficient to enable the successful collection of all data from the CH, we can disregard it in the formulation for the objective function and focus on minimizing the average time required to gather the data. The number of CH and UAVs may be limited due to financial constraints, (P.b)–(P.c) and (P.d)–(P.e) ensure that the CHs and SNs, and the UAVs and CHs, can successfully communicate and associate with another.

It has been established that this is an NP-hard problem. Throughout the specific scenario of a single UAV that has currently planted CHs concerning (P.d), the issue is limited to determine the shortest possible route for the UAV. One possible interpretation of the latter is the TSP. A salesman in TSP must visit several places while cutting down on overall travel time. It only makes sense for UAVs to replace salespeople and CHs to replace cities. For the same reason that TSP is NP-hard, we can safely assume that our problem is also intractable.

5 Proposed Approach

Directly solving (P) is difficult. So, we take two steps:

- We discover the locations of CHs to fulfill (P.b), (P.e), for given K' and V' , namely, sets $\mathcal{D}_j (j = 1, \dots, K')$ and $\mathcal{H} = \bigcup_{v=1}^{V'} \mathcal{H}_v$. This shows the clustering of nodes of the WSN.

- The corresponding CHs, trajectory, and specific aerial places to visit are then calculated for each UAV, that is, \mathcal{C}_v , \mathcal{H}_v and \mathcal{W}_v , $\forall v = 1, \dots, V'$. This step is when the planning of the trajectory takes place.

It should be emphasized that the suggested method produces a suboptimal result. Nevertheless, it has appealing implementation properties such as simplicity and complete parameter control.

5.1 Nodes Clustering in Wireless Sensor Networks

Data collection techniques that include clustering SNs and deploying CHs are popular due to the low power consumption they provide. It is true that low-power SNs are limited in their ability to interact with the collector or that the collector must go extremely close to the SNs to do so, resulting in a waste of energy. Moreover, CHs enable centralized data collection and filtering before transmission to the network. This ensures the scalability of the WSN and allows CHs to analyze massive volumes of data. Using techniques like filtering, backup CHs, and re-clustering, clustering can withstand malfunctions and breakdowns.

In this work, we choose K-means for SN clustering and the placement of specialized CHs. Consequently, the corresponding clustering issue is

$$\min_{\mathcal{H}, \mathcal{D}_j} \sum_{j=1}^{K'} \sum_{s \in \mathcal{D}_j} \|\mathbf{g}_s - \mathbf{h}_j\|^2$$

s.t. P.a, P.b, P.e. (P')

where $\mathbf{g}_s = [x_s, y_s, z_s]$, for every $s = 1, \dots, N$, is the position of SNs. It has been established that this is an NP-hard issue. K-means updates the positions of CHs, which are the positions of SNs that have been averaged inside the same cluster iteratively to solve (P').

5.2 Trajectory Planning

(P') is simplified to linking CHs to UAVs and determining data-collection paths with the least number of CHs deployed. After visiting all CHs, each UAV returns to its dock station. mTSP is this problem. Many techniques have been suggested for NP-hardness. First, we optimize trajectory when UAVs hover directly over CHs. Next, UAVs hovering near CHs will be examined.

5.2.1 UAVs Fly in a Straight Line Directly Over CHs

Connecting a modest number of CHs to UAVs and planning the most effective routes for data collection allow us to achieve this (P). UAVs depart their bases, go to their assigned CHs, collect information while hovering above, and eventually return to their bases. Let us assume, for the sake of argument that a UAV remains at the same altitude during its entire flight and is never further than a safe distance from the related CHs (P.e). In addition, we assume that the hovering time is constant and long enough to capture all the data from the CH properly and that the UAVs all fly at the same average speed $S_v = S, \forall v = 1, \dots, V'$. This allows us to write as an expression for the trajectory planning problem [15].

$$\begin{aligned}
 & \min_{\mathcal{X}} \sum_{v=1}^{V'} \sum_{i=0}^{K'} \sum_{\substack{j=0 \\ j \neq i}}^{K'} \| \mathbf{h}_i - \mathbf{h}_j \| x_{ij}^v \quad (P'') \\
 & \text{s.t.} \sum_{v=1}^{V'} \sum_{i=0}^{K'} x_{ij}^v = 1, \forall j = 0, \dots, K', j \neq i \quad (P''.a) \\
 & \sum_{i=1}^{K'} x_{ip}^v - \sum_{j=1}^{K'} x_{pj}^v = 0, \forall p = 1, \dots, K', \forall v = 1, \dots, V' \quad (P''.b) \\
 & \sum_{j=1}^{K'} x_{0j}^v = 1, \forall v = 1, \dots, V' \quad (P''.c) \\
 & n_i - n_j + K' \sum_{v=1}^{V'} x_{ij}^v \leq K' - 1, x_{ij}^v \in \{0, 1\}, \quad (P''.d)
 \end{aligned}$$

With $\| \mathbf{h}_i - \mathbf{h}_j \|$ represents the distance between CHs i and j , $\mathbf{h}_0 = \mathbf{w}_0 = \mathbf{w}_{v,0}$ is the first and the same docking station for all UAVs, $\mathcal{X} = \{ x_{ij}^v | i = 1, \dots, K', j = 1, \dots, K', v = 1, \dots, V' \}$ with x_{ij}^v corresponds to a binary index of UAV v flying from CH $_i$ to CH $_j$, $x_{ij}^v = 1$ if the trajectory of UAV v includes the route, else, $x_{ij}^v = 0$. In addition to this, both n_i and n_j are positive integers [15]. It should be stated in (P''.a) that each CH is only visited once. (P''.b) refers to flow conservation restrictions, which state that when a UAV visits a CH, it must then leave from the same CH it has already visited. (P''.c) guarantees that a UAV is only used once, and (P''.d) are the MTZ-based sub-tour removal restrictions, which ensure that degenerate trips that are not linked to the dock station are not taken into account.

Alternatives like approximate and metaheuristic algorithms are appealing if they can provide nearly optimum paths. In this article, we use methods such as ACO, SA, PSO, and GA that use an approximation to plan a trajectory.

5.2.2 UAVs Fly in a Range of CHs (TSP-M)

A UAV may rest in a distant location but still, be close enough to do the job. This may be seen as evidence that the CHs have a coverage region in the sky where UAVs can reliably send and receive data. Specifically, we solve $P_{jv}^{UAV} = P_c - \Lambda_{jn}$. to obtain the largest possible radius for this coverage region at the specified UAV altitude.

Lemma 1 As indicated by Eq. (7), the Cluster Head’s maximum coverage radius may be represented when a UAV is flying at a height of z .

$$\mathcal{R}^*(s) = \frac{s}{\tan[f^{-1}[P_h - (L_{FS}(f_c) + P_{th} + 20\log(s))]]} \tag{7}$$

In which f^{-1} and \tan represents respectively the reciprocal of f (given by Equation (8)) and the tangent function.

$$f(\phi) = \frac{\gamma_{LoS} + \gamma_{NLoS}ae^{-b(\phi-a)}}{1 + ae^{-b(\phi-a)}} - 20\log(\sin(\phi)), \phi \in [0, \pi] \tag{8}$$

Remark 1 The proof of the previous result in Lemma 1, can be easily adapted using similar techniques to those of [14].

UAVs may now optimize their flight paths by traveling to the periphery of the territories covered by CHs, given \mathcal{R}^* of every CH. This issue is known as mTSPN (multiple-TSPN). For this issue, we recommend modifying the original mTSP solution. Indeed, after determining the trajectory among the CH to be visited, we alter the hovering locations by changing them to be the nearest locations on the edges of the coverage area. For example, suppose that UAV v is located at the coordinates $\mathbf{w}_v = [x_v, y_v, z_v]$ and that the coordinates of the projected position of the next CH₁ to be visited in the UAV plane are $\overline{\mathbf{w}}_c = [x_c, y_c, z_v]$, and the projected position of the next CH₂ to be visited on the UAV plane are $\overline{\mathbf{w}}_{c'} = [x_{c'}, y_{c'}, z_v]$, and we define \mathbf{w}_t that we use in the following, has coordinates $\mathbf{w}_t = [x_t, y_t, z_v]$ where $x_t = x_v + x_{c'} - 2x_c$ and $y_t = y_v + y_{c'} - 2y_c$. The function in Eq. (9) can represent the perimeter of coverage of the related CH1’s.

$$(\mathcal{R}^*(z_v))^2 = (y - y_c)^2 + (x - x_c)^2 \tag{9}$$

When the UAV travels in the direction of CH₁ that will be investigated, the point on the perimeter of the coverage that is closest to the UAV is the first place where a line made between the locations $\overline{\mathbf{w}}_c$ and \mathbf{w}_t intersects with the perimeter of the coverage.

Lemma 2 Let $y = p_1x + p_2$ be the line’s function, where $p_1 = (y_t - y_c)/(x_t - x_c)$ and $p_2 = y_c - p_1x_c$.

The hovering location's coordinates \mathbf{w}_v' can then be supplied by

$$\mathbf{w}_v' = \begin{cases} \mathbf{w}_v^0 = [x_0, y_0, z_v], & \text{if } \|\mathbf{w}_v^0 - \mathbf{w}_v\| \leq \|\mathbf{w}_v^1 - \mathbf{w}_v\| \\ \mathbf{w}_v^1 = [x_1, y_1, z_v], & \text{otherwise} \end{cases} \quad (10)$$

where $x_0 = x_c + \sqrt{q_1}$, $x_1 = x_c - \sqrt{q_1}$, $y_0 = p_1x_0 + p_2$, $y_1 = p_1x_1 + p_2$ and $q_1 = (\mathcal{R}^*(z_v))^2 / (p_1^2 + 1)$.

Proof The points that satisfy both Eqs. (9) and $y = p_1x + p_2$ are those found at the intersection of the coverage perimeter and the line that is drawn through $\bar{\mathbf{w}}_c$ and \mathbf{w}_t . After making a few adjustments, the result that we get when we insert the y variable from the equation of the line into Eq. (9) is

$$\begin{aligned} (p_1x + (p_2 - y_c))^2 + (x - x_c)^2 &= (\mathcal{R}^*(z_v))^2 \\ \Leftrightarrow (p_1x - p_1x_c)^2 + (x - x_c)^2 &= (\mathcal{R}^*(z_v))^2 \\ \Leftrightarrow (x - x_c)^2 &= \underbrace{\frac{(\mathcal{R}^*(z_v))^2}{p_1^2 + 1}}_{=q_1} \end{aligned} \quad (11)$$

Thus, Eq. (10) is a polynomial that can be solved as in Eq. (11).

5.2.3 Modification of TSP-M (TSP-M1)

We will modify our method, and for that, we will define another trajectory. We suppose the UAV v_1 has coordinates $\mathbf{w}_{v_1} = [x_{v_1}, y_{v_1}, z_v]$ and that the projection of the next CH_1 to visit on the UAV's plane has coordinates $\bar{\mathbf{w}}_c = [x_c, y_c, z_v]$, and $\mathbf{w}_v'' = [x_{v_1}'', y_{v_1}'', z_v]$ is the intersection point of the trajectory of the UAV v_1 and the perimeter of coverage for CH_2 and we define \mathbf{w}_{t_1} that we use in the following, as having coordinates $\mathbf{w}_{t_1} = [x_{t_1}, y_{t_1}, z_v]$ where $x_{t_1} = x_{v_1} + x_{v_1}'' - 2x_c$ and $y_{t_1} = y_{v_1} + y_{v_1}'' - 2y_c$. Note that the nearest point outside the coverage area is the first place where a line drawn through locations $\bar{\mathbf{w}}_c$ and \mathbf{w}_{t_1} intersects the coverage perimeter because the UAV flies toward the CH_1 that should be visited.

5.2.4 Modification of TSP-M1 (TSP-M2)

For the second modification, we will take another trajectory, we assume that UAV v_2 is at coordinates $\mathbf{w}_{v_2} = [x_{v_2}, y_{v_2}, z_v]$ and $\mathbf{w}_{v_1}'' = [x_{v_1}'', y_{v_1}'', z_v]$ is the intersection point of the UAV trajectory v_1 and the perimeter of coverage for CH_2 and we define $\mathbf{w}_{t_2} = [x_{t_2}, y_{t_2}, z_v]$ where $x_{t_2} = x_{v_2} + x_{v_1}'' - 2x_c$ and $y_{t_2} = y_{v_2} + y_{v_1}'' - 2y_c$. The nearest location outside the coverage region is where a line drawn through the $\bar{\mathbf{w}}_c$ and \mathbf{w}_{t_2} locations intersect the coverage perimeter because the UAV flies towards the CH_1 that should be visited.

6 Numerical Results & Discussion

We assume $M = 100$ SNs distributed randomly and evenly throughout a $10 \times 10 \text{ km}^2$ geographical region. Unless otherwise indicated, we utilize the simulation settings from [14].

Our K-means clustering is employed to get CH locations that meet (P.b), shown in Fig. 2. In the case where $K' = 7$, the hypothesis holds. In Fig. 3, we investigate the distances traveled by UAVs when they hover immediately above (TSP), (TSPN [14]) and our proposed TSPN-M. Let the lengths of the linked UAV's trajectories be denoted by the symbols d_{TSP} , d_{TSPN} and d_{TSPN-M} respectively. It has been demonstrated that TSPN-M is capable of achieving a shorter trajectory while simultaneously guaranteeing the collection of data from all 10 CHs. In this scenario, the trip distance gain, which is determined by the equation $\rho_{TSPN} = 1 - \frac{d_{TSPN}}{d_{TSP}}$, may be estimated as $\rho_{TSPN} = 13.49\%$ and $\rho_{TSPN-M} = 18.95\%$ respectively. That is, hovering within a particular distance of every CHs is beneficial because it reduces the length traveled, as a result, the overall amount of time required for the data-gathering operation.

Assuming $V' = 1$, we evaluate the distance traveled by the UAV versus the required number of CH plotted for the proposed TSP-ACO, TSP-GA, TSP-SA, and TSP-PSO algorithms in Fig. 4. For this discussion, let us pretend the UAV is hovering precisely above each CH. Therefore, the other techniques, such as TSPN, TSPN-M, TSPN-M1, and TSPN-M2, extend to the coverage limit of CH. Each strategy requires a longer travel as K' increases. This is to be expected, as increasing the number of CH visited will correspondingly increase the total distance covered. Given its near-optimal performance, TSP-ACO is recommended for use in TSPN, TSPN-M, TSPN-M1, and TSPN-M2.

Fig. 2 K-means (100 iterations)

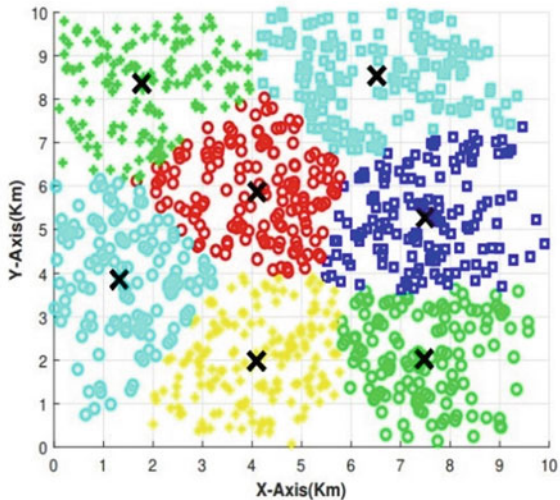


Fig. 3 Trajectories of UAVs

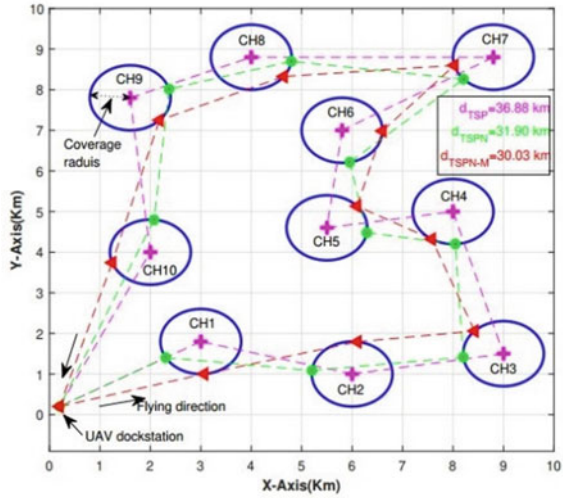
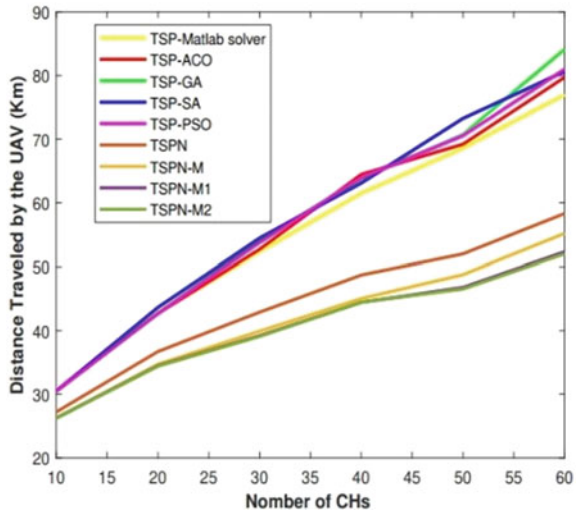
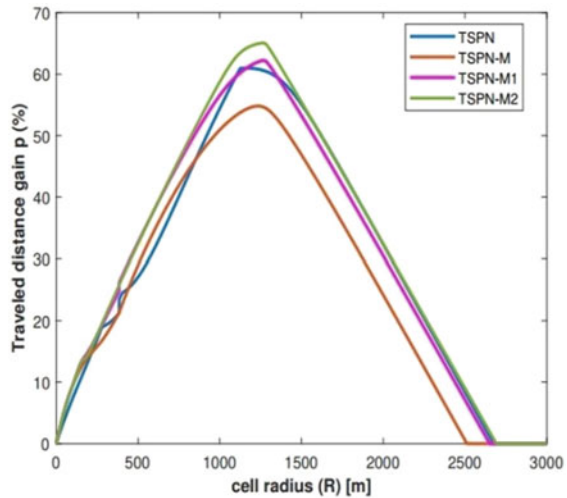


Fig. 4 The distance traveled by the UAV in relation to the number of CHs



In Fig. 5, we illustrate the relationship between Cell radius (R) and the traveled distance gain ρ . The scenario has a value of 10 CHs. The latter corresponds to the optimal radius (R). ρ increases rapidly up to 1200m and then degrades rapidly as R increases.

Fig. 5 Gain in average traveling distance versus cell radius (R)



7 Conclusion

In this article, we tackled the issue of collecting data from several UAVs in densely connected IoT networks where time-sensitive data must be gathered. We suggested a two-stage approach to minimize the energy costs associated with data collection deployment. In the first stage, a tailored K-means technique is used to optimize the number and placement of CHs, which collect data from related IoT sensors. Then, a methodology for collecting data while conserving energy and deploying a minimum number of UAVs is described, in which the trajectories of the UAVs are determined with regard to data collection deadlines and energy availability. Our modified version of K-means clustering outperforms the status quo in simulations. A large energy gain was also realized by relocating the CHs closer to the dock station. Extensive experiments are carried out on a set of instances with up to 100 IoT devices. The experimental results demonstrate that ACO achieves better performance compared with other algorithms. Future research might benefit from combining two of the competing methods (ACO & SA) since ACO provides the shortest distance while SA speeds up the process. So, they balance each other out and make up for their weaknesses in other optimization problems related to the TSP.

References

1. S. El Kafhali, K. Salah, Performance modeling and analysis of Internet of Things enabled healthcare monitoring systems. *IET Networks* **8**(1), 48–58 (2019)
2. S. El Kafhali, C. Chahir, M. Hanini, K. Salah, Architecture to manage internet of things data using blockchain and fog computing, in *Proceedings of the 4th International Conference on Big Data and Internet of Things*, pp. 1–8 (2019)
3. S. El Kafhali, I. El Mir, M. Hanini, Security threats, defense mechanisms, challenges, and future directions in cloud computing. *Arch. Comput. Methods Eng.* **29**(1), 223–246 (2022)
4. J. Chen, P. Chen, Q. Wu, Y. Xu, N. Qi, T. Fang, A game-theoretic perspective on resource management for large-scale UAV communication networks. *China Commun.* **18**(1), 70–87 (2021)
5. K.P. Valavanis, G.J. Vachtsevanos, (Eds.), *Handbook of unmanned aerial vehicles* (Vol. 1). Dordrecht: Springer Netherlands (2015)
6. B. Li, Z. Fei, Y. Zhang, UAV communications for 5G and beyond: recent advances and future trends. *IEEE Internet Things J.* **6**(2), 2241–2263 (2018)
7. F. Valdez, F. Moreno, P. Melin, A comparison of ACO, GA and SA for solving the TSP problem, in *Hybrid Intelligent Systems in Control, Pattern Recognition and Medicine*, pp. 181–189. Springer, Cham (2020)
8. K. Chaudhari, A. Thakkar, Travelling salesman problem: an empirical comparison between ACO, PSO, ABC, FA and GA, in *Emerging Research in Computing, Information, Communication and Applications*, pp. 397–405. Springer, Singapore (2019)
9. G. Pajares, Overview and current status of remote sensing applications based on unmanned aerial vehicles (UAVs). *Photogramm. Eng. Remote. Sens.* **81**(4), 281–330 (2015)
10. G.N. Muchiri, S. Kimathi, A review of applications and potential applications of UAV, in *Proceedings of the Sustainable Research and Innovation Conference*, pp. 280–283 (2022)
11. Y. Zhang, Z. Mou, F. Gao, L. Xing, J. Jiang, Z. Han, Hierarchical deep reinforcement learning for backscattering data collection with multiple UAVs. *IEEE Internet Things J.* **8**(5), 3786–3800 (2020)
12. X. Diao, J. Zheng, Y. Cai, Y. Wu, A. Anpalagan, Fair data allocation and trajectory optimization for UAV-assisted mobile edge computing. *IEEE Commun. Lett.* **23**(12), 2357–2361 (2019)
13. S. Jeong, O. Simeone, J. Kang, Mobile edge computing via a UAV mounted cloudlet: Optimization of bit allocation and path planning. *IEEE Trans. Veh. Technol.* **67**(3), 2049–2063 (2017)
14. S. Alfattani, W. Jaafar, H. Yanikomeroglu, A. Yongacoglu, Multi-UAV data collection framework for wireless sensor networks, in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6. IEEE (2019)
15. T. Bektas, The multiple traveling salesman problem: an overview of formulations and solution procedures. *Omega* **34**(3), 209–219 (2006)

Synergistic Fibroblast Optimization Algorithm for Solving Knapsack Problem



T. T. Dhivyaprabha and P. Subashini

Abstract Knapsack Problem (KP) is a classical combinatorial optimization problem in operation research. The characteristics of knapsack problems are widely applied in diverse sort of engineering fields. Various algorithms and techniques were proposed by many researchers to resolve knapsack problems. But, it still remains a challenging task due to the tremendous increase in problem complexity in a dynamic environment. The objective of this paper is to introduce the novel Synergistic Fibroblast Optimization (SFO) algorithm for solving knapsack problems, particularly 0–1 single knapsack problems and multiple knapsack problems. The efficiency of SFO is tested with benchmark instances and compared with other most popular metaheuristic algorithms such as Particle Swarm Optimization (PSO), Differential Evolution (DE) and Cuckoo Search (CS). The computational results demonstrate that the proposed SFO produces highly qualitative results and it outperforms other global optimization algorithms, in terms of, finding the optimal solution, reliability, fitness evaluation, convergence analysis and statistical measurements.

Keywords Combinatorial optimization · Friedman test · Global optimal solution · Knapsack problem · Synergistic Fibroblast Optimization (SFO)

1 Introduction

Knapsack problem is a typical combinatorial optimization problem proposed by Dantzing in the 1950s [1]. During the past few decades, many algorithms and techniques were introduced by researchers to resolve knapsack problem effectively. They offer several practical applications in various domains such as logistics and

T. T. Dhivyaprabha (✉) · P. Subashini

Centre for Machine Learning and Intelligence, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
e-mail: ttdhivyaprabha@gmail.com

P. Subashini

e-mail: subashini_cs@avinuty.ac.in

shipping, project selection, resource allocation, scheduling, portfolio optimization, cutting stock, investment decision making and so on [2, 3]. In this paper, Synergistic Fibroblast Optimization (SFO), a novel metaheuristic algorithm is applied to solve 0:1 single and 0–1 multiple knapsack problems, and the performance of SFO is compared with other popular nature-inspired computing paradigms, namely, Particle Swarm Optimization (PSO), Differential Evolution (DE) and Cuckoo Search (CS) [4–6]. Let n be the distinct items where each j th item has associated weight (w_j), profit (p_j) and knapsack capacity (C). The goal of knapsack problem is to maximize the profit value with respect to most of the items packed into the knapsack in all possible ways while satisfying the weight constraints that should not exceed the whole knapsack capacity. The most appropriate combination of item set determines the optimal profit value. Knapsack problem is classified into different types based on the distribution of items and quantity of knapsack used which is enlisted below.

- a. 0–1 single knapsack problem: Each given item may be chosen, at most once.
- b. 0–1 multiple knapsack problem: Each item from subset may choose to different knapsacks when the capacity of knapsack must be greater than the weight of items in the subset.
- c. Multiple choice knapsack problem: Items are partitioned into subsets and each item may choose per subset.
- d. Bounded knapsack problem: Each item may choose from subset until the upper bound limit of the item of each type has been attained.
- e. Unbounded knapsack problem: Each item may be chosen from the unlimited number of items of each type from subset available.

Among these, 0–1 single and 0–1 multiple knapsack problems have experimented in this study. Let n be the number of items and C denotes the capacity of knapsack. Each j th item has its own weight w_j and profit p_j . The mathematical representation of 0–1 single knapsack problem is denoted in Eq. (1).

$$\text{Maximize } f(x) = \sum_{j=1}^N p_j x_j \text{ subject to } \begin{cases} \sum_{j=1}^N w_j x_j \leq C \\ x_j = 0 \text{ or } 1, j = 1, 2, \dots, N \end{cases} \quad (1)$$

The binary decision variable x_j is used to indicate whether the item is included in the knapsack or not. It is assumed that the profit and weight of all the items are positive integer values, and the total weight of all the items should be less than the predefined capacity of knapsack.

The 0–1 multiple knapsack problem consists of n items and m knapsacks ($m \leq n$) and C denotes capacity of knapsack. Each n disjoint subset of items can be potentially placed into different knapsacks whose capacity should be greater than the weight of items. The mathematical representation of 0–1 multiple knapsack problem is given in Eq. (2).

$$\text{Maximize } f(x) = \sum_{i=1}^m \sum_{j=1}^n p_j x_{ij} \quad (2)$$

subject to

$$\sum_{j=1}^n w_j x_{ij} \leq C_i, j \in N = \{1, \dots, n\}$$

$$x_{ij} = 1 \text{ or } 0, i \in M = \{1, \dots, m\}, j \in N = \{1, \dots, n\}$$

where

$$x_{ij} = \begin{cases} 1, & \text{if item } j \text{ is assigned to knapsack } i \\ 0, & \text{otherwise} \end{cases}$$

The remainder of the paper is structured as follows. Section 2 describes the classification of the knapsack problem. The analysis of SFO modeling to solve combinatorial optimization problems is formulated in Sect. 3. The procedural steps of SFO to solve knapsack problems are presented in Sect. 4. Results and discussions are given in Sects. 5, and 6 draws the conclusion and future works.

2 Literature Review

Shabet et al. [7] proposed a variant of the discrete Artificial Bee Colony (ABC) optimization algorithm by introducing a hybrid probabilistic mutation scheme in the stochastic search strategy of traditional ABC for solving multiple knapsack problems. It was tested with knapsack instances and compared with Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) algorithm. Experimental results demonstrated that the proposed ABC achieved better results and high convergence speed than other metaheuristic algorithms [7]. Zoheir Ezziene et al. [8] implemented an adaptive genetic algorithm (GA) to solve 0–1 knapsack problem. GA was used with a penalty function consisting of a constant penalty function and an adaptive penalty function. Investigational results show that the self-adaptive penalty function fine-tunes the control parameters to yield the improved offspring (solution space) to knapsack problem and it gives an optimized solution better than the Greedy Method (GM) and Heuristic Algorithm (HA) [8]. Fidanova [9] applied Ant Colony Optimization (ACO) algorithm to solve multiple knapsack problems. Here, the problem is represented by graph and the solutions are represented by paths (arcs) through the graph. Two pheromone models, namely, pheromone on the arcs of the graph and pheromone on the nodes of the graph are compared. When the pheromone is laid on the nodes of the graph, the pheromone concentration becomes very high and ants choose the best solution with high probability [9].

Yampolskiy and EL-Barkouky [10] implemented a novel Wisdom of Artificial Crowd (WoAC) algorithm for solving knapsack problem. It was validated with knapsack instances and the proposed algorithm achieves higher performance than genetic

algorithm-based approach [10]. Kulkarni and Shabir [11] proposed Cohort Intelligence (CI) algorithm for solving 0–1 knapsack problem. It was applied to solve twenty distinct test cases of knapsack problem and the obtained results demonstrate its effectiveness [11]. Feng et al. [12] proposed a new variant hybrid encoding Cuckoo Search (CS) algorithm to solve 0–1 knapsack problem. The concept of Confidence Interval (CI) is introduced to improve the global best solution space rapidly at each generation. Experiments with a large number of instances show that the proposed algorithm does not affect the local optimum and it has the ability to achieve good solutions [12]. Bhattacharjee and Sarmah [13] introduced Binary Cuckoo Search Algorithm (BCSA) for solving the 0–1 knapsack problem. Experiments with several benchmark knapsack problem instances and comparison with PSO show that BCSA is able to attain the best solution over variants of PSO [13]. Moosavian [14] proposed a new metaheuristic Soccer League Competition (SLC) algorithm for solving 0–1 knapsack problems. It was tested with some benchmark knapsack test problems and the examined results demonstrated that SLC is an efficient algorithm, which produces promising results compared with other algorithms [14]. Layeb [15] presented Quantum Inspired Harmony Search (QIHS) hybrid approach for solving 0–1 knapsack optimization problem. This hybridization framework achieves better exploration of HS algorithm and exploitation capabilities of quantum-inspired computing. Examined results on the knapsack problem demonstrate the effectiveness of the proposed framework and its capability to achieve good quality solutions [15].

The potential literature studies demonstrated that Nature Inspired Computing [NIC] paradigm, which is inspired by nature as a source of metaphor for problem-solving, is considered the most appropriate method for solving this type of problem. To perform a better evaluation of SFO that resolves combinatorial optimization problems, knapsack problem has been selected in this work. The problem-solving strategy, adopted by knapsack, is used in various kinds of engineering problems such as trajectory detection, graph colouring, resource allocation and so on. The core idea of this problem is to find optimal or near-optimal solutions while satisfying the criteria which are subjected to certain constraints. The best combination of object selection determines the fitness of yielded solution. Henceforth, choosing the optimal set of objects is a non-deterministic process and it is considered a non-linear optimization problem.

3 Description of SFO to Solve Combinatorial Optimization Problems

Synergistic Fibroblast Optimization (SFO) is a population-based global search algorithm that has been developed by the inspiration obtained from collaborative and self-adaptive characteristics of fibroblast organism role in the dermal wound healing process [16, 17]. The promising results produced by SFO on solving different kinds

of problems, emphasized that the collaborative and self-adaptive characteristics, followed by SFO algorithm, could be well suited to solve non-linear complex problems [18–20]. This section gives a description of the transformation of SFO algorithm suitable for solving methods of combinatorial optimization problems.

3.1 Sample Generation

Each individual in the cell population is composed of D -dimensional vectors, where D is the dimensionality of the search space, x_i denotes position, v_i represents velocity and ECM designates the matrix representation of collagen protein deposition. The number of knapsack items is encoded in the cell population. The parameters such as collagen (C_i), weight (W_i) and profit (P_i) of each items are preset in Extracellular Matrix (ECM) which is represented as follows.

$$Population = \begin{bmatrix} D_1 \\ \cdot \\ D_n \end{bmatrix} = \begin{bmatrix} C_1 & W_1 & P_1 \\ \cdot & \cdot & \cdot \\ C_n & W_n & P_n \end{bmatrix}$$

3.2 Individual Assessment

Next, an objective function, including a profit valuation relating to each solution vector, is calculated. For every iteration, cell with the best candidate solution is chosen as local optima, and the rest of the cells are reorganized to follow the trajectory of outperforming cell. The velocity and position of cells are updated at each cycle to migrate toward the global optimum. Based on the obtained results, the best candidate solution is upgraded in the ECM and it is selected to seed the next generation of population evolution in the problem space. SFO employs significant interaction of population, that is, the swarm has shared collagen value as a message to interact with other cells, irrespective of their location in the evolutionary region. It reveals that SFO does not often suffer in the steep decline of gradients, i.e., quick convergence problem in the search space when compared to other metaheuristics. The parameters consist of cell speed (s), and diffusion coefficient (ρ) maintains an equilibrium state of divergence and convergence of candidate solution, which directs towards the optimum in the search space.

3.3 Check the Stopping Criterion

The evolutionary process of fibroblast to find global optima is repeated either the maximum number of executions or predetermined conditions are to be met. The process of Synergistic Fibroblast Optimization algorithm applied to solve Knapsack problem is described as follows.

Step 1: Initialize the discrete set of items as population of cells (n size), with their random generation of position (x_n) and velocity (v_n) in the two-dimensional problem space. The profit and weight values of the corresponding items are reserved as collagen deposition present in ECM. Initialization of parameters includes diffusion coefficient (ρ) = 0.8, cell speed ($s = \frac{s}{k_{ro}L}$), number of cells (L) and capacity of Knapsack (M) are predefined.

Step 2: Evaluate the fitness of cells using the following 0–1 single and 0–1 multiple objective functions.

$$Max C = \sum_{k=1}^N p_k w_k \text{ subject to } \sum_{i=1}^N w_k x_k \leq W \tag{3}$$

$$Max C = \sum_{k=1}^N p_k w_k \text{ subject to } \sum_{i=1}^N w_k x_k \leq W_k \tag{4}$$

where

C = objective function; p_k = profit of N items;

w_k = weight of N items;

Max = Maximization;

$x_k = (1, 0)$ indicates whether an item is included in the knapsack or not;

W = capacity of knapsack;

N = total number of items;

Step 3: The fitness value of each cell is compared with global best solution (Cbest). If the current cell has better fitness value ($Cbest_i$) than previous value ($Cbest_{i-1}$), it is substituted with the current value. Otherwise, no further operations are carried out in this step.

if ($cbest_i < Cbest$).

Cbest = $cbest_i$;

else

Cbest = $cbest_{i-1}$;

Step 4: The position and velocity of cell is updated at each cycle according to the following Eqs. (5) and (6).

$$v^{i(t+1)} = v^i(t) + (1 - \rho)c(f^i(t), t) + \rho * \frac{f^i(t - \tau)}{\|f^i(t - \tau)\|} \tag{5}$$

where

t = current iteration t ; τ = time lag; v_i = velocity of i th cell; f_i = i th cell;
 c = collagen chosen by i th cell;

$$x^{i(t+1)} = x^i(t) + s * \frac{v^i(t)}{\|v^i(t)\|} \quad (6)$$

where

$s = 15\mu\text{mh}^{-1}$; $K_{ro} = 10^3\mu/\text{min}$; $L = 10$; x_i = position of i th cell

Step 5: Each collagen is considered as a candidate solution in SFO algorithm. Cell with best collagen is updated in the ECM.

Step 6: Repeat the steps from 2 to 5 either the predefined Fitness Calculation Number (FCN) attained to 1000 or the predetermined capacity of knapsack is to be met.

4 Experimental Settings and Numerical Examples

The efficacy of Synergistic Fibroblast Optimization (SFO) algorithm to solve single and multiple knapsack problems is extensively investigated in this section. Six benchmark instances with a varied size of items and knapsacks taken from operation research library and are considered to testify the efficiency of the SFO algorithm [21, 22]. All the computations are implemented using MATLAB, JAVA programming language runs on Netbeans 7.1 environment and R studio open source statistical tool executed on Intel (R) Core (TM) i7-4790 CPU running 3.60 GHz with 4 GB RAM. The operating system platform was Windows 7 professional 64-bit machine. The parameter settings of SFO, namely, diffusion coefficient and cell speed, are kept constant for all selected test problems to validate the capability of SFO on solving knapsack problems with different sizes.

The details of test instances such as knapsack problem (f), dimensionality (Dim), profit, weight and capacity are listed in Table 1. Here, f1, f2 and f3 represent single knapsack problem; f4, f5 and f6 denote multiple knapsack problem. To investigate the efficiency of SFO for solving complex problems, it is tested with benchmark instances having diverse sort of complexity i.e., number of items are differing in f1 (5:1), f2 (15:1) and f3 (24:1) as well as the quantity of items and knapsacks are varied in f4 (35:4), f5 (60:5) and f6 (105:2).

The examined results of Friedman test obtained by PSO, DE, CS and SFO for f1, f2, f3, f4, f5 and f6 is listed in the last column F value of Table 2 in which the lower F value gives better results. It shows that the significance of global optimization algorithms is probably equivalent and the sensitivity and adaptability of SFO with different datasets is characterized by its collaborative and co-evolutionary behavior of cells. It is evident that SFO is comparatively good and competitive over other metaheuristics.

Table 1 Parameters for the 0–1 Knapsack Problem

f	Dim n × m	Parameter
f1	5 × 1	Weight = (12, 7, 11, 8, 9), Capacity = 26, Profit = (24, 13, 23, 15, 16)
f2	15 × 1	Weight = (70, 73, 77, 80, 82, 87, 90, 94, 98, 106, 110, 113, 115, 118, 120), Capacity = 750, Profit = (135, 139, 149, 150, 156, 163, 173, 184, 192, 201, 210, 214, 221, 229, 240)
f3	24 × 1	Weight = (382, 745, 799, 601, 909, 247, 729, 069, 467, 902, 44, 328, 34, 610, 698, 150, 823, 460, 903, 959, 853, 665, 551, 830, 610, 856, 670, 702, 488, 960, 951, 111, 323, 046, 446, 298, 931, 161, 31, 385, 496, 951, 264, 724, 224, 916, 169, 684), Capacity = 6,404,180, Profit = (825, 594, 1,677, 009, 1,676, 628, 1,523, 970, 943, 972, 97, 426, 69, 666, 1,296, 457, 1,679, 693, 1,902, 996, 1,844, 992, 1,049, 289, 1,252, 836, 1,319, 836, 953, 277, 2,067, 538, 675, 367, 853, 655, 1,826, 027, 65, 731, 901, 489, 577, 243, 466, 257, 369, 261)
f4	35 × 4	Weight = (560, 1125, 620, 68, 328, 47, 122, 196, 41, 25, 115, 82, 22, 631, 132, 420, 86, 42, 103, 215, 81, 91, 26, 49, 316, 72, 71, 49, 108, 116, 90, 215, 58, 47, 81), Capacity = (163 165 239 168) Profits: 3186 constraint 1 = (40, 91, 30, 3, 12, 3, 18, 25, 1, 1, 8, 1, 1, 49, 8, 21, 6, 1, 5, 10, 8, 2, 1, 0, 42, 6, 4, 8, 0, 10, 1, 8, 3, 2, 4) constraint 2 = (16, 92, 16, 4, 18, 6, 0, 8, 2, 1, 6, 2, 1, 70, 9, 22, 4, 1, 5, 10, 6, 4, 0, 4, 8, 4, 3, 0, 10, 0, 6, 22, 0, 2, 2) constraint 3 = (38, 39, 71, 5, 40, 8, 12, 15, 0, 1, 20, 3, 0, 40, 6, 8, 0, 6, 4, 22, 4, 6, 1, 5, 8, 2, 8, 0, 20, 0, 0, 13, 6, 1, 2) constraint 4 = (38, 52, 42, 7, 20, 0, 3, 4, 1, 2, 4, 6, 1, 18, 15, 38, 10, 4, 8, 6, 0, 0, 3, 0, 6, 1, 3, 0, 3, 5, 4, 18, 3, 4, 0)

(continued)

Table 1 (continued)

f	Dim n × m	Parameter
f5	60 × 5	Weight = (360,83,59,130,431,67,230,52,93,125,670,892,600,38,48,147,78,256,63,17,120,164,432,35,92,110,22,42,50,323,514,28, 87,73,78,15,26,78,210,36,85,189,274,43,33,10,19,389,276,312,94,68,73,192,41,163,16,40,195,138) Capacity = (1024,1700,1850,510,1310) Profits: 4554 constraint 1 = (7,0,30,22,80,94,11,81,70,64,42,47,52,32,26,48,55,6,29,84,2,4,18,56,7,29,93,44,71,3,86,66,31,65,0,79,20,65,52,13,48, 14,5,72,14,39,46,27,11,91) constraint 2 = (8,66,98,50,30,0,88,15,37,26,72,61,57,17,27,83,3,9,66,97,42,2,44,71,11,25,74,90,20,0,38,33,14,9,23,12,58,6,14,78, 0,12,99,84,31,16,7,33,20,5,18,96,63,3,1,0,70,4,66,9) constraint 3 = (3,74,88,50,55,19,0,6,30,62,17,81,25,46,67,28,36,8,1,52,19,37,27,62,39,84,16,14,21,5,60,82,72,89,16,5,29,7,80,97,1, 46,15,92,51,76,57,90,10,37,25,93,5,39,0,97,6,96,2,81) constraint 4 = (21,40,0,6,82,91,43,30,62,91,10,41,12,4,80,77,98,50,78,35,7,1,96,67,85,4,23,38,2,57,4,53,0,33,2,25,14,97,87,42,15, 65,19,83,67,70,80,39,9,5,1,31,36,15,30,87,28,13,40,0) constraint 5 = (94,86,80,92,31,17,65,51,46,66,44,3,26,0,39,20,11,6,55,70,11,75,82,35,47,99,5,14,23,38,94,66,64,27,77,50,28,25,61, 10,30,15,12,24,90,25,39,47,98,83,56,36,6,66,89,45,38,1,18,88)
f6	105 × 2	Weight = (41,850,38,261,23,800,21,697,7074,5587,5560,5500,3450,2391,367,24,785,47,910,30,250,107,200,4235,9835,9262,15,000, 6399,6155,10,874,37,100,27,040,4117,32,240,1600,4500,70,610,6570,15,290,23,840,16,500,7010,16,020,8000,31,026,2,568,365,4350, 1972,4975,29,400,7471,2700,3840,22,400,3575,13,500,1125,11,950,12,753,10,568,15,600,20,652,13,150,2900,1790,4970,5770,8180, 2450,7140,12,470,6010,16,000,1,100,11,093,4685,2590,11,500,5820,2842,5000,3300,2800,5420,900,13,300,8450,5300,750,1435, 2100,7215,2605,2422,5500,8550,2700,540,2550,2450,725,445,700,1720,2675,220,300,405,150,70) Capacity = (3000,3000) Profits: 1,095,445 constraint 1 = (75,40,365,95,25,17,125,20,22,84,75,50,5,0,0,12,0,10,0,50,0,0,10,0,0,50,60,150,0,0,0,75,0,102,0,0,0,40,60,0,165,0,0,0,45, 0,0,0,25,0,150,0,0,158,0,85,95,0,89,20,0,0,0,0,0,80,0,110,0,15,0,60,5,135,0,0,25,0,300,35,100,0,0,25,0,0,225,25,0,0,0,0,0,0,5,0, 60,0,100,0,0,0,0) constraint 2 = (0,0,0,0,0,0,0,0,0,0,5,10,10,50,2,5,5,10,5,6,11,41,30,5,40,2,6,100,10,25,39,30,13,30,15,60,5,5,10,5,15,91,24,10, 15,90,15,60,5,5,60,50,75,100,65,15,10,30,35,50,15,45,80,40,110,80,80,36,20,90,50,25,50,35,30,60,10,150,110,70,10,20,30,104,40, 40,94,150,50,10,50,50,16,10,20,50,90,10,15,39,20,20)

Table 2 Comparison Results of 0–1 Knapsack Problems

f	Dim $n \times m$	Method	AVPFT	MAXPFT	WHTGP	F value
f1	5×1	PSO	40.00	46	0	3.9669
		DE	39.90	37	15	3.9130
		CS	25.30	31	19	4.0336
		SFO	51	51	0	1.9682
f2	15×1	PSO	1020.1	1388	8	14.0006
		DE	483.60	690	7	14.0001
		CS	458.90	920	14	14.0000
		SFO	1457.8	1458	0	13.8888
f3	24×1	PSO	1,251,152	10,427,637	9	23.0000
		DE	329,660.9	98,124	27	23.0000
		CS	469,660.9	1,022,340	45	23.0000
		SFO	12,206,734	13,549,094	0	22.8175
f4	35×4	PSO	3126.5	3180	1	34.0045
		DE	1631.2	2631	3	34.0000
		CS	1059	2436	2	34.0000
		SFO	3160.1	3186	0	34.0108
f5	60×5	PSO	6000	6569	3	1.0400
		DE	5565.6	6130	4	1.0407
		CS	4566.1	5499	4	1.0400
		SFO	6712.4	6912	1	1.0000
f6	105×2	PSO	625,162.1	102,885	4	1.3400
		DE	1,558,138	931,223	8	1.0300
		CS	779,585.4	1,029,757	8	1.0301
		SFO	1,059,697	1,087,687	1	1.0100

5 Conclusion and Future Works

In this paper the efficacy of SFO algorithm is tested for solving a combinatorial optimization problem, specifically, 0–1 single and 0–1 multiple knapsack problems in this case. SFO is utilized to solve benchmark instances with varied sizes such as, f1 (5), f2 (15), f3 (24), f4 (35), f5 (60) and f6 (105). Investigational results signify that SFO exhibit great potential in solving 0–1 knapsack problems. SFO demonstrates a strong advantage over PSO, DE and CS algorithms to find the global optimum with high reliability. It delivers a highly qualitative solution than the other three metaheuristic algorithms in terms of, packing a large number of items to yield maximum profit. This study can be further extended to validate SFO in solving large

volumes of dataset, bounded and unbounded knapsack problems and other types of combinatorial optimization problems in the future.

References

1. Y. Zhou, X. Chen, G. Zhou, An improved monkey algorithm for a 0–1 knapsack problem. *Elsevier J. Appl. Soft Comput.* **38**, 817–830 (2016)
2. D. Zou, L. Gao, S. Li, Wu., Jianhua, Solving 0–1 knapsack problem by a novel global harmony search algorithm. *Elsevier J. Appl. Soft Comput.* **11**, 1556–1564 (2011)
3. J. John Bartholdi, The knapsack problem building intuition: insights from basic 19 operations management models and principles. Chapter 2, Springer Science + Business Media, pp. 19–31 (2008), <https://doi.org/10.1007/978-0-387-73699-0>
4. Jr. Iztok Fister, D. Fister, A comprehensive review of cuckoo search: variants and hybrids. *Int. J. Math. Model. Numer. Optim.* **4**, 387–409 (2013)
5. R. Poli, J. Kennedy, T. Blackwell, Particle swarm optimization an overview. *Springer Sci. Swarm Intell.* **1**, 33–57 (2007)
6. S. Das, Ponnuthurai Nagaratnam Suganthan, “Differential evolution: a survey of the state-of-the-art.” *IEEE Trans. Evol. Comput.* **15**, 4–31 (2011). <https://doi.org/10.1109/TEVC.2010.2059031>
7. S. Sabet, M. Mohammad Shokouhifar, F. Farod Farokhi, A discrete artificial bee colony for multiple knapsack problem. *Int. J. Reasoning-Based Intell. Syst.* **5**, 88–95 (2013)
8. Z. Ezziane, Solving the 0/1 knapsack problem using an adaptive genetic algorithm. *Artif. Intell. Eng. Des. Anal. Manuf.* **16**, 23–30 (2002). <https://doi.org/10.1017/S0890060401020030>
9. S. Fidanova, Ant colony optimization and multiple knapsack problem. IGI, Chapter **33**, 1–12 (2007)
10. A. Roman Yampolskiy, Ahmed EL-Barkouky, “Wisdom of artificial crowds algorithm for solving NP-hard problems.” *Int. J. Bio-Inspired Comput.* **3**, 358–369 (2011)
11. J. Anand Kulkarni, H. Shabir, Solving 0–1 knapsack problem using cohort intelligence algorithm. *Int. J. Mach. Learn. Cyber* **7**, 427–441 (2016), <https://doi.org/10.1007/s13042-014-0272-y>
12. Y. Feng, K. Jia, Y. He, An improved hybrid encoding cuckoo search algorithm for 0–1 knapsack problems. *Hindawi Publishing Corporation Comput. Intell. Neurosci.*, 1–9 (2014), <https://doi.org/10.1155/2014/970456>
13. K.K. Bhattacharjee, S.P. Sarmah, A binary cuckoo search algorithm for knapsack problems. *International Conference on Industrial Engineering and Operations Management*, pp. 1–5 (2015), <https://doi.org/10.1109/IEOM.2015.7093858>
14. N. Moosavian, Soccer league competition algorithm for solving knapsack problems. *Elsevier J. Swarm and Evolut. Comput.* **20**, 14–22 (2015)
15. A. Layeb, A hybrid quantum inspired harmony search algorithm for 0–1 optimization problems. *Elsevier J. Computat. Appl. Math.* **253**, 14–25 (2013). <https://doi.org/10.1016/j.cam.2013.04.004>
16. P. Subashini, T.T. Dhivyaprabha, M. Krishnaveni, Synergistic fibroblast optimization, in: S.S. Dash, Vijayakumar, K., B.K. Panigrahi, S. Das (Eds.), *Springer Artificial Intelligence and Evolutionary Computations in Engineering Systems*, first ed. Chapter 26, 517, pp. 293–302 (2017)
17. T.T. Dhivyaprabha, P. Subashini, M. Krishnaveni, Synergistic fibroblast optimization: a novel nature-inspired computing algorithm. *Frontiers Inf. Technol. Electron. Eng.* **19**, 815–833 (2018). <https://doi.org/10.1631/FITEE.1601553>
18. T.T. Dhivyaprabha, P. Subashini, M. Krishnaveni, Computational intelligence based machine learning methods for rule-based reasoning in computer vision applications. *IEEE Symposium Series on Computational Intelligence*, pp. 1–8 (2016)

19. T.T. Dhivyaprabha, P. Subashini, Performance analysis of synergistic fibroblast optimization (SFO) algorithm. IEEE International Conference on Current Trends in Advanced Computing, pp. 1–6 (2017)
20. M. Krishnaveni, P. Subashini, T.T. Dhivyaprabha, A new optimization approach – SFO for denoising digital images. IEEE International Conference on Computational Systems & Information Technology for Sustainable Solution, pp. 34–39 (2016)
21. http://people.sc.fsu.edu/~jburkardt/datasets/knapsack_01/knapsack_01.html
22. <http://people.brunel.ac.uk/~mastjjb/jeb/orlib/files/mknap2.txt>

Acceptance of Job Access with Speech (Jaws) as an Assistive Computer Application Software



Lihle Ndlovu, Anass Bayaga, and Sylvan Blignaut

Abstract South Africans acceptance and use of Job Access with Speech (JAWS) remain under-researched, and concern particularly with Technical and Vocational Education and Training (TVET) college educators and students with visual impairments (SWVI) in teaching and learning of computer practice. A review of literature is thus conducted to examine TVET college educators' and SWVIs' acceptance of JAWS as an assistive computer application software (ACAS). The study employed a literature review search on TVET educators' perceived ease of use, perceived usefulness, and self-efficacy on TVET educators' and SWVIs' acceptance of JAWS in response to the concerns. The examination of the literature revealed that there is insufficient evidence of studies using TAM theory and determinants such as the perceived usefulness, the perceived ease of use, and computer self-efficacy as determinants in understanding JAWS acceptance in South Africa. Future research is still required to establish how the acceptance of JAWS in teaching and learning of computer practice using JAWS is affected by the perceived usefulness and the perceived ease of use in South Africa as well as how self-efficacy influence TVET college educators' and SWVIs' acceptance of JAWS in teaching and learning of computer practice.

Keywords JAWS · Perceived ease of use · Perceived usefulness · Self-efficacy

L. Ndlovu · A. Bayaga (✉) · S. Blignaut
Nelson Mandela University, Gqeberha, South Africa
e-mail: Anass.Bayaga@mandela.ac.za

L. Ndlovu
e-mail: lihlendlovu955@gmail.com

S. Blignaut
e-mail: Sylvan.Blignaut@mandela.ac.za

1 Introduction

Like the term used in many other assistive technology studies, assistive computer application software (ACAS) in the current study refers to “any software program used to enhance the functional capabilities of any disabled person” [1, 2]. The JAWS software facilitates the interaction between SWVI and computing applications and enhances SWVI’s acquisition of computer skills, self-reliance, and independence [3]. Regardless, some persisting influencers include the perceived ease of use and the perceived usefulness including ACAS. In special education, perceived ease of use is the users’ perception of effortlessness in using an ACAS due to the little cognitive effort needed during the use in teaching and learning of students experiencing barriers to education [4]. On the other, perceived usefulness characterises the degree of usefulness end-users associate with the information systems, thus enhancing the user’s job performance [5]. Self-efficacy is a person’s judgment of their capabilities that they will be likely to succeed in a particular situation or task [6].

Nonetheless, the argument is that with heightened scholarly attention on ACAS specialised for enhancing SWVI’s educational experiences, there is still very little to no consistent attention in the South African context [7], particularly the explicitness of JAWS usage in the TVET sector.

Noteworthy increase of studies outside of South Africa have mentioned JAWS as an enabler of digital information access to SWVI who prefer reading through screen readers [8-12]. For example, JAWS was the enabler for library book access in academic libraries for 14 universities worldwide [9]. Additionally, JAWS was one of the screen readers used by academics and professionals with visual impairment to perform collaborative writing with sighted colleagues [13]. In another study, JAWS was an enabler for SWVI to access information in high institutions of learning in Pakistan [10]. Moreover, the unavailability of JAWS was one of the constraints and complexities related to the learning experiences of SWVI in higher education in Eswatini access [14]. JAWS has been a comparative screen reader to measure the task performance of screen-readers prototypes [15]; there was also an exploration of JAWS’ possibilities to read other foreign languages [16].

Outside of South Africa, ‘acceptance’ studies relating to various digital products and services used by people with visual impairments have proven the use of JAWS as a primary application required for the facilitation and accessibility of these services. For instance, determining factors that affect visually impaired users’ acceptance of audio and music websites [17], understanding of factors influencing accessibility to a website and its acceptance by university students with visual impairments [18], and factors enabling the users with visual impairment to use websites [20]. A common finding from these studies is the inaccessibility of some of the products’ application features due to screen readers’ incompatibility, and JAWS was no exception. As such, a consistent recommendation from these acceptance studies is a need for further research on the accessibility and compatibility of JAWS and other screen readers to enhance access and acceptance of the pre-mentioned products and services for people with visual impairment. Noticeably, the pre-mentioned acceptance studies examining

perceived ease of use and perceived usefulness of these products as behavioural contracts have less to no relevance to teaching and learning. Hence, the view is that perhaps the start would be a need to examine the implications of perceived ease of use and perceived usefulness in the acceptance of JAWS with particular focus on users' computer self-efficacy in the seamless application of JAWS as a fundamental requirement for accessibility of any other digital application, including applications for teaching and learning purposes.

Regardless of the increased use of ACAS, often referred to as assistive technologies, most technology studies related to the education of students experiencing barriers to learning, there has been the inability to re-direct attention towards a consideration of self-efficacy as a central behavioural construct which is a contributing enabler of successful implementation of inclusive practices and ACAS usage within inclusive institutes of learning in South Africa [8, 21, 22]. Thus, [22, 23] put an emphasis on the importance of educators of SWVI to be well-informed and fluent with high self-efficacy in such specialised computer skills to train their students in using ACAS such as JAWS. Other constructs attached to self-efficacy include but are not exclusive to educators' ACAS knowledge, ACAS training, and basic computer literacy [8].

1.1 Research Objective

Consequently, the objective is to examine ease of use, perceived usefulness, and self-efficacy on TVET college educators' and SWVIs' acceptance of JAWS in teaching and learning computer practice.

1.2 Research Questions

- (1) What are the effects, if any, of the perceived ease of use on TVET educators' and SWVI's acceptance of JAWS?
- (2) What are the effects, if any, of the perceived usefulness on TVET educators' and SWVI's acceptance of JAWS?
- (3) What are the effects, if any, of self-efficacy on TVET educators' and SWVI's ac-acceptance of JAWS?

2 Methodology

To respond to the following objects:

The study located articles relating to implications of perceived ease of use, perceived usefulness, and self-efficacy as determinants of TVET educators' and

SWVIs' acceptance of JAWS as an ACAS. The following journals were accessible to the researchers: Journal on Multimodal User Interfaces, British Journal of Visual Impairment, Education and Information Technologies, Universal Access in the Information Society, Journal of Visual Impairment & Blindness, Journal of Computer Education, Teacher Education and Special Education, International Journal of STEM Education, Journal of Special Education Technology, and Google Scholar. The search was enabled through the use of Boolean Operators: AND and "". While in some cases, it was single entity, in others it was combined with other entities. Due to the scarcity of published research about the intended study, there was no limit on the number of years for the available published research. Therefore, the literature identified was published in the past 14 years (2006–2020). Additionally, specific articles relevant to the topic were purposively sampled and studied through in-depth literature. Consequently, the literature search yielded eight studies related to the research themes (see Table 1).

2.1 Criteria for Inclusion and Exclusion

We were interested in recent (2021 to 2022) articles which met the following criteria: The study is a peer-reviewed article published in English, included the use and acceptance of JAWS software in the TVET sector, included students with visual impairments and or educators or lecturers of students with visual impairments as participants. However, due to the scarcity of published research about the intended study, we expanded the search by including publications for conference proceedings and literature published in the past 16 years (2006–2022) on the acceptance of SWVIs' digital services and products, which are accessible through JAWS. We also excluded articles that reported on the education of SWVI primary, high schools, and universities. We had to search and find "JAWS" on all articles which used "assistive technologies" and excluded all those that did not specify the type of assistive technology.

3 Background and Theoretical Examination of Jaws as an Assistive Computer Application Software

In the past decade, there has been a slight interest in understanding educators' and students' perceptions of using ACAS for teaching and learning purposes within a few universities and schools [8, 24, 25]. The approach has been on understanding either students' or educators' feelings about the general use of JAWS and other related screen readers in teaching and learning of SWVI under the umbrella of assistive technologies. However, there have been attempts to comprehend the factors influencing the acceptance of ACAS [4, 5]. Unfortunately, their attention was only on

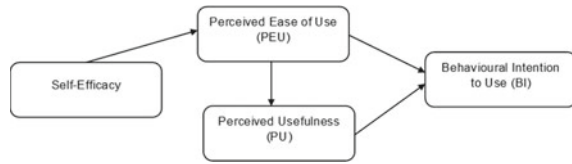
Table 1 Surveyed studies on implications and determinants of TVET educators’ and SWVIs’ acceptance of JAWS

Guiding principles	Sources	Theory used	Application and influencing factors discovered	Participants and country
Perceived ease of use	[26]	TAM	Web accessibility: perceived ease of use, perceived usefulness	Students with visual impairments. The country was not specified
Perceived usefulness	[20]	UTAUT	Web accessibility: accessibility, vision impairment level	People with and without visual in the United States
Self-efficacy	[17]	TAM	Audio and music websites: accessibility as well as its ease of use and usefulness, convenience	People with and without visual. The country was not specified
JAWS	[4]	TAM	Assistive technologies: facilitating condition, perceived ease of use, computer self-efficacy, result demonstrability, perceived usefulness	Special education teachers for the blind and the deaf in the United States
Students with visual impairments	[27]	TAM	Information communication technology: perceived usefulness and perceived easiness	Persons with visual impairments in Bangladesh
Educators of students with visual impairments	[18]	UTAUT	Website accessibility: perceived convenience, perceived reliability, accessibility, vision impairment level	Students with visual impairments in Thailand
Computer practice	[5]	TAM	Technologies: positive attitudes, self-efficacy, time, access	Special education teachers in the United Arab Emirates
TVET colleges	[19]	UTAUT	Mobile applications: performance expectancy, attitude functionality, accessibility	People with visual impairments in USA, India, Australia, Africa, and Europe

educators’ acceptance, not students. To the researcher’s knowledge, there have not been studies examining factors affecting SWVIs’ acceptance of ACAS.

Additionally, it is worth mentioning that it was more than two decades ago when Djasabi and Tullis [26] raised a concern of less attention given to the use of technology acceptance theories in including acceptance of systems for users with visual impairments. A similar lament was in Siyam’s [5] work due to the non-availability of literature on the acceptance of assistive systems used in special education through the guidance of behavioural theories. Till today, the scarcity persists. As such, there is a slow-paced growth body of research. For instance, [27] enjoyed the guidance of Technology Acceptance Model (TAM) to determine the acceptance of information

Fig. 1 Behavioural intention and determinants



technology by people with visual impairments; and [5] in exploring factors impacting special education teachers' acceptance and actual use of technology. However, other noted literature has been on the use of Unified Theory of Acceptance and Use of Technology (UTAUT) on website acceptance by people with visual impairments [17, 18, 20].

Noticeably, little evidence exists relating to the acceptance of JAWS as an ACAS. In other words, the acceptance of JAWS amongst SWVI and educators remains unexplored, thus unknown, particularly in South African TVETs. This dearth of literature related to users' acceptance of JAWS warrants a need for the extension of TAM in adding literature to technology acceptance studies. Hence the implications is that behavioural intention may be rooted in perceived ease of use, usefulness as well as self-efficacy as demonstrated in Fig. 1.

Perceived Usefulness, And Self-Efficacy on TVETs' Acceptance of Jaws as An ACAS adapted from The Acceptance Model of [28, 29].

Although behavioural theories such as the theory of planned behaviour (TPB) [28], the theory of reasoned action (TRA) [29] the technology acceptance model (TAM) [30], and the unified theory of acceptance and use of technology (UTAUT) [31] are commonly used in educational settings to expand user's understanding and acceptance of information technology. Thus far, the literature review for the current study has observed that the few acceptance studies for the applications used by people with visual impairments tried to use TAM and UTAUT. However, the results have left the current study inconclusive because of their simplicity and generalisation of the "assistive technologies" or "screen readers."

3.1 Effects of Perceived Ease of Use on TVET Educators' and SWVIs Acceptance of Jaws in Teaching and Learning of Computer Practice

Perceived ease of use is defined as "the degree to which a person believes that using a particular system (ACAS) would be free of effort" [30, p. 320]. Using two behavioural theories (TAM and UTAUT) it was noteworthy to understand acceptance of such ACAS among people with visual impairments. Some of the items for measuring the products' ease of use included but are not limited to the ease of keypad navigation, the ease of access to the required information via a screen reader, the ease of content

navigation, and the ease of reading graphical content, the ease of executing the commands with less cognitive efforts.

Fluency in the above-listed actions is equally important as the basics of JAWS navigation in computer literacy. For educators, teaching computer-related content to SWVI at an allocated period in class can be a tedious exercise leading to SWVI being passive receivers due to their learning nature [32, 33]. SWVI, on the other hand, are expected to recall a series of unseen keyboard commands in navigating their learning content to be interactive and participative in the computer practice lesson [34]. TVET colleges are still finding their footing with the inclusion of SWVI, and so is the use of JAWS in their classrooms. Also, there is a lament about ACAS proficiencies scarcity among educators and SWVI due to limited to no preparation or training on their usage. The expectation for TVET computer practice SWVI and educators to fluently use JAWS without training would be worst if the system is not user-friendly. When computer practice SWVI and educators from TVET colleges find it effortless to use JAWS, their perceived ease of use and intention to use it will increase. However, the studies of [8] and [22] did not use any of the behavioural theories to understand educators' perceptions of the adoption of assistive technologies. Their findings regarding the relevance of previously acquired computer proficiencies or even special education professional qualifications were contradictory. Lamond and Cunningham [8] found that knowledge and skills to use an assistive technology predicted perceived ease of use. Atanga et al. [22], on the other hand, discovered that educators might have the technology and special education professional qualifications and computer skills; however, that had no link to their readiness for technology acceptance.

Nonetheless, in the context of special education, two studies in the United States used TAM to investigate how determinants educators' acceptance, perceived ease of use, perceived usefulness, computer self-efficacy, and other constructs have on educators' behavioural intentions to adopt the use of assistive technologies in special education [4, 5]. In agreement, the findings for both studies were that perceived ease of use significantly affected educators' computer self-efficacy, subsequently positively affecting their behavioural intention to use the technology. Nam, Bahn and Lee [4], however, attached self-efficacy to a demonstration factor that resonates with prior findings of [8]. The conclusion is that there is a positive correlation between perceived ease of use and self-efficacy; thus, perceived ease of use is a fundamental determinant of special educators' use of assistive technologies.

3.1.1 Potential Implication 1

The implication is that accepting JAWS as an ACAS in teaching and learning computer practice in TVET colleges would fundamentally depend on educators' and SWVIs' perception that using JAWS would be free of effort in addition to factors such as self-efficacy.

3.2 Effects of Perceived Usefulness on TVET Educators' and SWVIs Acceptance of Jaws in Teaching and Learning of Computer Practice

Consistent with [30], perceived usefulness and hence users' perception of enhanced job performance due to a particular system's usage is one of the fundamental constructs that tend to characterise users' technology acceptance. In this study, perceived usefulness is the perception that using JAWS as an ACAS would enhance SWVIs' academic performance and learning experience and encourage students' dependence in learning computer practice. Also, using JAWS, educators will better understand their students learning process, consequently adapting their instruction and assessment processes to best meet their students' learning needs in JAWS usage. Furthermore, research suggested that perceived usefulness strongly influences the user's attitude, increasing the user's intention to use the system [30]. Indeed, in a recent study, perceived usefulness was a good predictor for both perceived attitude and behavioural intention to use an assistive system [5]. In another study, the conclusion was that as much as perceived usefulness was a dominant factor affecting the use of assistive systems in special education, demonstrability and self-efficacy significantly influenced perceived usefulness [4].

3.2.1 Potential Implication 2

The deduction is that TVET educators' and SWVIs' use of JAWS in teaching and learning computer practice would also depends significantly on the perceived usefulness. However, the usefulness of JAWS would only be significant provided a provision of demonstration or capacitation that will influence or heighten users' self-efficacy. Consequently, TVET educators' and SWVIs' use of JAWS will be a function of perceived usefulness.

3.3 Effects of Self-Efficacy on TVET Educators' and SWVIs Acceptance of Jaws in Teaching and Learning of Computer Practice

Self-efficacy is from social cognitive theory [35]. In this study's context, self-efficacy is the belief that educators' and SWVIs' have in their ability to use JAWS in teaching and learning given that "AT for the disabled has its unique characteristics that can be fully utilized if an easy and intuitional way of using it is secured by both teachers and students with a disability" [4, p. 368]. According to Schlebusch [36], computer self-efficacy has a strong motivating effect on the user's completion of a computer-related task. Siyam [5] declares self-efficacy as the significant construct in accepting ACAS in special education, and inclusive classrooms are no exception. As much as

inclusive educators are responsible for practicing inclusive practices by integrating ACAS to teach SWVI, it can be overwhelming for them to keep up to date with the ever-developing trends of technologies [37]. Thus Singleton and Neuber [35] highlighted the importance of providing educator training in computer applications.

Thus far, the implication of perceived ease of use, perceived usefulness, and self-efficacy in accepting JAWS as an assistive computer application software has found that self-efficacy has a certain degree of effect on perceived ease of use and perceived usefulness. Thus, self-efficacy mediates perceived ease of use, usefulness, and behavioural intention.

3.3.1 Potential Implication 3

Conclusively, it would be essential to explore further the relationship or correlation between perceived ease of use, perceived usefulness, self-efficacy, accessibility, facilitating conditions, and visual impairment level among SWVI and their educators of TVET colleges in South Africa and determine if there is a significant difference between their acceptance of JAWS.

3.3.2 Potential Implication 4

While plenty of research has examined behavioural factors via behavioural technology acceptance models, all aided in expanding theorists' and practitioners' knowledge in various contexts. There has been a paucity of research relating to JAWS acceptance examination via behavioural technology acceptance models such as TPB, TRA, TPB, TAM, and UTAUT. There is still no evidence regarding extending these models to include other factors related to the teaching and learning context of SWVI, particularly in mainstream TVET colleges. The examination using extended models considering educators' readiness to use JAWS, availability of resources, or facilitating conditions resources in supporting the integration of JAWS, both educators' and SWVIs' JAWS proficiencies thus self-efficacy, different levels of SWVIs' levels of visual impairments, and accessibility of JAWS.

Conclusively, assessment of the behavioural factors including perceived ease of use, perceived usefulness, perceived readiness, perceived resources, facilitating conditions, computer self-efficacy, visual impairments levels, and accessibility, and their effect on TVET college educators' and SWVIs' acceptance of JAWS in teaching and learning still need to be conducted.

4 Limitations

First, due to scarcity in the availability of published peer-reviewed articles, there needed to be more articles that could meet the inclusion criteria. There were no articles specifically relating to our intended study. Due to that, we had to expand our search scope to begin as further as 2006 as opposed to 2021. Second, most articles focused on the education of SWVI in primary and high schools and very few in universities, not TVET colleges. Third, behavioral theories such as Technology Acceptance Model have yet to be applied in JAWS-related studies for educational purposes, especially for TVET.

We recommend that future research consider empirical studies involving JAWS usage in TVET education to examine users' acceptance through technology acceptance theories.

5 The Conclusion and Potential Contribution

Although the above examination and four (4) potential implications, the effects of the perceived ease of use, usefulness, and computer self-efficacy on TVET educators' and SWVIs' acceptance of JAWS in teaching and learning computer practice still need to be confirmed, the study will extend more discourses on other possible ways to encourage and support the inclusion and participation of SWVI, particularly the employment of JAWS in TVET colleges in South Africa. The study would hopefully encourage more practical research in JAWS usage and determine educators' and SWVIs' perceived ease of use, perceived usefulness, and self-efficacy on the JAWS in TVET colleges, thus extending technology acceptance theories.

6 Future Work

Future work is required to examine empirically, whether JAWS' adoption in the teaching and learning of computer practice in TVET colleges is dependent on TVET educators' and SWVI's perceived ease of use of JAWS. Whether JAWS' adoption in the teaching and learning of computer practice in TVET colleges is dependent on TVET educators' and SWVI's perceived usefulness of JAWS. Whether JAWS' adoption in teaching and learning computer practice in TVET colleges depends on TVET educators' and SWVI's self-efficacy with JAWS. Lastly, from a theoretical perspective, an extension of behavioural technology acceptance theories (TRA, TPB, TAM, and UTAUT) may improve the scope of JAWS adoption among TVET educators and SWVI.

References

1. K. Manjari, M. Verma, G. Singal, A survey on assistive technology for visually impaired. *Int. Things* **11**(100188), 1–17 (2020)
2. M. Shoaib, I. Hussain, H.T. Mirza, Automatic switching between speech and non-speech: adaptive auditory feedback in desktop assistance for the visually impaired. *Univ. Access Inf. Soc.* **19**(4), 813–823 (2020)
3. M.M. Waqar, M. Aslam, M. Farhan, An intelligent and interactive interface to support symmetrical collaborative educational writing among visually impaired and sighted users. *Symmetry* **11**(2), 238 (2019)
4. C.S. Nam, S. Bahn, R. Lee, Acceptance of assistive technology by special education teachers: A structural equation model approach. *Int. J. Human-Comput. Interaction* **29**(5), 365–377 (2013)
5. N. Siyam, Factors impacting special education teachers' acceptance and actual use of technology. *Educ. Inf. Technol.* **24**(3), 2035–2057 (2019)
6. V. Tinto, Through the eyes of students. *J. College Student Retention. Res. Theory Pract.* **19**(3), 254–269 (2017)
7. R. Tekane, M. Potgieter, Insights from training a blind student in biological sciences. *S. Afr. J. Sci.* **117**(5–7), 1–7 (2021)
8. B. Lamond, T. Cunningham, Understanding teacher perceptions of assistive technology. *J. Spec. Educ. Technol.* **35**(22), 97–108 (2020)
9. A. Alabi, S. Mutula, Digital inclusion for visually impaired students through assistive technologies in academic libraries. *Library Hi Tech News* **37**(2), 14–17 (2022)
10. M. Ahmed, M. Naveed, Information accessibility for visually impaired students. *Pakistan J. Inf. Manag. Libraries* **22**, 16–36 (2021)
11. L.F. da Paixão Silva, A.A. de O. Barbosa, E.R.C.G. Freire, P.C.F. Cardoso, R.S. Durelli, A.P. Freire, Content-based navigation within mathematical formulae on the web for blind users and its impact on expected user effort. in *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion* (2018)
12. J. Matoušek, Z. Krňoul, M. Campr, Z. Zajíc, Z. Hanzlíček, M. Grüber, M. Kocurová, Speech and web-based technology to enhance education for pupils with visual impairment. *J. Multimodal User Interfaces* **14**(2), 219–230 (2020)
13. M. Das, D. Gergle, A.M. Piper, “It doesn’t win you friends” understanding accessibility in collaborative writing for people with vision impairments. *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–26 (2019)
14. R. Ferreira, M.M. Sefotho, A.B. Du Plessis, J. Erwee, A. Heard, H. Mokgolodi, H. Viljoen, Teaching Learners with Visual Impairment, *AOSIS*, p. 420 (2020)
15. A. Vtyurina, A. Fourny, M.R. Morris, L. Findlater, R.W. White, Bridging screen readers and voice assistants for enhanced eyes-free web search, in *The world wide web conference* (2019)
16. G. Kapperman, S.M. Kelly, E. Koster, Using the JAWS screen reader and the focus braille display to read foreign language books downloaded from the bookshare accessible online library. *J. Vis. Impairment Blindness* **112**(4), 415–419 (2018)
17. E.T. Loiacono, S. Djamasbi, T. Kiryazov, Factors that affect visually impaired users' acceptance of audio and music websites. *Int. J. Hum Comput Stud.* **71**(3), 321–334 (2013)
18. P. Sirikitsathian, S. Chaveesuk, C. Sathitwiriya Wong, A conceptual framework for better understanding of factors influencing accessibility to a website and its acceptance by university students with visual impairments, in *In 2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Phuket, Thailand (2017)
19. H. Moon, J. Cheon, J. Lee, D.R. Banda, N.A.P.M. Griffin-Shirley, Factors influencing the intention of persons with visual impairment to adopt mobile applications based on the UTAUT model. *Univ. Access Inf. Soc.*, 1–15 (2020)
20. R. Saqr, A. Bhattacharjee, Web accessibility: factors enabling the visually impaired to using websites, in *Proceedings of the Eighteenth Americas Conference on Information Systems*, Seattle, Washington (2012)

21. P. Engelbrecht, Inclusive education: developments and challenges in South Africa. *Prospects* **49**(3), 219–232 (2020)
22. C. Atanga, B.A. Jones, L.E. Krueger, S. Lu, Teachers of students with learning disabilities: assistive technology knowledge, perceptions, interests, and barriers. *J. Spec. Educ. Technol.* **35**(4), 236–248 (2020)
23. C.M. Baker, L.R. Milne, R.E. Ladner, Understanding the impact of TVIs on technology use and selection by children with visual impairments, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland (2019)
24. E. Al-Zboon, Perceptions of assistive technology by teachers of students with visual impairments in Jordan. *J. Vis. Impairment Blindness* **114**(6), 488–501 (2020)
25. M. Abed, T. Shackelford, Saudi Faculty Perspectives on Accommodations for Students with Visual Impairment. (2020). [Online]
26. S. Djamasbi, T. Tullis, Web accessibility for visually impaired users: extending the technology acceptance model (TAM) work in progress. *Web Accessibility and Users*, pp. 1–5 (2006)
27. M.J.A. Zahid, M.M. Ashraf, B.T. Malik, M.R. Hoque, Information communication technology (ICT) for disabled persons in Bangladesh: Preliminary study of impact/outcome, in *International Working Conference on Transfer and Diffusion of IT*, Berlin (2013)
28. I. Ajzen, The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* **50**(2), 179–211 (1991)
29. I. Ajzen, M. Fishbein, Attitude-behavior relations: a theoretical analysis and review of empirical research. *Psychol. Bull.* **84**(5), 888 (1977)
30. F. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quart.* 319–340 (1989)
31. V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: toward a unified view. *MIS Quart.* 425–478 (2003)
32. S. Amoor, Z. Magaji, Comparative analysis of using computer tutor guide and demonstration methods of teaching on students' skills acquisition in word processing in Nigerian Universities. *KIU J. Soc. Sci.* **3**(1), 251–256 (2017)
33. C.G. Ogbonna, N.E. Ibezim, C.A. Obi, Synchronous versus asynchronous e-learning in teaching word processing: an experimental approach. *S. Afr. J. Educ.* **39**(2), 1–15 (2019)
34. K.J. Singleton, K.S. Neuber, Examining how students with visual impairments navigate accessible documents. *J. Vis. Impairment Blindness* **114**(5), 393–405 (2020)
35. A. Bandura, Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* **84**(2), 191–215 (1977)
36. C.L. Schlebusch, Computer anxiety, computer self-efficacy and attitudes towards the internet of first year students at a South African University of Technology. *Africa Educ. Rev.* **15**(3), 72–90 (2018)
37. C.N. Thomas, K.N. Peeples, M.J. Kennedy, M. Decker, Riding the special education technology wave: policy, obstacles, recommendations, actionable ideas, and resources. *Interv. Sch. Clin.* **54**(5), 295–303 (2019)

Experimenting with Polymorphic Creativity Support Tools to Support Innovation in Participatory Ideation



Muhammad Mustafa Hassan , Imran Arshad Choudhry, Markku Tukiainen, and Adnan N. Qureshi

Abstract The importance of creativity support tools for designers has been recognized in human computer interaction community for some time. However, there is limited research on tools specifically targeted on novice end-user designers in participatory design and innovation contexts. The creativity support needs of such designers are atypical, as well as multifold, often not fully addressed by ordinary tools built for professional designers. To cater to the full spectrum of these requirements such as domain, design, feasibility and scope knowledge, as well as creativity and divergence triggers, the authors propose blending multiple tool forms into a single polymorphic tool. Each part of the polymorphic tool shall base upon a subset of user requirements identified via end-user population research and user modelling. The authors tested the proposition in a series of participatory ideation and participatory evaluation workshops. First, the authors developed three tools, two polymorphic and one regular, and applied each tool to a distinct user group in a separate participatory ideation workshop. The workshops had 30, 23, and 26 participants respectively, and the creative output of each group was collected. A series of participatory evaluation workshops with 58 distinct participants followed who evaluated creative output. The analysis of evaluation data showed statistical superiority of polymorphic tools over regular tool. Based on these findings, the authors recommend developing polymorphic creativity support tools tailored to the needs of respective end-user population when working in participatory design and innovation context.

Keywords Design aids · Creativity support tools · Creative ideation · Innovation ideation · Participatory design and participatory innovation

M. M. Hassan (✉) · M. Tukiainen
University of Eastern Finland, 80101 Joensuu, Finland
e-mail: mustafa.hassan@ucp.edu.pk; musth@uef.fi

M. M. Hassan · I. A. Choudhry · A. N. Qureshi
University of Central Punjab, Lahore 54000, Pakistan

1 Introduction

Participatory Design (PD) and innovation are coming together. Their marriage has given birth to Participatory INnovation (PIN). Ranging from Healthcare to Education, and industry to the government, the role of HCI, in general, is rapidly increasing in the design and development of innovative systems. As Schneiderman notes [1], the HCI community is now challenged with brining better Creativity Support Tools (CST). Nonetheless, this challenge is even harder in PIN contexts where the problem of limited creativity worsens due to several reasons.

Contrary to typical PD context, the PIN end-users may not have enough domain/task knowledge when ideating. The insufficiency of domain knowledge affects their confidence negatively, inculcating the feel that they are not creative [2]. Further, the lack of design expertise hinders expression even if good designs have been formulated in mind [3]. Both deficiencies contribute in further narrowing of the solution search space [4]. Moreover, naïve designers may extend beyond the desirability and feasibility scope [5, 6]. It is thus desirable to build and employ such design aids¹ that provide basic design and feasibility knowledge, trigger divergent thinking as well as foster creativity in naïve designers of innovations [7, 8]. Nonetheless, covering all these requirements in one typical design aid is not possible. To solve this problem, the authors introduce polymorphic design aids.

1.1 *Polymorphic Creativity Support Tools*

A number of CSTs (or design aids, design inspirations, creativity triggers etc.) have been reported in cognition, psychology, innovation, engineering, design, and related fields. Design aids have several attributes of interest, including level of abstraction, structure, relation to the problem domain, and bias factor. Each attribute influences observers' entry into the solution search space via some mechanism. In general, abstract and unstructured tools (like storyboards) ignite broader divergence but provide little to no design/domain knowledge [9]. Contrastingly, concrete and structured sources (e.g., exploratory prototypes) bring domain and design knowledge but do not spawn divergence. The lack of abstraction translates into guided entry points for observers resulting in narrower search space [9]. Metaphors and analogies (e.g., interactive illustrations and design games) trigger solution search but induce bias and may result in solutions not equally applicable to target domain [5].

For PD/PIN contexts, a careful balance of tool attributes is important to supplement participants' design/domain knowledge, foster creativity and divergence, minimize bias, and to keep the design activity in scope and feasibility boundary [5]. Nonetheless, such a blend is difficult to achieve with one typical design aid. The authors, hence, break the stereotypical design approach and mix several types

¹ In this work, the terms **design aids** and **creativity support tools** are used interchangeably.

together into one [polymorphic] tool to realize the benefits of all. However, such approach needs careful design to mitigate the risks associated with each type.

1.2 Background

The current work is a part of a larger project aiming at developing an ideation framework for innovation in participatory contexts. The authors use a case tool for their experiments, namely Jeliot Mobile (JM) [10]. Jeliot Mobile belongs to Jeliot family of Software Visualization (SV) tools. It is aimed at teaching Java through animations on a mobile platform in a context-free socio-constructivist way. The project primarily employs design-based research methodology to create knowledge in HCI, PD and PIN, as well as to materialize the concept of this tool.

In this part, the authors report on designing, developing, deploying and evaluating polymorphic tools. These tools were applied to the participants in three distinct ideation workshops in which they generated design ideas for JM. The design output of these workshops was used in the second phase of the research, in which a distinct group of participants evaluated the design output for creativity.

The rest of the paper is organized in 5 further sections. The next section provides a brief discussion on notable related works. The authors detail their research methodology in Sect. 3, with analysis and results presented in Sect. 4. Discussion and Implications follow in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related Work

The authors find several studies comparing different forms of design aids to promote participants' creativity in PIN/PD contexts. For example, Kwiatkowska et al. [9] compared structured and unstructured cards used as design aids. They found that unstructured cards worked better due to their abstraction, did not lead participants' creative process, and hence resulted in a broader creative output. Similarly, Lopez et al. [11] reported on the differences in impact of visual and sentential stimuli given to design teams. They found that one method was not wholly superior to the other. It was instead the situation and the objectives of the design session that defined which stimulus shall be used.

Goldschmidt and Sever [4] experimented with textual stimuli to foster creativity in novice designers in controlled laboratory settings. They found that novice designer who were given textual stimuli performed better than others. Their assessment criteria included evaluations of design ideas for originality and practicality. In another study, Goldschmidt and Smolkov [12] experimented with visual stimuli to promote creativity in design teams. The visual stimuli varied in intensity for three independent designer groups, whose design outputs were later judged for creativity by experts.

They found that the variation in visual stimuli was accounted for variance in their creative output.

Wilson et al. [13] studied the impact of different kind of examples given to designer during a design activity. They found that biological examples (surface dissimilar but structurally similar) worked better than human-engineered surface similar examples. They performed the experiment with engineering students from an undergraduate program and evaluated the resulting design ideation for novelty and variety. Nonetheless, all these studies presented interesting facts. However, the design aids reported were all based on single CST, which is in contrast to this report, where authors experiment with multi-part polymorphic CSTs.

3 Research Methodology

The authors conducted the study in three stages, namely Participatory Ideation, Participatory Assessment, and analysis stage. In participatory ideation stage, three groups of end-users treated with different design aids participated to co-ideate the case tool. In participatory assessment stage, another distinct end-user group evaluated creative ideation generated in participatory ideation stage. Finally, in Analysis stage, authors analyzed how much variance can be accounted for varying design aids.

3.1 *Participants*

The participants of all stages belonged to real users' population of the tool under design, complying with the notion of PIN/PD. All the groups were uniform in all characteristics of interests. They belonged to the same program at the faculty of Information Technology of University of Central Punjab Pakistan, studied same courses, learned same skills, and shared a similar background.

The participatory ideation group—12 females and 68 males—was divided into three treatment groups, namely PA, PB and PC with 30, 23, and 26 participants respectively. A further division in treatment groups followed to form working groups of 3 to 6 participants each. The participatory assessment group—8 females and 50 males—had 15 working groups of 3 to 5 participants, with one exception of a singleton. The facilitators allowed him to work alone, however dropped his evaluation from analysis. In both cases, the participants were allowed group formation by themselves to facilitate a comfortable in-group working environment.

3.2 Procedure

There were three stages, (1) a quasi-experimental Participatory Ideation Stage to apply various interventions and collect creative output, (2) a Participatory Assessment Stage to evaluate creativity in the work of various groups, and (3) an Analysis Stage.

3.2.1 Participatory Ideation Stage

The participatory ideation stage comprised of three ideation workshops, intervened with different design aids, but proceeding in a similar fashion. The design aids used were (1) a blend of exploratory prototype and use scenarios, (2) storyboards, and (3) a blend of exploratory prototype and storyboards, for workshops PA, PB and PC respectively. Figures 1 and 2. provide snapshots of exploratory prototype and storyboards respectively.

The authors used the framework of [2] to structure their participatory protocols. The framework divides a PD session into four broader activities, namely Probing (understanding participants), Priming (preparing participants), Understanding (understating impact of interventions) and Generating (ideating, designing, or evaluating etc.). Table 1 details the protocol used identically at workshops PA, PB and PC.

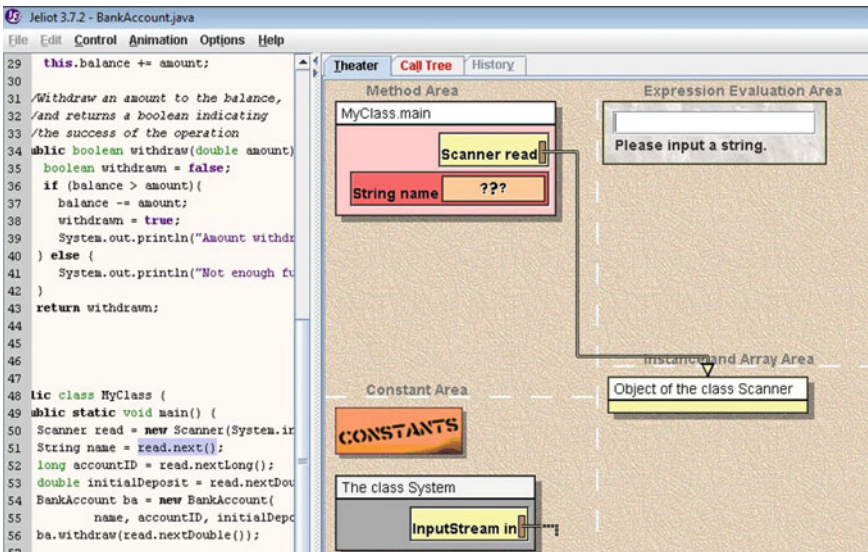


Fig. 1 A snapshot of exploratory prototype used at participatory ideation workshops

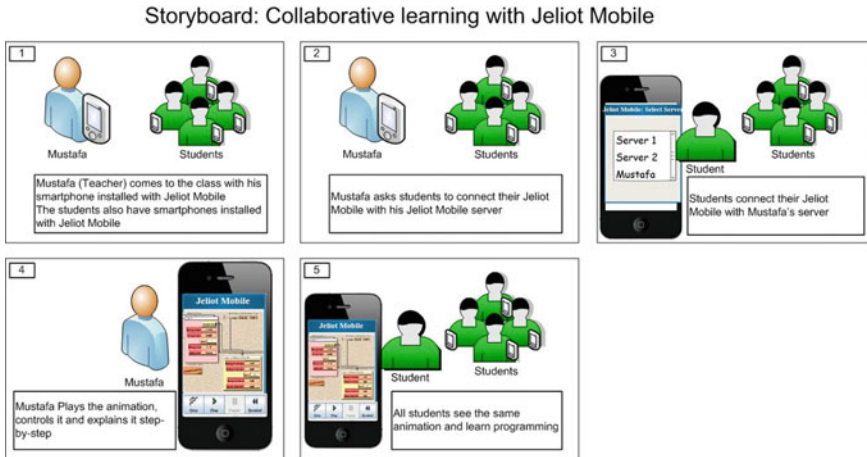


Fig. 2 One of the storyboards (Collaborative Learning) used at participatory ideation workshops

Table 1 Participatory protocol to be used at participatory ideation workshops

Activity	Procedure	Minutes
Probing: starting the workshop	Introduction of participants/ facilitators	10
	Briefing about upcoming activity and jeliot	
Priming: intervention	Applying design aid interventions	30
	Break	20
Understanding: assessment of interventions' impact	Questioning participants	10
	Receiving participants' questions	15
Generating: To generate innovative design ideas	Brainstorming	30
	Design ideas generation	85
	Total duration	200

3.2.2 Participatory Assessment Stage

The participatory assessment stage comprised of two workshops, E1 and E2, both attended by all 58 participants of the assessment group. The workshop E1 trained the participants individually—groups were not formed yet—for upcoming assessment activity in workshop E2. Table 2 details the protocol used at workshop E1. The workshop E2's time was largely used for assessments. The workshop opened with introduction and subgroup formation, followed by priming and understanding, and assessment. Table 3 shows the protocol used at workshop E2.

The assessment activity took two sessions, first with 145 min, and then with 90 min. The activity was divided because the facilitators felt that the participants were

Table 2 Participatory protocol to be used at participatory assessment workshop E1

Activity	Procedure	Minutes
Probing: to assess participants' current state of knowledge and expertise	Introduction	10
	Facilitators probe participants on participatory design, participatory innovation, and design tools	
Priming: to prepare the participants for the workshop by employing interventions	Facilitators explain the purpose of activity framing it in participatory assessment and design	15
	A brief situated design activity and the case tool (JM) introduction	40
Understanding: to assess the impact of interventions	Q/A session to assess participants understanding and inviting participants to resolve ambiguities	10
	Total duration	75

Table 3 Participatory protocol to be used at participatory assessment workshop E2

Activity	Procedure	Minutes
Probing: initializing and familiarizing	Introduction of participants, facilitators, and the activity. Forming groups to work	10
Priming: preparing for participatory assessment	Providing JM Design ideas and discussing assessment	15
Understanding: assessment of interventions' impact	Q/A session to assess participants understanding, and inviting participants to resolve ambiguities	10
Assessing: assessing JM design ideas produced earlier in PIN	Brainstorming	145 + 90
	Assessing and documenting jeliot mobile conception for innovation	
	Submitting the results of assessment	
	Total	270

not able to concentrate and work any longer. All 15 subgroups, assessed all design ideas (of PA, PB, and PC) for novelty and meaningfulness on a Likert type scale, as illustrated in Fig. 3. The questionnaire was an electronic document, and the participants used their computers to fill in the data. They returned filled-in questionnaire to the facilitators once the activity was completed.

3.2.3 Analysis Stage: Variables and Measurements

The independent variable was categorical. It coded three interventions used at participatory ideation workshops as *PA*, *PB*, and *PC* (reflecting the name of each workshop for easy traceability). The dependent variable, namely Absolute Innovation Score (*AIS*), was created by summation of the ratings given to each design idea by 13

Fig. 3 Snapshot of the questionnaire used for idea assessment

assessment groups on a 5-anchors unipolar Likert type response format. The response anchors included extremely meaningful (*em*), highly meaningful (*hm*), Meaningful (*m*), somewhat meaningful (*sm*), and not meaningful (*nm*). Responses of two assessment groups were not included into the analysis. One group was excluded because it had only one participant. The other excluded groups’ responses were found to be inconsistent.

The authors did not include negative anchor points, because they believed that a product/idea could not be negatively innovative/creative with respect to meaningfulness to the target community, though it can have zero or no meaningfulness in the opinion of a user. The measure of meaningfulness, if exists, varies in positive degrees, as in the case of most of the psychological measurements [14].

Once, all 15 assessment groups responded on each Likert type item, the authors calculated *AIS* of each design idea (Likert-type item) using the formula given in (1). Consider idea *A2-4* (fourth idea generated by productive Subgroup *A2*), for example. Two evaluator groups thought it was not meaningful at all, one took it as somewhat meaningful, two group thought it was meaningful, five groups rated it to be highly meaningful, and only one group believed it was extremely meaningful. Using (1), the said idea received an *AIS* of 28, as demonstrated in Table 4.

$$AIS_i = \sum_{j=1}^4 (T(p)_j \times nr_j) \tag{1}$$

where, AIS_i = Absolute Innovation Score of design idea i ,
 $T(P)_j$ = scale weight of anchor point j , and
 nr_j = no of responses anchored to point j

Table 4 The Score received by, and the sum calculated for an example Likert type item representing idea *A2-4*

Id	em	hm	m	sm	nm	<i>AIS</i>
A2-4	1	5	4	1	2	$(4 \times 1) + (3 \times 5) + (2 \times 4) + (1 \times 1) + (0 \times 2) = 28$

The *AIS* produced three distributions, *PA*, *PB*, and *PC* each representing the respective productive group. The authors then compared distributions means for between-group differences with an omnibus F-Test, i.e., with a one-way parametric analysis of variance (ANOVA).

4 Analysis and Results

The descriptive statistics showed that the distributions were slightly kurtotic and skewed. The assumption of normality was not tenable in some distributions ($n_{PA} = 67$, $n_{PB} = 55$, $n_{PC} = 50$). *PB* was skewed (skewness = -0.72 , $SE = 0.32$, $z = -2.22$) and kurtotic (kurtosis = 0.98 , $SE = 0.63$, $z = 1.54$). *PC* slightly approached non-normality [(skewness = -0.31 , $SE = 0.34$, $z = -0.93$), (kurtosis = -1.02 , $SE = 0.66$, $z = -1.53$)]. *PA* was approximately normal [(skewness = -0.17 , $SE = 0.29$, $z = -0.59$), (kurtosis = 0.07 , $SE = 0.58$, $z = 0.12$)]. The assumption of homoscedasticity was tenable, confirmed with Levene's test ($p = 0.23$). The smallest to largest variance ratio was 1.33—[*PA*: ($M = 28.79$, $SD = 6.94$, $SE = 0.85$), *PB*: ($M = 26.44$, $SD = 6.81$, $SE = 0.92$), *PC*: ($M = 31.18$, $SD = 7.84$, $SE = 1.11$)]. ANOVA results showed a statistically significant difference between groups [$F(2, 169) = 5.74$, $p = 0.004$, $\alpha = 0.05$] with moderate effect size ($\eta^2 = 0.064$, $\omega^2 = 0.052$).

Since the assumption of normality was not tenable, the authors crosschecked ANOVA with Kruskal–Wallis H-test. The H-test produced similar results and showed a statistically significant difference between treatment groups [$\chi^2(2) = 10.43$, $p = 0.005$, $R_{PA} = 87.05$, $R_{PB} = 71.22$, $R_{PC} = 102.57$].

The pairwise comparison of effect size was analyzed with Common Language (CL) effect size statistic. The CL calculations showed the probabilistic inferiority of the group *PB* to other groups. There were 64% chances that a randomly drawn sample from *PB* had lower *AIS* than a randomly drawn sample from (*PA U PC*). The group *PC* performed generally better than other groups with 64% probability of observing superior *AIS* in *PC* than in (*PA U PB*). The pairwise probabilities of superiority were $p(PC > PA) = 59\%$, $p(PC > PB) = 68\%$, $p(PA > PB) = 60\%$. These CL results corresponded consistently with η^2 , as checked against the correspondence of η^2 and CL provided by [15].

The consistency of parametric and non-parametric statistics provided evidence that the model developed for study was valid. Both parametric and non-parametric analysis failed to accept the null hypothesis that the groups were equivalent with respect to *AIS*, i.e., failed to accept that all three design aids influenced participants' creativity/innovation in a similar way. This implied that some design aids worked better than others with respect to participants' creativity/innovation.

The magnitude of positive influence was speculatively moderate ($\eta^2 = 0.064$). The same medium effect was exhibited by ω^2 , however with a smaller number since it adjusted for any bias effect in population. Similarly, all pairwise comparisons between groups aggregated into a 62% stochastic heterogeneity. That implied

a medium effect size explained by the choice of a particular design aid [16]. Further analysis revealed that the group with design aid composed of exploratory prototype and storyboards of the future system (*PC*) performed the best, the group with exploratory prototype and use scenarios (*PB*) stood in middle, and the group having storyboards (*PB*) performed the worst.

5 Discussions and Implications

Every single design aid has its own dynamics. Some tools are abstract, and hence promote a wider search space. However, the resulting ideas may be far from solutions, infeasible to implement, or simply out of the scope of the target system. Other tools are best at defining system boundaries, however guide the participants' creativity, and hence keep the person's creative process bound to a narrower solution space. The perfection of a design aid is to keep the bias minimum and trigger the creativity maximum. Using a single instrument, this balance seems difficult to achieve. For example, storyboards/scenarios provide good triggers to initiate broader thinking process, however unable to inform participants of the system boundaries/limits. Similarly, interactive illustrations/exploratory prototypes inform participants of the system concept, however, unable to initiate a divergent process sometimes. Moreover, not all tools are created equal, for example, both use scenarios and storyboard have intrinsic abstraction, but their impact was seemingly different on the participants.

The abstraction of storyboards required participants to spend more time in discovery than exploratory prototype. The authors understand that it was because they had to fill in missing gaps by themselves to create a big picture of the target domain. However, the boundaries of the target domain were still unclear, because everyone relied upon their own interpretation. Nonetheless, once primed, the participants asked fewer questions as compared to other groups. They were able to diverge and produce design ideas that were considered innovative both by the assessment group and by the authors. However, this group sometimes generated ideas which seemed too futuristic, infeasible or unrealistic. The authors attributed this diversity to the abstraction of the storyboards.

The exploratory prototype worked better for delivering the domain knowledge. The participants were able to understand about the problem domain. However, it brought strong and concrete visual cues binding the participants to a limited search space. The facilitators observed during the workshops that the participants were not sure what they were required to do or how they were expected to generate design ideas. Once exposed to exploratory prototypes, they mostly tried to correct any perceived problems, or considered ideas similar to that presented. This behavior showed that their thought process did not diverge.

Since, exploratory prototype did not help much in triggering divergent thinking process, the groups PA and PC were given use scenarios and storyboards respectively as the second part of their design aid. The facilitators did not note any differences in the behavior of the two groups, like clarity/confusion, questions asked etc., i.e., they

developed a similar level of understanding. However, the creative thinking process was influenced differently as exhibited by the difference in output of both groups. The authors attribute it to the medium of presentation. Nonetheless, the storyboards and the use scenarios offered the same story, however with different medium. The storyboards were pictorial and the use scenarios were verbal.

6 Conclusions and Future Work

Creativity support tools are used as design aids to trigger creativity and divergence in designers of innovations. From storyboards to sketches, and from stories to prototypes, they come in various forms, all focused on triggering divergence and creativity. When designing with professionals, the participants usually come with some amount of domain knowledge, adequate amount of design knowledge, and are aware of scope/feasibility boundaries. Hence, they only need creativity and divergence triggers.

The creativity support needs of end-user designers in PD/PIN are different. First, they need supplementation of design and expression knowledge, as well as domain knowledge in some cases. Second, they need clues on scope and feasibility boundaries of the intended innovations. Finally, they need triggers to push start divergent thinking and creativity. This spectrum of requirements is often not addressed by typical CSTs developed to promote creativity.

The authors' findings suggest that developing CSTs for PD/PIN requires detailing and separating user needs through careful user research, and then using this information to develop polymorphic CSTs, where each part addresses a subset of the users' requirements. The findings also highlight the importance of assessing user generated ideas for novelty, meaningfulness, and feasibility. Moreover, the authors emphasize on using a user-centric participatory assessment approach to ensure that the exercise remains user-centric.

Summarizing, the use of polymorphic CSTs tailored to the end-user population's needs is a promising strategy for innovation ideation in PD/PIN. It shall be combined with a round of participatory assessment to evaluate innovative ideas for novelty, meaningfulness, and feasibility. The finding is not domain specific, and can be generalized to any context involving end-user designers.

In the future, the authors want to develop more polymorphic CSTs and run comparisons via experimentation. It will help in finding the best blend of polymorphic tools as per the requirement of end-users and the current context.

References

1. B. Schneiderman, Creativity support tools: a grand challenge for HCI researchers. in *Engineering the User Interface*, M. Redondo, C. Bravo and M. Ortega, Eds., London, Springer (2009)
2. E.B.-N. Sanders, E. Brandt, T. Binder, A framework for organizing the tools and techniques of participatory design, in *Proc PDC'10*, Sydney, Australia (2010)
3. A. Pommeranz, U. Ulgen, C.M. Jonker, Exploration of facilitation, materials and group composition in participatory design sessions, in *Proc ECCE'12*, Edinburgh, Scotland (2012)
4. G. Goldschmidt, A.L. Sever, Inspiring design ideas with texts. *Des. Stud.* **32**(2), 139–155 (2011)
5. M.M. Biskjaer, P. Dalsgaard, K. Halskov, Creativity methods in interaction design, in *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design (DESIRE'10)* (2010)
6. M.M. Hassan, A.N. Qureshi, A. Moreno, M. Tukiainen, Participatory refinement of participatory outcomes: students iterating over the design of an interactive mobile learning application, in *In 2017 International Conference on Learning and Teaching in Computing and Engineering (LaTICE)* (2017)
7. B.A. Hennessey, T.M. Amabile, Creativity. *Annu. Rev. Psychol.* **61**, 569–598 (2010)
8. G. Pahl, W. Beitz, J. Feldhusen, K.-H. Grote, *Engineering Design: A Systematic Approach*, 3rd edn. (Springer, London, 2007)
9. J. Kwiatkowska, A. Szostek, D. Lamas, (Un)structured sources of inspiration: comparing the effects of game-like cards and design cards on creativity in co-design process, in *Proc PDC'14*, Windhoek, Namibia (2014)
10. M.M. Hassan, A. Moreno, E. Sutinen, A. Aziz, On the participatory design of Jeliot mobile: towards a socio-constructivist mlearning tool, in *2015 International Conference on Learning and Teaching in Computing and Engineering (LaTiCE)*, Taipei, Taiwan (2015)
11. B. López-Mesa, E. Mulet, R. Vidal, G. Thompson, Effects of additional stimuli on idea-finding in design teams. *J. Eng. Des.* **22**(1), 31–54 (2011)
12. G. Goldschmidt, M. Smolkov, Variances in the impact of visual stimuli on design problem solving performance. *Des. Stud.* **27**(5), 549–569 (2006)
13. J.O. Wilson, D. Rosen, B.A. Nelson, J. Yen, The effects of biological examples in idea generation. *Des. Stud.* **31**(2), 169–186 (2010)
14. S. Stevens, On the theory of scales of measurement. *Science* **103**(2684), 677–680 (1946)
15. C.O. Fritz, P.E. Morris, J.J. Richler, Effect size estimates: current use, calculations, and interpretation. *J. Exp. Psychol. General* (2011)
16. A. Vargha, H.D. Delaney, A critique and improvement of the CL common language effect size statistics of McGraw and wong. *J Educ Behav Stat* **25**(2), 101–132 (2000)