Princess Sumaya University for Technology
جامعة الأميرة سميّة للتكنولوجيا

صندوق دعم البحث العلمي والابتكار
Scientific Research and Innovation Support Fund

Jordanian Journal of Computers and Information Technology

September 2024 VOLUME 10 NUMBER 03
ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

JJCIT

www.jjcit.org jjcit@psut.edu.jo

An International Peer-Reviewed Scientific Journal Financed by the Scientific Research and Innovation Support Fund

Jordanian Journal of Computers and Information Technology (JJCIT)

The Jordanian Journal of Computers and Information Technology (JJCIT) is an international journal that publishes original, high-quality and cutting edge research papers on all aspects and technologies in ICT fields.

JJCIT is hosted and published by Princess Sumaya University for Technology (PSUT) and supported by the Scientific Research Support Fund in Jordan. Researchers have the right to read, print, distribute, search, download, copy or link to the full text of articles. JJCIT permits reproduction as long as the source is acknowledged.

AIMS AND SCOPE

The JJCIT aims to publish the most current developments in the form of original articles as well as review articles in all areas of Telecommunications, Computer Engineering and Information Technology and make them available to researchers worldwide. The JJCIT focuses on topics including, but not limited to: Computer Engineering & Communication Networks, Computer Science & Information Systems and Information Technology and Applications.

INDEXING

JJCIT is indexed in:

EDITORIAL BOARD SUPPORT TEAM

JJCIT ADDRESS

WEBSITE: www.jjcit.org
EMAIL: jjcit@psut.edu.jo
ADDRESS: Princess Sumaya University for Technology, Khalil Saket Street, Al-Jubaiha
B.O. BOX: 1438 Amman 11941 Jordan
TELEPHONE: +962-6-5359949
FAX: +962-6-7295534

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

CSGA: A DUAL POPULATION GENETIC ALGORITHM BASED ON MEXICAN CAVEFISH GENETIC DIVERSITY

Esra'a Alkafaween[1], Ahmad Hassanat[1], Ehab Essa[2] and Samir Elmougy[2]

(Received: 11-Feb.-2024, Revised: 5-Apr.-2024, Accepted: 27-Apr.-2024)

ABSTRACT

Genetic algorithms (GAs) are search algorithms based on population genetics and natural-selection concepts. Maintaining population variety in GAs is critical for ensuring global exploration and mitigating the risks of premature convergence. Rapid convergence to local optima is one such challenge in the application of genetic algorithms. To address this issue, we provide Cave-Surface GA (CSGA), an alternative method based on the Dual Population GA and inspired by the genetic variety observed in Mexican cavefish. Through inter-population cross-breeding, CSGA increases diversity via a secondary population (cave population) and facilitates the exchange of information between populations, effectively counteracting premature convergence. Several experiments are carried out utilizing benchmark instances of the Traveling Salesman Problem (TSP) obtained from TSPLIB, a well-known TSP problem library. Our experimental results over many TSP instances show that CSGA outperforms both classic GAs and other GAs that use diversity-preservation techniques, such as Multipopulation GA (MPGA). CSGA has the potential to give promising solutions to challenging optimization issues like TSP.

KEYWORDS

Cave-surface GA, Diversity, GAs, Premature convergence.

1. INTRODUCTION

GAs are one of the most well-known types of evolutionary algorithms [44]. The GA is based on the principles of biological evolution, which were first devised by John Holland [25] at the University of Michigan in the 1970s [19]. GA was created to investigate processes in natural systems and to construct artificial systems that preserve the adaptability and resilience of natural systems [18], [37].

The GA is regarded as an optimization technique, since it has demonstrated its durability and effectiveness in solving many challenges, such as: image recognition, combinatorial optimization, machine learning, computer networks, neural networks, …etc. [29]. Many combinatorial optimization problems in engineering and sciences have been effectively handled using GAs. There are many recent examples of the use of GA for combinatorial optimization. Examples include: TSP [30],[7] where the TSP is widely considered as a standard testbed of numerous combinatorial optimization strategies [8], routing problems [10], location problems [51] and scheduling problems [47][16].

The GA employs a set of solutions represented by a unique encoding. During the GA-implementation process, each solution or individual is assigned a fitness value that serves as a measure of the GA's performance. Each individual's fitness is directly related to the objective function of the optimization issue under consideration. The present individual population can be adjusted to form a new population utilizing three operations described by Holland: selection, crossover and mutation operators [38], [40],[22] and [5]. The selection operator selects which chromosomes in the population are permitted to reproduce, and generally, the chromosomes with the highest fitness are picked to generate more children than the others [6]. Sub-parts of two chosen chromosomes are exchanged by crossover operators [21],[24]. On the other hand, mutation operators randomly alter the allele values at certain chromosomal regions [23], [4] and [27]. The Ątness of the latest recent parent generation serves as an iterative guide for the searches in GAs. Every time we use GAs to solve an optimization problem, thousands of unique solutions are generated and to create offspring, the resulting solutions are assessed and recombined [27].

It is critical to provide a diversified population in order to achieve the best overall solution and

1. E. Alkafaween and A. Hassanat are with Computer Science Department, Mutah University, Al-Karak, Jordan. Emails: esra_ok@mutah.edu.jo and hasanat@mutah.edu.jo
2. E. Essa and S. Elmougy are with Department of Computer Science, Mansoura University, Mansoura, Egypt. Emails: ehab_essa@mans.edu.eg and mougy@mans.edu.eg

extensively explore the search space. According to the study of Osuna et al. [34], the amount of population diversity is a crucial factor contributing to premature convergence. In a GA, premature convergence occurs when a few highly ranked individuals dominate the population, forcing it to converge to a local optimum rather than the global optimum. According to the studies [33, 36], the main cause of premature convergence is a decline in population variety. This happens when the GAŠs population reaches a point where the genetic operators are unable to produce offspring who outperform their parents. It is crucial to protect population diversity throughout evolution in a GA in order to prevent premature convergence [32] and [31]. for maintaining diversity in GAs and avoiding the risk of premature convergence, numerous previous works used various methods, which include [35]: improvement of the genetic operators (mutation, crossover and selection) ([48], [6]-[4]), dynamic parameter control [42], crowding method [12], MPGAs [45], a multi-objective evolutionary algorithm, dual population GA (DPGA) [36], primal-dual GA [49], …etc.

This research provides additional significant contributions, mostly to the area of enhanced GA performance using a new approach in the form of a dual-population GA inspired by the genetic variety discovered in Mexican cavefish, fittingly termed the 'Cave-Surface Genetic Algorithm' (CSGA). Due to its use of an additional population, CSGA falls under the category of MPGA. The study's specific contributions are:

- Encouraging diversity of the GA: Permitting the insertion of important and contextually relevant features into an individual's chromosome, CSGA allows the natural introduction of genetic variation.
- Mitigation of premature convergence: By keeping a diversified population throughout the evolutionary process, CSGAŠs novel design seeks to prevent early convergence and allow for the exploration of a larger solution space.
- Enhancement of solution quality: By virtue of its distinct characteristics and structures, CSGA shows gains in the quality of solutions produced, advancing the performance of genetic algorithms. These contributions reflect substantial advances in the field, providing useful insights and prospective applications for both practitioners and researchers.

Our methodology will be tested by applying it to instances of the Traveling Salesman Problem (TSP) supplied by the TSPLIB, a well-known resource of TSP problems [39]. To evaluate its performance and effectiveness, we will conduct a comparison analysis, pitting CSGA against the standard GA, MPGA as well as Particle Swarm Optimization (PSO).

The remaining part of this paper is structured as follows: Section 2 goes over similar works. The proposed algorithm is shown in Section 3. Section 4 presents experimental data to demonstrate the efficacy of the proposed approach. Section 5 concludes with a conclusion and discussion of future work.

2. RELATED WORK

Many approaches have emerged in recent years to enhance and sustain population diversity and thus reduce premature convergence. This helps improvement by giving global exploration assistance and getting access to different global and local optima [18].

Du et al. [14] suggested the use of elitism and distance to reduce genetic drift. Elites remain in place. Candidates for selection who are farthest from each elite are also retained to preserve diversification. In their studies, three EAs are used, including a GA, which is called every generation to maintain diversity. The second algorithm, DE/rand/2/bin, is a fundamental Differential Evolution (DE) algorithm. The third EA uses CoBiDE, a cutting-edge DE algorithm.

In [11], Osuna gave a great number of robust experimental and theoretical studies for EA to show how and why diversity plays a crucial role. Different diversity methods have been compared in a number of test functions within the framework of various EAs. The results obtained from the study shed light on how factors and mechanisms related to apparent diversity influence the search behavior of evolutionary algorithms, both in the presence and absence of diversity. These studies particularly point out which diversity strategies work for particular issues and which don't. Most significantly, they describe how to create the best evolutionary algorithms for the issues at hand.

To address the problems of exploration and exploitation, an enhanced GA-based new selection strategy, stairwise selection (SWS), was introduced. Its overall performance was compared to those of many other selection methods by employing 10 well-known benchmark functions across multiple dimensions. Furthermore, the study compared the statistical significance of the proposed SWS. The empirical results, supported by the graphical representation, showed that the SWS outperformed other competing systems in terms of stability, efficiency and durability, as evidenced by the authentication of a performance index [20].

Hassanat et al. [21] proposed two innovative deterministic crossover and mutation rate-control strategies: Dynamic Decreasing of High Mutation/Dynamic Increasing of Low Crossover (DHM/ILC) and Dynamic Increasing of Low Mutation/Dynamic Decreasing of High Crossover (ILM/DHC). These methods are dynamic, allowing for linear changes in both crossover and mutation operator rates as the search advances. Experiments on 10 instances of the Traveling Salesman Problem (TSP) were carried out to assess the efficiency of the suggested techniques. These experiments' results confirmed the efficacy of the proposed techniques.

Shojaedini et al. [41] used an adaptable genetic operator to choose high-fitness individuals as parents while mutating low-fitness ones. During the mutation phase, a training technique was used to gradually learn which gene is the best replacement for the mutant gene. By learning about genes, the suggested technique adaptively balances exploration and exploitation. The algorithm uses this information to enhance the final outcomes during the last iterations.

Hussain and Muhammad [26] presented a new split-ranked selection operator that provided a solid trade-off between exploration and exploitation. The proposed solution solves the fitness-scaling problem by ranking individuals from poorest to strongest depending on the calculated fitness scores. A series of experiments was carried out of some conventional operators and simulation studies using TSPLIB instances.

Inspired by the theory of natural selection, Albadr et al. [3] proposed a novel GA based on natural selection (GABONST), to better control over exploitation and exploration in optimization problems. According to the study, GABONST has outperformed the regular GAs in fifteen different standard test objective functions based on implementation and results. The algorithm's efficacy is ascribed to its capacity to focus on the more promising portions of the search space, which is accomplished by a well-balanced combination of exploration and exploitation.

Koohestani [28] proposed a permutation-based GA for tackling combinatorial optimization issues to improve the effectiveness of permutation-based GAs and to aid in developing high-quality solutions. A new edition of the so-called Partially Mapped Crossover is the main component of this GA. To evaluate the usefulness and efficiency of this crossover operator, two sets of experiments were carried out on popular benchmark problems.

The Population Diversity Controller-GA (PDC-GA) technique was devised as a distinctive feature-selection approach to reduce the search space during the construction of a machine-learning classifier. To effectively manage population diversity during the exploration phase, the PDC-GA combines GA with k-means clustering. When approximately 90% of the solutions become concentrated in a single cluster, an injection approach is employed to redistribute the population, ensuring a controlled level of diversity within the population [2].

A multi-objective binary GA with an adaptive operator-selection mechanism (MOBGA-AOS) was proposed by [48]. MOBGA-AOS employs five crossover operators, each with a unique set of search criteria. Each of them is allocated a probability based on how they perform during the evolution process. The proposed approach was compared against five well-known evolutionary multi-objective algorithms using ten datasets. MOBGA-AOS can remove a significant number of attributes while keeping a low classification error, according to the experimental results. Furthermore, it can handle high-dimensional feature-selection applications.

To solve the difficulties of readily slipping into a local optimum, low solution quality and sluggish convergence speed when solving TSP using GA, a GA incorporating jumping gene and heuristic operators (GA-JGHO) was presented by [50]. This algorithm features several improvements: a bidirectional heuristic crossover operator, enhanced roulette selection, a combination mutation operator and a jumping gene operator to prevent the formation of many similar individuals in the

population. To avoid the development of nimiety identical to those within the population, a unique operator was included. In addition, the local search operator was added to boost exploitation potential.

According to the preceding analysis of the literature, different approaches and algorithms have been proposed to handle distinct issues in their respective sectors. While each of these approaches has made substantial contributions and breakthroughs, it is crucial to remember that none of them is perfect and there is still much opportunity for development. The wide range of proposed algorithms and methodologies emphasizes the importance of the ongoing study and development in this sector. Each method has advantages and disadvantages and it is critical to identify areas where improvements can be made to improve their performances. Subsequently, while the literature study highlights the availability of methods and algorithms, it also underscores the importance of further breakthroughs and improvements. Researchers can help build more effective and more efficient algorithms in the future by addressing the inadequacies of existing approaches.

3. METHODS

Many new concepts and ideas were incorporated into GAs. Following these concepts, we propose a dual population for GAs inspired by cavefish. Cave-dwelling species have offered scientists valuable knowledge regarding the evolutionary modifications of traits in response to distinct environmental and ecological limitations, as highlighted since Darwin's publication of "Origin of Species" [43]. Among such species, the Mexican blind cavefish stands out as a powerful research model due to its well-documented evolutionary lineage, clear ecological context and the presence of independently evolved cave populations. This species provides researchers with an excellent opportunity to investigate the factors contributing to convergent evolution [17]. Many of the cave-derived characteristics of cavefish, such as eye loss, loss of schooling and sleep loss have evolved repeatedly through independent origins and frequently by using various genetic pathways across caves [9] and [15], see Figures 1 and 2. This recurring evolution is a powerful feature of the Mexican blind cavefish system.

Figure 1. Surface fish. Adapted from (Bradic, Martina et al., 2012, with permission) [9].

Figure 2. Cave fish. Adapted from (Bradic, Martina et al., 2012, with permission) [9].

This work is based on the study carried out by Bradic et al. [9] on cavefish, which concluded the following points:

1) Many cavefish often get migrant fish from the surface and researchers continue to note that as one descends further into the cave, the frequency of surface fish in the pools rises.
2) Many fish in caves often accept migratory fish from the surface.
3) Estimates of migration rates and population sizes verified the concept that the influx of genes from surface populations and their effective population sizes are linked to the genetic diversity of cave populations.
4) Several of the cave populations were distinct and had increased genetic diversity, which was associated with rather high levels of migration from the surface. There was a significant gene flow in both ways between surface and cave populations.

Based on these conclusions, we proposed a mechanism for the GA, called: Cave-Surface GA (CSGA), in the hope of producing good individuals, increasing the diversity of the population and thus improving the efficiency of the GA. This mechanism uses two distinct groups of populations: The primary population is called the Cave, while the secondary population is called the Surface. The Cave population plays the same role as the GA, while the Surface population only provides diversity to the Cave population through cross-breeding.

CSGA begins with two populations that are generated at random, the Cave population and the Surface population. Individuals in each population are assessed using the same fitness functions. The Cave population undergoes evolutionary changes through a combination of inbreeding within the same population and cross-breeding with individuals from Surface populations. On the other hand, the surface population evolves primarily through inbreeding between parents from the same population. Next, we will describe the fitness function that was utilized for both populations. Subsequently, we will explain the evolutionary mechanism, including the processes of reproduction and survival selection, for both populations.

A) **Fitness function**: The population fitness function is just the objective function of the specific problem. In CSGA, the two populations use the same fitness function.
B) **Evolutionary procedure**: Traditional MPGAs reproduce simply by inbreeding among parents belonging to the same population. Their populations interact by exchanging certain individuals in accordance with a predetermined policy. Because MPGA populations share the same evolutionary goals and fitness function, excellent migrants are quickly absorbed into the new population. However, since the CSGA populations have similar fitness functions, there is no migration process and we will use cross-breeding as a process to enrich diversity in the populations.

The CCSGA generates offspring through both inbreeding and cross-breeding processes to facilitate the exchange of information between populations. Cross-breeding takes place when an individual from the Cave population reproduces with an individual from the Surface population, resulting in offspring that possess genetic material from both populations. These cross-bred offspring often exhibit fitness values that enable their survival in either population due to the combination of advantageous traits from both sources.

A standard GA chooses two parent chromosomes for a cross-over operation and produces two offspring. CSGA, on the other hand, has two extra parameters than GAs: the cross-breeding interval (CI) and the cross-breeding rate (CR). The cross-breeding interval (CI) is the number of generations between each cross-breeding and cross-breeding rate (CR) is the number of individuals selected from each population at the time of cross-breeding. These factors have an impact on the accuracy of the results, as well as on the computation time.

Based on the crossbreeding rate, CSGA randomly selects a number of parents from the two populations for recombination and generates two offspring through each cross-over operator between two parents. Then, through selection for local survival, one resulting offspring is selected to be a member of the next generation of Cave population, whereas the other offspring is sent to the Surface

population, as shown in Figure 3. This procedure is repeated for each of the two candidate parents and in addition, this process does not take place in every generation, but takes place through specific generations, based on CR rate. The CSGA psudocode is shown in Algorithm1.

Algorithm 1: CSGA Algorithm

Begin
step1: initialize input parameters of problems: crossover rate, mutation rate, crossbreeding rate (cr), crossbreeding interval (ci), max generation.
step2: initialize two subpopulations, cave population (cp) and surface population (sp).
step3: For each subpopulation, repeat the following steps until the termination criterion is met.
step4: Calculate fitness value;
step5: inbreeding:
a. Selection
b. Crossover
c. Mutation
step6: crossbreeding (based on ci and cr do):
a. choose individual from cp and choose individual from sp.
b. Crossover
c. Move offspring one to
cp and the second to sp.
step8: output the final best solution.
End.

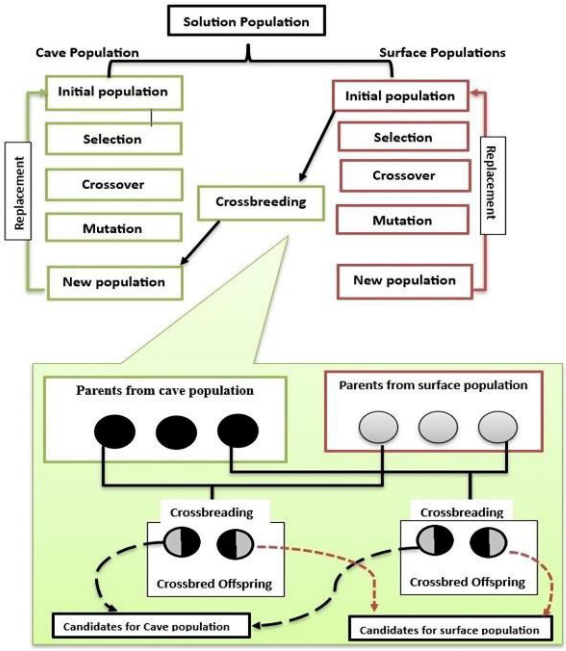


Figure 3. CSGA method; candidate offspring for cave and surface populations.

3.1 CSGA Procedure

The method begins by initializing the Cave population and Surface population, as well as the cross-over rate, mutation rate, cross-breeding rate, cross-breeding interval and maximum generation. The size of both populations is the same. Initially, the Cave population goes through the traditional GA evolution cycle. In Step 2, the fitness of the cave and surface populations is calculated using an objective function. Step 3: Inbreed both the Cave and Surface populations. The proper number of parents is then determined for reproduction based on the cross-over rate. The Cave population and the Surface population are then provided with diversity *via* mutation.

Cross-breeding between separate populations is performed in Step 4. Cross-breeding on the Cave and Surface populations produces a particular number of offspring. In cross-breeding, the intermediate Cave population, the intermediate Surface population and their offspring form a candidate set for the Cave and Surface populations' next generation. Step 5 evaluates created candidate sets using a fitness

function for both populations and CSGA develops until either the maximum number of generation sets is attained or the algorithm provides the best possible solution to the present problem.

Table 1. TSP benchmark dataset.

Class	Instance size	TSP instances
1	size < 200	a280, att48, berlin52, bier127, ch130, ch150, eil51, kroA100, lin105, pr76, pr144
2	size ≥ 200	Lin318, ali535, rat783, kroB200

4. EXPERIMENTAL SETTINGS, RESULTS AND DISCUSSION

To verify the performance of the proposed algorithm, we conducted two sets of experiments on different TSP instances, which are given by the TSPLIB [39], which contains between 40 and 800 vertices. It is crucial to note that the TSP is used in this study only to compare the proposed CSGA to other methods in terms of diversity, rather than to seek a superior solution to the TSP problem. Simulation experiments are performed in the Microsoft Visual Studio 2022 environment and the system's hardware and software specifications are as follows:

- 11th Gen Intel(R) Core (TM) i7-1165G7 @ 80GHz 2.80 GHz
- 8.00 GB of RAM
- Windows 11 Pro, 64-bit operating system.

In the conducted experiments, each algorithm was applied 10 times to multiple instances of the TSP. The results obtained from each execution were averaged to provide a comprehensive assessment. Our GA utilized the reinsertion strategy, specifically the expansion sampling method introduced by Dong et al. [13]. This strategy involves selecting only the best half of individuals, including both new individuals and individuals from the previous generation, to form the population for the next generation. In other words, during the production of a new generation, the old generation competes with the newly generated individuals and only the fittest individuals are retained.

4.1 First Set of Experiments

The proposed method is compared to established GA algorithms utilizing fifteen TSP instances and varied numbers of vertices. Table 1 shows how the selected TSP instances were divided into two classes based on TSP size. Table 2 displays the selected GA parameters used in our experimental setup. The results of the proposed algorithms evaluated on TSP instances are summarized in Table 3.

As illustrated in Table 3, the CSGA performed better than the GAs in 7 out of 11 instances, in the first class. As for the second class, CSGA achieved the best performance over the GAs in the four TSP instances belonging to that category. Furthermore, when we look at the table, we can see in the Min column that the proposed CSCA had the lowest cost in 12 of the 15 TSP instances. It is important to mention that the simple GA-algorithm parameters were utilized as shown in Table 2; we did not use any sophisticated parameter control procedures, since the main purpose of this paper is to evaluate the efficacy and to demonstrate the goodness of the proposed method compared to the GAs in terms of diversity, regardless of the parameters employed, neutralizing parameters' tuning effect.

Table 2. The selected GA parameters used in our experimental setup.

Parameter	Value
Population size	200
Generation limit	3000
Initialization method	Random
Cross-over	One-point modified
Cross-over rate	0.85
Mutation	Exchange
Mutation rate	0.08
Selection	Truncation selection
CR	5
CI	7
Termination criteria	Generation limit

Figure 4 depicts each algorithm's convergence to the shortest route. Again, CSGA outperforms GA on a280 in terms of convergence to the minimal value, indicating that the population diversity provided by the CSGA allows for better convergence. This is also evident in Figures 5 and 6 that show the average convergence of the GA and CSGA, in small and large TSP instances, respectively.

Table 3. The results achieved by the GA and CSGA algorithms for TSP instances after 3000 iterations.

Class No.	Optimal Solution	Instance	GA		CSGA	
			Min.	Average	Min.	Average
Class1	2579	a280	6952	7587.1	5914	6541.2
	10628	att48	35843	41766.4	35704	40468.3
	7542	berlin52	8253	9123.1	8497	9572.5
	118282	bier127	152453	170944.7	146855	161837.4
	6110	ch130	8865	10000.7	8768	9777.2
	6528	ch150	10114	10914.66667	9965	10906
	426	eil51	465	478.3	476	502.5
	21282	kroA100	27555	32230.6	27175	31655.4
	14379	lin105	20153	24129.1	20006	23199.1
	108159	pr76	134438	133195.8182	130101	143111.4
	58537	pr144	112926	119005.8	117911	134050.4
Class2	42029	lin318	125686	139947.1	112936	119163.4
	202339	ali535	9484	10249.5	8418	8827.6
	8806	rat783	51871	53879.1	46348	47547
	29437	kroB200	58704	67404.5	57500	61319.4

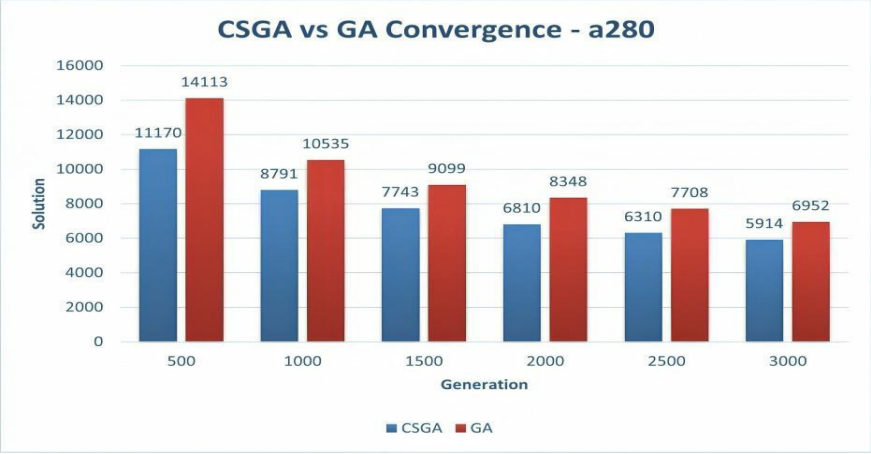


Figure 4. Average convergence of GA and CSGA for a280.

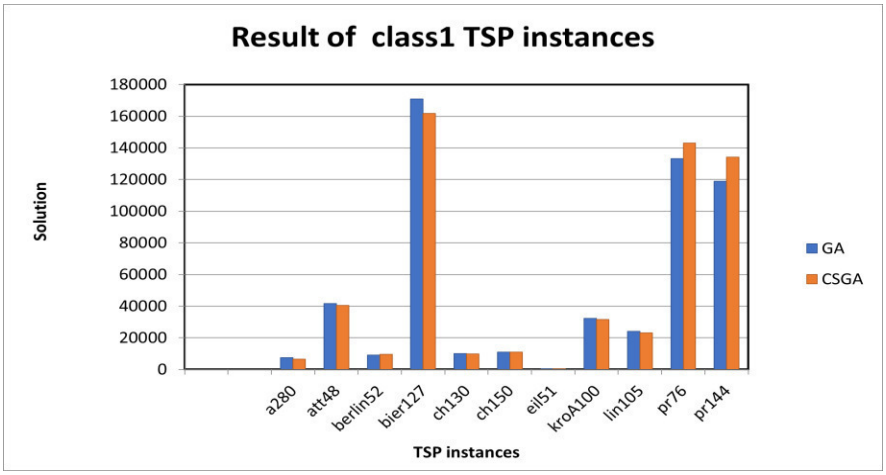


Figure 5. Average convergence of GA and CSGA for small instances from TSP (class1).

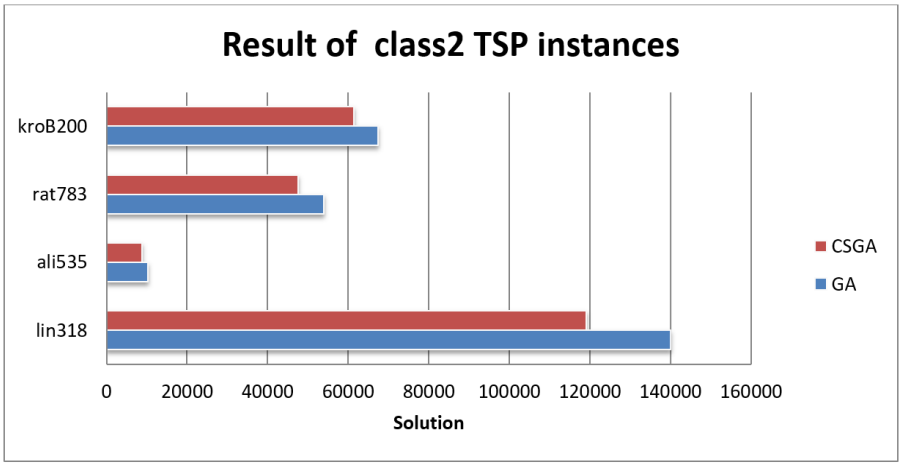


Figure 6. Average convergence of GA and CSGA for big instances from TSP (class2).

Figure 7 and Figure 8 show the route of two cities from TSBLIB, which are lin318 and kroA100, respectively. The resulting solution when applying the GA for the lin318 city is 124600 and the result when applying CSGA is 105240. As for the city of kroA100, the result of applying the GA was 32901 and as for the CSGA method, the result for this city was 26293. Note that these solutions and figures are the result of applying the two methods (GA and CSGA) to the same parameters found in Table 2.

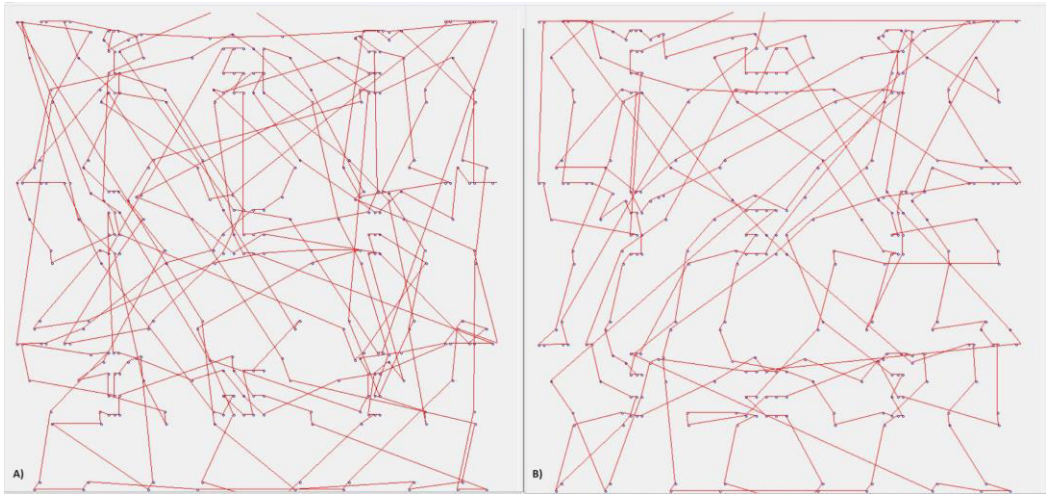


Figure 7. The resultant routes after applying the GA and CSGA methods on lin318 instance; (a) GA, (b) CSGA.

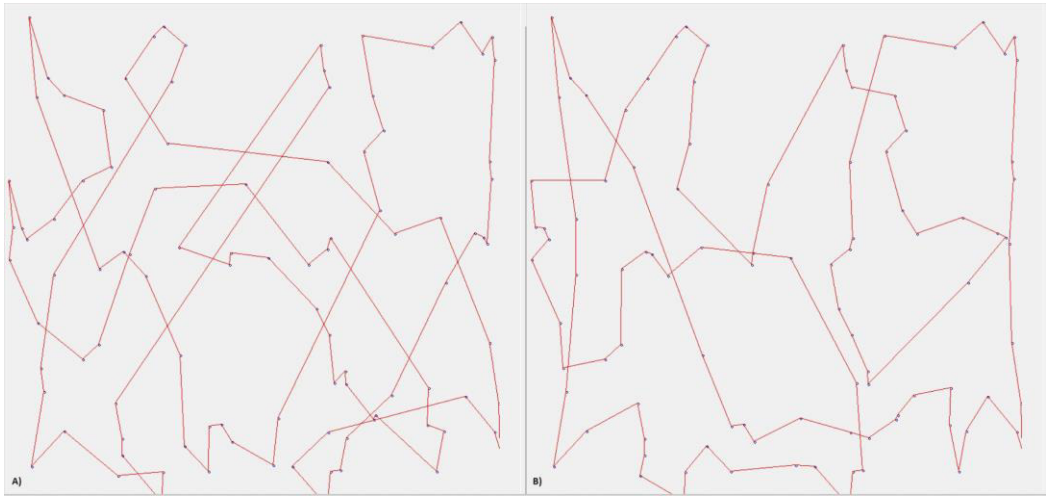


Figure 8. The resultant routes after applying the GA and CSGA methods on kroA100 instance; (a) GA, (b) CSGA.

4.2 Second Set of Experiments

The second group of experiments aims to study the effects of cross-breeding rate and interval on the proposed algorithm. It also aims to compare the proposed algorithm with one of the most famous methods that help diversify the population, which is MPGA. The results were also compared with those of a well- known optimization method; namely, PSO. We choose PSO, because GA and PSO are heuristic-based optimization methods and have many similarities in their inherent parallel characteristics [1]. In other words, PSO shares many similarities with evolutionary computation techniques, such as GA.

The parameters for the PSO were as follows: number of population =200, number of iterations =3000, cognitive component =1.5, social component=1.5 and inertia weight=0.7.

Table 4. Performance comparison of the CSGA, GA, PSO and MPGA.

Optimal	Instance	GA	CSGA	MPGA	PSO
2579	a280	7587.1	6496.2	7306.3	30814.28302
10628	att48	41766	39417	41766	124827.769
7542	berlin52	9123.1	9411.7	8955.3	26070.21051
118282	bier127	170945	165368	173151	581653.293
6110	ch130	10001	9312.8	9354.6	41792.24209
6528	ch150	10915	10650	10859	50091.11352
426	eil51	478.3	497.9	476.2	1367.987448
21282	kroA100	32231	30701	30957	146503.0298
14379	lin105	24129	22508	21860	107403.6209
29437	kroB200	67405	57890	62478	303044.6882

Table 5. P-values of Wilcoxon signed-rank test for each pair of the methods reported in Table 4.

	GA	CSGA	MPGA	PSO
GA	-	0.0137	0.0756	0.0020
CSGA	-	-	0.1309	0.0020
MPGA	-	-	-	0.0020
PSO	-	-	-	-

The parameters of the GA are the same as those used in the first set of experiments, except for the cross-breeding rate and cross-breeding interval, where the following values are set: CI=50, CR=10. As these factors have an impact on the proposed algorithm. However, the number of populations of the MPGA was 2. We performed experiments on ten TSP instances, each of which has a known optimal solution. These instances include att48, eil51, berlin52, KroA100, lin105, bier127, ch130, ch150, kroB200 and a280. Table 4 displays the results of the CSGA in comparison to the traditional methods: GA and MPGA. We increased the number of individuals for cross-breeding and at the same time reduced the cross-breeding interval. The aim is to study the effect of these variables on the diversity of the solutions. Consequently, these variables influence the quality of the results.

As seen in Table 4, by contrasting the results found in this table, the proposed CSGA has clearly been able to give more satisfactory outcomes than GA, which is evident from the improvement in the quality of the solutions obtained for 8 instances.

Table 4 demonstrates that the proposed CSGA outperforms the MPGA. Specifically, the CSGA algorithm yielded better results than the MPGA algorithm in 7 instances: a280, att48, bier127, ch130, ch150, kroA100 and kroB200.

The Wilcoxon signed-rank test is a non-parametric statistical test that is used to compare two related samples and determine whether there are statistically significant differences between them. Table 5 shows the p-values from the Wilcoxon signed-rank test for every method pair. In this context, we're comparing the performance of several optimization algorithms on TSP instances.

A p-value is a measure of evidence against a null hypothesis, which in our case is that there are no statistically significant differences in the performance of each pair of methods compared. Researchers

typically employ a significance level (alpha) to evaluate whether a p-value is statistically significant or not. The most common alpha levels are 0.05 and 0.01. Here, we considered statistically differences significant if the p-value is less than or equal to alpha=0.05.

Table 4 shows that the proposed CSGA surpasses PSO in all TSP instances tested. The p-value of 0.0020, which is less than 0.05, supports this conclusion, suggesting statistically significant differences between CSGA and PSO. In most cases, the proposed CSGA outperforms classic GA and somewhat outperforms MPGA. The p-value between CSGA and GA is 0.0137, which is smaller than (alpha = 0.05), indicating that the differences are statistically significant. The p-value between CSGA and MPGA, on the other hand, is 0.1309, which is not statistically significant at alpha = 0.05. This is owing to the difference being insignificant, despite being in favour of the proposed CSGA.

It is worth noting that when addressing the TSP problem, the GA produced much better results than the PSO. This conclusion is confirmed further by references [29, 46], which compare the performances of PSO and GA. According to their findings, PSO has quicker computing performance, although GA produces shorter optimized pathways. Also, GA is a better option for dealing with TSP, particularly when time is not a big concern according to [29, 46]. As a result, when handling the TSP issue, the proposed approach outperforms typical GAs and MPGA, as well as one of the well-known optimization methods (PSO). This highlights the ability of the proposed approach to improve the efficiency of GAs in solving the TSP problem, which is a typical example of an optimization problem. Furthermore, this enhancement may be relevant to a broader range of optimization problems; however, further work is required, which is beyond the scope of this paper. The average convergence of GA, CSGA, PSO and MPGA on 10 TSP instances is also shown in Figure 9.

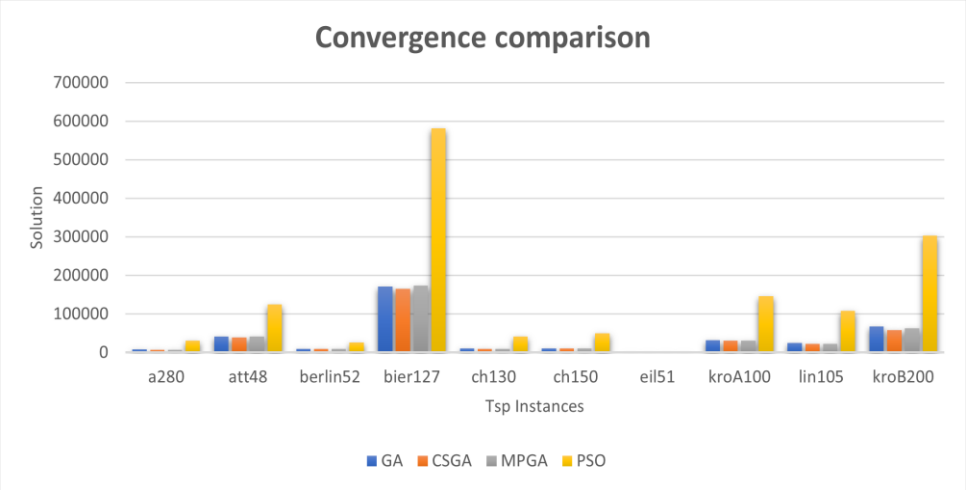


Figure 9. Average convergence of GA and CSGA for big instances from TSP (class2) [9].

It is commonly known that the parameters of GAs have a substantial impact on the results of earlier studies. There isn't a single, best option for every parameter that may be used in every TSP instance. As a result, adjusting these settings becomes problem-specific [22]. However, the initialization of the starting population is one of the most important parameters. This procedure guarantees that the GA has a good starting point instead of starting from scratch, which entails initializing random solutions. When solving a TSP issue, applying an approximation technique or a heuristic solution during the initialization stage improves the GASs performance and speeds up its convergence to better solutions [5].

Consequently, we carried out several experiments utilizing Iterative Approximate Methods for Solving TSP (IAMTSP+), a recent initialization technique primarily intended to offer a heuristic solution for TSP, as detailed in [8]. We also initialized the MPGA and the GA to ensure a fair comparison. This excellent result from IAMTSP+ functions as the Surface population for the proposed CSGA method. We choose to use a different, less advanced initialization method that is based on linear regression (LG) [22] simultaneously. The Cave population responds effectively to this comparatively less sophisticated initialization technique, because the solutions that it provides seem to devolve and are of lower quality than those offered by IAMTSP+. Examples of both techniques applied to cities created at random are shown in Figure 10.

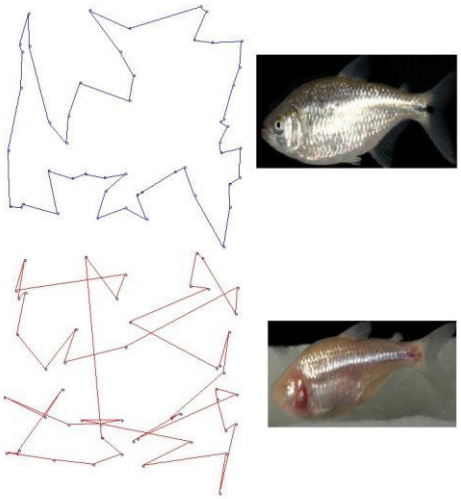


Figure 10. Visualization of the performance of the initialization methods. Top/Surface: IAMTSP+) and Bottom/Cave: LG. Both applied on the same randomly generated TSP.

As anticipated, all methods performed better when initialized with the IAMTSP+ method, as Table 7 illustrates. Of them, CSGA is the most effective method, outperforming the other methods in five TSP instances, yielding outcomes that are on par with the other approaches and obtaining the least average approximation to the optimal solutions. It is worth noting that we used the same parameters in Table 2, except for the CR and CI, which were set to 1. This adjustment helped prevent the Surface population from completely dominating the Cave population, thereby avoiding premature convergence.

This exceptional result can be described to CSGA's novel methodology, which makes use of two separate populations: fully evolved surface fish and less evolved cave fish allowing for more diversity. Diversity is introduced by cross-breeding such populations and then separating their offspring, sending one to the surface and the other to the cave. Through this process, some less-developed genes taken to the surface by the offspring have the opportunity to breed with those in the cave. With time, superior genes from the surface benefit the population residing in caves, while useful genes from the surface can accelerate solution development in the cave, but may reach a local optimum solution without the diversity provided by the offspring that newly inhabited the Surface population. This approach fosters a broader exploration of the search space, contributing to the algorithm's effectiveness. Figure 11 illustrates the performance of the proposed CSGA with IAMTSP+ and LG compared to the same TSP instance (kroA100) shown in Figure 8.

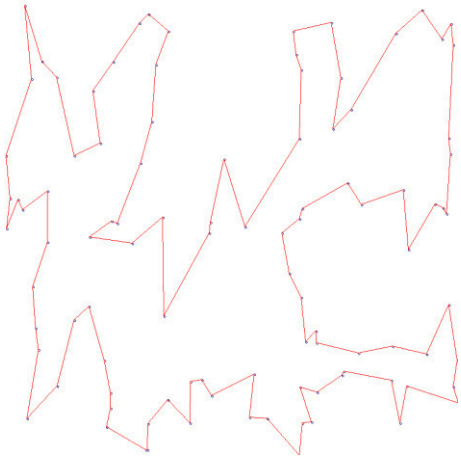


Figure 11. Visualization of the performance of the CSGA when using both initialization methods, IAMTSP+ and LG, applied on kroA100.

Although the proposed CSGA performed reasonably well, in five instances it just marginally outperformed other approaches and in five other cases it was not even close to the best. As seen from the p-values of the Wilcoxon test in Table 7, these show no significant differences. As shown by [8],

this phenomenon can be linked to the effectiveness of the initialization process, which autonomously produces near-optimal solutions without requiring a meta-heuristic.

Table 6. Performance comparison of the CSGA, GA and MPGA using initial population (IAMTSP+).

Instance	MPGA	GA	CSGA	Optimal	App.MPGA	App.GA	App.CSGA
a280	2984	2872	**2803**	2579	1.157037611	1.11361	1.086855
att48	34414	34334	**33607**	10628	3.238050433	3.230523	3.162119
berlin52	8065	7946	**7748**	7542	1.069345001	1.053567	1.027314
bier127	**128142**	129215	128743	118282	1.083360106	1.092432	1.088441
ch130	6521	**6418**	6491	6110	1.067266776	1.050409	1.062357
ch150	7208	7007	**6954**	6528	1.104166667	1.073376	1.065257
eil51	438	**435**	437	426	1.028169014	1.021127	1.025822
kroa100	**22057**	**22057**	22274	21282	1.03641575	1.036416	1.046612
lin105	18062	15523	**15485**	14379	1.256137423	1.07956	1.076918
krob200	32149	**32025**	33390	29437	1.092128953	1.087917	1.134287
Sum	260040	257832	257932	217193	13.13207773	12.83894	12.77598
Avg.	26004	25783.2	25793.2	21719.3	1.313207773	1.283894	1.277598

Table 7. P-values of Wilcoxon signed-rank test for each pair of the methods reported in Table 6. Differences are shown in the upper diagonal and p-values are shown in the lower diagonal.

	GA	CSGA	MPGA
GA		-10	-220.8
CSGA	0.6953125		-210.8
MPGA	0.09720109	0.4921875	

Although the CSGA shows promise for use, not only in the TSP, but also in several related fields, including urban planning, networking, transportation planning, and location-based services, it has the following limitations:

- Solution quality: Using more diverse selection techniques, such as tournament selection, might improve the quality of the solutions even further. More gains could come by tailoring cross-over and mutation operators to the unique features of the proposed CSGA. Addressing these areas of enhancement could pave the way for future research aimed at boosting the performance of the proposed method.
- Separated offspring: In this version of the proposed method, the appearance of offspring on the surface might significantly increase the number of the Surface population, which could lead to memory issues. By keeping the Surface population at a consistent size, this restriction can be lessened. Appropriate actions must be taken as these offspring increasingly converge towards the initial IAMTSP+ solutions. Examples of workable solutions would be coming up with random new solutions or initiating the RG process over again.
- Manual parameter selection: The CSGA's settings were determined by hand without optimization, allowing for future enhancements *via* the application of adaptive parameter-management approaches. The exploration of adaptive parameter-tuning procedures points to a promising future-research direction.

All of these difficulties highlight the need for additional studies to overcome them and improve the efficacy of the proposed CSGA approach.

5. CONCLUSIONS

In this study, we introduced a novel multi-population method for the GAs called the Cave Surface Genetic Algorithm (CSGA). Inspired by the evolutionary mechanism observed in cavefish, this algorithm incorporates a secondary population to maintain diversity in the primary population. The cross-breeding mechanism ensures the preservation of a diversified population. The CSGA was applied to various instances of the Traveling Salesman Problem (TSP).

The experimental results show that the proposed CSGA outperforms classical GAs, MPGAs and PSOs in terms of solution quality across the majority of benchmark TSP instances. Nonetheless, limitations must be acknowledged. The parameter choices for the CSGA were selected by hand without optimization, giving the possibility for prospective improvements through the use of adaptive parameter-management techniques. The investigation of adaptive parameter-tuning strategies indicates a promising future research direction.

Extending the scope of our testing to include varied issue domains will not only provide significant insights, but will also further validate the efficacy of the CSGA approach. Furthermore, digging into multi-objective optimization issues has the potential to greatly expand the applicability of our approach. These enhancements and extensions will be the key points of our future-research efforts.

ACKNOWLEDGEMENTS

The authors genuinely appreciate the reviewers' voluntary efforts and are grateful for their valuable insights.

REFERENCES

[1] W. F. Abd-El-Wahed, A. A. Mousa and M. A. El-Shorbagy, "Integrating Particle Swarm Optimization with Genetic Algorithms for Solving Nonlinear Optimization Problems," Journal of Computational and Applied Mathematics, vol. 235, no. 5, pp. 1446-1453, 2011.

[2] N. Al-Milli, A. Hudaib and N. Obeid, "Population Diversity Control of Genetic Algorithm Using a Novel Injection Method for Bankruptcy Prediction Problem," Mathematics, vol. 9, no. 8, p. 823, 2021.

[3] M. A. Albadr, S. Tiun, M. Ayob and F. Al-Dhief, "Genetic Algorithm Based on Natural Selection Theory for Optimization Problems," Symmetry, vol. 12, no. 11, p. 1758, 2020.

[4] E. Alkafaween and A. B. A. Hassanat, "Improving TSP Solutions Using GA with a New Hybrid Mutation Based on Knowledge and Randomness," Communications-Scientific Letters of the University of Zilina, vol. 22, no. 3, pp. 128-139, 2020.

[5] E. Alkafaween, A. B. A. Hassanat and S. Tarawneh, "Improving Initial Population for Genetic Algorithm Using the Multi Linear Regression Based Technique (MLRBT)," Communications-Scientific Letters of the University of Zilina, vol. 23, no. 1, pp. E1-E10, 2021.

[6] E. O. Alkafaween, "Novel Methods for Enhancing the Performance of Genetic Algorithms," arXiv preprint, arXiv: 1801.02827, 2018.

[7] E. Alkafaween, S. Elmougy, E. Essa and A. Hassanat, "An Efficiency Boost for Genetic Algorithms: Initializing the GA with the Iterative Approximate Method for Optimizing the Traveling Salesman Problem - Experimental Insights," Applied Sciences, vol. 14, no. 8, p. 3151, 2024.

[8] E. Alkafaween, S. Elmougy, E. Essa, S. Mnasri, A. S. Tarawneh and A. Hassanat, "IAM-TSP: Iterative Approximate Methods for Solving the Traveling Salesman Problem," Int. J. of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 11, DOI: 10.14569/IJACSA.2023.0141143, 2023.

[9] M. Bradic, P. Beerli, F. J. García-de León, S. Esquivel-Bobadilla and R. L. Borowsky, "Gene Flow and Population Structure in the Mexican Blind Cavefish Complex (Astyanax Mexicanus)," BMC Evolutionary Biology, vol. 12, Article no. 9, pp. 1-17, 2012.

[10] C.-M. Chen, S. Lv, J. Ning and J. Ming-Tai Wu, "A Genetic Algorithm for the Waitable Time-varying Multi-depot Green Vehicle Routing Problem," Symmetry, vol. 15, no. 1, p. 124, 2023.

[11] E. C. Osuna, Theoretical and Empirical Evaluation of Diversity-preserving Mechanisms in Evolutionary Algorithms: On the Rigorous Runtime Analysis of Diversity-preserving Mechanisms in Evolutionary Algorithms, PhD Thesis, University of Sheffield, 2018.

[12] K. A. De Jong, "An Analysis of the Behavior of a Class of Genetic Adaptive Systems," Technical Report, University of Michigan, 1975.

[13] M. Dong and Y. Wu, "Dynamic Crossover and Mutation Genetic Algorithm Based on Expansion Sampling," Proc. of the Int. Conf. on Artificial Intelligence and Computational Intelligence (AICI 2009), Proceedings 1, pp. 141-149, Shanghai, China, Springer, November 7-8, 2009.

[14] H. Du, Z. Wang, W. Zhan and J. Guo, "Elitism and Distance Strategy for Selection of Evolutionary Algorithms," IEEE Access, vol. 6, pp. 44531-44541, 2018.

[15] Y. Elipot, H. Hinaux, J. Callebert and S. Rétaux, "Evolutionary Shift from Fighting to Foraging in Blind Cavefish through Changes in the Serotonin Network," Current Biology, vol. 23, no. 1, pp. 1-10, 2013.

[16] Y. Fu, G. Tian, Z. Li and Z. Wang, "Parallel Machine Scheduling with Dynamic Resource Allocation *via* a Master-Slave Genetic Algorithm," IEEJ Transactions on Electrical and Electronic Engineering, vol. 13, no. 5, pp. 748-756, 2018.

[17] J. B. Gross, "The Complex Origin of Astyanax Cavefish," BMC Evolutionary Biology, vol. 12, no. 1, pp. 1-12, 2012.
[18] D. Gupta and S Ghafar, "An Overview of Methods Maintaining Diversity in Genetic Algorithms," Int. J. of Emerging Technology and Advanced Engineering, vol. 2, no. 5, pp. 56-60, 2012.
[19] S. Han and L. Xiao, "An Improved Adaptive Genetic Algorithm," Proc. of the 2022 Int. Conf. on Information Technology in Education and Management Engineering (ITEME2022), SHS Web of Conf., vol. 140, p. 01044, EDP Sciences, 2022.
[20] E.-ul Haq, I. Ahmad, A. Hussain and I. M. Almanjahie, "A Novel Selection Approach for Genetic Algorithms for Global Optimization of Multimodal Continuous Functions," Computational Intelligence and Neuroscience, vol. 2019, Article ID 8640218, pp. 1-14, 2019.
[21] A. Hassanat et al., "Choosing Mutation and Crossover Ratios for Genetic Algorithms: A Review with a New Dynamic Approach," Information, vol. 10, no. 12, p. 390, 2019.
[22] A. B. Hassanat et al., "An Improved Genetic Algorithm with a New Initialization Mechanism Based on Regression Techniques," Information, vol. 9, no. 7, p.167, 2018.
[23] A. B. A. Hassanat, "Enhancing Genetic Algorithms Using Multi Mutations: Experimental Results on the Travelling Salesman Problem," Int. Journal of Computer Science and Information Security, vol. 14, no. 7, p. 785, 2016.
[24] A. B. A Hassanat and E. Alkafaween, "On Enhancing Genetic Algorithms Using New Crossovers," Int. Journal of Computer Applications in Technology, vol. 55, no. 3, pp. 202-212, 2017.
[25] J. H. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence, no. 53, ISBN: 9780262581110, MIT Press, 1992.
[26] A. Hussain and Y. S. Muhammad, "Trade-off between Exploration and Exploitation with Genetic Algorithm Using a Novel Selection Operator," Complex & Intelligent Systems, vol. 6, no. 1, pp. 1-14, 2020.
[27] S. Katoch, S. S. Chauhan and V. Kumar, "A Review on Genetic Algorithm: Past, Present and Future," Multimedia Tools and Applications, vol. 80, pp. 8091-8126, 2021.
[28] B. Koohestani, "A Crossover Operator for Improving the Efficiency of Permutation-based Genetic Algorithms," Expert Systems with Applications, vol. 151, p. 113381, DOI: 10.1016/j.eswa.2020.113381, 2020.
[29] L. Kou, J. Wan, H. Liu, W. Ke, H. Li, J. Chen, Z. Yu and Q. Yuan, "Optimized Design of Patrol Path for Offshore Wind Farms Based on Genetic Algorithm and Particle Swarm Optimization with Traveling Salesman Problem," Concurrency and Computation: Practice and Experience, vol. 36, no. 2, p. e7907, DOI: 10.1002/cpe.7907, 2023.
[30] S. Mahmoudinazlou and C. Kwon, "A Hybrid Genetic Algorithm for the Min-max Multiple Traveling Salesman Problem," Computers & Operations Research, vol. 162, p. 106455, DOI: 10.1016/j.cor.2023.106455, 2024.
[31] S. Malik and S. Wadhwa, "Preventing Premature Convergence in Genetic Algorithm Using DGCA and Elitist Technique," Int. Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 6, pp. 410-418, 2014.
[32] E. Simona Nicoară, "Mechanisms to Avoid the Premature Convergence of Genetic Algorithms," Petroleum-Gas University of Ploiesti Bulletin, Mathematics-Informatics-Physics Series, vol. 61, no. 1, 2009.
[33] R. Ohira, M. S. Islam, J. Jo and B. Stantic, "LCS Based Diversity Maintenance in Adaptive Genetic Algorithms," Proc. of the 16th Australasian Conf. on Data Mining (AusDM 2018), pp. 56-68, Bahrurst, NSW, Australia, November 28-30, 2018, Springer, 2019.
[34] E. C. Osuna and D. Sudholt, "On the Runtime Analysis of the Clearing Diversity-preserving Mechanism," Evolutionary Computation, vol. 27, no. 3, pp. 403-433, 2019.
[35] H. M. Pandey, A. Chaudhary and D. Mehrotra, "A Comparative Review of Approaches to Prevent Premature Convergence in GA," Applied Soft Computing, vol. 24, pp. 1047-1077, 2014.
[36] T. Park and K. Ryel Ryu, "A Dual Population Genetic Algorithm with Evolving Diversity," Proc. of the 2007 IEEE Congress on Evolutionary Computation, pp. 3516-3522, Singapore, 2007.
[37] P. V. Paul et al., "A New Population Seeding Technique for Permutation-coded Genetic Algorithm: Service Transfer Approach," Journal of Computational Science, vol. 5, no. 2, pp. 277-297, 2014.
[38] B. R. Rajakumar and A. George, "APOGA: An Adaptive Population Pool Size Based Genetic Algorithm," AASRI Procedia, vol. 4, pp. 288-296, DOI: 10.1016/j.aasri.2013.10.043, 2013.
[39] G. Reinelt, "TSBLIB, 1996" ftp to softlib.rice.edu, 2023.
[40] X. Shi, W. Long, Y. Li, D. Deng and Y Wei, "Research on the Performance of Multi-population Genetic Algorithms with Different Complex Network Structures," Soft Computing, vol. 24, pp. 13441-13459, 2020.
[41] E. Shojaedini, M. Majd and R. Safabakhsh, "Novel Adaptive Genetic Algorithm Sample Consensus," Applied Soft Computing, vol. 77, pp. 635-642, 2019.
[42] M. Srinivas and L. M. Patnaik, "Adaptive Probabilities of Crossover and Mutation in Genetic

Algorithms," IEEE Transactions on Systems, Man and Cybernetics, vol. 24, no. 4, pp. 656-667, 1994.
[43] B. A. Stahl et al., "Manipulation of Gene Function in Mexican Cavefish," Journal of Visualized Experiments (JoVE), vol. 146, p. e59093, 2019.
[44] P. A. Vikhar, "Evolutionary Algorithms: A Critical Review and Its Future Prospects," Proc. of the 2016 Int. Conf. on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), pp. 261-265, Jalgaon, India, 2016.
[45] D. Whitley, S. Rana and R. B. Heckendorn, "The Island Model Genetic Algorithm: On Separability, Population Size and Convergence," J. of Computing and Inf. Technolo., vol. 7, no. 1, pp. 33-47, 1999.
[46] Z. Wu, "A Comparative Study of Solving Traveling Salesman Problem with Genetic Algorithm, Ant Colony Algorithm and Particle Swarm Optimization," Proc. of the 2020 2nd Int. Conf. on Robotics Systems and Vehicle Technology, pp. 95-99, Xiamen, China, 2020.
[47] W. Xu et al., "Optimization Approaches for Solving Production Scheduling Problem: A Brief Overview and a Case Study for Hybrid Flow Shop Using Genetic Algorithms," Advances in Production Engineering & Management, vol. 17, no. 1, pp. 45-56, 2022.
[48] Y. Xue, H. Zhu, J. Liang and A. Sşowik, "Adaptive Crossover Operator Based Multi-objective Binary Genetic Algorithm for Feature Selection in Classification," Knowledge-based Systems, vol. 227, p. 107218, DOI: 10.1016/j.knosys.2021.107218, 2021.
[49] S. Yang, "PDGA: The Primal-dual Genetic Algorithm," Design and application of Hybrid Intelligent Systems, pp. 214-223, Chapter in Book: Design and Application of Hybrid Intelligent Systems, IOS Press, 2003.
[50] P. Zhang et al., "A Genetic Algorithm with Jumping Gene and Heuristic Operators for Traveling Salesman Problem," Applied Soft Computing, vol. 127, p. 109339, DOI: 10.1016/j.asoc.2022.109339, 2022.
[51] G. Zhou et al., "Location Optimization of Electric Vehicle Charging Stations: Based on Cost Model and Genetic Algorithm," Energy, vol. 247, p. 123437, DOI: 10.1016/ j.energy.2022.123437, 2022.

ملخص البحث:

الخوارزميات الجينية هي خوارزميات بحثٍ مبنية على الجوانب الجينية للمجتمعات ومفهوم الانتخاب الطّبيعي. ويعدّ الحفاظ على تنوّع المجتمع أمراً حاسماً في الخوارزميات الجينية لضمان الفحْص الشّامل والتّخفيف من خطر الالتقاء قبل الأوان. وتجدر الإشارة إلى أنّ الالتقاء السّريع في اتجاه القيم المثالية المحلية يشكّل أحد أبرز التّحدّيات التي تواجه تطبيق الخوارزميات الجينية.

ولعلاج هذه المسألة، نقدّم في هذه الورقة خوارزميةً جينيةً تسمّى خوارزمية "الكَهْف/السَّطْح"، وتمثّل طريقةً بديلةً قائمةً على الخوارزمية الجينية مزدوجة المجتمع ومستوحاة من التّنوّع الجيني الملاحَظ في سمكة الكهف المكسيكية. ومن خلال الإكثار المتبادل بين المجتمعات، تعمل الطّريقة المقترحة على زيادة التّنوّع عبر مجتمع ثانوي (مجتمع الكهف) وتسهيل تبادل المعلومات بين المجتمعات؛ مقاوِمةً بذلك مشكلة الالتقاء قبل الأوان.

وقد تمّ إجراء العديد من التّجارب مع الاستفادة من حالاتٍ مرجعيةٍ لما يُعرف بـ "مشكلة البائع المتجول" (TSP) جرى الحصول عليها من مكتبةٍ معروفةٍ لحالاتٍ تتعلّق بتلك المشكلة. وقد بينت نتائج التّجارب أنّ الخوارزمية الجينية المقترحة في هذه الدّراسة والمسماة خوارزمية "الكَهْف/السَّطْح" تفوّقت على الخوارزميات الجينية الكلاسيكية وغيرها من الخوارزميات الجينية الّتي تستخدم تقنيات الحِفاظ على التّنوّع، من حيث إعطاء حلول واعدة للتّحدّيات المتعلّقة بتطبيقات الخوارزميات الجينية.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

AN IMPROVED AND EFFICIENT RSA-BASED AUTHENTICATION SCHEME FOR HEALTHCARE SYSTEMS

Fatty M. Salem[1], Nisreen F. Zaky[2], Elsayed M. Saad[2] and Hadeer A. Hassan Hosny[2]

(Received: 14-Feb.-2024, Revised: 5-Apr.-2024 and 26-Apr.-2024, Accepted: 8-May-2024)

ABSTRACT

Owing to the fast advancements of wireless communication, the telehealthcare platform makes it possible for patients to access healthcare services online. However, creating a secure and efficient authentication scheme for healthcare systems still presents a challenge. Several solutions have been introduced, but the majority of them are shortly found to be unable to meet some essential security standards. In this paper, we first revisit Dharminder et al.'s scheme and prove its failure to provide mutual authentication and pacient's untraceability and its vulnerability to impersonation attacks. Furthermore, we suggest an improved RSA-based authentication scheme to mitigate the deficiencies observed in Dharminder et al.'s schema. The proposed scheme can provide mutual authentication, patients' anonymity and untraceability and resist various types of attacks. Extensive evaluation on AVISPA proves the safeness of the proposed scheme against both passive and active attacks. Additionally, the proposed scheme is computationally and communicationally more efficient in comparison to other existing schemes.

KEYWORDS

Healthcare system, Authentication, Anonymity, Untraceability, Privacy.

1. INTRODUCTION

The Internet widespread utilization leads to creating a healthcare services' platform. The health records are transitioning from paper-based to electronic medical records that connect doctors and patients *via* medical servers. One widely used healthcare service is Telecare Medical Information System (TMIS) which provides healthcare services that are delivered to people at home [1].

In TMIS, all participants are registered by an established medical server, including nurses, doctors and patients. In general, it keeps electronic medical records of registered patients where a patient who has been registered can access the remote service whenever he/she wants and from any place. This system provides many advantages. First, due to the immediate communication offered by TMIS, healthcare services could be made available whenever and wherever. TMIS can overcome the limitations of time and distance. Second, TMIS servers allow access to patients' medical histories, which is extremely helpful for doctors in providing patients with the best possible medical care. Third, it requires less maintenance and can consolidate patient medical records from different healthcare providers.

Despite these benefits, TMIS may be susceptible to well-known attacks [2] due to its reliance on public networks. Therefore, a hacker can access messages sent over public channels and intercept, record, modify, delete and replay them. Furthermore, patient private information should be safeguarded to ensure user privacy. Therefore, to ensure the safety of TMIS, several security measures need to be achieved, such as mutual authentication, patient anonimity and untraceability and forward secrecy. Additionally, several attacks must be resisted, including insider attacks, attacks using stolen smart cards, offline password guessing attacks and attacks impersonating patients or servers.

For example, wearable health devices, such as fitness trackers and medical monitoring devices, represent a wealth of health data that can be vital for patient care. These devices often need to connect and transmit data to various healthcare platforms, ranging from personalized health applications to comprehensive healthcare systems. The challenge is to authenticate these devices to ensure transmitting data securely,

1. F. M. Salem (Corresponding Author) is with Department of Electronics and Communications Engineering, Faculty of Engineering, Helwan University, Helwan, Cairo, Egypt and with Faculty of Computing and Information Sciences, Egypt University of Informatics, New Administrative Capital, Cairo, Egypt. Email: faty_ahmed@h-eng.helwan.edu.eg
2. N. F. Zaky, E. M. Saad and H. A. Hassan Hosny are with Department of Computers and Systems Engineering, Faculty of Engineering, Helwan University, Helwan, Cairo, Egypt.

accurately attributed to the correct patient and smoothly integrating with the patients' health records in different systems.

1.1 Challenges and Basic Idea

The focus of this paper is on designing a healthcare authentication scheme and highlighting several challenges that must be addressed in developing a secure and efficient authentication model for healthcare systems [3]. The main challenges include:

- **Mutual Authentication and User Anonymity:** Previous schemes, including the one proposed by Dharminder et al. [4], have not succeeded in delivering mutual authentication and user anonymity. This is a significant gap, as mutual authentication ensures that both the user and the server verify each other's identity and user anonymity protects the identity of users accessing the system.
- **Resistance to Impersonation Attacks and Other Attacks:** This paper emphasizes the vulnerability of existing authentication systems to impersonation attacks, wherein an attacker masquerades as a legitimate user or server, along with other cyber threats, like impersonation attacks, replay attacks, insider attacks, stolen smart-card attacks and offline password-guessing attacks. These vulnerabilities expose sensitive patient data and system integrity to significant risks.
- **Efficiency in Computation and Communication Costs:** Another highlighted challenge is the need for an authentication scheme that prioritizes both security and efficiency in terms of computation and communication. Previous schemes suffer from high computational and communication costs, making them less practical for real-world applications.
- **Comprehensive Security Measures:** The review and analysis of various authentication schemes revealed a consistent challenge in developing an authentication system that effectively protects against diverse forms of attacks while maintaining the privacy and anonymity of users. This includes safeguarding against insider attacks, theft of smart cards and offline password guessing, as well as ensuring the freshness of session keys.

1.2 Motivation

The motivation behind this work stems from the critical analysis of existing authentication schemes for securing healthcare systems, especially the scheme proposed by Dharminder et al. [4]. The surge in healthcare services, facilitated by advancements in wireless communication, has highlighted significant security challenges, including the protection of sensitive patient data and the integrity of medical consultations. Existing systems, as exemplified by Dharminder et al.'s approach [4], failed in offering comprehensive security, lacking mutual authentication, user anonymity and resisting various types of attacks. Our motivation is driven by the urgent need to solve these vulnerabilities by proposing an authentication scheme for healthcare systems that is secure, efficient and feasible. Hence, we aim to contribute to the development of healthcare systems to ensure patient privacy, data integrity and service reliability, thus fostering trust and wider adoption of online healthcare services.

1.3 Our Contribution

This paper proposes an improved RSA [5] (Rivest-Shamir-Adleman)-based authentication scheme for healthcare systems. Specifically, the proposed scheme allows both patient and server to authenticate mutually; hence, they can agree on a secure shared session key. This paper's key contributions can be summarized as follows:

- We have looked over Dharminder et al.'s scheme in [4] and discovered that it lacks the ability to offer mutual authentication and untraceability and is incapable of withstanding both user and server impersonation attacks.
- An improved authentication scheme utilizing RSA encryption is proposed to address the security vulnerabilities present in Dharminder's scheme [4].
- The proposed scheme is created with two messages of exchange to ensure and speed up the session key agreement.
- A brief comparison is given between the proposed scheme and other current schemes. When

compared with similar schemes, the comparison suggests that the proposed scheme offers superior benefits in terms of security, computational costs and communication overheads.

- The proposed scheme has been subjected to security verification using the AVISPA (Automated Validation of Internet Security Protocols and Applications) tool, confirming its security against both passive and active attacks.

1.4 Roadmap of This Article

The structure of this article is outlined as follows: Related work on the authentication of healthcare systems has been reviewed in Section 2. Some preliminaries used in the proposed scheme are provided in Section 3. The threat model is defined in Section 4. Dharminder et al.'s scheme is reviewed and cryptanalyzed in Section 5. Furthermore, Section 6 presents the proposed scheme. In Section 7, the analysis of the proposed scheme's security through both formal and informal methods is presented. Section 8 includes a performance evaluation of the proposed scheme. Finally, conclusions are prsenetd in Section 9.

2. RELATED WORK

Numerous authentication schemes have been suggested for healthcare systems to achieve authentication and preserve the privacy of patients. For example, in 2016, Li et al. [6] introduced an authentication scheme for an e-healthcare system based on the chaotic map. However, Madhusudhan & Nayak [7] explained that Li et al.'s scheme couldn't withstand user impersonation attacks, server impersonation attacks, password guessing attacks and man-in-the-middle (MIMT) attacks. In addition, Madhusudhan [7] elucidated that Li et al.'s scheme [6] established an insecure session key and can't guarantee user anonymity.

In [8], the authors introduced a secured data access/sharing system based on chaotic maps for an Internet of Things (IoT)-enabled cloud storage environment. During the data-storage phase, a secret key derived from the user's biometric information is utilized to encrypt the data, enabling the data owner to store it in encrypted form on a cloud server. After receiving consent from both parties, users can retrieve the data from the cloud server during the data-sharing phase. Additionally, the authors in [9] introduced a Chaotic Map-based Authentication Protocol for Crowdsourcing IoT (CMAP-IoT). In CMAP-IoT [9], the medical server keeps patient health information and the users must authenticate with the medical server before being able to access the data contained in the medical server. Nevertheless, this system lacks the ability to offer user anonymity while being susceptible to user impersonation attacks.

In 2017, Ankita Chaturvedi et al. [10] reviewed several authentication schemes [11]-[14], but they found that these schemes [11]-[14] are unable to detect input correctness, leading to Denial of Service (DoS) scenarios and any error made during the password-changing process prevents the user from logging in with the same smart card. Chaudhry et al. introduced an enhancement to the two-factor authentication protocol [15], yet Shuming Qiu et al. [16] pointed out that Chaudhry et al.'s scheme [15] is unable to withstand offline password-guessing attacks, MIMT attacks and user/server impersonation attacks. To address these shortcomings, Qiu et al. [16] developed a scheme to defend against all known attacks and presented an elliptic curve-based mutual authentication scheme for TMIS.

Li et al. [17] have introduced a mutual-authentication scheme that is both secure and anonymous for two-hop wireless body-area networks. However, upon reviewing Li et al.'s scheme [17], still Koya and Deepthi [18] found vulnerabilities, including critical escrow issues and susceptibility to sensor node impersonation attacks. Additionally, they noted that the assumption regarding the reliability of the hub node is unrealistic. In addition, Kompara et al. [19] explained that Li et al.'s scheme [17] is ineffective and can't offer untraceability. Therefore, a new authentication scheme has been proposed in [19], employing only two types of operations: the cryptographic hash function and the XOR operation. This is aimed at reducing computational complexity and transmission costs. However, Rehman et al. [20] elucidated that Kamparaa et al.'s scheme [19] can't achieve anonymity and untraceability and can't resist impersonation attacks and sensor-node impersonation attacks. On the other hand, Rehman et al.'s scheme [20] is still susceptible to desynchronization attacks.

Lee et al. introduced a mutual authentication scheme for remote users using smart cards [21]. However, Radhakrishnan et al. [22] clarified that Lee et al.'s scheme [21] is susceptible to identity-guessing

attacks, password-guessing attacks, insider attacks, stolen smart-card attacks and DoS attacks. Mishra et al. [23] proposed an authentication method for TMIS utilizing biometrics and nonce. Still, Zhang et al. [24] elucidated that Mishra et al.'s scheme [23] is vulnerable to replay and MIMT attacks and can't achieve forward secrecy. Additionally, A three-factor authenticated TMIS technique based on chaotic maps was presented by Zhang et al. [24] to address the shortcomings of Mishra et al.'s scheme [23]. It is obvious that in schemes [9][22][24], users are at risk of identity and password guessing, as well as stolen smart-card attacks. Kim et al. [25] introduced a lightweight authentication method that ensures anonymity through biometric-based authentication, aiming at maintaining the freshness of the message requests' keys.

In 2020, Dharminder et al. [4] reviewed Radhakrishnan et al. [22] and found that it can't achieve unlinkability and resist identity-guessing attacks, password-guessing attacks and stolen smart-card attacks. Hence, an authentication method was presented by Dharminder et al. [4] to solve the weakness of the schemes in [9][22][24]. However, our findings indicate that Dharminder et al.'s scheme [4] fails to offer mutual authentication and untraceability and is susceptible to user and server impersonation attacks, as we will demonstrate later. In addition, Soni et al. [26] elucidated that Dharminder et al.'s scheme [4] requires more computation cost and communication overhead. To guarantee low latency, a mutual-authentication architecture based on fog computing for healthcare was described by Singh et al. [27] using RSA; however, Singh et al.'s scheme [27] is susceptible to Greatest Common Divisor (GCD)-based attacks.

In [28], an authentication protocol utilizing Radio Frequency Identification (RFID) and El-Gamal cryptosystem has been proposed to safeguard patients' privacy in TMIS. The scheme can defend against several types of attacks, including tag impersonation, location tracking, replay, de-synchronization and DoS attacks. In [29], Khan et al. introduced a lightweight RFID protocol to protect patient privacy by employing pseudonyms instead of real IDs. An anonymous patient monitoring system has been proposed by Amin et al. [30] using wireless medical-sensor networks. Still, Ali et al. [31] found that Amin et al.'s scheme [30] can't resist offline password-guessing attacks, user impersonation attacks and known session-key temporary-information attacks. Moreover, to solve these problems, Ali et al. [31] suggested a mechanism with three-factor authentication of Amin et al.'s scheme [30] using symmetric encryption and hash functions to offer authentication. However, Ali et al.'s scheme [31] can't offer perfect forward secrecy nor resist desynchronization attacks, guessing attacks and insider attacks.

Sharma et al. [32] recommend implementing an authentication scheme for cloud-based healthcare systems; however, Canetti and Krawczyk [33] demonstrated that Sharma et al.'s scheme [32] is susceptible to privileged insider attacks. Recently, an anonymity-preserving user-authentication method has been presented by Masud et al. [34] and its primary goal is to ensure users' anonymity, but Masud's scheme [34] can't achieve anonymity of communicating parties and can't resist insider attacks and desynchronization attacks. In [35], a three-factor authentication protocol has been proposed for consumer USB mass-storage devices. However, Renuka et al. [36] identified that the scheme in [35] fails to guarantee user anonymity and forward/backward secrecy and is unable to withstand session-specific temporary attacks. Moreover, the scheme in [35] can't enable users to change their password. Sonia et al. [37] suggested a remote authentication scheme consisting of three factors for a patient monitoring system; however, Sonia et al.'s plan [37] contains serious faults, according to Xu et al. [38], including sensor-node capture assault, lack of forward secrecy and loss of three-factor security. Table 1 summarizes the existing relevant schemes on the security of healthcare systems.

A lot of authors claimed that their schemes are secure and proved the security of their proposed schemes using various methods of formal security proof; however, Wang et al. [39]-[40], found that formal security proof has its limitations and for the formally proven schemes, informal analysis of these schemes proved their failure of offering several essential security requirements and their vulnerabilities to various types of attack. This is why we first reviewed and analyzed Dharminder et al.'s scheme and then proposed an improved authentication scheme to overcome the shortcomings of Dharminder et al.'s scheme.

3. PRELIMINARIES

This section will review some basic information essential for this study.

Table 1. Summary of related work.

Ref.	Countermeasure	Advantages	Disadvantages
[4]	RSA-based Authentication	Can achieve anonymity, forward secrecy and withstand replay attacks and man-in-middle attacks.	Can't achieve mutual authentication and untraceability and can't withstand impersonation attacks.
[9]	Chaotic-map Encryption	Can achieve mutual authentication and withstand DoS attacks, replay attacks and insider attacks.	Unable to offer user anonymity and susceptible to user impersonation attacks.
[10]	Identity-based Key Establishment	Can withstand offline password-guessing attacks and DoS attacks.	Can't resist man-in-the-middle attacks, replay attacks, insider attacks and stolen-verifier attacks.
[15]	Elliptic-curve Cryptography	Can offer mutual authentication while maintaining user anonymity.	Unable to withstand offline-password-guessing attacks, user/server impersonation attacks and MIMT attacks.
[16]	Elliptic-curve Cryptography	Capable of achieving mutual authentication and forward secrecy.	Can't resist insider attacks.
[19]	Hash Function and XOR Operation	Can achieve mutual authentication and reduce computation complexity and transmission costs.	Can't achieve anonymity and untraceability and can't resist impersonation attacks and sensor-node impersonation attacks.
[22]	Diffie-Hellman Problem and Hash Functions	Can offer mutual authentication, forward secrecy and guarantee user's anonymity and untraceability.	Can't resist man-in-the-middle attacks, replay attacks and insider attacks.
[30]	Hash Function and XOR Operation	Can offer mutual authentication and user's anonymity and untraceability.	Unable to withstand offline password-guessing attacks, user impersonation attacks and known session-key temporary-information attacks.
[32]	Hash Function and XOR Operation	Can offer mutual authentication, forward secrecy, data integrity and user's anonymity and untraceability.	Unable to withstand privileged inside attacks.
[34]	Hash Function and XOR Operation	Can provide mutual authentication and data privacy and resist against impersonation attacks, replay attacks and MIMT attacks.	Can't achieve anonymity of communicating parties and can't resist insider attacks and desynchronization attacks.
[36]	Fuzzy Extractor	Can offer mutual authentication, forward secrecy, anonymity, robustness to replay attacks and impersonation attacks.	Can't withstand offline password-guessing attacks, insider attacks and smart-card loss attacks.
[37]	Rabin Cryptosystem and Chaotic Maps	Can offer mutual authentication, forward secrecy and user's anonymity and untraceability.	Can't resist against sensor-node capture attacks and suffer lack of forward secrecy and loss of three-factor security.

3.1 RSA-based Cryptosystem

This sub-section will concentrate on revisiting the fundamental definition and properties of the RSA Cryptosystem [5].

First, choose two primes, p and q and compute the modulus $n = p \times q$ and the Euler totient function $\varphi(n) = (p-1) \times (q-1)$. To generate the public and private keys, select an integer e such that $gcd(e, \varphi(n)) = 1$ and compute $d \equiv e^{-1} \, mod \, \varphi(n)$. The public key is denoted by e, while the private key is represented by d. Encryption and decryption using the RSA algorithm are outlined as follows:

1. In RSA encryption, the sender encrypts a message m using the receiver's public key e as follows: $c = m^e \, mod \, n$.

2 . When the receiver receives the cipher text c, the receiver decrypts it using the receiver's private key d as $m = c^e \ mod \ n$.

3.2 Hash Function

An algorithm is a hash function that maps various-length inputs into a fixed-length output known as a hash value or hash code. The mathematical function condenses the message to a predetermined size and is a one-way function, making it challenging for attackers to recognize and decrypt the message. Cryptographic hashes come in various forms; for instance, SHA-1 (Secure Hash Algorithm 1) is a commonly utilized hash function that accepts an input and produces a 160-bit hash value. The cryptographic hash function provides many security services like message authentication and integrity and implements digital signatures. There are numerous aspects of the cryptographic hash function:

- Non-reversibility or one-way function: A strong hash should make reconstructing the original message from the hash output complex.
- Diffusion or avalanche effect: A half of the hash should also be changed if one piece of the original password is changed. In other words, if one bit in the initial message changes, the encrypted message should also vary.
- Determinism: The hash value or enciphered text generated by a particular message must always be the same.
- No collision: It would be difficult to find two distinct messages resulting in the same ciphertext.

4. THREAT MODEL

As the authentication protocol is executed over a public channel, the attacker utilizes several advantages or capabilities during the execution of the authentication protocol. We present some widely accepted assumptions as follows:

- An attacker can intercept, modify, delete and replay the exchanged messages.
- The smart card can be stolen or lost.
- An attacker may get the smart card and obtain its stored information.
- An attacker can run an impersonation attack if the user's password on the smart card is revealed.
- An attacker could be a legitimate user or a legitimate server.

5. REVIEW OF DHARMINDER ET AL.'S SCHEME

A review and cryptanalysis of Dharminder et al.'s scheme [4] are illustrated below.

5.1 Dharminder et al.'s Scheme

Three phases of the scheme introduced by Dharminder et al. are reviewed below:

- **Registration Phase**

User U_i registers to server S_j through secure channel as follows:

1.U_i selects Id_i and PW_i, generates a random b and computes $A_i = h(PW_i||b)$. Afterwards, U_i sends $\{Id_i, A_i\}$ to S_j.
2. After receiving $\{Id_i, A_i\}$, S_j calculates:

$Cid_i = h(Id_i||d)$

$A_u = Cid_i \oplus h(A_i||Id_i)$

$B_u = h(Cid_i||A_i||Id_i)$

3. U_i stores $\{A_u, B_u, e, n, h(.), h(Id_i \oplus PW_i) \oplus b\}$ in its Smart Card (SC).

- **Login Phase**

A legal U_i user starts in the login phase as:

1. U_i enters SC, inputs Id_i and PW_i and recovers $b = h(Id_i \oplus PW_i) \oplus b \oplus h(Id_i \oplus PW_i)$. Then, U_i calculates $A_i = h(PW_i||b)$ and $Cid_i = A_u \oplus h(A_i||Id_i)$ and verifies $h(Cid_i||A_i||Id_i) = B_u$.

2. After performing the verification, U_i selects a random $N_u \in Z_n^*$ and calculates $V_1 = h(A_i || Id_i)$, $V_2 = (Cid_i || A_u)^e$ and $V_3 = h(Cid_i || N_u || T_u)$, where T_u is the current timestamp of U_i.
3. Finally, U_i sends $M_1 = \{V_1, V_2, V_3, T_u\}$ to S_j.

- **Authentication Phase**

In the authentication phase, S_j obtains $M_1 = \{V_1, V_2, V_3, T_u\}$ from U_i and proceeds as follows:

1. S_j verifies T_u and computes:
   $V_2^d = (Cid_i || A_u)$
   $h(A_i || Id_i) = Cid_i \oplus A_u$
   $V_1 \oplus h(A_i \| Id_i) = N_u$
2. S_j calculates $V_3^* = h(Cid_i || T_u)$; if $V_3^* = V_3$, it authenticates U_i.
3. S_j selects a random value N_u and calculates:
   $SK_s = h(N_s || N_u || T_u || T_s)$
   $V_4 = N_s \oplus N_u \oplus Cid_i$
4. Then, S_j sends $M_2 = \{V_4, T_s\}$ to U_i.
5. After receiving $M_2 = \{V_4, T_s\}$, U_i verifies T_s and calculates $N_S^* = V_4 \oplus N_u \oplus Cid_i$.
6. Then, U_i computes $V_4^* = N_S^* \oplus N_u \oplus Cid_i$, if $V_4^* = V_4$, it authenticates S_j.
7. Finally, U_i computes a common session key $Sk_U = h(N_s || N_u || T_u || T_s) = Sk_s$.

5.2 Cryptanalysis of Dharminder et al.'s Scheme

Dharminder et al.'s scheme [4] will be analyzed below.

- **Impersonation Attack**

The attacker can potentially impersonate both the user and the server, as analyzed below**:**

User Impersonation Attack: To impersonate the user, let's consider that the adversary possesses the Id_i of a legal user U_i and also has access to the stored smart-card values $\{A_u, B_u, e, n, h(.), h(Id_i \oplus PW_i) \oplus b\}$, the adversary can generate any A_i^*, A_u^* and N_u^* and compute $Cid_i^* = A_u^* \oplus h(A_i^* || Id_i)$, $V_1^* = h(A_i^* || Id_i)$, $V_2^* = (Cid_i^* || A_u^*)^e$ and $V_3^* = h(Cid_i^* || N_u^* || T_u)$ and send $\{V_1^*, V_2^*, V_3^*, Tu\}$ to the sever.

After receiving $M_1 = V_1^*, V_2^*, V_3^*, Tu\}$, the server S_j computes:

$V_2^d = (Cid_i^* || A_u^*)$
$h(A_i^* || Id_i) = Cid_i^* \oplus A_u^*$
$N_u^* = V_1^* \oplus h(A_i^* || Id_i)$
$V_3' = h(Cid_i^* || N_u || T_u)$

It is evident that $V_3' = V_3^*$, which makes the server S_j identify the adversary as a legal user; hence, the adversary succeeds in masquerading to be a legal user.

Server Impersonation Attack: Furthermore, the attacker can act as the legal server by sending $M_2 = \{V_4^*, T_s\}$ and the user U_i will verify T_s and compute $N_S' = V_4^* \oplus N_u \oplus Cid_i$; then, U_i will compute V_4' as $V_4' = N_S^* \oplus N_u \oplus Cid_i$ using N_S^* which is computed using V_4^* itself; hence, for any value for V_4^*, when U_i compares V_4 with V_4', they will be equals and U_i will accept communication with the adversary as the legal server. Therefore, Dharminder et al.'s protocol can't achieve user authentication or server authentication.

- **Lack of Authentication**

Before providing any service, mutual authentication requires the server to verify the user and the user to authenticate the server. Furthermore, no illegal users or servers will be allowed to impersonate legal users or servers.

Mutual authentication can't be attained in Dharminder et al.'s scheme, leading to the problem of impersonating the user and the server. An attacker A can spoof a legal user and a legal server through the login process, as shown previously. This contradicts the process of mutual authentication for the scheme and limits the proper mutual authentication.

- **Lack of Unlinkability (Untraceability)**

Achieving linkability means that the attacker can't relate two messages to the same user; hence, the user's activities can't be traced in the system. As we see in equation $V_2 = (Cid_i||A_u)^e$, the value of V_2 is constant as it consists of Cid_i and A_u, which are encrypted by the fixed public key e and Cid_i consists of $\{A_u, A_i, Id_i, PW_i, b\}$ which are also fixed values. Consequently, V_2 remains constant in every communication applied between both the user and the server, allowing the equation mentioned above to link the user's message, thereby simplifying user identification.

6. THE PROPOSED SCHEME

To solve the flaws of Dharminder et al.'s scheme [4], we introduce a new authentication scheme utilizing RSA encryption with initialization to generate the server's public and private keys, as discussed in Section 2. The notations required for the proposed scheme are depicted in Table 2. The proposed scheme comprises four phases outlined as follows:

Table 2. Index of notations.

Notation	Interpretation
Id_i	User *i* identity
PW_i	User *i* password
e	Server *j* public key
d	Server *j* private key
T_u, T_s	Fresh timestamps
Sk_s	Session key computed by server *j*
S_{KU}	Session key computed by user *i*
$h(.)$	Hash function
$\oplus$	XOR operation
$\|\|$	Concatenation operation

6.1 Registration Phase

User U_i registers to server S_j through a secure channel as described in Figure 1 and illustrated as follows:

1. U_i selects Id_i, PW_i and a random number b, calculates $A_i = h(PW_i \,||\, b)$ and sends $\{Id_i, A_i\}$ to S_j.

2. S_j Computes:
$Cid_i = h(Id_i||d)$
$A_u = Cid_i \oplus h(A_i||Id_i)$
$B_u = h(Cid_i||A_i||Id_i)$
2. Then, S_j stores (Id_i, A_i, Cid_i) in its database and stores $\{A_u, B_u, e, n, h(.), h(Id_i \oplus PW_i) \oplus b\}$ in SC for U_i.

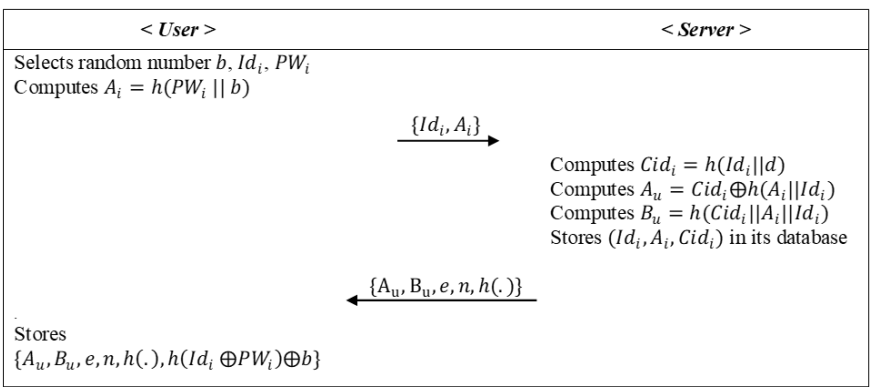


Figure 1. Registration phase.

6.2 Login Phase

In the login phase (described in Figure 2), a valid U_i progresses as follows:

1. First, a legal user U_i enters Sc, inputs Id_i and PW_i and recovers $b = h(Id_i \oplus PW_i) \oplus b \oplus h(Id_i \oplus PW_i)$.
2. U_i calculates $A_i = h(PW_i||b)$, $Cid_i = A_u \oplus h(A_i||Id_i)$ and performs verification for $h(Cid_i||A_i||Id_i) = B_u$.
3. After verification, U_i chooses a random number $N_u \in Z_n^*$ and calculates $V_1 = (Id_i||Cid_i||A_u||N_u||T_u)^e$, where T_u is the current timestamp of U_i.
4. Then, U_i Sends $M_1 = \{V_1, T_u\}$ to S_j.

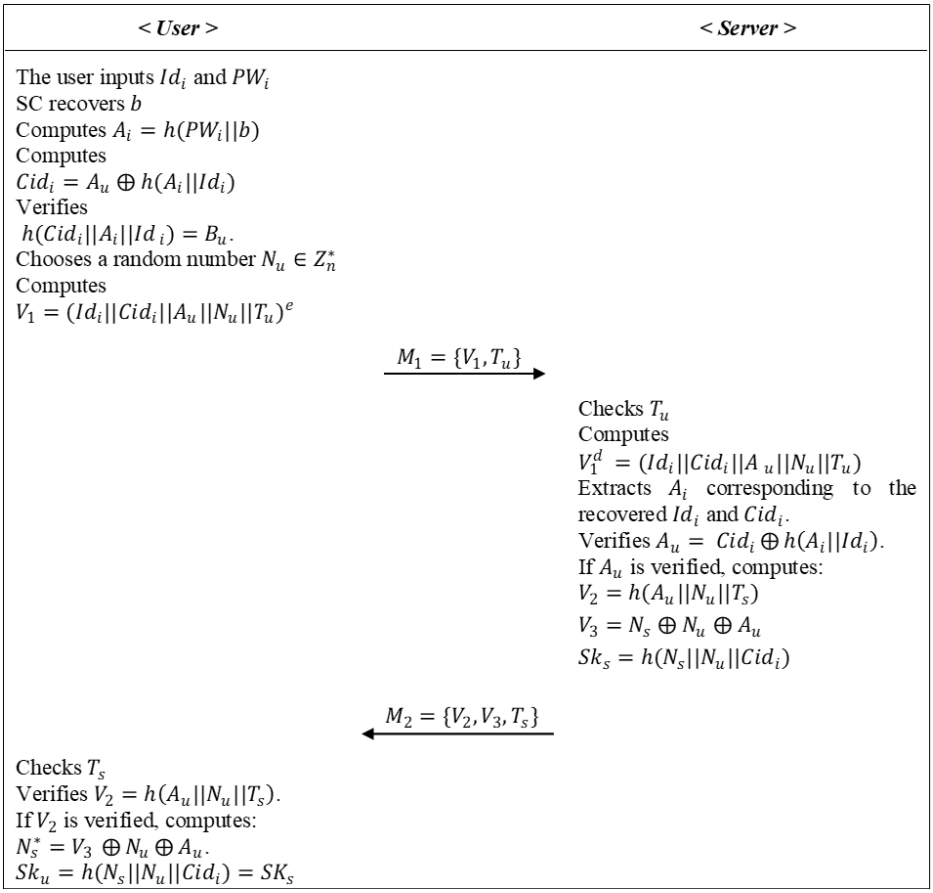


Figure 2. Login and authentication process.

6.3 Authentication Phase

Valid U_i proceeds in the authentication phase as follows:

1. Upon receiving $M_1 = \{V_1, T_u\}$ from U_i , S_j verifies T_u and calculates $V_1^d = (Id_i||Cid_i||A_u||N_u||T_u)$, extracts A_i corresponding to the recovered Id_i and Cid_i and verifies $A_u = Cid_i \oplus h(A_i||Id_i)$.
2. If A_u is verified, S_j computes:

   $V_2 = h(A_u||N_u||T_s)$
   $V_3 = N_s \oplus N_u \oplus A_u$
   $Sk_s = h(N_s||N_u||Cid_i)$
   Finally, S_j sends $M_2 = \{V_2, V_3, T_s\}$ to U_i.
3. Upon receiving $M_2 = \{V_2, V_3, T_s\}$, U_i checks T_s and verifies $V_2 = h(A_u||N_u||T_s)$.
4. If V_2 is verified, U_i computes $N_s^* = V_4 \oplus N_u \oplus A_u$.
5. Finally, U_i computes the session key $Sk_u = h(N_s||N_u||Cid_i) = SK_s$.

6.4 Password-change Phase

A legal user inputs Id_i, PW_i and a new password PW_i^{new}. Then, SC calculates $A_i = h(PW_i||b)$, $Cid_i = A_u \oplus h(A_i||Id_i)$ and verifies $h(Cid_i||A_i||Id_i) = B_u$. If verified, SC calculates the values $A_i^{new} = h(PW_i^{new}||b)$, $cid_i^{new} = A_u \oplus h(A_i^{new}||Id_i)$ and $h(cid_i^{new}||A_i^{new}||Id_i) = B_u^{new}$ and replaces B_u with B_u^{new}.

7. SECURITY ANALYSIS OF THE PROPOSED SCHEME

Here, we will conduct a brief security analysis of the proposed scheme. Furthermore, we utilized AVISPA [41] to simulate the proposed method for formal security verification.

7.1 Informal Security Analysis of the Proposed Scheme

This sub-section presents an informal security analysis of the proposed scheme. First, we will set up the achievement of the fundamental security needs and the suggested scheme's resistance to various forms of attacks in healthcare systems, as user impersonation, server impersonation, secure session key, insider attacks, stolen smart-card attacks and offline password-guessing attacks.

Mutual Authentication: When both the user and the server can authenticate each other successfully, the system is considered to have achieved mutual authentication.

On the server side, S_j computes $V_1^d = (Id_i||Cid_i||A_u||N_u||T_u)$ using its private key d to extract A_i correlating to the recovered Id_i and Cid_i and determines whether or not A_i already exists in its database from the registration phase and checks if $A_u = Cid_i \oplus h(A_i||Id_i)$. Suppose the extracted A_i exists in the server's database and A_u is verified. In that case, U_i is authenticated to S_j as only user U_i can compute valid Cid_i, which depends on A_i calculated using the user's password PW_i.

On the user side, U_i checks whether $V_2 = h(A_u||N_u||T_s)$. If V_2 is verified, S_j is then authenticated to U_i, as only server S_j can extract A_u and N_u from the transmitted message V_1 by decrypting V_1 using the server's private key d, known only to the server S_j.

Consequently, the proposed scheme enables the exchange of mutually-authenticated information between the server and the user. Hence, the suggested scheme can achieve mutual authentication.

User Anonymity: The proposed scheme can maintain the user's anonymity by concealing the user's identity, Id_i by concatenating it with the random number N_u and hiding it in $V_1 = (Id_i||Cid_i||A_u||N_u||T_u)^e$; therefore, the attacker is unable to access or obtain the user's identity Id_i as it requires the attacker to decrypt V_1 using the server's private key d. Additionally, the user's identity Id_i is not included in (V_2, V_3) in the transmitted message $M_2 = \{V_2, V_3, T_s\}$ from the server S_j.

User Untraceability: The proposed scheme can ensure the user's untraceability, as the user and server generate new random values (N_u, N_s) and fresh timestamps (T_u, T_s) in each session to compute $V_1 = (Id_i||Cid_i||A_u||N_u||T_u)^e$, $V_2 = h(A_u||N_u||T_s)$ and $V_3 = N_s \oplus N_u \oplus A_u$. As a result, in each session, the transmitted messages $M_1 = \{V_1, T_u\}$ and $M_2 = \{V_2, V_3, T_s\}$ provide new values which prohibit the attacker from relating the sent messages from the same user to each other or tracing its activities.

Forward Secrecy: Let's suppose that the attacker can acquire both the user's password PW_i-and the shared key between the user and the server $Sk_u = h(N_s||N_u||Cid_i) = SK_s$ and can intercept the transmitted message $M_1 = \{V_1, T_u\}$. Even in this situation, without knowledge of the random numbers (N_u, N_s), the attacker cannot disclose the previously transmitted messages. Furthermore, due to the RSA factorization issue, calculating the private key d of the server is difficult.

Secure Session Key: In this plan, the shared session key $Sk_u = h(N_s||N_u||Cid_i) = SK_s$ is computed based on random numbers N_u and N_s generated by the user and the server. Hence, the attacker can obtain the session key if he can retrieve the values of N_u and N_s; this needs disclosing the server's private key d, which is very hard due to the computational difficulty of factoring large integers. As a result, the session key in the proposed scheme is highly secure.

User Impersonation Attack: The attacker must create a legitimate value V_1 to successfully impersonate a legitimate user. For example, in order to succeed in generating $V_1 = (Id_i||Cid_i||A_u||N_u||T_u)^e$, it

requires the attacker to successfully compute $A_i = h(PW_i||b)$ and $Cid_i = A_u \oplus h(A_i||Id_i)$. This is because Cid_i depends on A_i, calculated using the password PW_i which is known only to the user U_i. Therefore, the proposed scheme is devised to mitigate user impersonation attacks.

Impersonation Attack: For the attacker to impersonate the server successfully, specific actions must be taken; it requires the attacker to generate valid values of V_2 and V_3. For instance, achieving $V_2 = h(A_u||N_u||T_s)$ and $V_3 = N_s \oplus N_u \oplus A_u$ necessitates the attacker obtaining A_u and N_u by decrypting V_1 using the server's private key d, which is exclusively known to the server S_j. However, it is known that getting the private key d of the server is a challenging process due to the factorization problem in RSA. Therefore, the suggested scheme can effectively prevent server impersonation attacks.

Replay Attack: The timestamps of the user and server (T_u, T_s) are inserted in each transmitted message. Additionally, with each login message $M_1 = \{V_1, T_u\}$, the user U_i selects a fresh random number N_u to provide $V_1 = (Id_i||Cid_i||A_u||N_u||T_u)^e$ and the server selects a fresh random number N_s to calculate $V_3 = N_s \oplus N_u \oplus A_u$. Moreover, these random numbers are used to compute the session key $Sk_u = h(N_s||N_u||Cid_i) = SK_s$. Hence, the proposed scheme can resist replay attacks.

Insider Attack: A malicious insider cannot obtain a server's password, since U_i sends $h(PW_i \,||\, b)$ instead of PW_i to server S_j through a secure channel. Hence, a malicious insider cannot acquire the PW_i of user U_i due to the one-way property of the hash function.

Stolen Smart-card Attack: As the smart card typically does not store A_i, the attacker can't elicit Id_i from Cid_i or PW_i from A_i. Nevertheless, the message V_1 is randomized using the random number N_u and protected by encrypting V_1 utilizing the server's public key e. Additionally, the identity is protected by hashing the identity Id_i concatenated with the private key d of the server, as $(Cid_i = h(Id_i||d))$.

Offline Password Guessing: To guess PW_i, we need to know A_i, which is not saved on the smart card and if the attacker is aware of the values that the user has saved on his smart card $\{A_u, B_u, e, n, h(.), h(Id_i \oplus PW_i) \oplus b\}$, the password can't be guessed, as the hash and xor functions protect the password. Additionally, the attacker can't obtain the password from the transmitted messages over the public channel, as (V_1, V_2, V_3) no longer include A_i. Therefore, the password is protected.

7.2 Simulation for Conducting Formal Security Verification Using AVISPA Tool

In this sub-section, the security properties of the proposed system will be verified using the AVISPA [41] simulation tool. The simulation is carried out using a language for high-level protocol specifications (HLPSL). The user's role is defined in HLPSL (High Level Protocol Specification Language) specification, as depicted in Figure 3. The user and server communicate by transmitting and receiving messages through channels using the Snd() and Rcv() operations. A channel(dy) type declaration indicates that the channel is designed for the DolevYao threat model. Lastly, Ui authenticates the server S, as per the declaration request (Ui, S, user_server_nu, Nu). Additionally, the server's role is specified in the HLPSL specification, as illustrated in Figure 4. The role specifications for session, goal and environment in HLPSL are depicted in Figure 5.

The simulation results are presented using the back-ends Constraint-logic-based Attack Searcher (CL-AtSe) and On-the-Fly Model-Checker (OFMC) [42]. Figures 6 and 7 validate that under both back-ends, CL-AtSe (Constraint-logic-based Attack Searcher) and OFMC (On-the-fly Model-Checker), the suggested scheme is deemed SAFE. This suggests that the proposed scheme is capable of resisting both passive and active attacks, including replay and MITM attacks.

8. PERFORMANCE EVALUATION

In this section, the security and performance of the proposed scheme are compared to those of other relevant schemes, including those proposed by Chaturvedi et al. [10], Qiu et al. [16], Radhakrishnan et al. [22], Renuka et al. [36] and Dharminder et al. [4]. Table 3 illustrates the estimated computation time [43] for each cryptographic function, facilitating performance comparison.

The process of logging in and authenticating is a critical phase of any authentication scheme. Therefore, we will focus on this phase. The proposed scheme relies on several hash functions, H, modular exponentiation for encryption/decryption using RSA, elliptic-curve point multiplication and a simple

```
role user(Ui:agent,S:agent,H,Mul:hash_func,SKas:symmetric_key,
SND,RCV:channel(dy))
played_by Ui
def=
local State:nat,
IDi,Ai,PWi,Cidi,Au,Bu,E,N,B,Nu,Ns,P,SKu,SKs,Q,D: text,
V1,V2,V3,Tu,Ts: message,
Inc: hash_func
const user_server, server_user,subs_1,subs_2,subs_3 : protocol_id
init State := 0
transition
1. State=0 /\ RCV(start)=|>
State':= 1 /\ B':=new() /\ Ai':=H(PWi.B) /\ SND({IDi.Ai}_SKas)
2. State=1 /\ RCV({Au'.Bu'.E'.N'.xor((H(xor(IDi,PWi))),B)}_SKas)
=|>
State':=2 /\ Nu':= new() /\ Tu':= new()/\ Ai':=H(PWi.B)
/\ Cidi':=xor(Au',H(Ai'.IDi))
/\ secret(PWi,subs_1,Ui)
/\ V1':=exp((IDi.Cidi'.Au'.Nu.Tu'),E)
/\ SND(V1'.Tu')
/\ witness(Ui,S,user_server_nu,Nu)
3. State= 2 /\ RCV(V2.V3.Ts)=|>
State':= 3 /\ Ns':=xor(V3,Nu,Au) /\ request(Ui,S,user_server_nu,Nu)
/\ SKu':=H(Ns.Nu.Cidi)
end role
```

Figure 3. Role of user Ui in HLPSL.

```
role server(Ui:agent,S:agent,H,Mul:hash_func,SKas:symmetric_key,
SND,RCV:channel(dy))
played_by S
def=
local State:nat,
IDi,Ai,PWi,Cidi,Au,Bu,E,N,B,Nu,Ns,P,SKu,SKs,Q,D: text,
V1,V2,V3,Tu,Ts: message,
Inc: hash_func
const user_server, server_user,subs_1,subs_2,subs_3 : protocol_id
init State := 0
transition
1. State=0 /\ RCV({IDi'.Ai'}_SKas) =|> State':=1 /\ P':=new() /\
Q':=new() /\ E':=new() /\ N':=Mul(P,Q) /\ Cidi':= h(IDi'.D) /\
Ts':=new()
/\ Au':=xor(Cidi,H(Ai)) /\ Bu':=H(cidi.Ai.IDi')
/\ secret(D,subs_2,S) /\ secret({Ai,Cidi,E,IDi,Ns},subs_3,{Ui,S})
/\ SND({Au'.Bu'.E'.N'}_SKas)
2. State=1 /\ RCV(V1,Tu) =|> State':=2 /\ IDi' :=exp(V1,D) /\
Au':=exp(V1,D) /\ Cidi':=exp(V1,D) /\ Nu':=exp(V1,D)
/\ witness(S,Ui,server_user_ns,Ns) /\ V2':=H(Au'.Nu'.Ts) /\
V3':=xor(Ns,Nu',Au') /\ SKs':=H(Ns.Nu'.Cidi')
end role
```

Figure 4. Role of server S in HLPSL.

```
role session(Ui:agent,S:agent,H,Mul:hash_func,SKas:symmetric_key)
def=
local SI,SJ,RI,RJ: channel(dy)
composition
user(Ui,S,H,Mul,SKas,SI,RI) /\ server(Ui,S,H,Mul,SKas,SJ,RJ)
end role
role environment()
def=
const ui:agent,s:agent,h,mul:hash_func,skas:symmetric_key,
v1,v2,v3,idi,ai,cidi,au,bu,e,n,p,q,sku,sks,b: text,
user_server_nu, server_user_ns, subs_1,subs_2,subs_3 : protocol_id
intruder_knowledge ={ui,s,h,mul,v1,v2,v3,idi,ai,cidi,au,bu,e,n}
composition
session(ui,s,h,mul,skas) /\ session(ui,s,h,mul,skas)
end role
goal
secrecy_of subs_1
secrecy_of subs_2
secrecy_of subs_3
authentication_on user_server_nu
authentication_on server_user_ns
end role
```

Figure 5. The proposed HLPSL scheme's role specifications for the session, goal and environment.

```
%OFMC
%Version of 2006/02/13
SUMMARY
 SAFE
DETAILS
 BOUNDED_NUMBER_OF_SESSIONS
PROTOCOL
/ home/span/span/testsuite/results/Tmis.if
GOAL
 as_specified
BACKEND
 OFMC
COMMENTS
STATISTICS
 parseTime: 0.00s
 searchTime: 0.03s
 visitedNodes: 13 nodes
 depth: 4 plies
```

Figure 6. The proposed scheme simulation results based on the OFMC back-end.

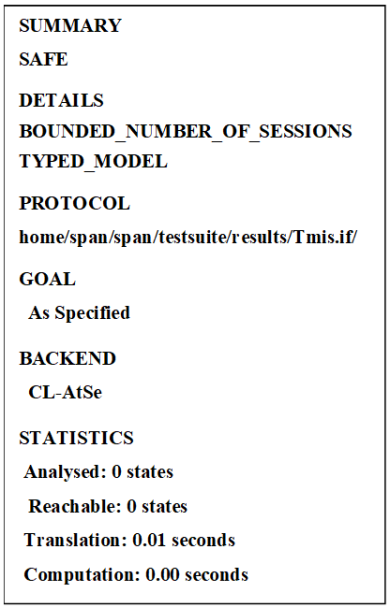


Figure 7. The proposed scheme's outputs based on the CL-AtSe back-end.

XOR function. However, compared to other cryptographic operations, the cost of concatenation (||) and XOR ($\oplus$) operations is negligible.

The scheme is convenient for low-power devices, as the cryptographic techniques are straightforward. In addition, the proposed scheme employs SHA-1 as a hash function for computation-cost evaluation. The SHA-1 algorithm receives messages of varying sizes divided into blocks; each block has an input of 512 bits and provides an output of 160 bits.

Table 3. Notations and approximate computation times.

Notation	Description	Approximate computation time (in seconds)
t_h	Time of hash function	0.00032
t_{me}	Time of modular exponentiation	0.0192
t_{ecm}	Time of elliptic-curve point multiplication	0.0172

8.1 Computation Cost

The computational expense of implementing the proposed scheme is detailed in Table 4.

Table 4. The proposed scheme's computation cost.

Computation Cost	User	Server	Total
Registration phase	$1t_h$	$3t_h$	$4t_h$
(ms)	0.32	0.96	1.28
Login and Authentication phase	$5t_h + 1t_{me}$	$3t_h + 1t_{me}$	$8t_h + 2t_{me}$
(ms)	20.8	20.16	40.96
Total computation cost	$6t_h + 1t_{me}$	$6t_h + 1t_{me}$	$12t_h + 2t_{me}$
(ms)	21.12	21.12	42.24

For user U_i : In the registration phase, user U_i executes one hash function, which takes 0.32 ms. Throughout the login and authentication phase, the user and server perform 5 hash functions and one modular exponentiation which require 20.8 ms. Thus, the user consumes a total of 21.12 ms.

For server S_j : In the registration phase, the server S_j executes 3 hash functions which need 0.96 ms. While in login and authentication, the server performs 3 hash functions and 1 modular exponentiation

which require 20.16 ms. Hence, the server consumes a total of 21.12 ms.

As depicted in Table 4, the total time needed to complete the registration phase is 1.28 ms, while the real-time needed to complete the login and authentication phase is 40.96 ms.

8.2 Security and Performance Assessment

In this sub-section, we compare the security and performance of our proposed scheme with those of recent existing schemes. Performance is evaluated based on communication and computational costs.

8.2.1 Security Comparison

For security comparison, we put forward a list of 10 independent criteria of security evaluation as follows:

- SEC1. Mutual authentication: The server must verify the user's identity before offering any service and the user should verify the server's identity to trust and accept the offered services.
- SEC2: User anonymity: The user must act or communicate in the system without stating his name or identity.
- SEC3: User untraceability: The attacker couldn't be able to relate any two messages to the same user; hence, the user's activities in the system can't be traced.
- SEC4: Forward secrecy: Even if one or more of the generated session keys are compromised, the data from other sessions is protected.
- SEC5: Session-key agreement: The user and server can agree on a shared session-key to ensure secure data transmission.
- SEC6: Resistance to impersonation attack: The scheme prevents the attacker from joining and/or accessing the system as a legitimate user or server.
- SEC7: Resistance to replay attack: The scheme prevents the attacker from retransmitting previously transmitted messages as a legitimate user or server.
- SEC8: Resistance to insider attack: The attack can be executed by an authorized user to access the system. However, when changing or obtaining the user's information from the server, the scheme prevents the attacker from compromising both the user's privacy and the integrity of the system.
- SEC9: Resistance to stolen smart-card attack: If the attacker captures the user's stolen smart card, the scheme prevents it from recovering the password or impersonating the user.
- SEC10: Resistance to offline password-guessing attack: The scheme prevents the attacker from guessing the password.

Table 5 presents a comparison between the proposed scheme and other related schemes, focusing on key aspects of online privacy and security, such as authentication, anonymity, untraceability and forward secrecy in healthcare systems. In Table 5, the symbols "√" and "X" indicate whether the scheme meets the relevant security features or not.

Table 5. Security comparison.

Scheme	Chaturvedi et al. [10]	Qiu et al. [16]	Radhakrihnan et al. [22]	Renuka et al. [36]	Dharminder et al. [4]	The Proposed Scheme
SEC1	√	√	√	√	×	√
SEC2	√	√	√	√	√	√
SEC3	√	√	√	√	×	√
SEC4	√	√	√	√	√	√
SEC5	√	√	√	√	√	√
SEC6	√	√	√	√	×	√
SEC7	×	√	×	√	√	√
SEC8	×	×	×	×	√	√
SEC9	√	√	√	×	√	√
SEC10	√	√	√	×	√	√

8.2.2 Computation-cost Comparison

Table 6 displays the computational cost of the proposed scheme in comparison with those of the relevant authentication schemes. The results indicate that the login and authentication phase of the proposed scheme requires $8t_h + 2t_{me}$, amounting to 40.96 ms. Consequently, it is suggested that the computational expense of the proposed scheme is lower than Chaturvedi et al.'s scheme [10] and Renuka et al.'s scheme [36], while it shares the same computation cost as Dharminder et al.'s scheme [4].

The proposed scheme exhibits a slightly higher computation cost compared to Qiu et al.'s scheme [16]. However, this slight increase is justified, as the proposed scheme is capable of resisting insider attacks, a vulnerability present in Qiu et al.'s scheme [16]. In contrast to Radhakrishnan et al.'s scheme [22], the proposed scheme incurs a higher cost, which is reasonable, given that the proposed scheme can defend against several types of attacks, including replay attacks, MITM assaults, stolen-verifier attacks and insider attacks.

Table 6. Computation-cost comparison.

Schemes	User computation cost	Server computation cost	Total cost
Chaturvedi et al. [10] (ms)	$7t_h + 1t_{me}$ 21.44	$4t_h + 1t_{me}$ 20.48	$11t_h + 2t_{me}$ 41.92
Qiu et al. [16] (ms)	$6t_h + 1t_{ecm}$ 19.12	$5t_h + 1t_{ecm}$ 18.8	$11t_h + 2t_{ecm}$ 37.92
Radhakrihnan et al. [22] (ms)	$9t_h$ 2.88	$7t_h + 1t_{me}$ 21.44	$16t_h + 1t_{me}$ 24.32
Renuka et al. [36] (ms)	$7t_h + 3t_{ecm}$ 53.84	$5t_h + 3t_{ecm}$ 53.2	$12t_h + 6t_{ecm}$ 107.04
Dharminder et al. [4] (ms)	$6t_h + 1t_{me}$ 21.12	$2t_h + 1t_{me}$ 19.84	$8t_h + 2t_{me}$ 40.96
The Proposed Scheme (ms)	$5t_h + 1t_{me}$ 20.8	$3\ t_h + 1t_{me}$ 20.16	$8\ t_h + 2t_{me}$ 40.96

8.2.3 Communication-cost Comparison

Let's consider the bit sizes as follows: 160 bits for a random number and the user's identity, 32 bits for the timestamp, 320 bits for the elliptic-curve point and 160 bits for the hash output (if SHA_1 is applied as h(·)). İn the proposed scheme, the message $M_1 = \{V_1, T_u\}$ needs 1024+32 = 1056 bits and the message $M_2 = \{V_2, V_3, T_s\}$ needs 160+160+32=352 bits. Thus, the proposed scheme requires a communication cost of 1056 + 352 = 1408 bits for the two messages sent between the user U_i and the sender S_j .

Table 7 demonstrates that the proposed scheme's and other related systems' communication costs are comparable. For example, the schemes in [10], [16], [22], [36] and [4] require 1248, 1280, 1888, 1184 and 1568 bits, respectively, whereas the proposed scheme takes 1408 bits. Furthermore, compared to Chaturvedi et al.'s scheme [10], Qiu et al.'s scheme [16] and Renuka et al.'s scheme [36], the communication cost of the proposed scheme is a little increased; the suggested scheme can meet all security standards, whereas the other schemes can't, hence this is justifiable.

Table 7. Communication-cost comparison.

Scheme	Communication Cost	
	No. of Messages	Cost (in bits)
Chaturved et al. [10]	2	1248
Qiu et al. [16]	3	1280
Radhakrihnan et al. [22]	2	1888
Renuka et al. [36]	2	1184
Dharminder et al. [4]	2	1568
The Proposed Scheme	2	1408

Based on the mentioned results and analysis, the proposed scheme provides enhanced security services and resilience against attacks compared to related schemes in [10][16][22][36][4]. Additionally, the proposed scheme demonstrates a comparable computational cost to other related schemes which lack the security services achieved by our proposed scheme. These advantages position the proposed scheme as a more suitable choice for healthcare systems.

From the results mentioned above and the analysis, the proposed scheme achieved more security services and resisted more attacks than the related schemes in [10][16][22][36][4]. Moreover, the proposed scheme has a comparable computation cost to other related schemes which failed to achieve the security services offered by our proposed scheme. These advantages enhance the suitability of the proposed scheme for healthcare systems.

9. CONCLUSIONS

In this paper, we conducted a thorough review and analysis of Dharminder et al.'s scheme, revealing its inability to ensure authentication, user untraceability and vulnerability to both user-impersonation attacks and server impersonation attacks. We developed an improved RSA-based authentication scheme to address these critical security flaws and provide authorized access to healthcare services. The proposed scheme offers mutual authentication, user anonymity, untraceability and forward secrecy. Additionally, it is resilient against a range of attacks, including stolen smart-card attacks, user and server impersonation attacks, password-guessing attacks, insider attacks and offline password-guessing attacks. The security analysis of the proposed scheme, supported by simulations using the AVISPA tool, confirms the proposed scheme's superior security over existing schemes. Additionally, through performance evaluation, we illustrated that the proposed scheme requires 61.7% and 2.3% lower computation costs than Renuka et al.'s scheme and Chaturvedi et al.'s scheme, respectively. This research not only addressed vital security gaps in the authentication schemes of healthcare systems, but also introduced an improved and secure scheme that is both computationally and communicationally more efficient. The proposed improvement is pivotal for the secure and effective implementation of healthcare systems, ensuring the protection of sensitive patient information and facilitating the reliable delivery of medical services.

REFERENCES

[1] J. Srinivas, D. Mishra and S. Mukhopadhyay, "A Mutual Authentication Framework for Wireless Medical Sensor Networks," J. of Medical Systems, vol. 41, pp. 1-10, 2017.

[2] D. Mishra et al., "Cryptanalysis and Improvement of Yan et al.'s Biometric-based Authentication Scheme for Telecare Medicine Information Systems," J. of Medical Systems, vol. 38, pp. 1-12, 2014.

[3] A. Kumar et al.,"A Novel Privacy-preserving Blockchain-based Secure Storage Framework for Electronic Health Records," Journal of Information and Optimization Sciences, vol. 43, pp. 549-570, 2022.

[4] D. Dharminder, D. Mishra and X. Li, "Construction of RSA-based Authentication Scheme in Authorized Access to Healthcare Services," J. of Medical Systems, vol. 44, pp. 1-9, 2020.

[5] R.L. Rivest, A. Shamir and L. Adleman, "A Method for Obtaining Digital Signatures and Public-key Cryptosystems," Communications of the ACM, vol. 21, pp. 120-126, 1987.

[6] C.T. Li et al., "A Secure Dynamic Identity and Chaotic Maps-based User Authentication and Key Agreement Scheme for E-Health Care Systems," J. of Medical Systems, vol. 40, pp. 1-10, 2016.

[7] R. Madhusudhan and C. S. Nayak, "A Robust Authentication Scheme for Telecare Medical Information Systems," Multimedia Tools and Applications, vol. 78, pp. 15255-15273, 2019.

[8] M. Tanveer et al., "CADF-CSE: Chaotic Map-based Authenticated Data Access/Sharing Framework for IoT-enabled Cloud Storage Environment," Physical Communication, vol. 59, p. 102087, 2023.

[9] M. Tanveer et al., "CMAP-IoT: Chaotic Map-based Authentication Protocol for Crowdsourcing Internet of Things," Arabian J. for Science and Engineering, vol. 49, pp. 3453–3466, 2024.

[10] A. Chaturvedi, D. Mishra and S. Mukhopadhyay, "An Enhanced Dynamic ID-based Authentication Scheme for Telecare Medical Information Systems," Computer and Information Sciences, vol. 29, pp. 54-62, 2017.

[11] Q. Jiang, J. Ma, Z. Ma and G. Li, "A Privacy-enhanced Authentication Scheme for Telecare Medical Information System," Journal of Medical Systems, vol. 37, pp. 1-8, 2013.

[12] D. Mishra, "On the Security Flaws in ID-based Password Authentication Schemes for Telecare Medical Information Systems," J. of Medical Systems, vol. 39, pp. 1-16, 2015.

[13] X. Xu, P. Zhu and Q. Wen, "A Secure and Efficient Authentication and Key Agreement Scheme Based on ECC for Telecare Medicine Information Systems," J. of Medical Systems, vol. 38, pp. 1-17, 2014.

[14] Z. Zhu, "An Efficient Authentication Scheme for Telecare Medicine Information Systems," J. of Medical

Systems, vol. 36, pp. 3833-3838, 2012.
[15] S. A. Chaudhry, K. Mahmoud, H. Naqvi and M. K. Khan, "An Improved and Secure Biometric Authentication Scheme for Telecare Medicine Information Systems Based on Elliptic Curve Cryptography," J. of Medical Systems, vol. 39, pp. 1-11, 2015.
[16] S. Qiu, G. XU, H. Ahmad and L. Wang, "A Robust Mutual Authentication Scheme Based on Elliptic Curve Cryptography for Telecare Medical Information Systems," IEEE Access, vol. 6, pp. 7452-7463, 2017.
[17] X. Li, M. H. Ibrahim, S. Kumari, A. K. Sangaiah and V. Gupta, "Anonymous Mutual Authentication and Key Agreement Scheme for Wearable Sensors in Wireless Body Area Networks," Computer Networks, vol. 129, pp. 429-443, 2017.
[18] A. M. Koya and P. P. Deepthi, "Anonymous Hybrid Mutual Authentication and Key Agreement Scheme for Wireless Body Area Network," Computer Networks, vol. 140, pp. 138-151, 2018.
[19] M. Komparaa, SK.H. Islam and M. Hölbl, "A Robust and Efficient Mutual Authentication and Key Agreement Scheme with Untraceability for WBANs," Computer Networks, vol. 148, pp. 196-213, 2019.
[20] Z. U. Rehman, S. Altaf and S. Iqbal, "An Efficient Lightweight Key Agreement and Authentication Scheme for WBAN," IEEE Access, vol. 8, pp. 175385-175397, 2020.
[21] T. F. Lee et al., "Secure and Efficient Password-based User Authentication Scheme Using Smart Cards for the Integrated EPR Information System," J. of Medical Systems, vol. 37, pp. 1-7, 2013.
[22] N. Radhakrishnan and M. Karuppiah, "An Efficient and Secure Remote User Mutual Authentication Scheme Using Smart Cards for Telecare Medical Information Systems," Informatics in Medicine Unlocked, vol. 16, p. 100092, 2019.
[23] D. Mishra et al., "Security Enhancement of a Biometric-based Authentication Scheme for Telecare Medicine Information Systems with Nonce," J. of Medical Systems, vol. 38, pp. 1-11, 2012.
[24] L. Zhang, S. Zhu and S. Tang, "Privacy Protection for Telecare Medicine Information Systems Using a Chaotic Map-based Three-factor Authenticated Key Agreement Scheme," IEEE J. of Biomedical and Health Informatics, vol. 21, pp. 465-475, 2016.
[25] K. Kim, J. Ryu, Y. Lee and D. Won, "An Improved Lightweight User Authentication for the Internet of Medical Things," Sensors, vol. 23, no. 3, p. 1122, 2023.
[26] M. Soni and D. K. Singh, "Privacy-preserving Secure and Low-cost Medical Data Communication Scheme for Smart Healthcare," Computer Communications, vol. 194, pp. 292-300, 2022.
[27] S. Singh and V. K. Chaurasiya, "Mutual Authentication Framework Using Fog Computing in Healthcare," Multimedia Tools and Applications, vol. 81, pp. 31977-32003, 2022.
[28] F. Salem and R. Amin, "A Privacy-preserving RFID Authentication Protocol Based on El-Gamal Cryptosystem for Secure TMIS," Information Sciences, vol. 527, pp. 382-393, 2020.
[29] M. A. Khan, S. Ullah, T. Ahmad, K. Jawad and A. Buriro, "Enhancing Security and Privacy in Healthcare Systems Using a Lightweight RFID Protocol," Sensors (Basel), vol. 23, no. 12, p. 5518, 2023.
[30] R. Amin et al., "A Robust and Anonymous Patient Monitoring System Using Wireless Medical Sensor Networks," Future Generation Computer Systems, vol. 80, pp. 483-495, 2018.
[31] R. Ali, A. K. Pal, S. Kumari, A. K. Sangaiah, X. Li and F. Wu, "An Enhanced Three-factor Based Authentication Protocol Using Wireless Medical Sensor Networks for Healthcare Monitoring," J. of Ambient Intelligence and Humanized Computing, vol. 15, pp. 1165-1186, 2018.
[32] G. Sharma and S. Kalra, "A Lightweight User Authentication Scheme for Cloud-IoT Based Healthcare Services," Iranian J. of Science and Technology Trans. of Electrical Eng., vol. 43, pp. 619–636, 2019.
[33] R. Canetti and H. Krawczyk, "Universally Composable Notions of Key Exchange and Secure Channels," Proc. of Advances in Cryptology, Part of the Book Series: Lecture Notes in Computer Science, vol. 2332, pp. 337-351, Springer Berlin Heidelberg, 2002.
[34] M. Masud et al., "Lightweight and Anonymity-preserving User Authentication Scheme for IoT-based Healthcare," IEEE Internet of Things J., vol. 9, pp. 2649 - 2656, 2021.
[35] D. He and N. Kumar, "Enhanced Three-factor Security Protocol for Consumer USB Mass Storage Devices," IEEE Trans. on Consumer Electronics, vol. 60, no. 1, pp. 30-37, 2014.
[36] K. M. Renuka, S. Kumari and X. Li, "Design of A Secure Three-factor Authentication Scheme for Smart Healthcare," J. of Medical Systems, vol. 43, pp. 1-12, 2019.
[37] P. Sonia, A. K. Pal and S.H. Islam, "An Improved Three-factor Authentication Scheme for Patient Monitoring Using WSN in the Remote Healthcare System," Computer Methods and Programs in Biomedicine, vol. 182, p. 105054, 2019.
[38] G. Xu et al., "Efficient and Provably Secure Anonymous User Authentication Scheme for Patient Monitoring Using Wireless Medical Sensor Networks," IEEE Access, vol. 8, pp. 47282-47294, 2020.
[39] D. Wang et al., "Anonymous Two-factor Authentication in Distributed Systems: Certain Goals are Beyond Attainment," IEEE Trans. on Dependable and Secure Computing, vol. 12, no. 4, pp. 228-442, 2015.
[40] Q. Wang and D. Wang, "Understanding Failures in Security Proofs of Multi-factor Authentication for Mobile Devices," IEEE Trans. on Information Forensics and Security, vol. 18, pp. 597-612, 2022.
[41] A. Armando et al., " AVISPA: Automated Validation of Internet Security Protocols and Applications," pp. 281-285, [Online], Available: http://www.avispa-project.org/, Accessed in January 2024.

[42] D. Basin, S. Mödersheim and L. Vigano, "OFMC: A symbolic Model Checker for Security Protocols," Int. J. of Information Security, vol. 4, pp. 181–208, 2005.

[43] J. Srinivas et al., "Anonymous Lightweight Chaotic Map-based Authenticated Key Agreement Protocol for Industrial Internet of Things," IEEE Trans. on Dependable and Secure Computing, vol. 17, no. 6, pp. 1133-1146, 2018.

ملخص البحث:

نتيجةً للتّقدُّم الهائل في الاتّصالات اللّاسلكية، فإنّ منصّات الرّعاية الصّحية عن بُعد جعلت من الممكن للمرضى تلقّي الرّعاية الصّحية اللّازمة عن بُعد دون الحاجة إلى الحضور إلى مراكز الرّعاية الصّحية. ومع ذلك، يبقى إيجاد خطّة تحقُّق آمنة وفعّالة لأنظمة الرّعاية الصّحية من أبرز التّحدّيات. وقد اقتُرحت حلول عديدة لهذه المشكلة، لكنها ظلّت في معظمها قاصرةً عن تأمين الأمان المطلق والحماية من الهجمات السّيبرانية.

في هذه الورقة، بدأنا أوّلاً بدراسة وتحليل أحد الحلول المقترحة في المرجع [4] لبيان إخفاقه في توفير التّحقُّق المتبادل وضمان سرّية معلومات المريض وعدم تتبُّعهِ، إلى جانب هشاشة ذلك الحلّ المقترح للهجمات. بعد ذلك، نقترح خطّةً محسّنة وفعّالة لمعالجة أوجُه القصور في الخطّة المقترحة في المرجع [4].

ويمكن للخطّة المقترحة في هذه الدّراسة أن تؤمّن التّحقُّق المتبادل، وأن تحافظ على خصوصية المريض وعلى سرّية معلوماته، إلى جانب ضمان عدم تتبُّعه من قبل المهاجمين والمتطفّلين الإلكترونيين، مع توفير الحصانة اللّازمة لنظام الرّعاية الصّحية من الهجمات والاختراقات على اختلاف أنواعها. من ناحيةٍ أخرى، تتميّز الخطّة المقترحة في هذه الدّراسة بالفعالية مقارنة بمثيلاتها القائمة من حيث تكلفة الحسابات وتكلفة الاتّصالات.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

A RESEARCH-BASED ONTOLOGY FOR COLLABORATIVE INNOVATION: A METHODOLOGY LEVERAGING AI AND DOMAIN EXPERT KNOWLEDGE

Faten Kharbat[1], Abdallah Alshawabkeh[2] and Mohammad Sharairi [2]

(Received: 17-Feb.-2024, Revised: 22-Apr.-2024, Accepted: 10-May-2024)

ABSTRACT

This paper introduces a method, for creating research-driven ontology to foster collaboration and innovation. The concept of collaborative innovation implies a process where multiple stakeholders work together to generate novel ideas, solutions or products. The suggested approach combines Artificial Intelligence (AI) and expert knowledge to build a comprehensive model encompassing various aspects of research, development and innovation. To demonstrate the feasibility of this method, the paper showcases its implementation in the field of accounting science. First, AI-powered machine-learning algorithms and text-mining techniques are used to extract the main ontological elements from a large corpus of accounting literature. Subsequently, expert knowledge is utilized to refine and validate these identified elements. The resulting ontology can be used as the foundation of a knowledge-based system to promote collaboration and analyze the state of innovation.

KEYWORDS

Artificial intelligence, Expert knowledge, Ontology, Research-based, Accounting, Text mining.

1. INTRODUCTION

The most commonly cited problem in developing expert systems is acquiring specific knowledge for a well-defined domain from experts and representing it in the appropriate digital format. Within the Artificial Intelligence (AI) field, this has been called the knowledge acquisition's problem and has been identified as a bottleneck in the process of building expert systems [1]-[2]. The exponential growth of data due to Industry 4.0 technologies necessitates capturing and transforming data into useful information for organizations [3]. It is worth mentioning that ontologies and knowledge graphs are used to represent complex knowledge. Knowledge graphs enable efficient querying and reasoning about complex data, supporting the development of intelligent agents that can learn from the represented knowledge [4]. A well-designed ontology describing the expert knowledge of a domain is at the core of many knowledge-based systems [5] to "support large-scale data/information interoperability, sharing of information and ontology-supported processes" [6]. According to Zhang and Li [7], a well-designed ontology is the key to building a successful knowledge-based system. However, [8] argued that systematic literature review highlighted the need for more structured development processes in knowledge-based systems, emphasizing knowledge elicitation and formalization, which could potentially involve ontology.

Ontology is a formal, explicit specification of a shared abstract representation of a real-world phenomenon by describing its relevant concepts, relations and axioms of a domain of interest [9]. At the beginning of the 1990s, computer science began to recognize ontology. It became an important area to investigate, especially in the area of artificial intelligence AI, because it was proposed as an effective method for creating representations of reality to be used later in the process of AI [10]-[11]. Ontologies enable inferences based on their content and relationships, simulating human inference capabilities [12].

Different ontology-development methods have been proposed, such as methods for new ontology development, new ontology alignment and merging ontology learning and re-engineering existing ontologies [13]-[14]. However, building and maintaining an ontology is considered a "labor-intensive process" [15] that costs time, money and effort. Different techniques and methods have been proposed for building new ontologies, including new-ontology development, alignment, merging, ontology

1. F. Kharbat is with Department of Computer Science, College of Engineering, Al Ain University, Abu Dhabi, UAE. Email: faten.kharbat@aau.ac.ae
2. A. Alshawabkeh and M. Sharairi are with College of Business, Al Ain University, Abu Dhabi, UAE. Emails: {Abdallah.Alshawabkeh, Mohammad.Sharairi}@aau.ac.ae

learning and re-engineering existing ontologies [3], [12]-[13], [15]-[16]. This research is developed to support collaborative innovation in ontology development by establishing a common language and conceptual framework that bridges interdisciplinary boundaries and enhances communication among diverse stakeholders. A well-designed ontology captures expert knowledge in a usable format, empowering stakeholders to leverage domain-specific insights, identify synergies and co-create innovative solutions. The ontology's capacity to organize and structure information improves information retrieval, idea generation and decision-making processes, fostering a collaborative environment conducive to innovation.

1.1 Current Status of Accounting Ontology

Organizational accounting knowledge is vital; therefore, Aparaschivei [17] outlined the academic and commercial advantages of creating an accounting ontology. The transactional accounting model with its associated value constraint as well as the asset liability equity resource types was examined in detail by Shannaq and Fatima [18], who conducted a hierarchical observation of accounting ontology. Developing a functional ontology for accounting is the first step in establishing a knowledge base for a discipline within an organization. A concept hierarchy was created to fully understand and simplify the accounting system [18].

Currently, there are several economic exchange ontologies, such as OntoREA [19], COFRIS [20] and ATE [21]. Most of them are derived from or refer to McCarthy's REA accounting model, which is commonly used as a reference for accounting ontologies focusing on economic exchanges [19]. The REA is designed to unify accounting and management perspectives on accounting information systems (AISs) [22]. However, it has been argued that accounting is more about reporting economic exchanges. The REA accounting model refers to AISs that record these exchanges. The processing that is done in AISs, the way data is aggregated into financial reports and how the quality of data is assured are not included in the REA model [23].

As part of REA business ontology, Geerts and McCarthy [24] added a policymaking architecture to the REA accounting framework. This way of thinking allows for the consideration of company policies regarding acquisition, transformation, revenues, banking and investment transactions by extending economic reasoning from a prospective to a context and setting viewpoint. The consequences for the evolution of EIS, EIS built using social-networking sites, EIS built on the network and EIS compatibility between different types of businesses are brought into sharp focus by the comprehensive general framework of REA ontology (REA2) [25].

Any occurrence in a company's operations that has a financial impact on its accounting information is considered as a financial-reporting transaction. This information can be found in a company's books. Financial accounting requires that, after adjustments have been made, an entity's revenues equal its total liabilities and its stakeholders' equity. The accountability equation is the foundation of the dual-entry accounting information system. The basic accounting formula is assets = liabilities + shareholders' equity [26]. To provide a uniform meaning for the subtraction and addition of assets, liabilities and equity inside the financial statement, the concepts of both debit and credit are required.

Accounting experts may utilize their expertise to automate accounting activities. To characterize the types, qualities and interactions of ideas in an area is the goal of the ontology method [17]. Intelligent applications can be used in the field of accounting to automate accounting information. The ontology-based automated transaction financial-reporting system was presented by Shen and Tijerino [27], with a particular emphasis on processing total count documentation, such as receipts, to satisfy the corporation requirements. This is time-consuming, because it requires physically sorting receipts into relevant categories for company-expenditure reports.

There are several clear benefits to developing an ontology for accounting transactions. Users of accounting records may benefit from this ontology by better comprehending the specific meaning of the accountancy-operation terminology. Developing an accounting-related understanding may also be grounded in the ontology of accounting transactions. Even though its theoretical roots are in accounting, REA economic ontology does not address core needs in that discipline. Schwaiger et al. [19] found that the REA commercial ontology lacked the necessary traditional accounting logic, such as documenting credited and crediting modifications in assets and obligations, all of which provided

appropriate categories as part of accounting records. A gap between scientific work and actual accounting practices has been noted [23]. By using domain-specific expertise in accounting, one may streamline the overall bookkeeping process. As an example of such an intelligent user's use in the field of accounting, we consider the automation of financial statements.

Therefore, the key challenges that specialists confront are: (1) locating a complete domain ontology and (2) locating domain specialists to enlist in collecting the necessary domain ontology. Cognitive computing and information-extraction methods allow for the semi-automatic identification of ontology when reading subject materials, providing insight into these topics. Similar semi-automatically and intelligently produced accountancy subjects and ideas are essential for something like an agile framework that develops and expands a global financial intelligent system. It also eliminates proprietary information that is either too costly to protect or forbidden by law. Consequently, this study aims to develop a low-effort, high-return strategy for extracting the most value from a previously specified domain ontology.

2. METHODOLOGY

Owing to the nature of the research, the developed ontology requires a methodology that supports the integration of several components from different sub-results within the project. This research adopts the agile methodology for building an ontology proposed by Abdelghany et al. [28]. This supports the core activities of building ontologies [28]. In a comparative analysis of methodologies for domain ontology development conducted by Sattar et al. [29], various aspects of ontology development and documentation were compared, including ontology construction strategies and support for integration and merging. The adopted methodology had a high rank among other methodologies to build an ontology. [29] highlighted the importance of selecting a methodology that adequately addresses both core ontology-building activities and project-management requirements.

The research-based ontology crafted for collaborative innovation acts as a foundational framework that amalgamates intelligence from AI algorithms, text-mining techniques and expert knowledge to facilitate knowledge sharing, interdisciplinary collaboration and innovation management. This ontology serves as a structured representation of domain-specific concepts, relationships and constraints, enabling stakeholders to access, interpret and contribute to a shared-knowledge repository.

The methodology process [28] consists of three main stages: the pregame stage, the development stage and the postgame stage, as shown in Figure 1. The three stages involved a team of experts consisting of five experts in the fields of knowledge management, artificial intelligence and three from the accounting. The first (pregame) stage allows the team to identify the goal of the ontology, tools and techniques to be used while building it and to select the data-collection methodology. The next step was to formulate ontology requirements. Because the main aim of ontology is to build a research-based ontology, the research dataset was identified and selected in this phase, as explained in the data collection part.

The second (development) stage executes an iterative process for multiple unsupervised cycles of development and evaluation. During the second stage, different meetings were held to discuss the ontologies produced. At each meeting, changes in the ontologies were verified. These steps led to the final set of ontologies approved by all participants.

The last (postgame) stage should be followed by another iterative process for multiple supervised cycles, per experts' feedback. This cyclic process and results are an essential step in contributing to improving knowledge acquisition for the representation of scientific knowledge in ontologies. The final stage (postgame stage) allows verification and validation processes, along with documentation.

To elucidate the relationship between lexicon elicitation and ontology development, it is crucial to emphasize how defining domain-specific terms and concepts contributes to building a robust ontology. Lexicon elicitation involves extracting and defining domain-specific terms, essential for constructing an ontology that accurately represents the domain's knowledge landscape. By structuring expert knowledge, the ontology can organize information, facilitate knowledge sharing and support collaborative innovation. Utilizing AI-powered machine-learning algorithms and text-mining techniques in lexicon elicitation helps extract ontological elements from domain literature, identifying key concepts and relationships foundational to the ontology. Expert-knowledge validation further

refines these elements, ensuring the ontology's accuracy and relevance in fostering collaboration and innovation.

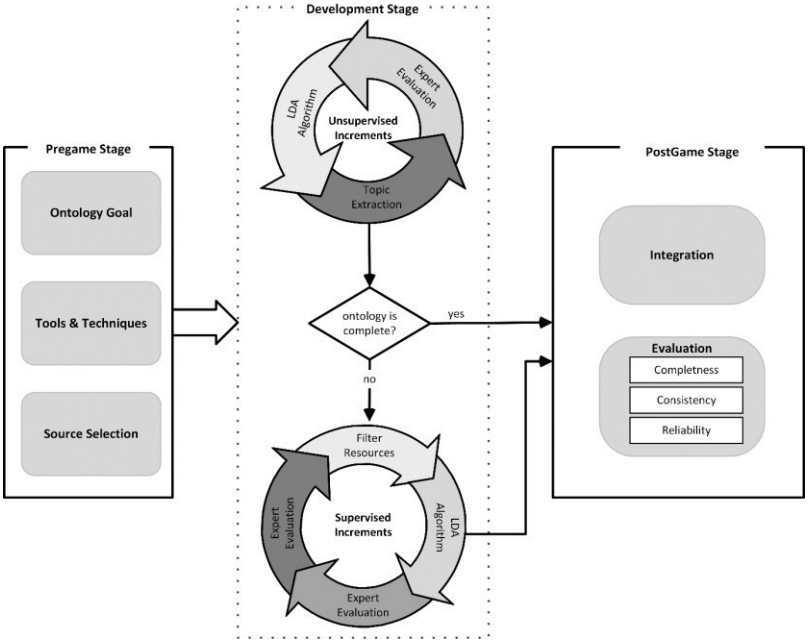


Figure 1. The research methodology as adopted from [28].

3. Results and Discussion

3.1 Pregame Stage

3.1.1 Goal & Scope

The fundamental scope of accounting has not changed over time; nevertheless, new innovative accounting directions have evolved [30]. The accounting domain was selected due to its inherent complexity, evolving nature and critical role in various industries. The compilation of reports, transaction recording, cyber-auditing and other technological advancements have grown. As a result, different specialized accounting topics have emerged. Thus, by developing an accounting ontology, the paper aims to establish a structured framework that captures and organizes domain-specific knowledge, facilitating collaboration and innovation within the accounting field. The goal of building the proposed accounting ontology is to establish a common understanding of the meaning of the terms used by researchers to support the knowledge-acquisition process of innovative directions related to accounting in the current-research directions. [31] explores the accounting dimensions of innovations, highlighting the interconnectedness of accounting practices with innovative processes. In addition, [32] discusses the identification of misstatement accounts in financial statements through ontology reasoning. It demonstrated the application of ontology-based decision-support systems in the accounting domain, emphasizing the importance of ontology reasoning in financial-statement analysis. In fact, the selected literature reinforces the importance of ontology-based approaches in addressing accounting challenges, thereby justifying the choice of the accounting domain.

The proposed accounting ontology can be used as a tool to map the concepts and services used by expert systems to align with the most recent information. The scope of accounting ontology focuses on the main specialized topics of accounting in the recent literature: financial accounting, management accounting [33], cost accounting [33], tax accounting [34] and auditing [34]. Other topics may have been included; however, the experts preferred not to expand the search for more detailed ones.

3.1.2 Tools

The Protégé ontology editor and framework in the OWL language were selected as the main tools to formalize the accounting ontology. The selection was based on previous research, such as [28] and its well-known ability to customize and extend.

3.1.3 Source Selection

To align with the research aims, the proposed ontology is based on a research-based source and the research dataset is identified and selected as follows: To collect all research written in "Accounting," three main sources were identified [33], [35]: ABDC, Scopus and Web of Science. All journals indexed in ABDC, Scopus and the Web of Science related to this field were identified and selected by September 2021. From the Scopus index, 154 journals related to the accounting discipline were published by 46 publishers. The journal citation scores ranged from 0 (e.g. Journal of Taxation, SJR=0.1) to 10.3 (e.g. Journal of Finance, SJR= 17.134). In the ABCD list, 2679 journals related to the accounting discipline were published by 744 publishers. The journal ranks ranged from A* (199 journals, e.g. Australian Tax Forum) to C (982 journals, e.g. Real Estate Taxation). In the Web of Science, 18 journals related to the accounting discipline were published by 12 publishers. Of course, there were duplicate findings between the three main sources, which will be discussed later.

All research published in journals from 1946 to 2021 was automatically collected using a script written in Python 3.8.5. The total number of papers is 209,345. However, the data collected was filtered, because some documents were not journal papers or their abstracts were not electronically available. After filtering, the total number of journal papers was 159,239, with available abstracts. All collected titles and abstracts were injected into an empty dataset to be processed *via* machine-learning and text-mining techniques.

3.2 Development Stage

The development stage aims to produce pieces for the accounting ontology from the final dataset processed in the previous stage to integrate them in the third stage. The development stage consisted of several increments to allow different levels of information extraction from the available dataset containing titles and abstracts. This process was introduced and implemented in [36], where the iOntoBioethics ontology for Bioethics Ontologies in Pandemics was proposed and evaluated.

The relation between the extracted ontology topic/concept and the dataset is generally based on proportion. For example, the topic of "cost accounting" is extracted as an important topic based on the number of times it is mentioned in different abstracts. If mentioned in one abstract twice, this does not mean that it is as important as if mentioned once in two abstracts. Each given abstract relates to the discovered accounting topic/concepts with different proportions. To work with these proportions, certain factors must be investigated: accounting topics per abstract and assignment of accounting concepts per accounting topic in an abstract. Hence, in the current development stage, the latent Dirichlet allocation (LDA) [37]-[38] algorithm in the field of machine learning (text mining) was utilized to illustrate the topics and their related concepts within the abstracts. The stage started with unsupervised increments, followed by supervised increments, as shown below. To illustrate the outcomes from the supervised and unsupervised iterations, Protégé was used, as mentioned in sub-section 3.1. Every topic is mapped to a class and every concept is mapped to a property associated with the class.

3.2.1 Unsupervised Increments

This stage consisted of three iterations, in which the LDA algorithm automatically identified accounting topics and their associated concepts. An iterative automated process was conducted to determine the maximum coherence value for the best number of topics to avoid bias in executing the LDA algorithm.

Iteration 1: Preparation. The first increment was a preparation increment, in which the dataset containing all abstracts and titles was fed into a customized Python tool, the Standardization Text Characteristics (SRC) with a reliance process [39] was directed and the LDA algorithm was executed to extract the general topics and their associated concepts. The SRC process converted all upper-case characters into lower-case characters, removed all stop words such as "the," "on," …etc., removed all the white spaces and removed punctuations. The output of the first iteration shows that six topics were extracted, as shown in Figure 2.

Three accounting-domain specialists were involved in investigating the six topic structures to validate the output of the first iteration, arriving at a consensus on the extent which these topics represent. Three experts evaluated the six topics separately and the research team collected their responses. The

main comments from the three experts focused on unrelated concepts within the extracted topics. For example, in topic 1, the words “research”, “new” and “case” were considered irrelevant. Topic 3 has the words “study” and “purpose” that “weaken the aggregation concept,” as one of the experts expressed.

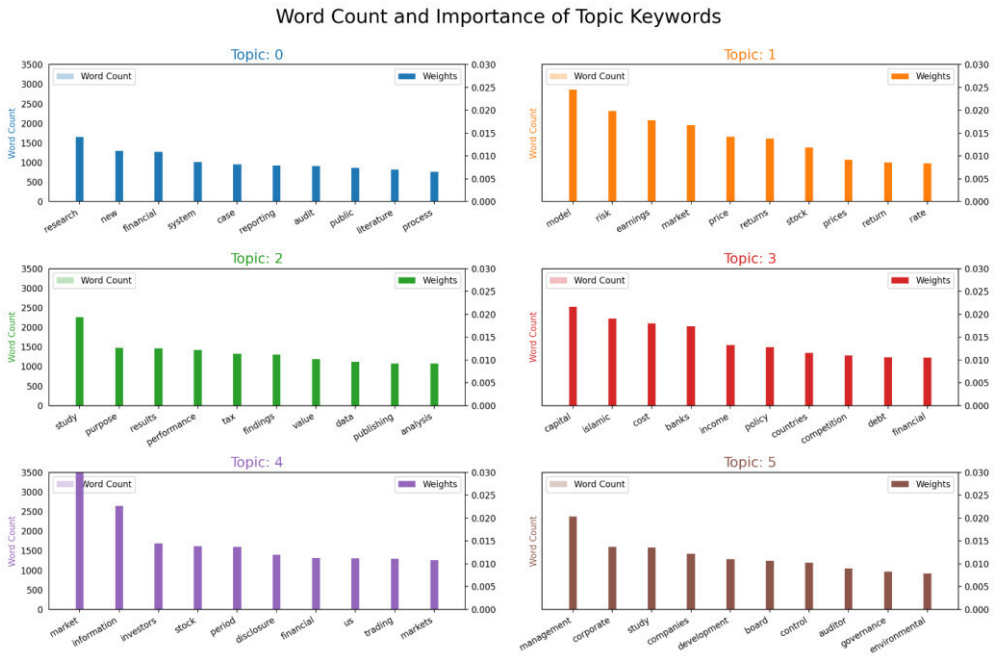


Figure 2. The topics extracted from the preparation interaction.

Iteration 2: Enhancement. The second iteration involved enhancing the basic environment for the dataset of abstracts and titles. The LDA algorithm was executed several times to detect new unrelated words and phrases to enhance the final results. All intermediate results were provided to the three experts for validation and recommendations. All unrelated words recommended by the experts from the first iteration were added to the set of stop words. The final stop-word list is presented in Table 1.

Table 1. The stop words included in the experiments.

```
'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
"you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself',
'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her',
'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this',
'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were',
'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does',
'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because',
'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after',
'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off',
'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there',
'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few',
'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only',
'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will',
'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm',
'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't",
'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't",
'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn',
"mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't",
'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn',
"wouldn't",'from','subject','astract','avaible','a','find','firm','firms'
,'accounting','paper','accountent','Research','literature', 'new',
'study', 'purpose','Findings','rights','reserved','well','methodology',
'design', research', 'approach','findings','finding', 'company','used',
'use','uses','also','de','part','parts','find','white','wisdom','elsevier
ltd', 'academic press limited', 'american accounting association', 'by de
la salle university','all rights reserved'
```

Iteration 3: Output Production. The execution of the third iteration results in four topics, as shown in Figures 3 and 4. In Figure 3, the number of documents related to the topics is shown, whereas in Figure 4, the concepts related to each topic are illustrated.

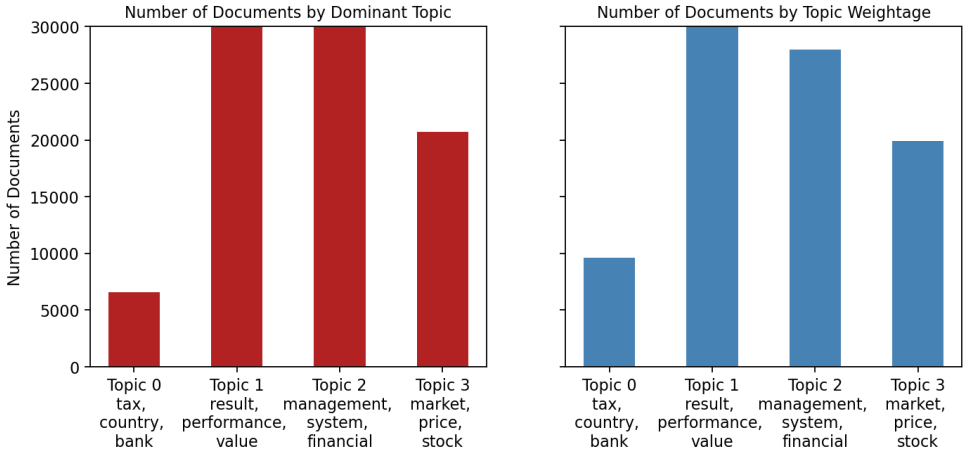


Figure 3. The number of documents related to each extracted topic (iteration 3).

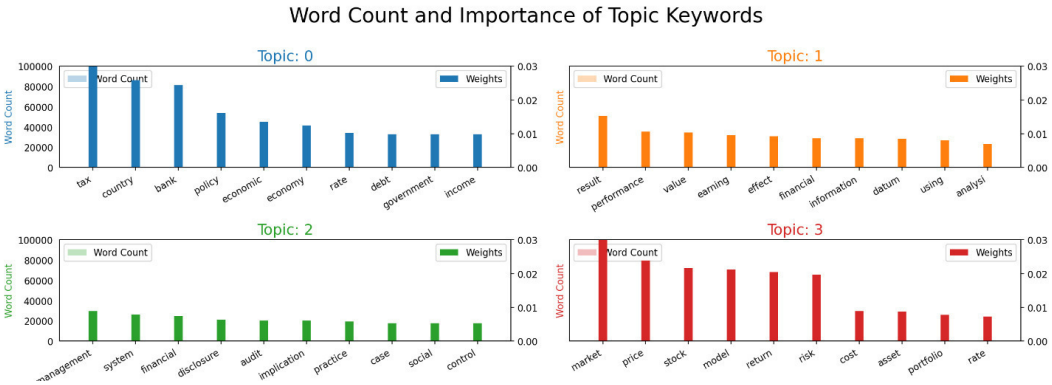


Figure 4. The main four extracted topics with their associated concepts (iteration 3).

It is worth noting that the LDA algorithm does not name the extracted topics, but assigns them numbers within the range of related topics. Therefore, the three experts were involved later in independently determining the titles of the topics that aggregated the shown concepts.

At the end of the third iteration, the three experts were given the generated topics and the related concepts to evaluate them independently. They were also asked to suggest a title for each topic. Consequently, the results of this process are as follows:

- ***Topic 0: Tax Accounting***

The terminology on this topic refers to tax accounting. Tax accounting checks whether businesses adhere to all rules set out by tax authorities. Economic data and other revenue data are subject to reporting by tax accountants to tax authorities. The accounting and tax standards used by each nation are different. Long-term economic advancement and resource allocation are susceptible to taxation considerations [40].

- ***Topic 1: Financial Accounting***

The terminology for this topic expresses the topic of financial accounting. Financial accounting is a sub-field of financial reporting that focuses on disseminating a firm's financial information to stakeholders, including stockholders, researchers, suppliers and regulatory agencies. Statement of income, financial statements, cash-flow statements and cash-position declarations are the four fundamental audited financials. This accounting information is the foundation for evaluating a corporation's economic health and competitive position. In addition, people use such data to decide whether or not to invest in a firm. Investors and traders utilize such accounting records in the financial sector to evaluate the health and potential returns of publicly-traded businesses and the stock market and brokerage houses [41].

- ***Topic 2: Auditing***

The terminology for this topic mainly indicates auditing. An audit examines and assesses the effectiveness of internal company controls and financial-analysis processes. External and internal accountants perform this function. The public relies on a firm's publicly released income statement when making a variety of financial choices. Conversely, internal audits directly report to managers and supervisors inside an organization. A company's auditors examine whether senior-management directives are being followed. When conducting an internal review, it is crucial to determine whether a company's actions are in accordance with its own stated objectives [42].

- ***Topic 3: Financial Accounting - Portfolio Investment***

The terminology of this topic points toward portfolio investment (finance), which relies on financial-accounting information. Sector-specific breakdowns in private investment, investment spending and foreign reserves are all part of a complete financial statement. When investing in stocks, bonds and other marketable securities to earn a profit, grow in valuation or combine them, you are creating a portfolio. This implies a less proactive managerial position than direct investment by way of the passive ownership of assets. When valuing a company or conducting a financial analysis, shareholders rely heavily on the data provided in its financial reports. As a result, it is crucial to understand the fundamentals of business accounting as well as the rules governing the creation of financial statements. Accounting is advantageous to investors, because it allows them to assess the worth of a firm's profits, learn about its financial products, measure its financial performance and gauge the risks inherent in the income statement [43].

Visual Illustration

To illustrate the topics and concepts from the third iteration of the unsupervised increments, the topics and concepts are implemented in Protégé as follows.

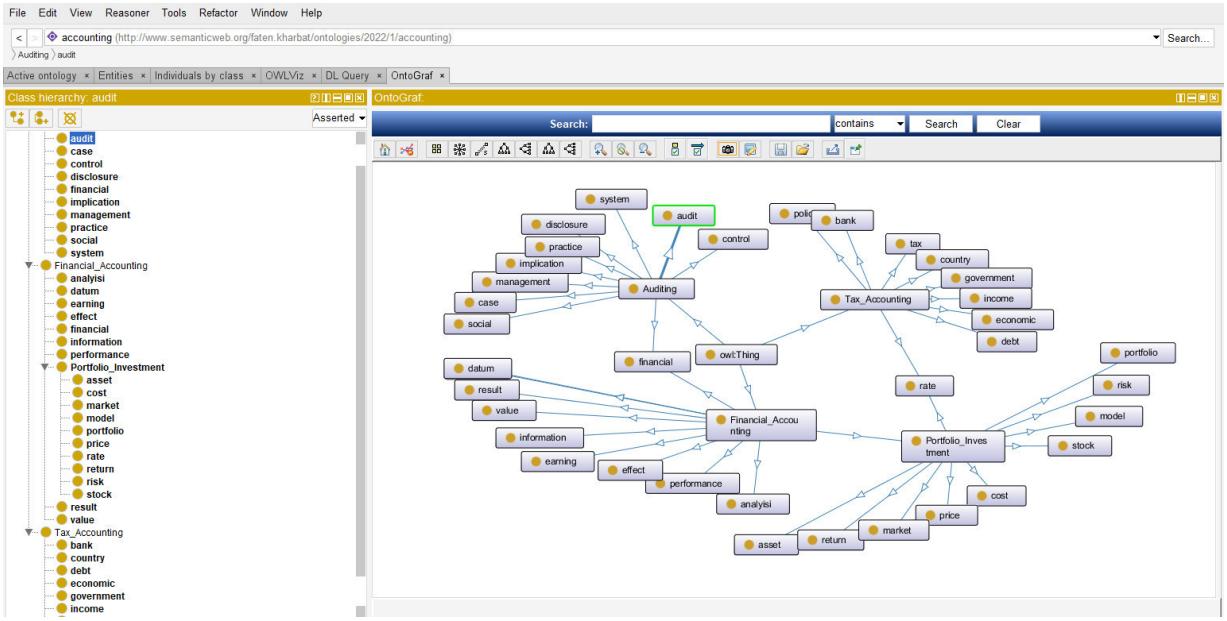


Figure 5. The draft of the ontology from the unsupervised increments.

3.2.2 Supervised Increments

Ontology validation was discussed by Quinn and McArthur [44] to evaluate whether the ontology matches the world model as demonstrated by the industry dataset [44]. Different methods have been used to measure the validity of an ontology [44]–[46]. However, a very strong approach was proposed and implemented by Quinn and McArthur [44], which included qualitative and quantitative approaches to express ontology completeness and expressiveness. Completeness indicates the sufficiency of the semantic relationship between the extracted topics/concepts and the accounting domain. To measure the completeness of the resulting concepts, an anonymous survey was distributed to the accounting experts that included a table with all the topics and related concepts to state whether

the topics/concepts were related to the accounting domain and to what extent (0 = not related and 10 = entirely related). The average number of experts was then calculated on the agreement that any average less than 7.0 would be excluded from the results.

Regarding the topics, the average number of experts was 9.4/10. The minimum average mark was given to the topic of portfolio investment (7.6/10). Regarding the concepts, the average from the experts was 8.4/10, indicating that the extracted concepts were highly related to the topics and accounting domain.

The next step was to measure the expressiveness of the current version of the ontology. The expressiveness of an ontology, the so-called "coverage percentage" [28], is defined as "quantifying the number of key relationships required" in a domain [44]. To conduct such measurements, a list of known accounting topics and some emerging ones were listed from different sources, such as Shkulipa [33] and the experts were asked whether they would think that any of them would be essential to be added. At least 70% of the experts agreed to manually add the main topics: cost, managerial accounting and forensic accounting.

Therefore, a supervised iteration was implemented to extract related concepts for missing topics. To conduct supervised iteration, a separate iteration for each topic was executed to extract the related concepts. For each iteration, the dataset, including all titles and abstracts, was filtered to include only that topic in particular. The results of each iteration are as follows.

- **Cost Accounting**

The three experts were again given the topics generated from the supervised iteration to independently evaluate the topics and their concepts and suggest a couple of titles for each topic. Consequently, the conclusion from the supervised iterations for the "cost accounting" was identified as follows:

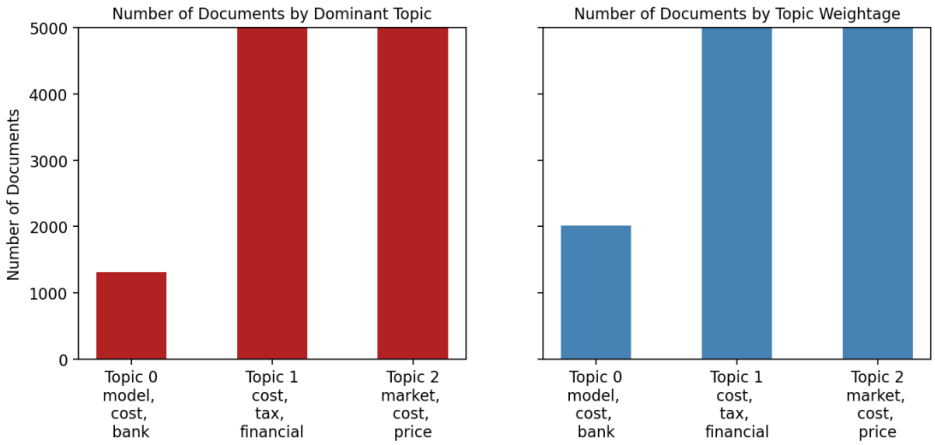


Figure 6. The number of documents related to each extracted topic of cost accounting.

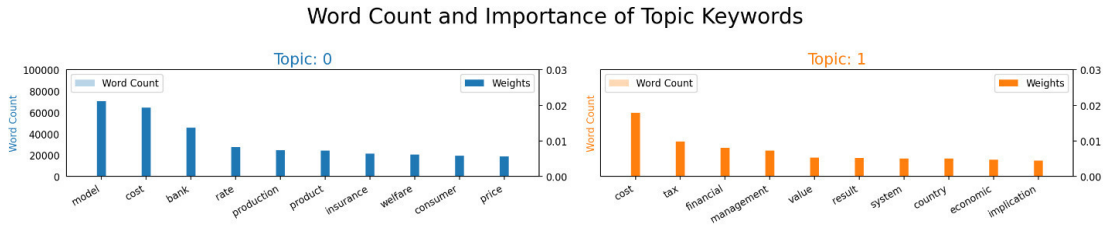


Figure 7. The generated topics for cost accounting.

- Topic 0: Cost Accounting

The terminology used here refers to cost accounting. In cost accounting, to maximize revenue and enhance the efficacy of business processes, standard costing (also known as control accounting) generates data to be utilized in the improvement of the organization. Cost accounting entails taking care of all the money that comes up in the operation of a company. Management uses cost information to prepare and manage various cost activities. Standard costing is concerned with collecting, categorizing and interpreting cost data in a quantifiable manner. The primary objective of cost

accounting is to collect and analyze a firm's variable and constant expenses. From an administrative perspective (marketing, transportation organization, protection, manufacturing, …etc.), both direct and indirect materials and direct and indirect employees are the main components of the costing system [47].

○ *Topic 1: Miscellaneous Topics*

It is difficult to give a specific title to this group, as the terminologies are related to different general accounting topics, such as cost, tax, financial and management.

- **Managerial Accounting**

The three experts were again given the topics generated from the supervised iteration to independently evaluate the topics and their concepts and suggest a couple of titles for each topic. As a result, the conclusions from the supervised iterations for "managerial accounting" were identified as follows:

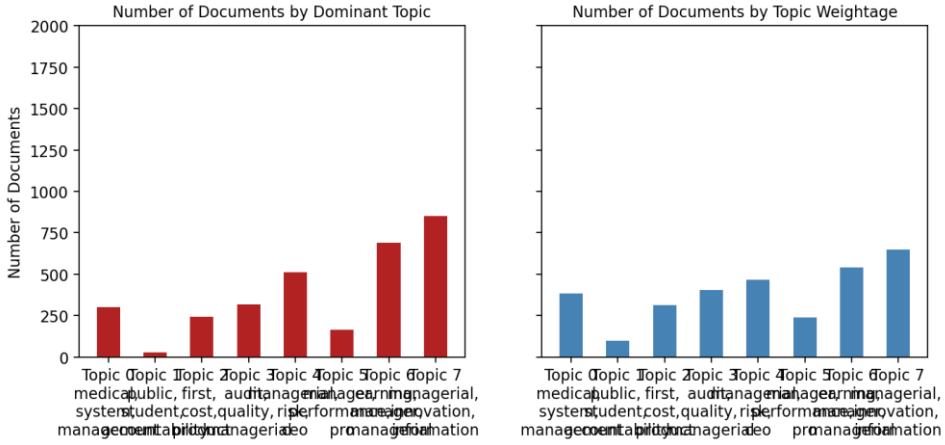


Figure 8. Number of documents related to each extracted topic for managerial accounting.

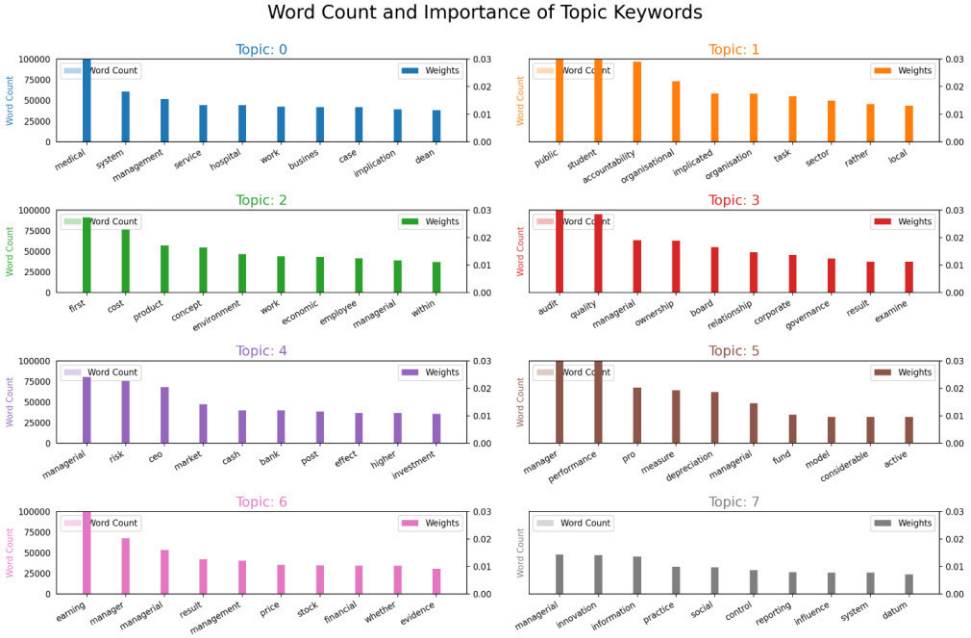


Figure 9. The topics extracted for managerial accounting.

○ Topics 0, 1, 4, 5, 6 and 7

It is difficult to give a specific accounting title for this topic, as the results are not accounting-related terminologies or are related to different general accounting topics.

○ *Topic 2: Managerial Accounting*

The terminology on this topic refers to managerial accounting. Managerial accounting, as opposed to financial accounting, includes the dissemination of financial information to a company's operational

units. It is a sub-set of accountancy that focuses on analyzing financial data to generate financial accounting documents and reports to aid in the judgment procedure of heads of departments as well as top management. Managerial accounting clarifies the company's monetary information and delivers meaningful numbers and statistics to higher management and decision-makers.

Executives and departments, such as sales, marketing and production, may request custom reports that meet their unique reporting requirements. These reports combine actual and predicted data to provide managers with a wealth of information to make better business choices. In contrast to financial accounts, which are made public as well as publicized, data packets are used internally to enhance procedures, including total profit appraisal, departmental planning and other similar activities [48].

- *Topic 3: Auditing*

Discussed previously. Adding some concepts: ownership, governance and quality.

- **Forensic Accounting**

Once again, the three experts were given the topics generated from the supervised iterations to independently evaluate the topics and their concepts and suggest a couple of titles for each topic. As a result, the conclusion from the supervised iterations for "forensics accounting" was identified as follows:

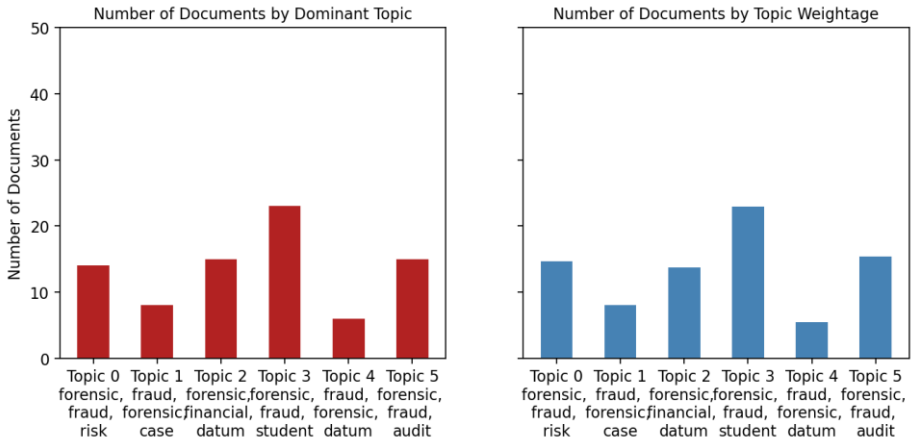


Figure 10. Number of documents related to each extracted topic for forensic accounting.

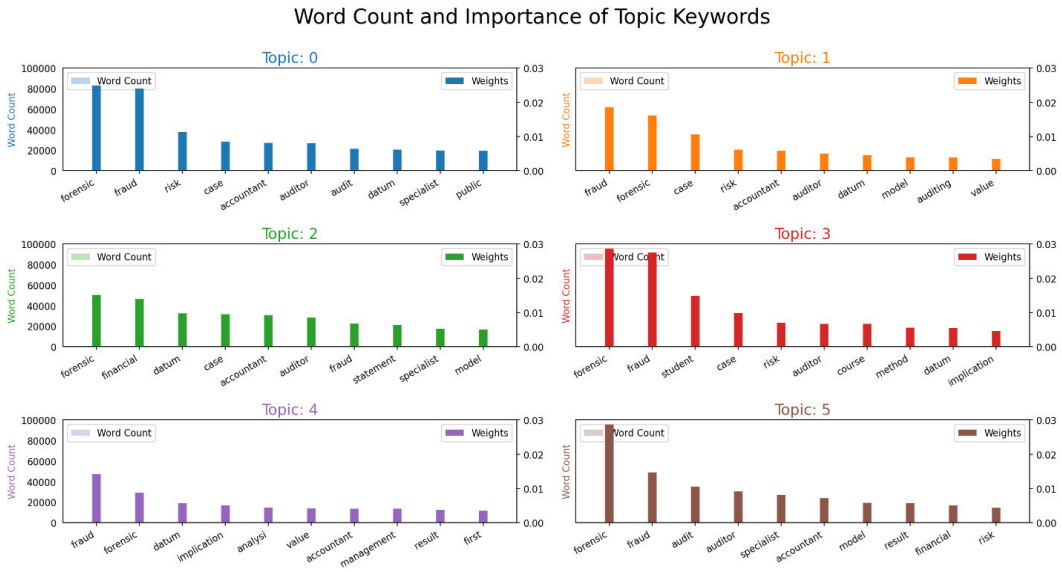


Figure 11. The topics extracted for forensic accounting.

- *Topic 0, 1 & 2: Forensics Audit*

The terminology for this topic is related to forensic auditing. During a forensic audit, experts apply their knowledge of accounting to investigate cases that may have criminal consequences. The scope of the full investigation exceeds that of standard forensic accounting. Forensic auditing investigates the

nature of the transactions themselves and identifies signs of possible asset theft [49].

- *Topic 3: Accounting Education-Forensics*

The terminology used here refers to accounting education, specifically in forensics, as course content. Accounting education aims to prepare students for jobs, as accounting professionals are the ultimate goal of financial reporting. There has been an increase in interest in the accounting profession from the ranks of professional accounting organizations, which seek to blend theoretical studies with hands-on experience as well as outline specifics of required coursework [50]-[51].

- *Topic 4 & 5: Forensic Accountant*

The terminology of this topic mainly refers to forensic accountants. Forensic accounting is used to investigate a person's or company's financial situation. Forensic accountants use a wide range of techniques in the accounting, reporting and investigation industries. Accounting information is frequently referred to as a sub-set of accounting. Experts witness that testimony is common for forensic accountants, who also conduct investigations of financial information that could be admissible in court. In court, professionals may demonstrate the monetary elements of crimes fraudulently. There is a growing need for forensic-accounting professionals in various fields, including law-enforcement agencies, auditing firms and insurance organizations [52].

3.3 PostGame Stage

The final stage of the adopted methodology (as described in Section 2) consists of two main steps: integration and evaluation. The integration process in this context is simple, because the outputs from the supervised and unsupervised iterations are based on the same dataset and for the same domain. It is unlikely that there is any inconsistency between topics or concepts. Every topic is mapped to a class and every concept is mapped to a property associated with the class. The only relation demonstrated was "is-a" from the main domain. The new version of the ontology is demonstrated using Protégé, as shown in Figure 12:

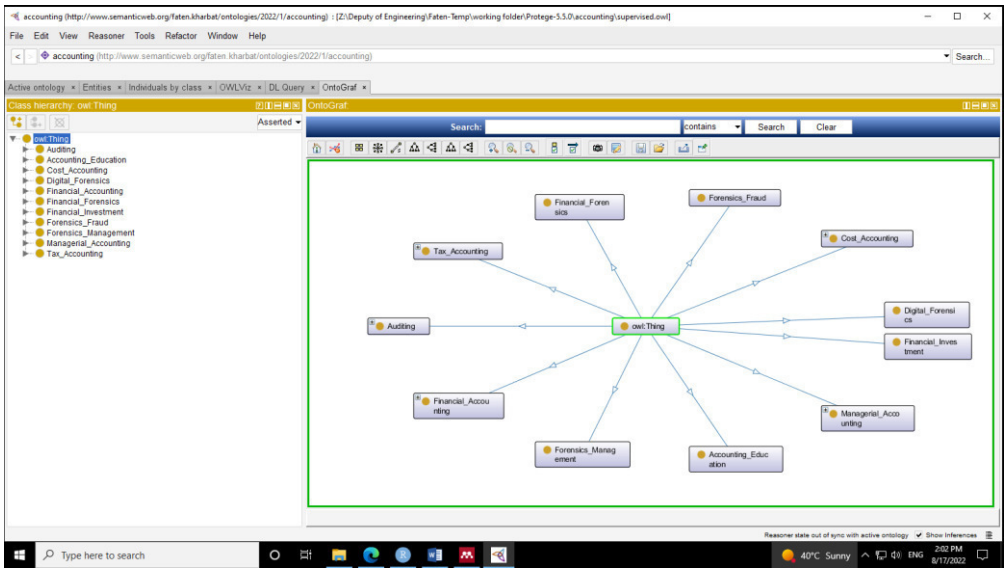


Figure 12. The general view of the new version of the ontology.

While the development stage is pivotal in establishing the ontology's foundation, the postGame stage plays a critical role in assessing the ontology's utility, identifying areas for improvement and ensuring its alignment with the intended objectives. Figure 13, created towards the end of the postGame stage, serves as a visual representation of key components, relationships or insights derived during the ontology-construction process. Drawing insights from the research-based ontology methodology for collaborative innovation, the postGame stage involves a comprehensive assessment of the ontology's utility in facilitating knowledge sharing, interdisciplinary collaboration and innovation management. Figure 13 encapsulates the culmination of the ontology-development process, showcasing the refined ontology structure, validated concepts and the integration of AI-powered machine-learning algorithms and expert knowledge to create a robust knowledge-based system.

The evaluation process is implemented again as in the previous stage: qualitative and quantitative approaches to express ontology completeness and expressiveness. To measure the completeness of the final draft of the ontology, a concise anonymous survey was distributed to a group of accounting experts that included a table with all the topics and related concepts to state whether the topics/concepts were related to the accounting domain and to what extent (0 was not related and 10 was entirely related). The average number of experts was then calculated. Regarding topics, the average number of experts was 8.8/10. Regarding the concepts, the average from the experts was 7.5/10, indicating that the extracted concepts were highly related to the topics and accounting domain.

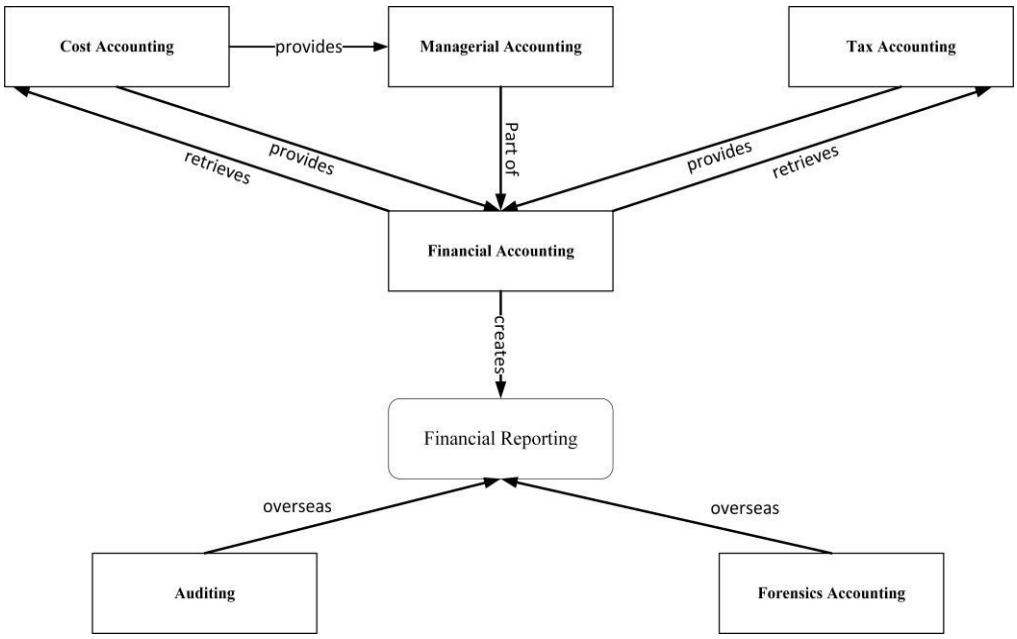


Figure 13. The final version of the generated ontology including the supervised and unsupervised iterations.

4. LIMITATIONS

While the integration of machine-learning techniques within the AMOD methodology presents promising opportunities for enhancing ontology development, several limitations warrant consideration. Firstly, as highlighted in previous research, the knowledge-acquisition process remains a critical challenge in ontology engineering. Despite leveraging machine learning for automated knowledge extraction, the reliance on existing data sources and expert input may limit the scalability and generalizability of the ontology across diverse domains.

Secondly, as discussed in the literature, the balance between automated machine learning -driven processes and expert validation is crucial for ensuring the ontology's relevance and reliability. The potential risk of overlooking domain difficulties or context-specific traces in favour of automated ML algorithms underscores the importance of maintaining a robust expert-driven validation process.

Lastly, the feasibility and effectiveness of the proposed methodology may vary across different domains and data-availability scenarios. While the research demonstrates the integration of machine learning within AMOD in a specific context, the transferability and performance of the approach in domains with limited data or distinct characteristics require further investigation. The adaptability of the methodology to diverse domains and the robustness of the machine-learning components in handling domain-specific complexities represent areas for future exploration and refinement.

5. CONCLUSIONS

Different ontology development methods have been proposed in the literature to build and maintain a comprehensive ontology, such as new ontology alignment, merging ontology learning and re-engineering existing ontologies. However, these methods are considered labor-intensive, and are incapable of describing the specific requirements of different new-research directions. Therefore, this research utilizes one of the Agile Methodologies for Ontology Development (AMOD) to develop an

ontology and integrate it with machine-learning techniques. This adaptation signifies an evolution of the existing methodology to leverage machine learning capabilities, highlighting the importance of clearly articulating the degree of innovation in the research. The incorporation of machine-learning components within AMOD can enhance the ontology-development process by introducing automation, predictive analytics or pattern recognition, thereby improving efficiency and accuracy. The developed ontology for collaborative innovation plays a pivotal role in facilitating knowledge sharing, interdisciplinary collaboration and innovation management within the research domain. By serving as a structured-knowledge repository, promoting communication among stakeholders and aligning with agile ontology-development principles, the ontology contributes to creating a dynamic ecosystem that nurtures collaborative innovation and propels research advancement.

To achieve this, as a proof of concept, the related literature on accounting was collected and analyzed from among the most influential and well-cataloged works in the accounting field since their inception in 1945. This has helped gain comprehensive coverage of the main concepts required for optimized text analysis and computational-modeling technology. The proposed method automatically detects accounting-related topics and their associated concepts, thereby empowering the derived ontology. As recent findings, regulations, laws, quality assessments, …etc. have been released, an ontology of this kind has to adapt and accommodate fresh accounting topics and concepts that arise. These intelligently generated accounting topics and concepts established by artificial intelligence are fundamental to developing an intelligent accounting system.

For future work, the research can explore the integration of knowledge graphs alongside ontologies to enhance the representation and utilization of complex, heterogeneous data in the accounting domain. Leveraging knowledge graphs can provide a more comprehensive and interconnected view of accounting concepts, enabling advanced querying, reasoning and knowledge discovery. Additionally, incorporating knowledge graphs can support the development of intelligent systems that can learn and reason from the rich knowledge encapsulated in the graph, thereby enhancing the paper's focus on ontology development for collaborative innovation. Lastly, the feasibility and effectiveness of the proposed methodology may vary across different domains and data-availability scenarios. While the research demonstrates the integration of machine learning within AMOD in a specific context, the transferability and performance of the approach in domains with limited data or distinct characteristics require further investigation. The adaptability of the methodology to diverse domains and the robustness of the machine-learning components in handling domain-specific complexities represent areas for future exploration and refinement.

REFERENCES

[1] J. A. Oravec, "Experts in a Box : Expert Systems and Knowledge-based Engineering (1984–1991)," Chapter in Book: Historical Instructional Design Cases, pp. 253–270, DOI: 10.4324/9780429330995-14, Nov. 2020.

[2] W. P. Wagner, J. Otto and Q. B. Chung, "Knowledge Acquisition for Expert Systems in Accounting and Financial Problem Domains," Knowledge-based Systems, vol. 15, no. 8, pp. 439–447, Nov. 2002.

[3] D. Spoladore, E. Pessot and A. Trombetta, "A Novel Agile Ontology Engineering Methodology for Supporting Organizations in Collaborative Ontology Development," Computers in Industry, vol. 151, p. 103979, DOI: 10.1016/j.compind.2023.103979, Oct. 2023.

[4] A. Holzinger et al., "Human-in-the-Loop Integration with Domain-knowledge Graphs for Explainable Federated Deep Learning," Proc. of the Int. Cross-Domain Conf. for Machine Learning and Knowledge Extraction, Part of the Book Series: Lecture Notes in Computer Science, vol. 14065, pp. 45–64, 2023.

[5] T. R. Gruber, The Acquisition of Strategic Knowledge, ISBN: 978-0-323-16258-6, Elsevier, 2013.

[6] K. I. Kotis, G. A. Vouros and D. Spiliotopoulos, "Ontology Engineering Methodologies for the Evolution of Living and Reused Ontologies: Status, Trends, Findings and Recommendations," Knowledge Engineering Review, vol. 35, p. e4, DOI: 10.1017/S0269888920000065, Jan. 2020.

[7] L. Zhang and J. Li, "Automatic Generation of Ontology Based on Database Trajectory Analysis View Project Automatic Generation of Ontology Based on Database," J. of Computer Information Systems, vol. 7, pp. 1148–1154, 2011.

[8] P. Kügler, F. Dworschak, B. Schleich and S. Wartzack, "The Evolution of Knowledge-based Engineering from a Design Research Perspective: Literature Review 2012–2021," Advanced Eng. Informatics, vol. 55, p. 101892, DOI: 10.1016/j.aei.2023.101892, Jan. 2023.

[9] R. Studer, V. R. Benjamins and D. Fensel, "Knowledge Engineering: Principles and Methods," Data

Knowl. Eng., vol. 25, no. 1–2, pp. 161–197, DOI: 10.1016/S0169-023X(97)00056-6, Mar. 1998.
[10] F. M. Mendonça, A. M. P. Cardoso and E. Drumond, "Ontology of Application into the Domain of Mortality: A Tool Support for Filling out the Death Certificate," Ciência da Informação, vol. 39, no. 3, pp. 23–34, DOI: 10.1590/S0100-19652010000300002, 2010.
[11] R. Kishore and R. Sharman, "Computational Ontologies and Information Systems I: Foundations," Communications of the Association for Information Systems, vol. 14, pp. 158–183, 2004.
[12] D. Spoladore et al., "A Review of Domain Ontologies for Disability Representation," Expert Systems with Applications, vol. 228, p. 120467, DOI: 10.1016/j.eswa.2023.120467, Oct. 2023.
[13] O. Corcho, M. Fernández-López and A. Gómez-Pérez, "Methodologies, Tools and Languages for Building Ontologies. Where Is Their Meeting Point?," Data Knowl. Eng., vol. 46, no. 1, pp. 41–64, DOI: 10.1016/S0169-023X(02)00195-7, Jul. 2003.
[14] A. Gómez-Pérez, "Ontology Evaluation," Chapter of Book: Handbook on Ontologies, pp. 251–273, DOI: 10.1007/978-3-540-24750-0_13, 2004.
[15] N.-W. Chi, Y.-H. Jin and S.-H. Hsieh, "Developing Base Domain Ontology from a Reference Collection to Aid Information Retrieval," Automation in Construction, vol. 100, pp. 180–189, Apr. 2019.
[16] S. Al-Fedaghi, "Toward Flow-based Ontology," Studies in Computational Intelligence, vol. 653, pp. 125–137, Springer, 2016.
[17] F. Aparaschivei, "The Importance of an Accounting Ontology," Economic Informatics, vol. 1, no. 1–4, 2007.
[18] B. Shannaq and K. Fatima, "Hierarchy Concept Analysis in Accounting Ontology," Asian J. of Computer Science and Technology, vol. 2, no. 2, pp. 13–20, 2012.
[19] W. S. A. Schwaiger et al., "The OntoREA©Accounting and Finance Model: Inclusion of Future Uncertainty," Proc. of IFIP Working Conf. on the Practice of Enterprise Modeling, Part of the Book Series: Lecture Notes in Business Information Processing, vol. 369, pp. 53–67, Nov. 2019.
[20] I. Blums and H. Weigand, "Towards a Core Ontology of Economic Exchanges for Multilateral Accounting Information Systems," Proc. of the 2020 IEEE 24th Int. Enterprise Distributed Object Computing Conf. (EDOC), pp. 227–232, Eindhoven, Netherlands, Oct. 2020.
[21] D. Porello, G. Guizzardi, T. P. Sales and G. Amaral, "A Core Ontology for Economic Exchanges," Proc. of the 39th Int. Conf. on Conceptual Modeling, pp. 364–374, DOI: 10.1007/978-3-030-62522-1_27, Vienna, Austria, Nov. 2020.
[22] W. E. McCarthy, "The REA Accounting Model: A Generalized Framework for Accounting Systems in a Shared Data Environment," Accounting Review, vol. 57, no. 3, pp. 554-578, 1982.
[23] H. Weigand, P. Johannesson and B. Andersson, "An Artifact Ontology for Design Science Research," Data Knowl. Eng., vol. 133, p. 101878, DOI: 10.1016/J.DATAK.2021.101878, May 2021.
[24] G. L. Geerts and W. E. McCarthy, "An Ontological Analysis of the Economic Primitives of the Extended-REA Enterprise Information Architecture," Int. J. of Accounting Information Systems, vol. 3, no. 1, pp. 1–16, DOI: 10.1016/S1467-0895(01)00020-3, Mar. 2002.
[25] W. Laurier, J. Kiehn and S. Polovina, "REA 2: A Unified Formalisation of the Resource-Event-Agent Ontology," Applied Ontology, vol. 13, no. 3, pp. 201–224, DOI: 10.3233/AO-180198, Jan. 2018.
[26] J. J. Wild, K. W. Shaw, B. Chiappetta, K. Dahawy and K. Samaha, Fundamental Accounting Principles, McGraw-Hill, 2007.
[27] Z. N. Shen and Y. Tijerino, "Ontology-based Automatic Receipt Accounting System," Proc. of the 2012 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT 2012), pp. 236–239, DOI: 10.1109/WI-IAT.2012.265, Macau, China, 2012.
[28] A. Abdelghany, N. Darwish and H. Hefni, "An Agile Methodology for Ontology Development," Int. J. of Intelligent Engineering and Systems, vol. 12, no. 2, pp. 170–181, Apr. 2019.
[29] A. Sattar, E. Salwana, M. Nazir, M. Ahmad and A. Kamil, "Comparative Analysis of Methodologies for Domain Ontology Development: A Systematic Review," Int. J. of Advanced Computer Science and Applications, vol. 11, no. 5, DOI: 10.14569/IJACSA.2020.0110515, 2020.
[30] S. Kruskopf, C. Lobbas, H. Meinander, K. Söderling, M. Martikainen and O. Lehner, "Digital Accounting and the Human Factor: Theory and Practice," ACRN J. of Finance and Risk Perspectives, vol. 9, pp. 78–89, DOI: 10.35944/jofrp.2020.9.1.006, 2020.
[31] A. Dyhdalewicz and U. Widelska, "Accouting and Marketing Dimensions of Innovations," e-Finanse, vol. 13, no. 2, pp. 1–13, DOI: 10.1515/fiqf-2016-0018, Dec. 2017.
[32] L. Chen, B. Xiu and Z. Ding, "Finding Misstatement Accounts in Financial Statements through Ontology Reasoning," IEEE Access, Early Access, DOI: 10.1109/ACCESS.2020.3014620, 2024.
[33] L. Shkulipa, "Evaluation of Accounting Journals by Coverage of Accounting Topics in 2018–2019," Scientometrics, vol. 126, no. 9, pp. 7251–7327, Sep. 2021.
[34] H. Khlif, "Hofstede's Cultural Dimensions in Accounting Research: A Review," Meditari Accountancy Research, vol. 24, no. 4, pp. 545–573, 2016.
[35] P. Mongeon and A. Paul-Hus, "The Journal Coverage of Bibliometric Databases : A Comparison of Scopus and Web of Science," Proc. of METRICS 2014 Workshop at ASIS & T, pp. 1751–1577, 2014.

[36] M. Odeh et al., "iOntoBioethics: A Framework for the Agile Development of Bioethics Ontologies in Pandemics, Applied to COVID-19," Frontiers in Medicine, vol. 8, DOI: 10.3389/fmed.2021.619978, 2021.
[37] D. M. Blei, "Probabilistic Topic Models," Communications of the ACM, vol. 55, no. 4, pp. 77–84, DOI: 10.1145/2133806.2133826, Apr. 2012.
[38] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," J. of Machine Learning Research, vol. 3, no. Jan, pp. 993–1022, 2003.
[39] T. Kwartler, Text Mining in Practice with R, ISBN: 9781119282013, John Wiley & Sons, 2017.
[40] F. Moisescu, "Issues Concerning the Relationship between Accounting and Taxation in Determining Financial Result," European J. of Sustainable Development, vol. 7, no. 1, pp. 287–297, Feb. 2018.
[41] R. M. Bushman and A. J. Smith, "Financial Accounting Information and Corporate Governance," J. of Accounting and Economics, vol. 32, no. 1–3, pp. 237–333, Dec. 2001.
[42] M. K. Power, "Auditing and the Production of Legitimacy," Accounting, Organizations and Society, vol. 28, no. 4, pp. 379–394, DOI: 10.1016/S0361-3682(01)00047-2, May 2003.
[43] R. A. G. Monks et al., Corporate Valuation for Portfolio Investment: Analyzing Assets, Earnings, Cash Flow, Stock Price, Governance and Special Situations, ISBN: 1576603172, John Wiley & Sons, 2010.
[44] C. Quinn and J. J. McArthur, "A Case Study Comparing the Completeness and Expressiveness of Two Industry Recognized Ontologies," Advanced Eng. Informatics, vol. 47, p. 101233, Jan. 2021.
[45] E. Rohrer, P. Severi and R. Motz, "Applying Meta-modeling to an Accounting Application," ceur-ws.org, vol. 2228, pp. 92–103, 2018, Accessed: Nov. 21, 2021.
[46] M. von Rosing and W. Laurier, "An Introduction to the Business Ontology," Chapter in Book: Sustainable Business: Concepts, Methodologies, Tools and Applications, pp. 1–24, IGI Global, 2020.
[47] C. M. Drury, Management and Cost Accounting, ISBN: 978-1-4737-4887-3, Springer, 2013.
[48] M. M. Mowen and Don R. Hansen, Managerial Accounting, ISBN: 1337116009, South-Western, 2007.
[49] K. Mircheska et al., "The Importance of Forensic Audit and Differences in Relation to Financial Audit," Int. J. of Sciences: Basic and Applied Research, vol. 54, no. 2, pp. 190–200, 2020.
[50] B. Nikkel, "Fintech Forensics: Criminal Investigation and Digital Evidence in Financial Technologies," Forensic Science Int.: Digital Investigation, vol. 33, p. 200908, Jun. 2020.
[51] K. Farmer, "A Qualitative Study: Forensics Coaches' Perceptions of Administrators' Leadership Styles and the Impact within Their Professional Learning Communities," Electronic Theses & Dissertations Center, [Online], Available: https://digitalcommons.acu.edu/etd/234, Apr. 2020.
[52] W. D. Huber, "Is Forensic Accounting in the United States Becoming a Profession?," J. of Forensic Investigation, vol. 4, no. 1, DOI: 10.1002/9781119488552, Apr. 2012.

ملخص البحث:

تقدّم هذه الورقة طريقةً لبناء نموذجٍ وجودي مبنيٍ على البحث لدعم التّعاون والإبداع. ويتضمّن مفهوم الإبداع التّعاوني عمليةً تسمح لأصحاب المصلحة المتعدّدين بالعمل معاً لإنتاج أفكار وحلول ومنتجات مبتكرة.

يدْمج النّموذج المقترح بين الذّكاء الاصطناعي ومعرفة الخبراء لبناء نظامٍ شاملٍ يضمّ جوانب متعدّدة من البحث والتطوير والإبداع. ولبيان جدوى الطّريقة المقترحة، تُبين هذه الورقة تطبيقها على مجال المحاسبة. فأوّلاً، يتمّ استخدام خوارزميات تعلُّم الآلة وتقنيات التّنقيب عن النُّصوص لاستخلاص العناصر الأساسية من مجموعةٍ ضخمةٍ من الأبحاث والدّراسات في علم المحاسبة. وبعد ذلك، تجري الاستفادة من معرفة الخبراء في حقل المحاسبة للتحقُّق من تلك العناصر وتنقيحها. ويُمكن استخدام النّموذج النّاتج كأساسٍ لنظامٍ قائمٍ على المعرفة لتعزيز التّعاون وتحليل حالة الإبداع.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

OVERVIEW OF MULTIMODAL DATA AND ITS APPLICATION TO FAKE-NEWS DETECTION

Nataliya Boyko

(Received: 29-Feb.-2024, Revised: 26-Apr.-2024, Accepted: 14-May-2024)

ABSTRACT

In the context of the growing popularity of social media over the past ten years, an urgent problem of fake news spreading has arisen, which underscores the research's relevance. The aim of this article is to assess the efficacy of multimodal approaches in detecting fake news, a pressing issue given the substantial impact misinformation can have on health, politics and economics. To achieve this goal, a multimodal approach was chosen that combines deep-learning frameworks and pre-trained models. This approach provides a comprehensive analysis of textual, visual and audio information, allowing for more accurate identification of disinformation sources. The use of various knowledge-transfer methods made it possible to process information efficiently, improving the quality of classification. The study conducted a thorough analysis of various data-collection strategies, as well as a comparative analysis of available multimodal approaches to fake-news detection and the datasets used. The results of this study included a detailed analysis of current research work in the field of fake-news detection and the development of a multimodal approach to this problem. Textual, visual and audio information was processed using pre-trained models and deep learning, achieving high accuracy in fake news detection. The results of the study indicated that the multimodal approach allows for more accurate identification of sources of disinformation and increases the efficiency of fake-news classification compared to other methods. A comparative analysis of various data collection strategies and datasets was also conducted, confirming the high efficiency of the approach under various conditions.

KEYWORDS

Technologies, Information environment, Neural networks, Testing approaches, Disinformation sources.

1. INTRODUCTION

The spread of false or unconfirmed information on the Internet is a pressing problem that covers not only the present, but also the early, stages of the network's development. Such information messages, whether true or not, are called "fake news" by G. Di Domenico and M. Visentin [1]. According to J.C. Culpepper [2], it is "deliberately false or misleading news." This phenomenon is commonly described as inauthentic information disseminated through various news platforms with the aim of misleading public opinion, as noted in the study by O. Ajao et al. [3].

The problem of fake news deserves attention in both political and social contexts, as well as in psychology. In the past, the main sources of news for society were traditional channels of information exchange, such as newspapers and television. However, as technology and the Internet have developed, the role of these channels has diminished and online content has become the most accessible way to obtain up-to-date information. In this regard, social media became the most effective platforms for receiving news, along with traditional media, such as television and newspapers. In addition to covering social movements that may be overlooked by mainstream media, social-media users are also actively engaged in a variety of issues, including politics, business, arts and entertainment.

Today, social-media platforms play a significant role in the spread of fake news. These platforms provide a wide audience reach, which contributes to the further spread of false information, as noted by H. Allcott and M. Gentzkow [4]. In the modern age, the spread of disinformation on social-media platforms has become an alarming phenomenon, including even the fabrication of data on COVID-19 pandemic remedies [5]. This creates serious difficulties in determining the true information in the online community. Under the influence of fake news, the society faces various negative consequences. These false-information messages spread among readers can affect various aspects of life. Examples of such impacts include changes in healthcare plans and strategies, as noted in the study by C.M. Greene and G.

N. Boyko is with Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Lviv, Ukraine. Email: boykonataliyaa@ukr.net

Murphy [6]. This can lead to poor decisions and the insufficient preparation of the society for various health challenges. Another serious aspect is the increase in scepticism about vaccination, as argued by M.S. Islam et al. [7]. Fake news can contain false claims about vaccines' harmfulness and ineffectiveness, which can undermine public trust in vaccination and contribute to the spread of infectious diseases. In addition, the economic impact is also an important consideration. Fake news can cause panic and unreasonable market reactions, leading to significant losses, as noted by E. Brown [8]. Inaccurate information about financial and economic events can mislead investors and entrepreneurs, affecting their investment and business decisions.

O.P. Prosyanyk and S.G. Holovnia [9] made a significant contribution to the study of fake-news detection on social media. The authors offer a detailed analysis and comparative evaluation of different methods for identifying disinformation in the online environment. They explore the quality and effectiveness of different approaches, including text analysis, network structure and the use of machine learning algorithms. D.O. Tatarchuk's work [10] is an important addition to the study of fake news and disinformation in social media. The author of this study concentrates on the tools and methods available for confirming the accuracy of information and identifying fake news on social media. The study provides an analysis of innovative approaches that can be useful for fact-checking and combating the spread of disinformation in the online environment. This work is significant, because it contributes to the development of practical strategies and tools for detecting fake-news messages, which will help preserve the quality and reliability of information on social media.

I. Ivanova and O. Lysytskaia [11] examined the use of postmodernism as a manipulative tool in Ukrainian advertising. They explored how advertising campaigns use postmodern elements to create consumer culture and support advertising strategies and highlighted the impact of these artistic approaches on consumers and contemporary culture. The study emphasizes the role of art and artistic expression in contemporary advertising, helping reveal the manipulative potential of advertising and its impact on shaping consumer habits and identity. The paper by Y.F. Shtefaniuk et al. [12] analyzes the methods of detecting fake news and their applicability to counteract information propaganda. The paper explores the possibility of using existing techniques to identify and control information manipulation, which is important in the context of the modern information environment and the spread of disinformation. V. Bazylevych and M. Prybytko [13] discuss the creation of a fake news detection system using data-science methods. The study presents a practical approach to combating the problem of fake news, that uses data analysis and machine-learning algorithms to automatically detect false information, which can be an important tool for ensuring the reliability of information in the digital world.

The aim of this article is to assess the efficacy of multimodal approaches in detecting fake news, a pressing issue given the substantial impact misinformation can have on health, politics and economics. Specifically, the research seeks to answer two questions: How can multimodal methods be optimized to handle multilingual data effectively, ensuring accurate detection across diverse linguistic environments? And, what advancements can be made in the synchronization of text and image data to enhance the accuracy and reliability of fake-news detection? The study contributes significantly to the field of fake-news detection by presenting novel insights into the integration of complex datasets and the refinement of multimodal techniques. It proposes innovative strategies for improving data processing and analysis, which are crucial for developing more robust systems capable of adapting to the evolving nature of misinformation across different media and languages.

2. LITERATURE REVIEW

In recent years, the scientific community has shown considerable interest in developing methods for the automatic detection of fake news. Researchers such as A. Thota et al. [14] have dedicated several studies to this problem. Researchers have proposed various approaches to fake-news detection, depending on the type of data. We can divide this classification into two categories: unimodal and multimodal. To perform fake-news detection, unimodal methods use only one type of input to perform the task of fake-news detection. For instance, we can use text or images separately to verify the authenticity of a news item.

In multimodal approaches, fake-news information is identified by analyzing several types of data, such as audio, video, images and text, as noted in L. Donatelli et al. [15]. This allows us to consider different

aspects of the content and create a more complete picture. Researchers are actively researching the use of multimodality to detect fake news, with numerous attempts to enhance its effectiveness. Many researchers, such as V.K. Singh et al. [16], A. Giachanou et al. [17] and Y. Khimich [18], have achieved significant results in this area. They have shown that multimodality-based approaches can achieve greater accuracy than unimodal methods.

There are several approaches to classifying fake news. Some researchers consider news as a binary classification, dividing news into real and fake news, as shown by H. Ahmed et al. [19], D. Kumar Sharma and S. Sharma [20] and S. Garg & D. Kumar Sharma [21]. Other researchers consider this task as a multi-class classification or use of regression and clustering methods, as shown in the studies of H. Karimi et al. [22] and R. Oshikawa et al. [23]. Researchers have developed a variety of single-modal and multimodal methods based on the current state-of-the-art in fake-news detection. The review of existing multimodal approaches in this article allows us to present important results and directions for further research. Multimodality, which includes textual and visual characteristics, can indeed increase the effectiveness of fake-news detection, given the semantic features of the data. The diversity of disinformation materials has prompted many scholars to focus on the development of multimodal methods and these efforts promise interesting results for the future.

Currently, there are many single-modal and multimodal methods for detecting fake news. A review of existing multimodal approaches allows us to identify important results and directions for further research. A multimodal approach that combines the analysis of textual and visual characteristics promises to increase the effectiveness of fake-news detection by considering the semantic features of the data [24]-[25]. This integration makes it possible to analyze information more comprehensively and create a more complete picture for classification. In the context of widespread misinformation and data diversity, many researchers are actively focusing on the development of multimodal approaches. These efforts promise interesting results for the future, helping combat the problem of fake news and providing greater reliability in determining the authenticity of information in the digital age.

3. MATERIALS AND METHODS

In today's information environment, the spread of fake news is an urgent problem that requires immediate solution. This article explores approaches and methods for detecting disinformation. The first stage of the study includes an analysis of literature sources relevant to this problem. The scientific community pays particular attention to the various strategies and methodologies developed to detect disinformation. This stage not only helps get an overview of different approaches to the problem, but also identify best practices in the field. After that, researchers study and analyze the methods used to detect fake news. We aim to develop more effective strategies to combat disinformation and ensure the accuracy of information in our information environment.

The study pays special attention to the analysis of machine-learning methods used to detect fake news using multimodal data. Modern technologies allow working with different types of information, such as text, images, sounds and videos. Combining these modalities significantly increases the accuracy of disinformation detection. This article discusses various algorithms and approaches, including deep learning and content analysis based on multimodal data. These methods help identify characteristic patterns and anomalies in information, which contributes to a more accurate identification of fake news. The study of multimodal data and its application in fake-news detection is an important step in the fight against disinformation. An extensive literature review and analysis of machine-learning methods allow us to develop effective strategies based on modern technologies for more accurate detection of fake news in various multimodal data. An important aspect of the study is to assess the reliability of the developed models in different contexts of information noise. The results' applicability in real-world information environments is also taken into account.

This study bases its data-collection methodology on a systematic literature review. This method allows us to cover an extensive database of scientific articles available in leading online repositories, such as IEEE Xplore and ScienceDirect, among others. The presence of publications from various fields allows for a more complete and comprehensive understanding of multimodal data and its impact on effective fake-news detection. The next important stage is the analysis of the collected articles. This stage entails a thorough and careful examination of each article, selecting those that are most relevant to the study's topic and objectives. We exclude papers that do not meet the research objectives. We then subject the

selected articles to in-depth analysis. This stage is important, because it allows us to distinguish from many studies the approaches that are most successful in the field of multimodal fake-news detection.

4. RESULTS

The advantage of deep learning lies in its ability to automatically extract features from raw data, unlike classical machine learning, which requires the intervention of specialists. The process of creating more general features contributes to the development of more specific characteristics. Deep learning is applied using Deep Neural Networks (DNNs) structured in three layers: convolutional, pooling and full connected. In the context of detecting fake news, the most commonly used deep-learning algorithms are Convolutional Neural Networks (CNN), bidirectional Long Short-Term Memory (LSTM) and ResNet50. A variety of datasets effectively employ them. As shown in Table 1, a variety of datasets are available to evaluate which fake-news detection method performs best compared to other approaches.

Table 1. Fake-news datasets with their characteristics.

Dataset	Processing	Expansion	Preliminary processing	Fake news	Truthful news	Total
ISOT fake-news dataset	Not conducted	Not completed	Not conducted	23502	21417	44919
Fake-news data	Not conducted	Not completed	Not conducted	10413	10387	20800
News (fake or real)	Not conducted	Not conducted	Not conducted	3154	3161	6315
Fake-news detection	Not conducted	Not conducted	Not conducted			
Research dataset	Not processed	Not expanded	Could not be preliminarily processed	2135	1870	4005

Source: compiled by the author.

Table 1 provides details on the various datasets utilized in research and analysis pertaining to the identification of fake news and false information. Let's consider each row of the table in more detail:

1. ISOT fake-news dataset: This database contains information about fake news. It consists of 44919 records, of which 23502 are fake news and 21417 are true news. No processing, enhancement or pre-processing has been performed on this dataset.
2. Fake-news data: This dataset contains data on false news. It contains 20,800 records, of which 10413 are fake news and 10387 are true news. There is also no processing, enhancement or pre-processing in this dataset.
3. News (fake or real): This dataset contains 6315 entries that can be either fake or real news. We have not processed or enhanced it, nor have we performed any pre-processing.
4. Fake-news detection: There is no specific data in this row. This is likely a place where data on fake-news detection could be included, but it is not.
5. Research dataset: The 4005 records in this research dataset have not undergone any processing, enhancement or pre-processing.

The overall context of this table is that it provides information on the scope and origin of various datasets that can be used to analyze fake and false news.

A Convolutional Neural Network (CNN) consists of input and output layers, a sequence of hidden layers and software for pooling and convolutional operations that transform the input data. Figure 1 illustrates the structure of this approach. CNN is a deep-learning method that trains each object in the image with weights and shifts to distinguish them from each other. A convolutional neural network is a deep-learning method that assigns importance (configured weights and biases) to different objects in an image, distinguishing them from each other. There has been a lot of research on the use of convolutional neural systems in various fields of computer science, including computer vision, where they are considered to be at the forefront. At present, the field of natural-language processing actively utilizes CNNs.

This architecture serves as an example of a typical deep-learning model for processing text data, such as in text-classification tasks. Let's look at each layer of the architecture and its purpose:

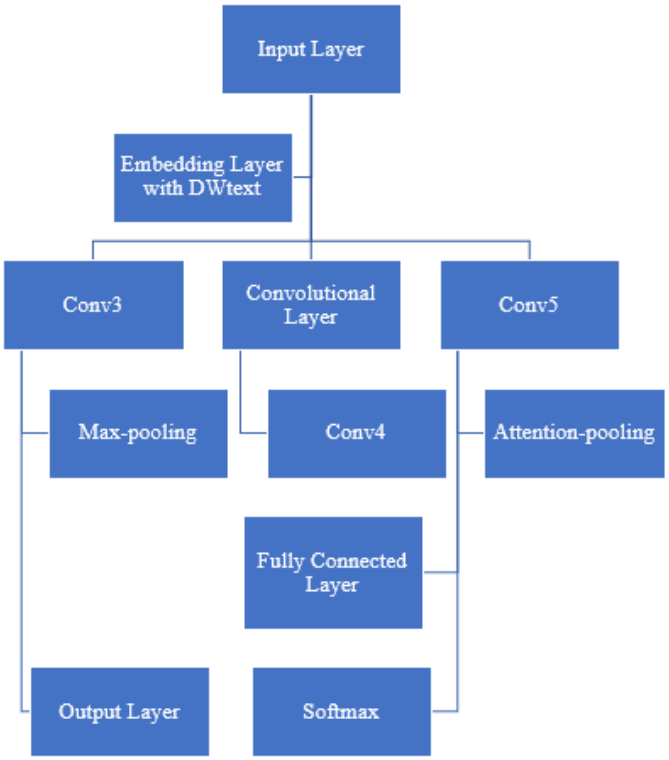


Figure 1. Structure of a convolutional neural network.
Source: compiled by the author.

1. Input Layer: This is the first layer that accepts input. This scenario likely involves processing and classifying textual content.
2. Embedding Layer with DWtext: This layer is responsible for converting text words (or tokens) into fixed-length vectors. This layer aims to generate a text representation suitable for subsequent processing.
3. Conv3: This layer employs convolutional filters of width 3 to detect local features in the text-vector representation.
4. Max-pooling: After the convolutional layer, the maximum pooling operation is used to reduce the dimensionality of the obtained features and highlight the most important ones.
5. Output Layer: This layer is probably responsible for classifying the text into several classes or categories.
6. Convolutional Layer: It follows this for further feature detection.
7. Conv4 i Conv5: These convolutional layers can use filters of other widths to detect different types of features in the text.
8. Attention-pooling: This operation is responsible for highlighting important parts of the text, considering their context and significance.
9. Fully Connected Layer: This layer further processes the obtained features and reduces their dimensionality before transferring them to the final layer.
10. Softmax: This layer is responsible for calculating the probabilities of the input text belonging to different classes or categories.

This architecture integrates various text-processing operations, such as convolution and pooling, and uses an attention function to improve the quality of text processing and classification. This analysis could be used to look more closely at how these layers work together and how they affect the model's results.

The behaviour of the passive-aggression algorithm compared to naive Bayes and the support-vector method was more efficient. It also coincided with the completeness index for naive Bayes and the support-vector method. S. Singhal et al. [26] used a two-channel convolutional neural network to identify news stories. They first factorized the text and then performed classification. The COVID-19 news data served as the basis for the analysis. We also applied this method to data from the Beijing Academy of Artificial Intelligence. The dataset includes 38471 records, of which 19285 are labelled as

“Distorted” and 19186 as “Genuine”. For the text, we used ultra-accurate neural networks, multi-channel convolutional neural networks, RCNN, DPCNN, transformer, BERT and DC-CNN. The results for the Covid-19 dataset are shown in Table 2.

Table 2. Results for the COVID-19 dataset.

Model	Completeness	Accuracy	F1
DC-CNN	0.947	0.948	0.948
Transformer	0.88908	0.8299	0.8586
Text CNN	0.8926	0.8926	0.8926
ERINE	0.9752	0.8771	0.9235
Multi-channel CNN	0.9243	0.9194	0.9218
BERT	0.882	0.9281	0.9045
DPCNN	0.7683	0.869	0.8155

Source: compiled by the author.

Table 2 shows the results of evaluating different models in the classification task and it appears to contain metrics for each model. Let’s consider each column of the table:

1. Model: This column contains the names of the different machine-learning or deep-learning models that were evaluated in the study.
2. Recall: This is a metric that determines what proportion of positive instances was correctly recognized by the model. High completeness implies that the model does not miss many positive cases.
3. Precision: This metric shows what proportion of the positive instances recognized by the model is positive. High accuracy indicates that the model makes few false positives.
4. F1 Score: This metric combines completeness and accuracy into one metric and provides an overall assessment of the model’s performance. A high F1 score indicates that the model has a good balance between completeness and accuracy.

Thus, this table provides information on the performance of different models in the classification task, focusing on metrics, such as completeness, accuracy and F1 score. The analysis aims to help choose the model that is best suited for a particular task.

This multimodal CNN uses two parallel convolutional structures in a single architecture to detect hidden features in text and visual information. This model is highly extensible, allowing for easy integration of other news components, including overt and covert features based on textual and visual information. In addition, training a convolutional neural network model is much faster than LSTM or many other recurrent neural network (RNN) models, as CNN sees all the input information at once. This method's process of processing fake news includes a textual branch (word embedding), which is the process of converting words into sequences of integers and a visual branch, which includes creating a feature vector from an image by extracting the resolution and number of faces. The vector is then converted into visually explicit features using a fully-connected layer, such as a convolutional layer, maximum pooling layer and so on. The process also uses rectified linear neurons, regularization and neural-network training [27].

Deep learning's main advantage is its capability to autonomously extract features from raw data, contrasting with traditional machine learning which depends heavily on manual feature engineering by experts. This capability is pivotal in developing nuanced characteristics from more general features. Key models used in the realm of fake-news detection include Convolutional Neural Networks (CNNs), bidirectional Long Short-Term Memory networks (LSTM) and ResNet50, which are applied across a spectrum of datasets. Discussing the application of these algorithms, a variety of datasets are utilized to assess the performance of each method relative to others. For instance, the ISOT fake-news dataset, encompassing 44,919 records with a nearly equal split between fake and real news, serves as a benchmark with no processing or preliminary operations applied. Similarly, other datasets like the fake-news data and News (fake or real) follow suit, providing a rich basis for comparison without prior manipulations. Such a setup ensures that the assessment is focused on the effectiveness of the deep-learning models themselves.

Furthermore, the role of CNNs in this field is significant. A typical CNN architecture involves multiple

layers including input and output layers, interspersed with several hidden layers that perform convolutional and pooling operations. This setup helps in identifying and distinguishing between various features in the data. The structure is adept, not only in handling visual data, but has found extensive use in processing text for tasks like text classification. For example, the architecture might start with an input layer taking in textual content, followed by an embedding layer that translates text into vectors. Subsequent convolutional layers with varying filter widths help in detecting different textual features which are then refined through pooling layers. Attention mechanisms are incorporated to emphasize significant textual components before reaching the output layer which classifies the content.

In terms of performance, various models exhibit distinct strengths. The experimental analysis on datasets such as those related to COVID-19 news highlights how different configurations and models fare. For instance, a DC-CNN model might achieve high scores across completeness, accuracy and the F1 metric, suggesting a strong ability to handle fake-news detection. Other models like BERT and Transformer also demonstrate commendable performance, though with variations across different datasets. The choice of a multimodal approach, integrating both textual and visual information, emerges as superior in certain contexts. By processing features through parallel convolutional structures within a single CNN architecture, this method enhances the model's ability to discern more complex patterns and interactions in the data, making it highly effective and adaptable for various types of news.

A new approach to machine learning can benefit fake-news classification, which is the main idea of this study. This empirical study conducted the fake-news detection process on five benchmark datasets, utilizing various combinations of machine-learning vectors and classifiers. Experimental results showed that the passive aggression classifier with TF-IDF vectorizer is the best combination for fake-news detection using machine learning methods. Since deep learning is now playing an important role in solving all practical real-world problems, our future work is aimed at finding the best deep-learning classifier for fake news detection. In addition, experimental analysis can be conducted with more effective features by improving the vectorization with augmentations and increasing the size of the dataset. Table 3 shows the results of the experiments.

Table 3. Experimental results driven by vectorization improvements through data expansion and data volume.

Model		Kaggle Fake_Recall	Covid Fake_Recall	Gossipcop Fake_Recall	ISOT Fake_Recall	Politifact Fake_Recall	Welfake Fake_Recall
Multinomial Naive Bayes	COUNT	89.3	92.1	83.7	95.5	84.2	81.1
	TF-IDF	85.7	90.2	82.6	92.2	84.5	81.3
Passive Aggressive	COUNT	89.4	92.7	79.1	98.7	80.2	83.8
	TF-IDF	93.5	92.9	79.7	99.2	81.8	83.9
Support-vector Machine	COUNT	86.2	93.1	84.8	99.1	74.8	82.5
	TF-IDF	92.2	93.7	85.1	98.8	83	83.5
Random Forest	COUNT	90.5	92	83	98.9	73.9	80.9
	TF-IDF	83.5	92.4	83.9	98	73.9	81.1

Source: compiled by the author.

Table 3 presents the results of the evaluation of different machine-learning models that were used to classify fake news and real news in different data sources. Table 3 includes the following components:

1. Model: these are the names of the machine-learning models that were used to classify the news.
2. Feature type: Two types of features are specified: “COUNT” (quantitative approach) and “TF-IDF” (term-weighted by document inversion and term frequency).
3. Kaggle Fake_Recall, Covid Fake_Recall, Gossipcop Fake_Recall: these are percentage values that indicate the quality of the models’ performance in detecting fake news in different datasets. “Recall” indicates the proportion of fake news that was recognized by the model among all fake news in the respective dataset. Higher recall values indicate a better ability of the model to detect fake news.

Table 3's overall purpose is to compare different machine learning models and different types of features in terms of their ability to detect fake news in different data sources. For example, it can be seen that the

support-vector machine model with TF-IDF features achieves a high recall for most data sources, which may indicate its effectiveness in detecting fake news.

In recent research, DNNs have shown impressive results in data representation. DNNs are powerful structures that allow you to model sequential data using loops within a neural network. LSTMs and gated recurrent units (GRUs), two variants of recurrent neural networks with permanent memory, effectively handle long-term dependencies. This simplifies the detection of long-term dependencies and avoids the problem of gradient vanishing, which is typical for conventional RNNs. The memory-management operation in LSTMs is based on input, output and forgetting gates. The study by A. Giachanou et al. [17] proposed an interesting use of RNNs. The authors noted that the unidirectional LSTM_RNN has a training accuracy of 0.997 and a validation accuracy of 0.89. The testing accuracy reached 0.92. Similarly, the bidirectional LSTM-RNN has a training accuracy of 1, a validation accuracy of 0.97 and a testing accuracy of 0.99. The RNN searches for the given structure in a topological order by always applying the same weights. This makes structured predictions for structures with different dimensions. J. Ma et al. [28] developed two recursive neural models, using bottom-up and top-down tree neural networks, to represent the structure of tweets. A. Giachanou et al. [17] introduced a tree-based LSTM that utilizes an attention mechanism to identify connections between image regions and descriptive words.

The bidirectional LSTM architecture arranges two LSTM memories, one for long-term dependencies and the other for short-term dependencies, taking into account their interaction [21]. This architecture is a type of RNN. In this study, we use a bidirectional LSTM architecture with an input layer of size 1000 and an embedding layer. This structure uses bidirectional LSTMs that are symmetric in both directions. The structure performs classification using a global maximum pooling layer, a fully-connected layer and an output layer. To minimize the problem of gradient vanishing, Q. You et al. [29] implemented a final block in deep learning. This block allows the signal to flow directly through it, bypassing the previous layers. Unlike the standard CNN architecture, ResNet uses final blocks instead of two consecutive sets of convolutional layers and pooling. This introduction of a finite structure increases the architecture's robustness to the problem of gradient vanishing and improves the ability to retrain the network and retain previously-learned information.

Implementing a multimodal approach to detect fake news in real-world settings can significantly enhance the robustness and accuracy of detection systems. This method, which integrates both textual and visual data, allows for a comprehensive analysis that leverages the strengths of different data types. One practical recommendation for effective implementation involves developing a streamlined data-ingestion pipeline that can handle diverse data formats efficiently. Organizations should focus on establishing robust pre-processing mechanisms that can cleanse and standardize incoming data to ensure consistency and reliability in the analysis. Additionally, training the models with a diverse and extensive dataset that reflects the complexity and variety of real-world data is crucial. This training should include examples from various sources and contexts to minimize bias and improve the generalizability of the models.

However, challenges such as data scarcity, especially in terms of labeled datasets for training, can hinder the effectiveness of a multimodal approach. To overcome this, organizations could collaborate to share resources and data, or leverage synthetic data-generation techniques to enrich their training datasets. Moreover, ensuring the privacy and security of the data while handling sensitive information must be a priority, requiring rigorous compliance with data-protection regulations. Another limitation is the computational cost and complexity of processing multiple data types simultaneously. This can be mitigated by optimizing neural network architectures, possibly through pruning techniques that reduce the model size without significantly impacting performance, or by implementing more efficient algorithms that can process data faster. Furthermore, there is the issue of integrating and synchronizing different types of data. Effective alignment techniques are necessary to ensure that textual and visual data complement each other appropriately in the analysis. This might involve developing advanced feature-extraction techniques that can accurately link features from text and images, enhancing the model's ability to detect discrepancies or manipulations.

Finally, the dynamic nature of news and information propagation requires these systems to continuously learn and adapt. Implementing feedback loops where the model's predictions are regularly reviewed and refined can help maintain the accuracy and relevance of the system. Regular updates to the model based

on new data and emerging trends in fake news are essential for sustaining performance over time. By addressing these challenges with strategic planning and innovative solutions, the practical application of a multimodal approach to fake-news detection can be effectively realized, leading to more resilient and accurate systems.

5. DISCUSSION

P. Kumar Verma et al. [30] proposed a new approach for detecting fake news using local and global text semantics, called Message Credibility (MCred). The authors demonstrate through experimental results on the popular Kaggle dataset that MCred is more accurate than state-of-the-art methods. Biased data fusion combines with the CNN classifier to classify fake news. This method is based on the analysis of both local and global text semantics. In their study, they demonstrated that this approach to the popular Kaggle dataset improves the accuracy of the existing model by 1.1%. In their approach, they defined a bilateral data fusion and combined it with the CNN classifier to classify fake news.

K. Sharifani et al. [31] extended the same idea to more generalized fake news-related data and used it to detect fake-news fragments. The naive Bayesian classifier achieved impressive accuracy, demonstrating an accuracy of 0.85, a precision of 0.89 and a completeness of 0.87. The same figures for the passive-aggression method were 0.93, 0.92 and 0.89. The support-vector method had an accuracy of 0.84 and a precision of 0.82. According to their experiment results, the support-vector method also has a completeness of 0.87.

Y. Wang et al. [32] stated that an analysis of the accuracy of different combinations was conducted to compare and determine the best combination of classifier and vectorizer. The data clearly shows that the combination of the Passive-aggression classifier and TF-IDF vectorizer provides an accuracy of 93.5% when analyzing Kaggle data related to fake and real news, 99.2% when analyzing the ISOT dataset and 83.9% when analyzing the Welfake dataset, which includes fake and real news.

Comparing this study with C. Song et al. [33], it is obvious that both approaches pay attention to analyzing the impact of various factors on the results under study. However, it should be noted that the researchers focus mainly on time-series analysis and long-term trends, while the present study additionally considers short-term factors and actively uses machine-learning methods to predict outcomes more accurately. This methodological difference allows for a more complete and accurate coverage of both long-term and short-term influences on the final results. Thus, the analysis and comparison of the results of both studies highlight the importance of considering both long-term and short-term factors when analyzing and predicting outcomes in this topic area.

Comparing the results of this study with the work of I. Goodfellow et al. [34], the following similarities and differences are revealed. Both approaches focus on analyzing the impact of various factors on the outcomes studied, but there is an important difference in the methodological approach. The researchers focused primarily on time-series analysis and long-term trends, while the present study additionally considers short-term factors and actively uses machine-learning techniques to more accurately predict outcomes. This methodological difference allows for a more complete and accurate coverage of both long-term and short-term influences on the final results. Thus, the approach of this paper complements the researchers' work by providing a more in-depth analysis of long-term trends, whereas their study focuses on time-series analysis. We can identify similarities and differences between the study conducted by J. Du et al. [35] and the present research in the field of multimodal data analysis and its application to fake-news detection. Both studies draw attention to the influence of various factors on the study's results, but they focus on different aspects. The researchers pay more attention to short-term factors and use machine-learning techniques to make accurate predictions in the current context, whereas this study focuses on long-term trends and time series, aiming to identify deep patterns and long-term influences on fake-news detection.

J. Wang et al. [36] conducted a similar study to this one; so it can be noted that both studies focus on analyzing multimodal data to detect fake news. However, they differ in their methodological approach. The researchers pay more attention to the use of deep-learning and ensemble methods to improve accuracy, while this study covers a wide range of aspects of multimodal data and provides a more comprehensive analysis of their impact on fake-news detection. Both studies provide valuable insights for addressing the challenges of countering disinformation and ensuring the accuracy of information by

revealing a variety of methods and approaches for analyzing multimodal data sources. The study by S. Xiong et al. [37] can be compared to the present work and similar or different aspects can be identified. Both studies discuss multimodal data and its role in detecting fake news. However, they approach this issue from different perspectives. The study by the researchers puts more emphasis on analyzing textual information and uses deep neural networks to detect patterns in the text. This study focuses on combining textual and visual information using multimodal data analysis methods. It is also important to note that both studies contribute to the development of disinformation-detection methods and provide a deeper understanding of the role of multimodal data in this context.

There are several similarities and differences between J. Xue et al. [38] and this study. The review and use of multimodal data to detect fake news is the focus of both articles. However, the approaches to this topic are slightly different. The study by the researchers focuses on the use of deep-learning algorithms, such as convolutional and recurrent neural networks, to process textual information. This study pays more attention to multimodal aspects, including the analysis of both textual and visual information. Both studies are important for the development of disinformation and fake-news detection methods and their combined contribution helps better understand the role of multimodal data in this context. This study also has similarities and differences with that of the researchers. Both articles focus on analyzing and using multimodal data to detect fake news. However, there are differences in the approaches. The researchers' work focuses more on analyzing textual information using natural language processing methods. This study actively uses visual data in addition to text, expanding the scope of analysis and potentially improving the accuracy of fake-news detection. These two studies complement each other, enriching the understanding of the impact of multimodal data on the effectiveness of countering disinformation.

We can identify some similarities and differences between this study and the work of J. Hua et al. [39]. Both studies aim to analyze multimodal data and use it to detect fake news. Both approaches are likely to use machine-learning and data-mining techniques to identify patterns and features that are characteristic of fake news. However, there may be differences in the choice of features or machine-learning algorithms that may affect the effectiveness of detection. It is also important to note that this study is likely to focus on the visual aspects of the data, as multimodal sources of information contain images in addition to text. This may add a layer of complexity to the analysis and help better detect fake news. Compared to the study by V.K. Singh et al. [16], this study also analyzes multimodal data and uses it to detect fake news. Both studies are likely to look at different features and characteristics in textual and visual information that can help separate fake news from real news. We may use similar machine-learning and data-analysis techniques to identify patterns typical of fake news. However, differences in the choice of algorithms, data processing and features used may lead to differences in detection efficiency [40]-[43]. Furthermore, this approach likely takes into account modern machine-learning methods and approaches that may have emerged since the researchers' publication, making this study more relevant than earlier work [44].

To summarize, it can be concluded that the analysis of the results of this study in comparison with the works of other authors has revealed similarities and differences in approaches to analyzing multimodal data to detect fake news. While many studies emphasize the use of machine-learning and data-mining techniques, our approach complements this by integrating short-term factors and actively utilizing machine-learning techniques for accurate predictions. This integrated approach allows for a deeper analysis of the impact of various factors on the results and has the potential to develop effective strategies for detecting disinformation and fake news.

6. CONCLUSIONS

Detecting fake news plays an important role, as it has a significant negative impact on various aspects of life, including health, politics and economics. That’s why most researchers focus on developing advanced algorithms to detect and identify fake news. These systems allow for the fast and reliable detection of fake news. Multimodal approaches overcome the limitations of using textual features alone. The present review study has examined state-of-the-art multimodal approaches. The review analysis concludes that there has been minimal advancement in the field of fake-news detection. The results demonstrate the widespread use of CNN-based models for image processing and RNN-based models for maintaining consistent information in text documents. However, the use of social media often limits

these models' ability to process multilingual data. In addition, there is a large amount of work on fake-news detection, while research on developing datasets available for public use remains limited. Such datasets are an important basis for the development of more effective methods for detecting disinformation and fake news.

The effectiveness of multimodal approaches lies in the fact that they combine information from different sources, such as text and images, which allows for a more complex and reliable fake-news detection model. In addition, one of the main challenges is the lack of high-quality and diverse training data for fake-news detection. It is important to continue developing and disseminating such datasets, so that researchers can test and improve their methods. To ensure more accurate fake-news detection in different language environments, we should also focus on developing methods that can handle multilingual data and take into account the various features of language structures.

In the field of fake-news detection, future research could focus on several key areas to refine and extend the capabilities of multimodal approaches. Enhancing the ability to process and analyze multilingual data is crucial, given the global nature of information dissemination. Developing algorithms that can adapt to different linguistic features and cultural contexts will be essential for improving detection accuracy across diverse populations. Another promising direction is the integration of more sophisticated natural-language processing techniques that leverage semantic analysis to understand the nuances and implied meanings within text. This could involve the use of advanced machine-learning models like transformers that have shown significant potential in understanding context and relationships within data. Additionally, improving the synchronization of textual and visual data in multimodal systems is vital. Research could explore more effective ways to correlate features from different modalities, ensuring that the combined data provides a clearer and more accurate picture of potential misinformation. There is also a pressing need to create and make publicly available more comprehensive datasets that include a variety of fake and real-news examples. These datasets should encompass a range of media formats, languages and cultural contexts to foster broader research and application of detection technologies. Finally, considering the rapid evolution of misinformation techniques, continuous updates to models and methods are necessary. Implementing adaptive-learning systems that can evolve with new trends in fake news and misinformation could provide more resilient and enduring solutions in the fight against fake news.

REFERENCES

[1] G. Di Domenico and M. Visentin, "Fake News or True Lies? Reflections about Problematic Contents in Marketing," International Journal of Market Research, vol. 62, no. 4, pp. 409-417, 2020.

[2] J. C. Culpepper, "Merriam-Webster Online: The Language Center," Electronic Resources Review, vol. 4, no. 1/2, pp. 9-11, 2000.

[3] O. Ajao, D. Bhowmik and S. Zargari, "Sentiment Aware Fake News Detection on Online Social Networks," Proc. of the ICASSP 2019 – 2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 2507-2511, Brighton, UK, 2019.

[4] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211-236, 2017.

[5] A. K. Dubey and M. Saraswat, "Fake News Detection through ML and Deep Learning Approaches for Better Accuracy," Proc. of Advances in Computational Intelligence and Communication Technology (CICT 2021), Part of the Book Series: Lecture Notes in Networks and Systems, vol. 399, pp. 13-21, 2022.

[6] C. M. Greene and G. Murphy, "Quantifying the Effects of Fake News on Behavior: Evidence from a Study of COVID-19 Misinformation," J. of Experimental Psychology: Applied, vol. 27, no. 4, pp. 773-784, 2021.

[7] M. S. Islam et al., "COVID-19 Vaccine Rumors and Conspiracy Theories: The Need for Cognitive Inoculation against Misinformation to Improve Vaccine Adherence," PLoS ONE, vol. 16, no. 5, Article no. e0251605, 2021.

[8] E. Brown, "Online Fake News is Costing us $78 Billion Globally Each Year," ZD NET, [Online], Available: https://www.zdnet.com/article/online-fake-news-costing-us-78-billion-globally-each-year/, 2019.

[9] O. P. Prosyanyk and S. G. Holovnia, "Methods of Detecting Fake News in Social Networks," Proc. of the Int. Conf. on Scientific and Practical Developments (The European Development Trends in Journalism, PR, Media and Communication), pp. 81-85, 2021.

[10] D. O. Tatarchuk, "Fact-checking Tools for Detecting Fake Information in Social Media," Proc. of the Int. Conf. on Scientific and Practical Developments (The European Development Trends in Journalism, PR, Media and Communication), pp. 84-86, 2020.

[11] I. Ivanova and O. Lysytskaia, "Postmodernism As a Manipulative Technology in Modern Ukrainian Advertising: The Artistic Dominant Characteristic," Int. J. of Philology, vol. 11, no. 1, pp. 108-113, 2020.

[12] Y. F. Shtefaniuk, I. R. Opirskyy and O. I. Harasymchuk, "Analysis of Application of Existing Fake News Recognition Techniques to Counter Information Propaganda," Ukrainian Scientific J. of Information Security, vol. 26, no. 3, pp. 139-144, 2020.

[13] V. Bazylevych and M. Prybytko, "Fake News Detection System Based on Data Science," Technical Sciences and Technologies, vol. 4, no. 22, pp. 91-95, 2020.

[14] A. Thota, P. Tilak, S. Ahluwalia and N. Lohia, "Fake News Detection: A Deep Learning Approach," SMU Data Science Review, vol. 1, no. 3, pp. 1-10, 2018.

[15] L. Donatelli, N. Krishnaswamy, K. Lai and J. Pustejovsky, Proc. of the 1st Workshop on Multimodal Semantic Representations (MMSR), Association for Computat. Linguist., Groningen, Netherlands, 2021.

[16] V. K. Singh, I. Ghosh and D. Sonagara, "Detecting Fake News Stories *via* Multimodal Analysis," Journal of the Association for Information Science and Technology, vol. 72, no. 1, pp. 3-17, 2021.

[17] A. Giachanou, G. Zhang and P. Rosso, "Multimodal Multi-image Fake News Detection," Proc. of the 2020 IEEE 7th Int. Conf. on Data Science and Advanced Analytics (DSAA), pp. 647-654, Sydney, Australia, 2020.

[18] Y. Khimich, "Formation of Information Culture of Higher Education Students in the Digital Age," Library Science-Record Studies-Informology, vol. 1, no. 1, pp. 86-95, DOI: 10.32461/2409-9805.1.2023.276773, 2023.

[19] H. Ahmed, I. Traore and S. Saad, "Detection of Online Fake News Using N-gram Analysis and Machine Learning Techniques," Proc. of the 1st Int. Conf. on Intelligent, Secure and Dependable Systems in Distributed and Cloud Environments (ISDDC 2017), Part of the Book Series: Lecture Notes in Computer Science, vol. 10618, pp. 127-138, 2017.

[20] D. Kumar Sharma and S. Sharma, "Comment Filtering-based Explainable Fake News Detection," Proc. of 2nd Int. Conf. on Computing, Communications and Cyber-security, Part of the Book Series: Lecture Notes in Networks and Systems, vol. 203, pp. 447-458, 2021.

[21] S. Garg and D. Kumar Sharma, "New Politifact: A Dataset for Counterfeit News," Proc. of the 2020 9th Int. Conf. System Modeling and Advancement in Research Trends (SMART), pp. 17-22, Moradabad, India, 2020.

[22] H. Karimi, P. Roy, S. Saba-Sadiya and J. Tang, "Multi-source Multi-class Fake News Detection," Proc. of the 27th IEEE Int. Conf. on Computational Linguistics, pp. 1546-1557, Santa Fe, New Mexico, USA, 2018.

[23] R. Oshikawa et al., "A Survey on Natural Language Processing for Fake News Detection," Proc. of the 12th Language Resources and Evaluation Conf., pp. 6086-6093, Marseille, France, 2020.

[24] T. Martyniuk, O. Voytsekhovska, M. Ochkurov and O. Voinalovych, "Properties of Unit Encoding of Information in the Context of Functional Control," ITKI, vol. 57, no. 2, pp. 43-49, 2023.

[25] V. Varenko, "Electronic Communications in Information and Analytical Activities," Library Science-Record Studies- Informology, vol. 1, no. 1, pp. 53-58, DOI: 10.32461/2409-9805.1.2023.276765, 2023.

[26] S. Singhal et al., "SpotFake: A Multi-modal Framework for Fake News Detection," Proc. of the 2019 IEEE 5th Int. Conf. on Multimedia Big Data (BigMM), pp. 39-47, Singapore, 2019.

[27] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li and P.S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection," arXiv: 1806.00749v3, DOI: 10.48550/arXiv.1806.00749, 2018.

[28] J. Ma, W. Gao and K. F. Wong, "Rumor Detection on Twitter with Tree-structured Recursive Neural Networks," Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 1980-1989, Melbourne, Australia, 2018.

[29] Q. You, L. Cao, H. Jin and J. Luo, "Robust Visual-textual Sentiment Analysis: When Attention Meets Tree-structured Recursive Neural Networks," Proc. of the 24th ACM Int. Conf. on Multimedia (MM'16), pp. 1008-1017, DOI: 10.1145/2964284.2964288, 2016.

[30] P. Kumar Verma, P. Agrawal, V. Madaan and R. Prodan, "MCred: Multi-modal Message Credibility for Fake News Detection Using BERT and CNN," Journal of Ambient Intelligence and Humanized Computing, vol. 14, pp. 10617-10629, 2022.

[31] K. Sharifani, M. Amini, Y. Akbari and J. A. Godarzi, "Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model," Int. Journal of Science and Information System Research, vol. 12, no. 9, pp. 20-44, 2022.

[32] Y. Wang, S. Qian, J. Hu, Q. Fang and C. Xu, "Fake News Detection *via* Knowledge-driven Multimodal Graph Convolutional Networks," Proc. of the 2020 Int. Conf. on Multimedia Retrieval (ICMR'20), pp. 540-547, DOI: 10.1145/3372278.3390713, 2020.

[33] C. Song, N. Ning, Y. Zhang and B. Wu, "A Multimodal Fake News Detection Model Based on Crossmodal Attention Residual and Multichannel Convolutional Neural Networks," Information Processing & Management, vol. 58, no. 1, Article no. 102437, 2021.

[34] I. Goodfellow et al., "Generative Adversarial Networks," Communications of the ACM, vol. 63, no. 11, pp. 139-144, 2020.

[35] J. Du et al., "Cross-lingual COVID-19 Fake News Detection," Proc. of the 2021 Int. Conf. on Data Mining

Workshops (ICDMW), pp. 859-862, DOI: 10.1109/ICDMW53433.2021.00110, 2021.
[36] J. Wang, M. Gao, Y. Huang, K. Shu and H. Yi, "FinD: Fine-grained Discrepancy-based Fake News Detection Enhanced by Event Abstract Generation," Computer Speech & Language, vol. 78, Article no. 101461, DOI: 10.1016/j.csl.2022.101461, 2023.
[37] S. Xiong, G. Zhang, V. Batra, L. Xi, L. Shi and L. Liu, "TRIMOON: Two-round Inconsistency-based Multi-modal Fusion Network for Fake News Detection," Information Fusion, vol. 93, pp. 150-158, 2023.
[38] J. Xue, H. Zhou, H. Song, B. Wu and L. Shi, "Cross-modal Information Fusion for Voice Spoofing Detection," Speech Communication, vol. 147, pp. 41-50, DOI: 10.1016/j.specom.2023.01.001, 2023.
[39] J. Hua, X. Cui, X. Li, K. Tang and P. Zhu, "Multimodal Fake News Detection through Data Augmentation-based Contrastive Learning," Applied Soft Computing, vol. 136, Article no. 110125, DOI: 10.1016/j.asoc.2023.110125, 2023.
[40] R. N. Kvyetnyy, O. N. Romanyuk, E. O. Titarchuk, K. Gromaszek and N. Mussabekov, "Usage of the Hybrid Encryption in a Cloud Instant Messages Exchange System," Proc. of SPIE – The Int. Society for Optical Engineering, vol. 10031, Article no. 100314S, DOI: 10.1117/12.2249190, 2016.
[41] E. Ginters and E. Dimitrovs, "Latent Impacts on Digital Technologies Sustainability Assessment and Development," Advances in Intelligent Systems and Computing, vol. 1365, pp. 3-13, DOI: 10.1007/978-3-030-72657-7_1, 2021.
[42] E. Ginters, "New Trends towards Digital Technology Sustainability Assessment," Proc. of the World Conf. on Smart Trends in Systems, Security and Sustainability, vol. WS4 2020, pp. 184-189, DOI: 10.1109/WorldS450073.2020.9210408, 2020.
[43] M. Vasylkivskyi, O. Horodetska, B. Klymchuk and V. Hovorun, "Strategies of Technological Development of Hardware of Infocommunication Radio Networks," ITKI, vol. 56, no. 1, pp. 83-91, 2023.
[44] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi and L. Wei, "Detecting Fake News by Exploring the Consistency of Multimodal Data," Information Processing & Management, vol. 58, no. 5, Article no. 102610, 2021.

ملخص البحث:

في سياق الشّيوع المتنامي لوسائل التّواصُل الاجتماعي على مدى السّنوات العشر الماضية، برزت مشكلة الأخبار الزّائفة، الأمر الّذي يُعدّ من مسوغات البحث في هذا الموضوع. يهدف هذا البحث إلى تقييم فعالية الطّرق متعددة الأنماط في الكشْف عن الأخبار الزّائفة؛ لما لتلك الأخبار من آثار صحية وسياسية واقتصادية. ولتحقيق هذا الهدف، تمّ اختيار طريقة متعدّدة الأنماط تجمع بين أُطر التّعلّم العميق والنّماذج المدرَّبة مسبقاً. وتعطي الطّريقة المقترحة تحليلاً شاملاً لمعلومات النّصوص والصّور والأصوات، يسمح بتحديدٍ أدقّ لمصادر الأخبار الزّائفة. والجدير بالذّكر أنّ استخدام طرقٍ مختلفة لنقل المعرفة قد أتاح معالجة المعلومات على نحوٍ فعّال، ممّا حسّن جودة التّصنيف.

لقد أجري تحليل معمَّق لاستراتيجياتٍ متنوعةٍ لجمع البيانات، بالإضافة إلى تحليل مقارن للطّرق متعدّدة الأنماط المتاحة للكشف عن الأخبار الزّائفة، ومجموعات البيانات الخاصّة بذلك. وقد تضمّنت مُخرجات هذه الدّراسة إجراء تحليل مفصّل للدّراسات والأبحاث المتعلّقة بالكشْف عن الأخبار الزّائفة، إلى جانب اقتراح طريقةٍ متعدّدة الأنماط للتّصدّي لهذه المشكلة. وقد تمّت معالجة معلوماتٍ تتضمّن نُصوصاً وصوراً وأصواتاً باستخدام أنظمةٍ مسبقة التّدريب وتقنيات التّعلّم العميق، والحصول على دقّة عالية في الكشف عن الأخبار الزّائفة. فقد بينت النّتائج أنّ الطّريقة المقترحة متعدّدة الأنماط مكّنت من تحقيق دقّة أعلى في تحديد مصادر الأخبار الزّائفة والمعلومات المضلّلة عند مقارنتها بطُرقٍ أخرى مماثلة.

كذلك تمّ إجراء تحليلٍ مقارن لاستراتيجيات متنوّعة لجمع البيانات إلى جانب عددٍ من مجموعات البيانات المتعلّقة بموضوع البحث، وقد تأكّدت فاعلية الطّريقة المقترحة تحت ظروفٍ مختلفة.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

A HYBRID MODEL FOR ARABIC SCRIPT RECOGNITION BASED ON CNN-CBAM AND BLSTM

Mohamed Dahbali, Noureddine Aboutabit and Nidal Lamghari

(Received: 4-Mar.-2024, Revised: 5-May-2024, Accepted: 21-May-2024)

ABSTRACT

Handwriting recognition, particularly for Arabic, is a very challenging field of research due to various complex factors, such as the presence of ligatures, cursive writing style, slant variations, diacritics, overlapping and other difficult problems. This paper specifically addresses the task of recognizing offline Arabic handwritten text lines. The main contributions include the pre-processing stage and the utilization of a deep learning-based approach with data-augmentation techniques. The pre-processing step involves correcting the skew of text-lines and removing any unnecessary white space in images. The deep-learning architecture consists of a Convolutional Neural Network and Convolutional Block Attention Module for feature extraction, along with Bidirectional Long Short-Term Memory for sequence modeling and Connectionist Temporal Classification as a decoder. Data-augmentation techniques are utilized on the images in the database to enhance the system's ability to recognize a wide range of Arabic characters and to extend the level of abstraction in patterns due to synthetic variations. Our suggested approach has the capability of precisely recognizing Arabic handwritten texts without the necessity of character segmentation, thereby resolving various issues associated with this aspect. The results obtained from the KHATT database highlight the effectiveness of our approach, demonstrating a Word Error Rate of 14.55% and a Character Error Rate of 3.25%.

KEYWORDS

Handwriting recognition, Arabic database, Data augmentation, CNN, BLSTM.

1. INTRODUCTION

The Arabic script is one of the most widely utilized scripts in the world. The Arabic script is a cursive script and is renowned for its complex and challenging behaviors in comparison to handwritten Latin script [1][2][3], as well as the presence of handwriting-style variations and position-dependent character shapes. Some difficulties of handwritten Arabic, which involves segmentation into optical units, such as words, sub-words and characters, the overlapping of characters, the inclination of text lines, historical documents and the lack of publicly accessible databases, make the development of a method for recognizing handwritten Arabic text a serious challenge. It will be advantageous to design systems capable of resolving these issues and profiting from their applications, such as postal code/zip recognition, form processing, automatic cheque processing in banks …etc. [4].

Hidden Markov Model (HMM) was initially used for speech recognition due to its ability to model sequences, but in recent years, it has been studied for use in handwriting-recognition systems. There are numerous reasons for the success of HMM in text recognition, including the fact that it provides the advantage of joint segmentation and recognition [5] and has mathematical and theoretical bases. In recent years, researchers have focused on modeling Arabic handwritten sentences using techniques of deep learning. Several deep-learning models, such as Recurrent Neural Network (RNN), Long Short-term Memory (LSTM), Bi-directional Long Short-term Memory (BLSTM) and Multi-dimensional Long Short-term Memory (MDLSTM), can be utilized to model a sequence of data. Among the shortcomings of deep-learning models is their reliance on a large quantity of image data to converge and generalize, as well as their tendency to overfit training data. This problem could be avoided by incorporating data augmentation into the training process. This method can enhance the presentation of optical units, such as characters, sub-words and words, in handwritten Arabic text-line images. The primary advantage of modeling text at the character level is that the system can recognize out-of-vocabulary optical units in the test set and generalize the recognition system for unseen data. The method adopted during feature extraction is crucial for any recognition system to effectively model sentences. Two main approaches exist for extracting features: hand-crafted features, such as HOG [6] and LBP [7] descriptors and learned

M. Dahbali, N. Aboutabit and N. Lamghari are with LIPIM laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, Beni Mellal, Morocco. Emails: dahbali.mohamed@gmail.com, n.aboutabit@usms.ma and n.lamghari@usms.ma

features, such as those extracted by CNN [8]. This research is driven by the need to examine the impact of refining features extracted from CNN. Convolutional Block Attention Module (CBAM) is used to achieve this objective. The module sequentially infers attention maps along two different dimensions, channel and spatial and then multiplies them with the input feature map for adaptive-feature refinement. Our study aims to explore how refining the features extracted from CNN using Convolutional Block Attention Module (CBAM) can improve the recognition of Arabic handwritten text lines. In addition, our goal is to evaluate the impact of incorporating data-augmentation techniques. It is worth mentioning that, as far as we know, no prior research in the field of handwriting text-recognition systems has suggested the integration of CBAM into their systems. This highlights the originality and importance of our approach.

The primary contributions of this work include the proposal of a new technique for the pre-processing stage and expanding the size of the database through the use of data-augmentation techniques. Additionally, the utilization of CBAM as an efficient attention module for feedforward convolutional neural networks. This module not only reduces the number of parameters and computational power required, but also enhances the performance of the proposed system. Furthermore, the extracted features are used as input for a BLSTM and the output is then passed to the CTC layer. Finally, the evaluation of the proposed system is conducted using Character Error Rate (CER) and Word Error Rate (WER).

The paper is organized as follows: we review related works in Section 2 and introduce our proposed system in Section 3. Section 4 provides a comprehensive description of the database utilized in this study, as well as the data-augmentation techniques, evaluation methods and implementation details. Section 5 presents the findings and analysis, along with a comparative evaluation of our approach with other existing techniques. The conclusion of the paper is presented in Section 6.

2. RELATED WORKS

The literature contains works on Arabic handwriting-recognition systems employing HMMs and Deep Learning methods. The authors of [9] presented a comprehensive Arabic offline handwritten-text database (KHATT) encompassing all Arabic character forms. They proposed an HMM-based recognition system. Image adaptive pixel density and horizontal and vertical edge derivatives using the sliding window technique were utilized during the phase of feature extraction. The recognition is accomplished using HMM and HTK tools. The work [10] evaluated the quality of three types of language models (LMs) based on full word, hybrid word/Part-of-Arabic-Word (PAW) and full PAW by utilizing the Maurdor and Khatt databases. For the recognition system, they utilized hybrid HMM/Multi-directional LSTM Recurrent Neural Networks (MDLSTM-RNNs). The authors of [11] employs the KHATT database and proposes a new architecture based on MDLSTM, which consists primarily of three MDLSTM hidden layers and tanh layers. Another work [12] discussing the application of a Multi-directional LSTM Recurrent Neural Network (MDLSTM-RNN) based deep-learning model. The authors proposed examining the effect of utilizing various optical units with a hybrid word/part-of-Arabic word language model by training the optical model on Fixed and Random paragraphs from the KHATT database. In [13], another attempt to recognize handwritten Arabic text lines was made using MDLSTM. In this work, the authors proposed to capitalize on the visual similarities between Arabic, Urdu and Pashto. For this purpose, they examined a variety of combinations of these languages utilizing diverse optical models. Due to the lack of available datasets with real data for the Urdu language, the suggestion made in this work is to generate synthetic images for the three languages, so that meaningful comparisons can be made between the results of the optical models. For Arabic language in this experiment, the KHATT database used. The authors of the paper [14] proposed a novel technique for recognizing handwritten Arabic text line images using the KHATT database. For the feature extraction phase, segment-based and distribution-concavity (DC) based features were used and a 3-gram language model was employed for post-processing. Low-level, mid-level and high-level combinations of the BLSTM network were proposed. Experiments were conducted on two scenarios: the first employs a lexicon consisting of all tokenized words extracted from the KHATT corpus and the second employs a lexicon limited to words occurring in the training corpus. The authors of [15] utilized a multi-stage HMM-based text-recognition system for handwritten Arabic. They separated the core shapes of characters from their diacritics first. These Arabic core shapes are then converted into sub-core shapes to reduce the number of required models for modeling

Arabic characters. The models for multi-stage recognition system are sub-core shapes and diacritics. The proposed system was evaluated using the KHATT and IFN/ENIT databases. Another attempt for the recognition of handwritten Arabic text line images from the KHATT database is examined in [16]. This work employs a MDLSTM as an optical model. The main purpose of this work is to investigate the effect of data augmentation applied to each instance of text line image on the training of this optical model. Cross-validation based on a statistical process is used to evaluate the performance of the proposed system. The authors of [17] proposed a new architecture that combines CNN and BLSTM. Experiments are conducted on full text line images from the KHATT database. In [18], a unified end-to-end model for paragraph text recognition using hybrid attention is proposed. This module can be divided into three parts: an encoder that extracts feature maps from the entire paragraph image. Next, an attention module iteratively produces a vertical weighted mask that allows for focusing on the features of the current text line. A decoder module then recognizes the character sequence, resulting in the recognition of an entire paragraph. The suggested approach, applied to three widely-used datasets (RIMES, IAM and READ2016), produced state-of-the-art character error rates at the paragraph. The authors of [19] emphasized the significance of the encoder representation in enhancing the efficiency of Handwritten Text Recognition (HTR) systems. The authors suggested an encoder-decoder architecture for HTR that merges the advantages of local and global cross-channel attention to enhance the representation of the encoder. The experimental results on the IAM dataset demonstrate a significant decrease in CER and WER when the proposed module is implemented in the state-of-the-art HTR Flor model and Puigcerver model, respectively. The authors of [20] presented an encoder-decoder model for recognizing offline handwritten text. The DenseNet encoder is used to extract multiscale features. A contextual attention localizer was implemented between two gated recurrent units in order to integrate the role of context in the reading process and focus on particular characters. The model was assessed using the KHATT database and demonstrated superior performance compared to both simple-attention and multi-directional LSTM models. In [21], a new method for offline Arabic handwritten-text recognition is presented. The proposed system combines a CNN and a BLSTM followed by a CTC. Additionally, the authors introduce an algorithm for data augmentation to enhance the quality of the data. Significant performance was attained when compared to other models on the KHATT database. The authors of [22] examined two different end-to-end architectures for the recognition of Arabic handwritten-text lines: The transformer transducer and the standard transformer architecture with cross-attention. They generated a synthetic dataset consisting of printed Arabic text-line images, along with their corresponding ground truth, by utilizing different open-source fonts. The models conducted training using the synthetic dataset and were subsequently fine-tuned using the original dataset in order to assess their performance. The researchers discovered that both models exhibit strong competitiveness, with the cross-attention transformer achieving superior accuracy and the transformer transducer demonstrating faster processing speed when using the KHATT database.

As previously stated, numerous approaches have been developed to enhance offline Arabic handwritten-text recognition systems by increasing the depth or width of neural networks. While these approaches have achieved favourable results, the neural networks are excessively deep or wide, posing a significant computational challenge for computers. Motivated by the shortcomings of previous studies, we introduce a new and efficient feature representation of a lightweight CNN using CBAM. The objective of utilizing CBAM is to improve the accuracy of the proposed system while simultaneously minimizing computational requirements.

3. PROPOSED SYSTEM

This section provides an overview of the proposed system, which primarily comprises three different stages, as presented in Figure 1. The initial stage encompasses the pre-processing of the input image. Subsequently, features are extracted from the images utilizing CNN and CBAM. As the final stage, sequence modeling is done utilizing BLSTM and CTC as decoder. The components of the proposed system are detailed below.

3.1 Pre-processing

Pre-processing is a critical stage in addressing the issues associated with the text-lines of KHATT database, particularly the correction of text line slant and the elimination of extra white space in images. To correct the slant of text lines, we have utilized a technique based on horizonal projection inspired by

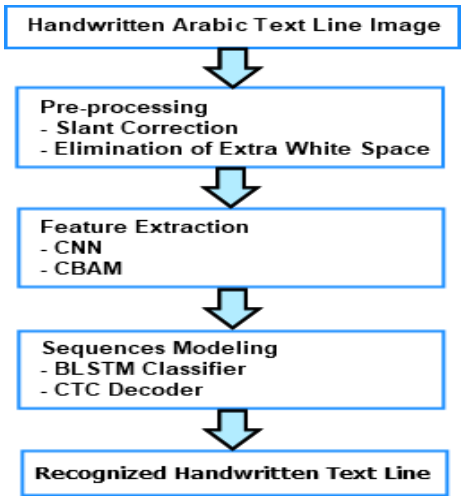
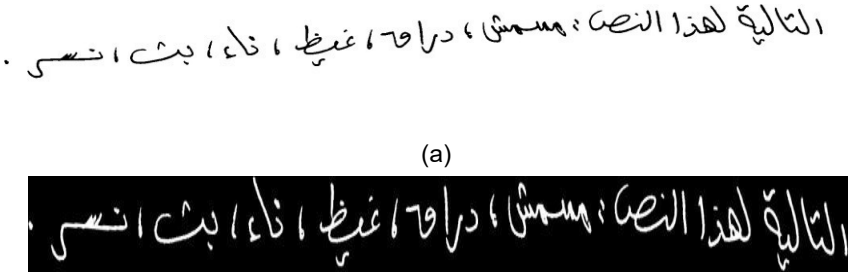


Figure 1. Proposed-system steps.

[23]. Diverse angle points are projected into an accumulator array, where the skew angle is defined as the projection angle within a search interval that maximizes alignment. To extract this angle, we first selected a testing range of angles, then rotated images using the selected angles and generated histograms and then computed a score based on the difference between peaks, with the angle with the highest score being the skew angle. During our experiments, we used a rotation angle ranging between -5 and 5. Pixels that do not belong to the text contribute to the existence of extra space, which degrades the performance of the system overall. We followed three steps based on morphological operations to resolve this issue. First, we dilated the handwritten text in images using a rectangular structuring element both vertically and horizontally and then we sorted all of the shapes in the images by contour space with their locations. Finally, we kept the shape with the most space and extracted text lines based on its location. Figure 2 depicts an example of a pre-processed text-line image.



(a)

(b)

Figure 2. (**a**) Original sample. (**b**) Image after pre-processing.

3.2 Feature Extraction Technique

Feature extraction is the most important element of any system designed to recognize handwritten text. CNN and CBAM were utilized in this study. CNN is known for its capacity to extract features from images using convolution layers. Typically, CNN consists of a series of convolution layers followed by down-sampling layers, such as max pooling. Depending on the size of the filter, the convolution layers divide the image in entry into partial images and extract features. Based on the number of filters utilized, a feature map is generated after each convolution operation. As the number of convolution layers increases, more robust features are obtained, which improves the performance of the recognition system. We used Rectified Linear Unit (RELU) activation function. There are three main advantages to using this function: the computational simplicity, because it is based on a max function, the representational sparsity to accelerate and simplify network learning, because it can output a true zero value, unlike other activation functions such as tanh and sigmoid and the linear behaviour, which aids in network optimization and prevents the vanishing of gradient.

CBAM [24] is a method for enhancing a recognition system by applying adaptive refining of CNN-extracted feature map. This method can be divided into two modules, specifically Channel Attention Module (CAM) and Spatial Attention Module (SAM). In this technique, CAM is implemented before

SAM. CAM is an architecture with minor differences to Squeeze Excitation Module (SEM) proposed in [25]. It is ideal to illustrate the procedure of SEM in order to comprehend the significance of CAM. All SEM operations can be broken down into three steps in order to extract channel weights. Initially, every channel is reduced to a single pixel. In the second step, a multi-layer perceptron (MLP) with a bottleneck is employed, followed by a sigmoid activation layer in the third step. To accomplish the first step, Global Average Pooling (GAP) operation is performed on the channels to aggregate spatial information into a single pixel. The outcome of the first step is a 1-D vector that will be utilized in the subsequent step. The second step is to use the vector output from the first step as input to a MLP network in which the bottleneck size is constrained by the reduction ratio R. The ratio of the number of channels N to the reduction ratio R is used to calculate the total number of neurons in a bottleneck. The greater the reduction ratio, the smaller the bottleneck. At the third step, the output vector from this MLP is passed to a sigmoid activation layer to maintain channel-weight values between 0 and 1.

The main distinction between SEM and CAM is the addition of Global Max Pooling (GMP). The output of the convolution operation is a feature map with the dimensions C×H×W, where H represents the height of each channel, W represents the width of each channel and C represents the total number of channels. Using GAP and GMP, the feature map generated by CNN is transformed into two consecutive vectors of size C×1×1. GMP preserves the contextual information in the form of edges presented in images. The combination of the two pooling operations provides more information than using GAP alone in SEM. The two vectors are successively passed to MLP. It is important to note that the same weights are used for both vectors in MLP. The final vector is constructed by applying an element-wise sum of the two MLP outputs and is then passed to the sigmoid layer to generate channel weights, which is used to distribute weights between channels in the feature map using an element-wise product. Figure 3 depicts the working process of CAM.

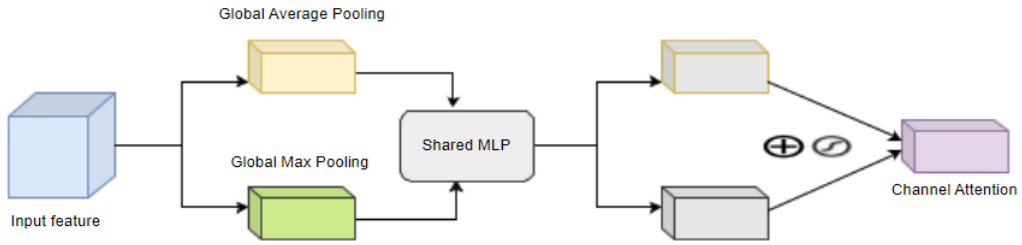


Figure 3. Diagram of channel-attention module (CAM).

The main aim of SAM is to generate an attention mask and apply it to the feature map to obtain more accurate features. The construction of attention mask consists of three sequential steps. The initial operation is a down-sampling operation. The objective of this step is to reduce the dimensions of feature map from C×H×W to 2×H×W by applying average-pooling and max-pooling operations along the channel axis and concatenate them. The resulting feature descriptor is then forwarded to the next step, which consists of a convolution-layer operation with a filter of size 7 × 7 and employs padding to do the same. In the third step, the output is sent to a sigmoid activation layer to map the mask values to the range from 0 to 1. To improve features, an element-wise product is performed between the current feature map and the resulting one channel output. Figure 4 depicts the working process of SAM.

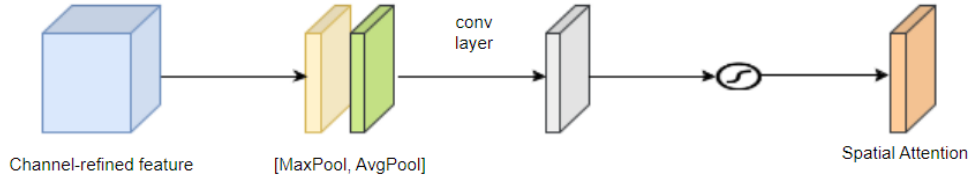


Figure 4. Diagram of spatial-attention module (SAM).

3.3 Optical Model

Diverse handwriting-recognition systems based on deep learning can be used for the recognition of handwritten Arabic text lines. BLSTM served as the optical model for our study. Because RNN is particularly susceptible to exploding/vanishing gradients, the LSTM model and its variants were

developed to address these issues. The addition of a new layer to maintain long-distance connections between optical units presented in sequence is the primary distinction between standard LSTM and BLSTM. In BLSTM, the input flow is bidirectional, which allows the system to model the relationship between sequence elements from the past to the future and *vice versa*. BLSTM consists of four main components: a forget gate, an input gate, an output gate and a cell state. These gates play a crucial role in the network, acting as filters to determine which useful data should be transmitted to subsequent steps. As its name suggests, the forget gate uses the previous hidden state and new input to determine what is relevant to keep from the prior cell state. Using the same parameters as the forget gate, the input gate identifies the information that should be updated in the current cell state. The output gate specifies the current hidden state that will be transmitted to the following LSTM unit. The output of BLSTM is input to a Softmax layer that contains a number of units equal to the size of characters in the vocabulary, plus a special character called blank to solve the problem of duplicate characters when encoding text and to create different alignments for the same text. The activation of L units is interpreted as the probability of observing the corresponding characters per time step and can be represented as a matrix. These output units specify the probabilities of all possible character-sequence alignments with the input sequence. The CTC layer comes at this point to compute the loss value for training the network. The advantage of CTC is that it avoids the explicit segmentation of Arabic text, which is notoriously difficult, particularly at the character level. Decoding is the process of determining the optimal alignment possible given the distribution over the output. Numerous approaches, such as best-path search and beam search, are proposed for approximating the most precise optimal alignment. Given the output probability matrix, the best-path search approach considers the character with the highest probability at each time step. The primary advantage of this strategy is the speed of decoding, while the disadvantage is a significant chance of failing to predict the correct ground truth. The beam-search method is based on a single hyper-parameter known as beam width, which specifies the number of characters with the highest predicted probabilities selected at each time step. Each predicted character in the first-time step will be the first character of output sequences, respectively. At each subsequent time step, based on the candidate output sequences from the previous time step, we continue to select candidate output sequences with the highest predicted probabilities until the final time step. The main advantage of this strategy is expanding space search and decoding the output close enough to the optimal sequence, while the primary drawback is the increased time required for decoding in comparison to greedy search. To extract the character sequence after decoding, all blanks and repeated characters are removed from the predictions.

4. EXPERIMENTAL SETUP

In this section, we initially present the database employed for this study. Next, we describe the methods utilized to increase the size of the database by adding extra synthetic images during the data-augmentation stage. Next, we outline the evaluation methods, followed by an explanation of the implementation details.

4.1 Database

KHATT is an Arabic offline handwritten-text database that is freely accessible for academic research. This database was created to address the lack of Arabic handwritten-text datasets. The database has 1000 forms written by different writers and each form has 4 pages [9]. It can be used for research in a variety of fields, including text recognition, writer identification and verification, form analysis, segmentation, …etc. One thousand writers from various countries were engaged to fill out four-page forms. The first page contains writer information, while the second page contains both fixed and randomly-selected paragraphs. Fixed paragraphs cover all Arabic character shapes and randomly-selected paragraphs were collected from a large corpus developed from 46 sources to create a database that is a statistical representation of the corpus, with each paragraph being unique across all forms. The third page includes another unique randomly-selected paragraph and the same fixed paragraph as the second page. The fourth page contains free writing on a variety of topics with ruled lines. For experimental purposes, we construct two partitions from the database. The first partition consists of a merged dataset that includes both unique and fixed-text lines. The second partition contains only unique text lines. Table 1 presents statistical information regarding the database partitions utilized in the evaluation of the proposed system.

Table 1. Statistical information pertaining to the data utilized in this study.

Partition	**Number of text lines**	**Number of text lines after data augmentation**
Unique text lines + Fixed text lines	Train: 9.470 Test: 2.001	Train: 28.410 Test: 6.003
Unique text lines	Train: 4825 Test: 960	Train: 14.475 Test: 2.880

4.2 Data Augmentation

The variability of data plays a significant role in training models for classification and sequence-modeling tasks, as well as improving the generalization capability of systems, particularly in the case of deep-learning models. To improve the performance of the proposed system, we used data augmentation to expand the size of the database by creating new image samples. Additive Gaussian Noise (AGN) and Salt and Pepper Noise (SPN) were utilized to add noise to the images. In the case of AGN, we used a normal distribution with a mean of 0 and a variance of 50 and for SPN, we covered 10% of all pixels in the image. Figure 5 illustrates an example of the applied augmentation method.

(a) (b) (c)

Figure 5. (**a**) Original sample. (**b**) AGN applied on image and (**c**) SPN applied on image.

4.3 Evaluation Methods

4.3.1 Cross-validation Method

The effectiveness of a deep-learning model is determined by its ability to generalize. It represents the ability of the system to recognize unseen images containing handwritten Arabic text lines. For this purpose, we utilized a statistical technique known as k-fold cross-validation to evaluate the performance of our proposed system. The proposed method divides the entire dataset into folds according to a parameter k that specifies the number of folds, using one fold for testing and the remaining folds for training. Each sample is eligible for use in training and evaluating the model. The primary benefit of using such a method is avoiding biased and overly optimistic results due to the nature of the utilized data. The most difficult aspect of this method is determining an adequate value for the parameter k to create a training and test-set partition that is statistically representative of the entire dataset. The performance metric derived from k-fold cross validation is summarized by the mean model skill scores.

4.3.2 Optical Model-evaluation Method

The evaluation of optical model performance requires the use of reliable evaluation metrics. CER and WER are two metrics that have been utilized previously in the literature. CER and WER represent the error of predicted labels at the character and word levels, respectively. These metrics are based on the Levenshtein distance [26] concept. In the case of CER, the error between the prediction and the ground truth is determined by calculating the ratio of insertions I, substitutions S and deletions D relative to the total number of characters NC in the ground truth, as shown in the following formula:

$$CER(\%) = \frac{S+I+D}{NC} \times 100 \quad (1)$$

In the case of WER, the error between the prediction and the ground truth is determined by calculating the ratio of insertions I, substitutions S and deletions D relative to the total number of words NW in the ground truth, as shown in the following formula:

$$WER(\%) = \frac{S+I+D}{NW} \times 100 \quad (2)$$

4.4 Implementation Details

The input to our system consists of grayscale images. All images are resized to dimensions of 64×1024, with 64 representing the height and 1024 representing the width. In the CNN architecture, we utilized 5 convolution layers with a filter size of 3×3. Additionally, we added 4 down-sampling layers with the max-pooling operation. The purpose of the first and second pooling operations is to reduce the vertical and horizontal size of the feature map. The last two pooling operations only reduce vertical size of the feature map. Figure 6 depicts the employed architecture. At the end of CNN architecture, we have 128 channels with a size of 4×256 that will serve as input to CBAM. The reduction ratio for CAM was defined as 2. The filter size employed by SAM is 7, identical to the original paper [24]. The features generated by the CBAM architecture are fed to the BLSTM. To model handwritten Arabic text lines, we utilized two distinct BLSTMs, each of which consisted of 256 LSTM units in both directions (i.e., 128 LSTM units in either direction). The output was decoded using the beam-search algorithm. We evaluated a number of beam-width values before settling on 10, because it offers the best performance. The Adam optimizer is employed for training with a batch size of 50. To assess the proposed system, we employed the k-fold cross-validation technique with a k-value of 5.

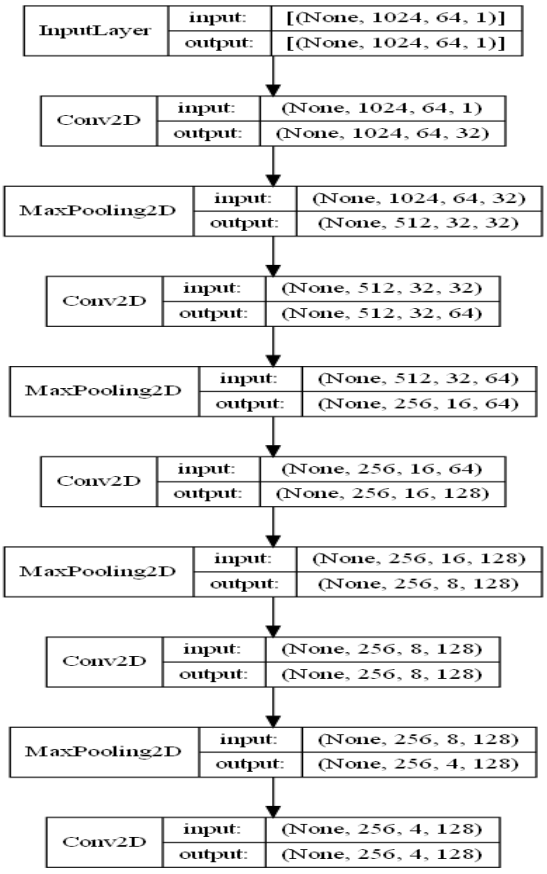


Figure 6. CNN architecture for feature-map extraction.

5. Results and Discussion

To assess the efficacy of the proposed system, we utilized two networks, one without CBAM and one with CBAM, employing two different approaches from KHATT database, with and without data augmentation. We performed two experiments during this study. The first experiment involved the merged dataset, which consisted of both fixed and unique text lines. The second experiment focused solely on the unique text lines.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

Table 2 demonstrates that in the first experiment, the CER decreased from 13.26% to 13.00% and the WER decreased from 45.33% to 44.49% as a result of incorporating CBAM into the proposed system. The use of data augmentation has resulted in a major enhancement of the proposed system, leading to a significant reduction in both CER and WER.

Table 2. Performance of the system in the first experiment using CER and WER.

	CER	**WER**
CNN	13.26%	45.33%
CNN+CBAM	13.00%	44.49%
CNN+CBAM+Data Augmentation	3.25%	14.55%

Table 3 demonstrates that, similar to the initial experiment, the CER decreased from 19.84% to 19.02% and the WER decreased from 67.78% to 64.67% when the CBAM was incorporated into the proposed system during the second experiment. Data augmentation was utilized in this experiment, leading to a significant improvement of the proposed system, akin to the initial experiment.

Table 3. Performance of the system in the second experiment using CER and WER.

	CER	**WER**
CNN	19.84%	67.78%
CNN+CBAM	19.02%	64.67%
CNN+CBAM+Data Augmentation	3.96%	18.57%

The primary distinction between fixed and unique text-line images is text content. Fixed text-line images consist of similar texts encompassing all Arabic character shapes, whereas unique text line images are made up of different texts. The presence of similar texts containing similar optical units in the dataset may allow the system to model sequences efficiently. However, the presence of distinct texts with distinct optical units in the dataset renders the system incapable of effectively modeling long-term dependencies. The distance between relevant information in the entire sequence and current time step information may be too great, thereby degrading the performance of the system, as in the case of unique text line images containing distinct texts. It is necessary to adopt a strategy to resolve this issue. We suggested employing data augmentation to increase the presence of optical units in the dataset and to enhance the level of abstraction in patterns due to synthetic variations. CBAM ensures that CNN focuses on useful features presented on images and avoids non-useful background information. This is achieved by exploiting the inter-channel and inter-spatial relationships of features through the integration of channel and spatial-attention modules, respectively. BLSTM was chosen as the optical model for our method. The primary distinction between the conventional LSTM and the bidirectional LSTM lies in the inclusion of an extra layer that handles the opposite direction of the sequence. This enables the modeling of sequential dependencies among characters and words in both the forward and backward directions of the sequence. CER and WER values for the unique text lines are higher than those for the merged dataset, demonstrating that similar text in the dataset improves the performance of the optical model. The inclusion of CBAM in the network resulted in a notable enhancement in the performance of the proposed system, leading to a reduction in both CER and WER by a significant margin when compared to the network without CBAM. Undoubtedly, the CBAM-CNN architecture produces superior quality features compared to those generated solely by CNN. To confirm this hypothesis, we can verify it by comparing the CER and WER results of our work with the results of recent studies [17][21] that employed a CNN-BLSTM architecture. Table 4 depicts a comparative study between our work and other works that utilized the same database. It is apparent that the proposed CNN-CBAM-BLSTM system is successful, as it outperforms other recent works that utilized the same database.

Table 4. Comparison with other state-of-the-art systems evaluated on the KHATT database.

Authors	**Features**	**Optical model**	**No. of text lines**	**Character Error Rate (CER)**	**Word Error Rate (WER)**
Mahmoud et al. [9]	Image adaptive pixel density features and horizontal and vertical edge derivatives	HMM using HTK tools	Train: 4808 Test: 966 Validation: 938	53.87%	
Ben Zeghiba et al. [10]	Raw pixels	HMM/MDLSTM-RNN	Train: 4428 Test: 959 Validation: 876		30.9%
Ahmed et al. [11]	Raw pixels	MDLSTM	Train: 4825 Test: 966 Validation: 937	24.25%	
Ben Zeghiba [12]	Raw pixels	MDLSTM-RNN	-Random paragraphs -Fixed paragraphs -Fixed and random paragraphs		41.4% 8.3% 23.0%
Ahmad et al. [13]	Raw pixels	MDLSTM	Train: 4825 Test: 966 Validation: 937	24.30%	
Jemni et al. [14]	Segment-based and distribution-concavity (DC)-based features	Combination of BLSTM	Train: 9475 Test: 2007 Validation: 1901	7.85%	13.52%
Ahmad and Fink [15]	Feature-based on ink pixels	Multi-stage HMM system	Train: 4808 Test: 966 Validation: 937	41.21%	
Ahmad et al. [16]	Raw pixels	MDLSTM	Train: 4825 Test: 966	19.98%	
Anjum and Khan [20]	DenseNet encoder	GRUs with Contextual Attention Localization	Train: 4825 Test: 966 Validation: 937	22.85%	64.17%
Momeni and Baba Ali [22]	Transformer encoder	Transformer with Cross-attention and Transformer Transducer	6742 of unique text lines	18.45%	
Noubigh et al. [17]	CNN	BLSTM	Train: 8505 Test: 1867 Validation: 1584	15.8%	35.6%
Lamtougui et al. [21]	CNN	BLSTM	Train: 4825 Test: 966		19.85%
Proposed system	CNN + CBAM	BLSTM	Train: 4825 Test: 960	3.96%	18.57%
			Train: 9.470 Test: 2.001	3.25%	14.55%

6. CONCLUSION

In this paper, we addressed the recognition of handwritten Arabic text lines, one of the most challenging problems associated with Arabic optical models. We used the database KHATT, which is renowned for the complexity of its handwriting, along with the data-augmentation technique to generate more synthetic images. Our study is the first to use CBAM to evaluate the effect of attention modules on CNN-generated feature map. Deep-learning architectures are the state-of-the-art systems used to model sequences in the literature over the past few years. BLSTM architecture was used to model Arabic

handwriting in our study. To evaluate our CNN-CBAM-BLSTM architecture, we conducted two experiments, the first with a dataset containing both unique and fixed text-line images and the second with unique text-line images. In the first experiment, the CER was 3.25% and the WER was 14.55%. In the second experiment, the CER was 3.96% and the WER was 18.57%. The reported results are comparable to those of recent studies that utilized the same database. Although our model performed well, the challenge of decreasing the WER presents an opportunity for future investigation. Furthermore, it will be essential to enlarge the database to include a wider range of handwritten Arabic text-line images in order to improve the model's generalizability and reduce any possible biases.

REFERENCES

[1] S. Ahmed, S. Naz, S. Swati, M. I. Razzak and A. I. Umar, "UCOM Offline Dataset: An Urdu Handwritten Dataset Generation," Int. Arab Journal of Information Technology, vol. 14, no. 2, pp. 239-245, 2017.

[2] A. Graves and J. Schmidhuber, "Offline Handwriting Recognition with Multidimensional Recurrent Neural NetWorks," Proc. of the 21st Int. Conf. on Neural Information Processing Systems (NIPS 2008), Red Hook, pp. 545-552, 2009.

[3] S. Naz et al., "The Optical Character Recognition of Urdu-like Cursive Scripts," Pattern Recognition, vol. 47, no. 3, pp. 1229–1248, DOI: 10.1016/j.patcog.2013.09.037, Mar. 2014.

[4] S. Faisal Rashid, M.-P. Schambach, J. Rottland and S. von der Null, "Low Resolution Arabic Recognition with Multidimensional Recurrent Neural Networks," Proc. of the 4th Int. Workshop on Multilingual OCR (MOCR '13), Article no. 6, pp. 1-5, DOI: 10.1145/2505377.2505385, Aug. 2013.

[5] D. Xiang, H. Yan, X. Chen and Y. Cheng, "Offline Arabic Handwriting Recognition System Based on HMM," Proc. of the 2010 3rd IEEE Int. Conf. on Computer Science and Information Technology, DOI: 10.1109/iccsit.2010.5564429, Chengdu, Jul. 2010.

[6] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recogn. (CVPR'05), vol. 1, pp. 886–893, 2005.

[7] T. Ojala, M. Pietikäinen and D. Harwood, "A Comparative Study of Texture Measures with Classification Based on Featured Distributions," Pattern Recognition, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[8] Y. LeCun and Y. Bengio, "Convolutional Networks for Images, Speech and Time Series," Part of Book: The Handbook of Brain Theory and Neural Networks, p. 3361, 1995.

[9] S. A. Mahmoud et al., "KHATT: An Open Arabic Offline Handwritten Text Database," Pattern Recognition, vol. 47, no. 3, pp. 1096–1112, DOI: 10.1016/j.patcog.2013.08.009, Mar. 2014.

[10] M. F. Ben Zeghiba, J. Louradour and C. Kermorvant, "Hybrid Word/Part-of- Arabic-Word Language Models for Arabic Text Document Recognition," Proc. of the 2015 13th IEEE Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 671-675, DOI: 10.1109/icdar.2015.7333846, Aug. 2015.

[11] R. Ahmad, S. Naz, M. Zeshan Afzal, S. Faisal Rashid, M. Liwicki and Dengel, "KHATT: A Deep Learning Benchmark on Arabic Script," Proc. of the 2017 14th IEEE IAPR Int. Conf. on Document Analysis and Recognition, pp. 10-14, DOI: 10.1109/icdar.2017.358, Nov. 2017.

[12] M. F. BenZeghiba, "A Comparative Study on Optical Modeling Units for Off-line Arabic Text Recognition," Proc. of the 2017 14th IEEE IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), DOI: 10.1109/icdar.2017.170, Nov. 2017.

[13] R. Ahmad, S. Naz, M. Zeshan Afzal, S. Faisal Rashid, M. Liwicki and A. Dengel, "The Impact of Visual Similarities of Arabic-like Scripts Regarding Learning in an OCR System," Proc. of the 2017 14th IEEE IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), DOI: 10.1109/icdar.2017.359, 2017.

[14] S. Khamekhem Jemni, Y. Kessentini, S. Kanoun and J.-M. Ogier, "Offline Arabic Handwriting Recognition Using BLSTMs Combination," Proc. of the 2018 13th IAPR Int. Workshop on Document Analysis Systems (DAS), DOI: 10.1109/das.2018.54, Apr. 2018.

[15] I. Ahmad and G. A. Fink, "Handwritten Arabic Text Recognition Using Multi-stage Sub-core-shape HMMs," Int. Journal on Document Analysis and Recognition (IJDAR), vol. 22, no. 3, pp. 329–349, 2019.

[16] R. Ahmad, S. Naz, M. Afzal, S. Rashid, M. Liwicki and A. Dengel, "A Deep Learning based Arabic Script Recognition System: Benchmark on KHAT," Int. Arab Journal of Information Technology, vol. 17, no. 3, pp. 299–305, DOI: 10.34028/iajit/17/3/3, May 2020.

[17] Z. Noubigh, A. Mezghani and M. Kherallah, "Contribution on Arabic Handwriting Recognition Using Deep Neural Network," Proc. of the Int. Conf. on Hybrid Intelligent Systems, Part of the Book Series: Advances in Intelligent Systems and Computing, vol. 1179, pp. 123–133, DOI: 10.1007/978-3-030-49336-3_13, 2020.

[18] D. Coquenet, C. Chatelain and T. Paquet, "End-to-End Handwritten Paragraph Text Recognition Using a Vertical Attention Network," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 508–524, DOI: 10.1109/tpami.2022.3144899, Jan. 2023.

[19] B. N. Shashank, S. Nagesh Bhattu and Krishna, "Improvising the CNN Feature Maps through Integration of Channel Attention for Handwritten Text Recognition," Communications in Computer and Information

Science, pp. 490–502, DOI: 10.1007/978-3-031-31417-9_37, Jan. 2023.
[20] T. Anjum and N. Khan, "CALText: Contextual Attention Localization for Offline Handwritten Text," arXiv.org, arXiv: 2111.03952, DOI: 10.48550/arXiv.2111.03952, 2021.
[21] H. Lamtougui, H. El Moubtahij, H. Fouadi and K. Satori, "An Efficient Hybrid Model for Arabic Text Recognition," Computers, Materials & Continua, vol. 74, no. 2, pp. 2871–2888, 2023.
[22] S. Momeni and B. Baba Ali, "A Transformer-based Approach for Arabic Offline Handwritten Text Recognition," Signal, Image and Video Processing, vol. 18, no. 4, pp. 3053–3062, Jan. 2024.
[23] J. Kanai and A. Bagdanov, "Projection Profile Based Skew Estimation Algorithm for JBIG Compressed Images," Int. J. on Document Analysis and Recognition, vol. 1, pp. 43–51, 1998.
[24] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional Block Attention Module," arXiv.org, arXiv: 1807.06521, DOI: 10.48550/arXiv.1807.06521, Jul. 2018.
[25] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 1–1, 2019.
[26] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," Soviet Physics-Doklady, vol. 10, no. 8, pp. 707-710, Feb. 1966.

ملخص البحث:

إنّ تمييز الكتابة بخطّ اليد، وبخاصّة باللّغة العربية، يُعدّ مجالاً بحثياً ينطوي على الكثير من التّحدّيات. ويرجع ذلك إلى عددٍ من العوامل المعقّدة، منها وجود علامات الشَّكْل، ونمط الكتابة، واختلاف اللّهجات، والتّداخل، إلى جانب العديد من المشكلات الصّعبة الّتي تكتنف تمييز الكتابة اليدوية.

تتناول هذه الورقة بالتّحديد أسطُراً مكتوبة بخطّ اليد باللّغة العربية، وتتلخّص مساهماتها الأساسية في مرحلة المعالجة الأوّلية، واقتراح طريقةٍ قائمةٍ على تقنيات التّعلُّم العميق، إلى جانب تقنيات زيادة البيانات. تتمثّل المعالجة الأوّلية بتصحيح الالتواء للأسطر وإزالة الفراغات غير الضّرورية فيها. أما بِنية التّعلُّم العميق فتتكون من شبكة عصبية التفافية، ووحدة لاستخلاص السِّمات، إلى جانب ذاكرة ثنائية الاتّجاه طويلة المدى وقصيرة المدى لنمذجة التّتابُع، ومصنِّف مؤقّت لفكّ التّرميز. وأمّا تقنيات زيادة البيانات فيستفاد منها في الصّور المتضمَّنة في قاعدة البيانات في تحسين قدرة النّموذج على تمييز طيْفٍ واسعٍ من الأحرف باللّغة العربية.

وتجدر الإشارة إلى أنّ النّموذج المقترح لديه القدرة على تمييز النّصوص المكتوبة بخطّ اليد دون الحاجة إلى تجزئتها إلى أحرف، مما يساعد في التّغلّب على العديد من المسائل المتعلقة بهذا الموضوع. وقد بينت النّتائج أنّ النّموذج المقترح أثبت عند تجربته على قاعدة البيانات (KHATT) فاعلية جيدة؛ إذ بلغ معدّل خطأ الكلمات 14.55% بينما بلغ معدّل خطأ الأحرف 3.25%.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

A MACHINE LEARNING BASED DECISION SUPPORT FRAMEWORK FOR BIG DATA PIPELINE MODELING AND DESIGN

Asma Dhaouadi[1], Khadija Bousselmi[2], Sébastien Monnet[3], Mohamed Mohsen Gammoudi[4] and Slimane Hammoudi[5]

(Received: 25-Mar.-2024, Revised: 13-May-2024, Accepted: 27-May-2024)

ABSTRACT

The data warehousing process requires an architectural revolution to settle big-data challenges and address new data sources, such as social networks, recommendation systems, smart cities and the web to extract value from shared data. In this respect, the pipeline-modeling community for the acquisition, storage and processing of data for analysis purposes is enacting a wide range of technological solutions that present significant challenges and difficulties. More specifically, the choice of the most appropriate tool for the user's specific business needs and the interoperability between the different tools have become primary challenges. From this perspective, we propose in this paper a new interactive framework based on machine learning (ML) techniques to assist experts in the process of modeling a customized pipeline for data warehousing. More precisely, we elaborate first (i) an analysis of the experts' requirements and the characteristics of the data to be processed, then (ii) we propose the most appropriate architecture to their requirements from a multitude of specific architectures instantiated from a generic one, by using (iii) several ML methods to predict the most suitable tool for each phase and task within the architecture. Additionally, our framework is validated through two real-world use cases and user feedback.

KEYWORDS

Big data, Data-warehousing modeling, Modeling assistance, Tools and technologies, ML methods.

1. INTRODUCTION

The technological revolution, the emergence of new Internet services, the blooming growth of smart devices and sensors, mobile and web applications and social media (Facebook, Twitter, Instagram, …etc.) generate a large amount of data daily, known as "Big Data". Notably, Big Data is facing several challenges labeled Vs, like: (i) the Volume presenting the massive amount of data collected by a company, (ii) the Variety which refers to the heterogeneity of data, including structured, semi-structured or unstructured types and (iii) the Velocity, which refers to the speed by which data is collected and needs to be taken into account for eventual processing and decision-making. To address these big-data challenges, numerous platforms, software systems and architectural frameworks have been developed for data warehousing and analysis. However, the diverse landscape of available solutions introduces additional challenges, including the deployment requirement, which addresses verifying the interoperability between the tools, their performance and the experts' technical constraints, such as the resources provided at the deployment phase of the architecture. Consequently, experts need help to select the most suitable tools from a wide range of options. Furthermore, modeling big-data pipelines is crucial before deployment and certain tools prioritize addressing some big-data challenges over others. For example, Apache Kafka does not focus on the veracity of data but rather on its transfer and, thus, on Volume and Velocity [1], while column-oriented tools dedicated to data storage, such as HBase, MongoDB and Cassandra, specifically favor data Variety by supporting different formats and data types [5]. Additionally, the need for adaptability and evolution of data-warehousing and analytics systems is pressing and currently needs to be a standardized architectural solution that guarantees the best selection of tools based on experts' requirements and constraints. To overcome these challenges, we propose in this paper ArchiTectAI: an AI-driven framework to assist experts in selecting the most appropriate tools to meet their particular requirements and constraints

1. A. Dhaouadi is with LISTIC, Savoie Mont Blanc University, France and RIADI, University of Tunis El Manar, Tunisia. Email: asma.dhaouadi74@gmail.com, ORCID 0000-0002-9832-5000.
2. K. Bousselmi is with LISTIC, IUT, Savoie Mont Blanc University, France. Email: khadija.arfaoui@univ-smb.fr
3. S. Monnet is with LISTIC, Polytech, Savoie Mont Blanc University, France. Email: sebastien.monnet@univ- smb.fr
4. M.M. Gammoudi is with RIADI Lab, ISAM La Mannouba, Tunisie. Email: gammoudimomo@gmail.com
5. S. Hammoudi is with ERIS, ESEO-TECH Angers, France. Email: slimane.hammoudi@eseo.fr

when elaborating big-data warehousing and analysis solutions. The main target is to assist the experts' pipeline modeling by considering the specificities of the data to be processed (i.e., the Volume, Velocity, Veracity and Variety) and respecting their preferences. This is achieved by selecting the best tools dedicated to their corresponding needs through the use of personalized ML models employing various ML methods, such as Decision Trees, Random Forest, Support Vector Machine and Gradient Boosting.

In this paper, the main contributions are outlined as follows:

- A hybrid, generic, two-level architecture that supports batch and stream-data processing is proposed to guide the instantiating of architecture models specific to big-data platforms and tools.
- ArchiTectAI: a decision-support framework based on ML that assists experts in modeling big-data pipelines, considering their specific needs and ensuring tool interoperability.
- An *ad-hoc* method for generating a base of tools that considers its alignment with big-data characteristics. This method employs an algorithm that ensures easy maintenance and scalability of the tool base.
- The use of several ML classifiers to predict the most suitable tool for each phase and task of the architecture.

This paper is structured as follows. Section 2 displays the proposed generic big-data pipeline architecture and its phases and tasks. Subsequently, in Section 3, we detail the proposed architecture to support the proposed framework for big-data pipeline modeling. Next, in Section 4, we validate and evaluate the introduced proposals. Section 5 provides a summary of the most prominent related works. Eventually, in Section 6, we conclude the whole work and offer new perspectives for future research.

2. TOWARDS A GENERIC BIG DATA PIPELINE ARCHITECTURE

In this section, we propose a generic architecture for end-to-end big-data warehousing and analysis. The specificities of this architecture are that: (i) it is a generic and Tools Independent Architecture (TIA) that supports two different scenarios. The first one is only for data collected using a single acquisition mode: batch or stream. The second one is the most complete hybrid scenario, which supports both batch and stream-data acquisition modes. As shown in Figure 1, the TIA generic architecture consists of the following four phases:

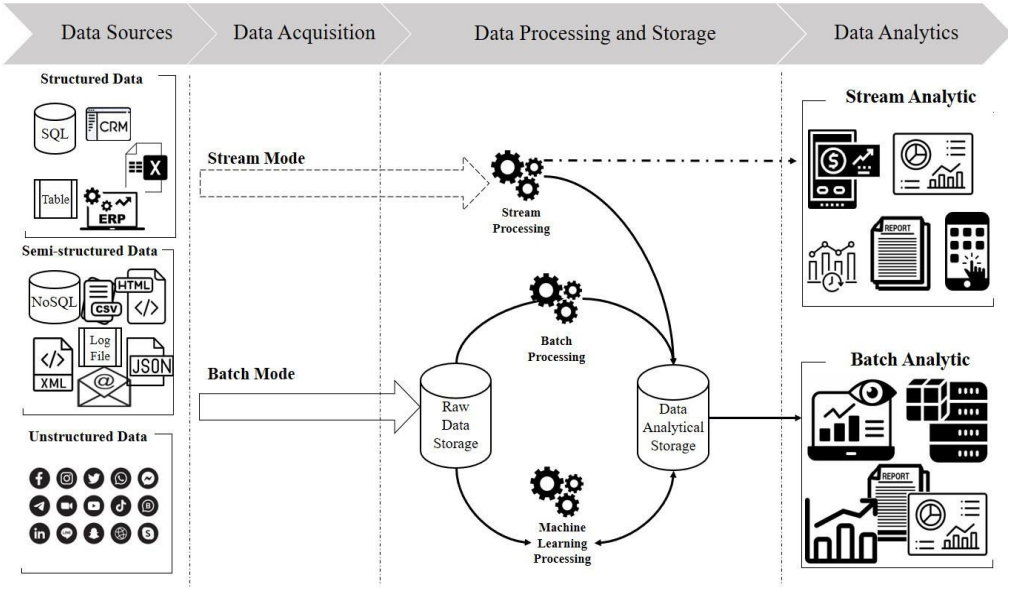


Figure 1. The generic architecture: tool independent architecture.

1) Data Sources phase: This layer gathers the different types of input data, which are classified into structured, semi-structured and unstructured data. The structured data is provided by the tabular data and traditional systems, like data warehouses, data mart, Customer Relationship Management (CRM) or Enterprise Resource Planning (ERP) systems. The semi-structured

data constitutes the NoSQL databases, HTML, JSON, XML and CSV files and e-mails. The unstructured data are generated from social networks, images, audio, videos and data streams, including chat messages shared on the Internet from WhatsApp and mobile SMS text and many other data types like geo-localization data and sensor data from smart devices [2].

2) Data Acquisition phase: Our architecture allows treating three different scenarios at this stage. The first scenario occurs when the data to be acquired, processed and stored is streaming data that necessitates real-time processing and must be ingested instantaneously as it is generated. This scenario is well-suited for mission-critical applications, such as cyber-security monitoring, financial-transaction processing and early-warning systems for natural disasters. The second scenario is when the data to be handled is ingested over a certain period and then processed all at once; this is known as the batch mode. Unlike real-time processing, this deferred processing concerns static data, such as relational tables and Hadoop files [3]. Finally, the third scenario combines streaming and batch data and the architecture will support both data types simultaneously.

3) Data Processing and Storage phase: In our architecture, we merged processing and storage tasks into a single layer to generalize as much as possible and cover stream and batch-processing modes. This middle layer is slightly divided horizontally into two sub-layers: the top layer is specific to stream processing, while the bottom layer corresponds to batch processing. For the batch mode, the data gathered is stored in its raw form and its processing is planned according to the experts' requirements. The processing involves tasks, such as data cleaning, outlier removal and preparation for storage in specific supports dedicated to analysis. Different methods for data cleaning and outlier removal have been outlined in [4]. ML type processing is also proposed at this level to perform increasingly sophisticated treatments in response to the business requirements of experts. In this phase, stream processing, batch processing, ML processing, raw-data storage and data analytic storage constitute the TIA tasks.

4) Data Analytics phase: Data analysis is based on creating dashboards, OLAP navigation, statistics, charts and reports. The stream or batch-data analysis type depends mainly on the data-processing type and data availability in the repositories. It also depends on the level of interactivity that the expert requires to be provided by the dedicated tool. Indeed, the analysis is adapted to experts' needs. Supposing that they require interactive queries, instantaneous processing answers on data collected in real-time or answers on batches of stored data. Each type of analysis is presented by a different icon. For instance, there are forecasting, real-time alerting, dashboards, mobile applications, reporting and visualization for streaming analysis. However, the most frequent business intelligence applications for batch analysis are dashboard, reporting, data visualization and data statistics on which recommendation engines, like Amazon and YouTube videos, can be based.

This generic architecture is designed to be invested for different deployment requirements, meeting the business constraints and the specificities of the experts' needs each time. From this perspective, we can derive different Tools Specific Architectures (TSAs) according to the use case. While, for lack of space in this paper, we will not mention details about the modeling approach formalizing the transition between the TIA level and the TSAs one, we highlight that a modeling approach has been proposed to establish the transition between the generic level and the applicative one by taking full advantage of the Model-Driven Architecture (MDA) approach for software design and development. The modeling details will be provided in a separate paper.

The following section presents the proposed framework for generating different concrete pipelines of TSAs, taking advantage of TIA's genericity.

3. ARCHITECTAI: PERSONALIZED BIG DATA WAREHOUSING ADVISOR

In this section, we present our interactive ML-based framework, assisting the expert in the composition of his/her big-data pipeline and allowing the automatic generation of several TSAs. Figure 2 shows the ArchiTectAI framework's architecture, including three modules: tools database generation, ML-model generation and architecture generation process. In the following part, we detail the role of each module.

3.1 Tools Database Generation

This module is based on an *ad-hoc* method for generating tool databases for big-data warehousing. In these databases, we specified the characteristics of each tool from the categorical variables defining the big data Vs (Volume, Velocity, Veracity and Variety). Among these categorical variables, we note the acquisition mode of data, the data type, the size of data, among others. As detailed in the pseudo-code of Algorithm 1, considering the specific list of tools for each phase and task of TIA and the different categorical variables, we peruse the TIA process and we determine all possible combinations between the different features and the corresponding tools based on predefined rules. The output of this method is a tools database for each TIA phase and task, containing the tools with their corresponding and valuable characteristics for the ongoing step. As depicted in Figure 2, there are distinct databases for data acquisition, raw-data storage, analytical data and data analysis. Regarding data processing, as previously mentioned, there are three types of processing: stream, batch and ML (see Figure 1). In classifying tools into databases, we utilize the criteria of "mode". Tools are categorized based on whether they support stream mode or batch mode, resulting in a single database for both processing modes. Conversely, for ML processing tools, we allocate a separate database, as we consider that ML is a type of treatment supported, rather than a processing mode. In all instances, there are tools performing more than one task, thus being shared by more than one base. For example, Spark, Python and Flume handle ML, batch and stream processing and are therefore present in both the ML data processing and data processing databases.

Algorithm 1 An *Ad-Hoc* Method for Generation of Tools Databases

Input 1: Tool Independent Architecture **TIA Input 2:**
List of Tools **ToolsList**
Input 3: categoricalVariables **catVars**
Output: Tools Databases **ToolsDBs**

```
TIA ← {TIAphs,TIAtsks}                                   ▷ The TIA phases and tasks T
oolsList ← TLacqui ∪ TLrawStor ∪ TLanalStor ∪ TLprocT ∪ TLprocMLT ∪ TLanal catV ars ←
{Acquisition_mode : {"Stream", "Batch"},
            SizeData: {"over1Peta","in1T1P","less1Tera"},
            D_type : {"UNS", "SEMI", "STRUC"},
            PotentielTimeAcqui: {"RT","NRT","Batch"},
            Quality: {"LossDup","LossData","DupData","NoLossNoDup"},
            Nb_DS : {"over20", "in5_20", "less5"},
            Latency : {"less5", "in5_15", "over10"},
            Complexity : {"one", "multiple"}}
foreach TIAphs AND TIAtsks in TIA do
    foreach T in ToolsList do
        T_Values ← Assign(Val)                    ▷ Assign characteristic values to each tool
        {Val is a value from catVars}
    end
end
foreach T in ToolsList do
▷ Generate specific combinations of each tool by using zip
    foreach t_Values in zip(T_Values) do
        csv_writer.writerow(t_Values + [T])          ▷ Write rows in the "ToolsDBs.csv" file
    end
end
return ToolsDBs
```

This proposed *ad-hoc* method for generating tool bases is conducive to easy updates and maintenance. The process automatically handles all updates by adding the tool with its characteristics expressed in categorical values and developing new data combinations. Then, the generated tool databases will be leveraged as input data by ML methods to predict the most suitable tools while ensuring adherence to the experts' constraints and preferences. As for the categorical variables, they are also used as features in the ML module (sub-section 3.2) and to express user constraints (Tab 2 in subsection 3.3).

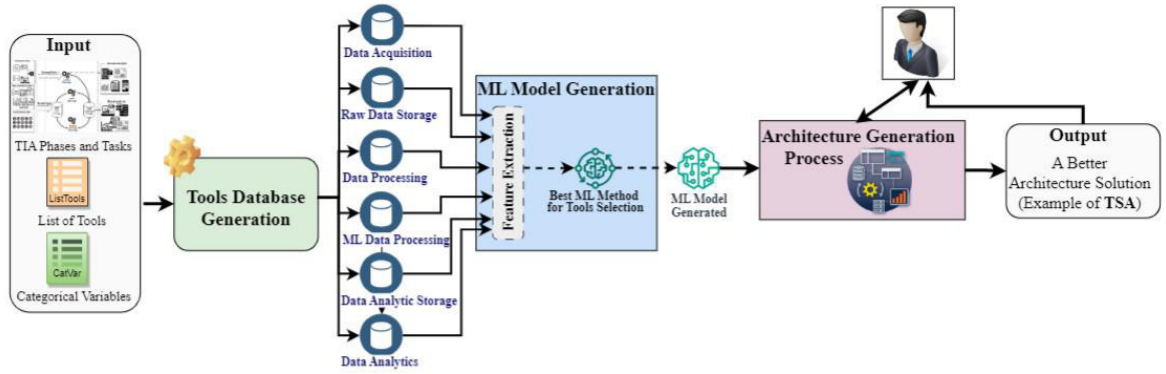


Figure 2. The architecture of ArchiTectAI framework.

3.2 ML Model Generation

This second module of the proposed framework, as shown in Figure 3, is based mainly on two steps: the first is for feature extraction and the second is for ML processing and determining the best ML method. The input data consists of the tool databases specific to each TIA phase and task. Each database contains around 15 tools and their big-data characteristic satisfaction values. Until now, the number of tools handled in the training dataset is 64 and is expected to evolve in line with the technological revolution. As for the result of this module, the generated ML models are to be used in predictions of the best tools for subsequent use by the experts. In the following part, we outline the two steps of this module and provide a comprehensive overview of the experiment conducted to motivate the chosen ML method.

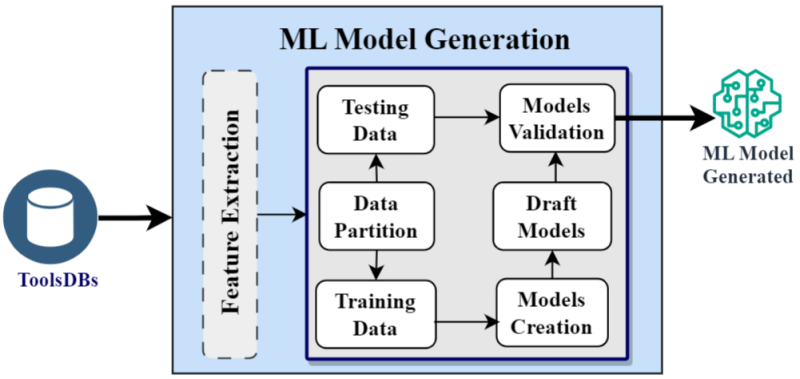


Figure 3. The machine-learning model generation process.

3.2.1 Feature Extraction

Since each phase and task of TIA has specific characteristics and must meet particular constraints, the features expressed as categorical variables in the above tool databases generation module are different for every task and phase of the pipeline. Moreover, each tool database has dedicated tools to perform the processing required for the corresponding task. For these reasons, for each TIA step, we have specified a particular ML processing with the corresponding tool database as input; the features are the tool characteristics and their target variables are the names of the tools.

3.2.2 ML Processing and Determining the Best ML Method

Our framework aims to predict the tool that best meets the experts' constraints and the specific data to be processed by the architecture. Leading to this prediction, we have opted for supervised learning, a branch of ML using algorithms to analyze the relationship between the constraints of each TIA phase and the tasks designated as features (e.g. SizeData, D_type, Latency, …etc.) and the tools stored in the corresponding database. However, in this instance, since each phase has its specific tool database and characteristics and to provide our framework with greater flexibility, we have opted to implement several ML methods and conduct performance tests, assuming that each ML method can perform differently and provide different results. The implemented ML methods in our process consist of Decision Trees, Random Forest, Support Vector Machine and Gradient Boosting.

The performance-evaluation measures examined in this module include Accuracy, Precision, Recall and F1-Score. In our context, we performed the following steps: Initially, we divided the input data into training, validation and testing sets. Then, we compared the performance of the different ML methods based on the specified performance metrics. Next, we trained each method on the training set using the extracted features. We optimized the hyperparameters of each classifier using grid search and cross-validation with a fold value of 5 (cv=5). The performance of each classifier was evaluated on the validation set utilizing the defined performance metrics. Finally, we choose the model generated by the best ML classification method in each phase to be used later in the architecture-generation process module.

As shown in Table 1, the evaluation of the implemented ML methods shows excellent results in the different phases. This asserts the effectiveness and reliability of the models generated, which are extremely useful for the framework's ability to provide precise, helpful advice on tool selection and pipeline implementation.

Despite the results on most phases (5/6) leading to the deduction that the decision tree has performed well, we have consolidated the evaluation by calculating the average of measures per metric over the whole pipeline. This confirms that the decision tree is the best method deployed to generate ML models and predict the best tool for each phase and task in the pipeline to the expert.

Table 1. Experimental results to determine the best ML method.

ML Method		**Decision Tree**				**Random Forest**				**Support Vector Machine**				**Gradient Boosting**			
Performance Metrics		**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Architecture Tasks	Data Acquisition	**0,83**	0,79	0,84	0,81	0,83	0,78	0,8	0,79	0,81	0,75	0,79	0,77	0,8	0,75	0,83	0,79
	Data Processing	**0,82**	0,79	0,84	0,81	0,8	0,74	0,83	0,78	0,8	0,75	0,79	0,77	0,79	0,76	0,79	0,77
	ML Data Processing	0,8	0,71	0,8	0,75	**0,84**	0,81	0,84	0,82	0,78	0,57	0,71	0,63	0,76	0,61	0,78	0,68
	Raw Data Storage	**0,83**	0,8	0,86	0,82	0,83	0,78	0,82	0,80	0,76	0,66	0,73	0,69	0,78	0,67	0,78	0,72
	Data Analytic Storage	**0,82**	0,76	0,87	0,81	0,78	0,76	0,81	0,78	0,77	0,68	0,78	0,73	0,76	0,69	0,78	0,73
	Data Analytics	**0,81**	0,79	0,88	0,83	0,8	0,74	0,82	0,78	0,79	0,69	0,81	0,75	0,77	0,77	0,85	0,81
Average Measures		**0,82**	0,77	0,85	0,81	0,81	0,77	0,82	0,79	0,79	0,68	0,77	0,72	0,78	0,71	0,80	0,75

Overall, in these experiments, we observed that the limited size of the tool databases slightly impacted the performance of the ML methods. However, the scalability and ease of maintenance and updating of the various modules of the framework allow for improving the performance of ML methods by expanding the number of supported tools. Indeed, the developed prediction models in the different pipeline phases and tasks can be easily updated when new data traces are available or the process model changes. This flexibility stems from the prediction model incorporating independent databases for each phase and task and we have appropriate characteristics for each corresponding tool. Finally, by leveraging ML methods and the relationships identified through supervised learning, we evaluated and validated the best ML method, allowing our proposed framework for effective tool selection at each pipeline phase and task.

In the next sub-section, we will detail the generation process of TSAs that meet the experts' needs and we will explain how they exploit the generated ML models.

3.3 Architecture Generation Process

The interactivity of our framework is based on multiple real-time exchanges between the expert and the framework during the various phases of the process of generating a tailored model of the TSA. As shown in Figure 4, this process consists of five phases: Predesign phase, Data Acquisition, Data Storage and Processing, Data Analytics and Data Consumption. After the authentication step, in the Predesign phase, the expert expresses the constraints related to his/her application case and the data specificities to be supported by the pipeline, such as data size, data type, acquisition mode and many other details like the intended analytic application, using a form. We have categorized the expert constraints in Table 2 according to the big-data V-challenges and the categorical variables previously used as features for generating ML models. Then, in each phase, the framework first displays all the tools available to perform the current task. Next, it runs the ML model corresponding to this task to predict the most suitable tools for the expert's needs, gathered from the form and the task's

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

specificities. Given the variety of tools studied in the ML phase, the prediction result may consist of several proposed tools. In this case, the expert selects the tool according to his/her preferences and validates his/her choice before proceeding to the next task.

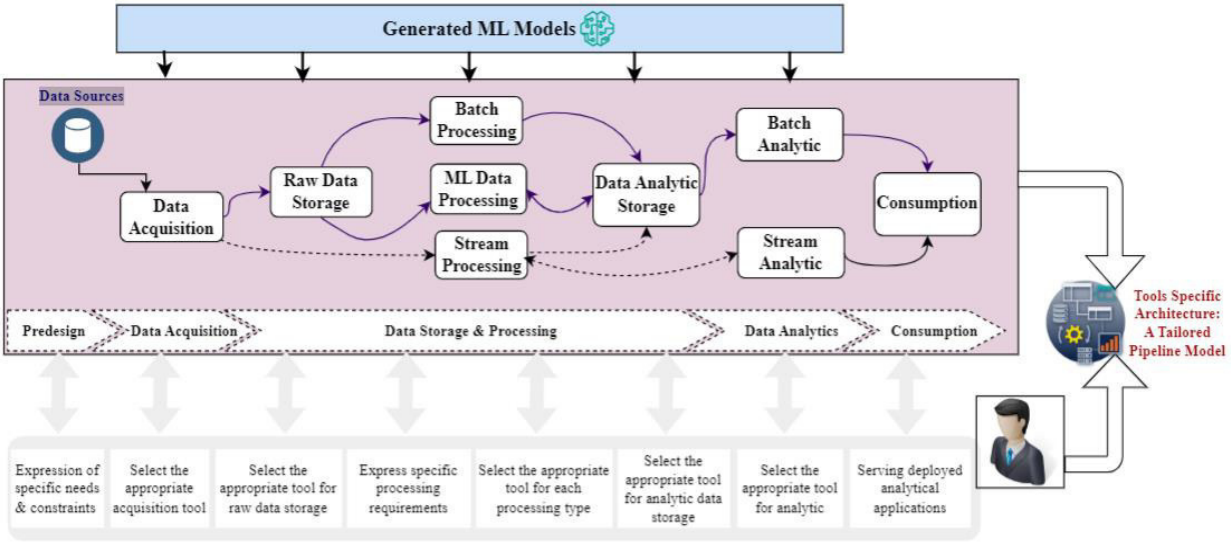


Figure 4. Architecture generation process's pipeline.

Moreover, given the importance of interoperability checking between tools of different phases and the subsequent issues arising during the deployment of a big-data architecture, we have conducted a detailed an interoperability study as detailed in [5] to which we refer interested readers. We have leveraged this study's results to deploy our framework. In this way, when the expert selects a tool that is non-interoperable with the tool already chosen in the previous phase, an alert message is displayed and consequently, the expert should choose another tool from the list predicted by the ML model. On the other hand, in some phases of the process, tasks require more details from the expert. For example, in the data processing task, to allow the process to predict the best tool, the framework requests the expert to express its requirements regarding the response time that the proposed tool should meet. The expert's interaction with the framework continues until the data consumption phase, which contains the data results to be consumed by the chosen tool in the data-analytics phase. We then find the analytical applications requested by the expert in the initial form, such as OLAP Navigation, dashboard, reporting, mobile application, forecasting, real-time analytics, statistics, visualization, …etc. Finally, at the end of the process, the expert recovers a tailored pipeline model that meets his/her specific constraints and the requirements of his/her application case, which he/she has been involved in composing step by step; this is its tailor-made end-to-end pipeline model for TSA.

Table 2. Experts' constraints expressed in terms of big-data features.

Big-data Features	**Volume**	**Velocity**	**Veracity**		**Variety**			
Expert's Constraint	**Size of Data**	**Time Acquisition**	**Response Time**	**Quality**	**Number of Data Sources**	**Data Type**		**Complexity**
Categorical Value	over1Peta	Real Time	less5	Loss and duplicate data	over20	Table, CSV, SQL DB, Spreadsheets, Log File	Structured	One model
	in1T1P	Near Real Time	in515	Loss data	in5_20	JSON, XML, Social Media	Semi-structured	Multiple models
	less1Tera	Batch	over15	Duplicate data	less5	Sensor Data, NoSQL DB, Video, Images, Geo-spacial Data	Unstructured	
				No loss nor duplicate				

In summary, our interactive, ML-based framework provides an entirely automatic process. The expert needs to express his/her requirements and is guided step by step to the final phase, where he/she obtains a pipeline model in which all the tools are interoperable. The following section will validate this framework by applying it to two use cases.

4. ARCHITECTAI: VALIDATION AND EVALUATION OF RESULTS

For ArchiTectAI evaluation, we opted for the ISO/IEC 25022 SQuaRE standard[1], in particular the user-satisfaction characteristic [6]. As defined in [7], satisfaction involves three sub-characteristics:

- Usefulness: "The degree of user satisfaction with achieving the objectives, including the results and consequences of use" [7].
- Trust: "The degree to which a user has confidence that a software product will perform as intended" [7].
- Comfort: "The extent of the user's satisfaction with physical comfort" [7].

We have developed a specific module for collecting and analyzing expert feedback to measure these criteria. As shown in Figure 5, we propose five levels of evaluation: Excellent, Good, Average, Poor and Very Poor. Next, we reached out to twenty experts who engage with big data for various purposes,requesting their participation in testing the framework against their respective requirements. Subsequently, we gathered their feedback, obtained upon the completion of the pipeline-generation process. Table 3 presents a comprehensive summary of these findings. Overall, the responses are very satisfactory for the majority and show that the framework has met the specific expectations of the experts.

For ArchiTectAI validation, we defined valuable criteria that were adapted to our context. For each criteria, we proposed the following evaluation questions:

- The consistency of the pre-design phase. Q1: Does the questions asked in the form cover all the experts' requirements and data specificities?
- Expert-framework interaction. Q2: Are the interactions satisfactory from an ergonomic perspective, particularly regarding usability and guidance?
- Clarity of the process for proposing and validating tool choice. Q3: Is it clear how the predicted tools are presented to the experts? Is the task of validating the choices simple?

Figure 5. ArchiTectAI evaluation by measuring the experts' satisfaction levels.

Table 3. Satisfaction evaluation.

Satisfaction	Usefulness	Trust	Comfort
Excellent	7	8	9
Good	10	10	9
Average	3	2	2
Poor	0	0	0
Very Poor	0	0	0

- Functional framework. Q4: When I implement the proposed TSA, do I have problems with tool interoperability?
- The expected results. Q5: After implementing the proposed TSA, do the analytical applications identified in the data-consumption phase meet the business needs?
- Technical constraints addressed. Q6: Did the ArchiTectAI consider the technical environment

[1] Systems and software-quality requirements and evaluation

of the experts when proposing the tools?

- Framework evaluation. Q7: Does the ArchiTectAI enable experts to express their satisfaction?

For the validation process, we have applied ArchiTectAI to two use cases, which have been addressed in previous works: [5] and [8]. For each use case, we proceeded as follows: We acted as the expert and followed the steps proposed by ArchiTectAI to generate the corresponding TSA. We checked how it assisted us and we assigned the symbol 'X' if ArchiTectAI validates the criteria expressed by the corresponding question. The results of this evaluation are displayed in Table 4.

4.1 Twitter Data Use Case Validation

In our previous work in [5], an interoperability and experimental study revealing the capabilities of the tools and their resource-consumption requirements were conducted. Two different pipelines were deployed for this experimental study to evaluate the popularity of teams and players before the start of the 2022 World Cup. In order to lead the validation process, we used ArchiTectAI to re-generate the two pipelines (examples of TSA). In the first pipeline, we acquired a dataset of approximately 80 million tweets in JSON format extracted from Twitter's general thread around November 2022, amounting to 500GB of data to be processed for batch-processing purposes. In the second one, we emulated a stream of approximately 1200 tweets per second for stream-processing purposes. In the pre-design phase, we selected in the form the options that addressed the specific characteristics of the data to be processed, e.g. data size, acquisition mode, data type, …etc. (Q1 validated). We then proceeded with all the steps guided by ArchiTectAI in continuous exchange throughout the process. For example, in the processing phase, ArchiTectAI demands the required response time as an additional request specific to the current phase (Q2 validated). At each step, the selection and validation of tool choices are simple and clear. Furthermore, in each phase of ArchiTectAI, a dynamic architecture diagram is generated using the Mermaid[2] technique, which assigns the tool to the current phase (Q3 validated). The previous work aimed to perform statistical reporting to achieve our analytical objective. The ArchiTectAI framework suggested uses Tableau tool, which we had already deployed in our architecture for generating reports; thus, the result was satisfactory (Q5 validated). Moving forward in the process, in the end, as shown in Figure 6, with the assistance of ArchiTectAI, we reproduced the pipeline-architecture model that we deployed in our case study. When implementing the architecture, we did not meet any issues of tool interoperability or a bottleneck for our processing engine. This proves that the proposed pipeline by ArchiTectAI is aligned with our technical constraints (Q4 and Q6 validated).

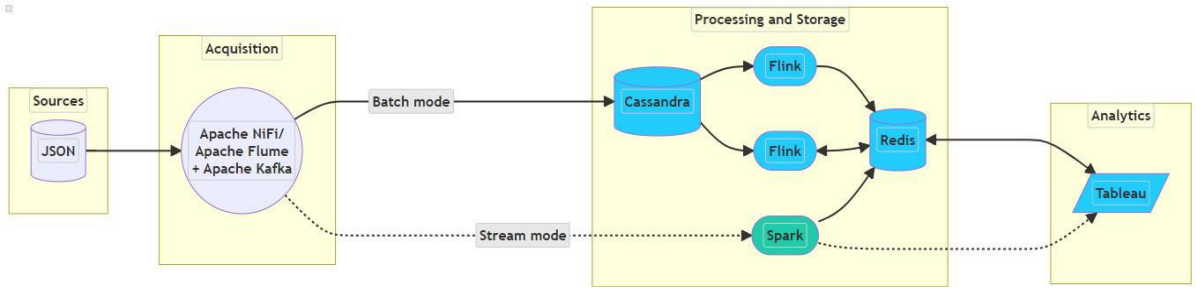


Figure 6. Tools specific architecture model generated for the Twitter data case study.

4.2 The Covid Pandemic Use Case Validation

In our previous work in [8], we introduced a multi-layer model for generating architectures for big-data warehousing. In this research study, we implemented an architecture for storing and analyzing multi-source data to examine the impact of the COVID-19 pandemic evolution on Twitter. This hybrid architecture supported streaming data from Twitter and batch data corresponding to the statistics collected on vaccination campaigns. To validate ArchiTectAI, we applied all its phases, initially, by specifying in the form all the business requirements and data specificities defined in this case study. Then, the responses to this form were analyzed and processed by ArchiTectAI in order to be available for the ML prediction model. As shown in Figure 7, particularly in the data-storage and processing phase, ArchiTectAI proposed a set of tools, in which we found those already deployed in the previous

[2] https://mermaid.js.org/

case study architecture [8]. Thus, we selected them, validated the choices and proceeded to the next phase (Q1 validated). Even for analysis tools, ArchiTectAI suggested tools already in previous use for dashboard creation, reporting and statistics to track the pandemic and vaccination campaigns. So, the expected result of the analysis would also be the same (Q5 validated). Regarding questions Q2 and Q3, we have encountered no problems. In fact, ArchiTechtAI's tool recommendations were clear, the choice was simple and the navigation to move from one phase to another was seamless. Moreover, during the implementation of our architecture, we had no interoperability problems. However, using Excel with a large amount of data caused a bottleneck (Q4 and Q6 validated). Finally, for both case studies, we completed the satisfaction survey proposed at the end of the ArchiTectAI process (Q7 validated).

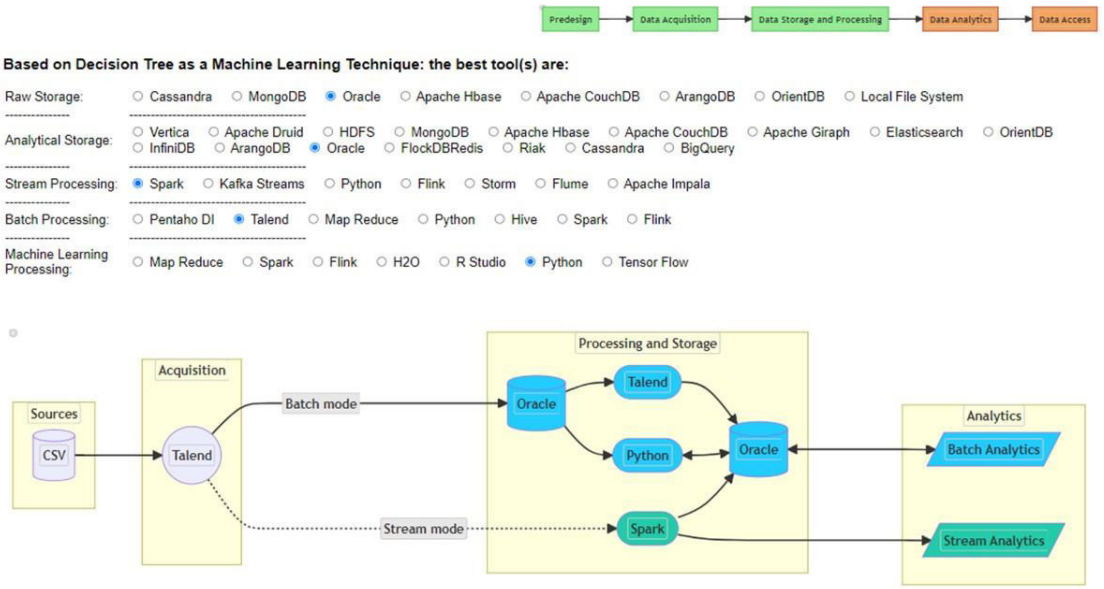


Figure 7. Tools specific architecture model generated for the Covid-19 pandemic case study: An overview of tool selection.

In summary, our ArchiTectAI interactive framework provided TSAs implemented in both use cases, proving that they are functional, consistent and support all data specifics and required constraints. They also proved that they met the professional requirements of both studies, with careful analysis of these requirements and supported by ML methods to predict the corresponding tools. Therefore, we have successfully validated ArchiTectAI, a decision-support framework, for big-data pipeline modeling, with particular emphasis on these two specific use cases on which we have extensively worked. However, the applicability of our framework extends far beyond these scenarios. Indeed, our overarching goal was to develop a highly generic framework that can be tailored to a diverse array of big-data application domains and use cases. For example, this includes processing streaming videos in real-time (e.g. ground traffic control), images (medical, satellite, …etc.), textual data (e.g., analyzing sentiment on social media) and other similar applications.

Table 4. Evaluation of validation questions by application of use cases.

Use Case	Validation Question						
	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Twitter Data	X	X	X	X	X	X	X
Covid Pandemic Data	X	X	X	X	X	X	X

5. Related Works and Ascertainments

This section is structured as follows. Sub-section 5.1 focuses on approaches addressing big-data challenges, while other approaches examine and evaluate big-data tools. In sub-section 5.2, we review a selection of research work that classifies some big-data tools within proposed architectures, with the purpose of facilitating a choice between them.

5.1 Big-data Warehousing and Analysis: Tools, Technologies and Big-data Features

Day to day, big-data features are raising new challenges for data-warehousing systems. Research studies have been conducted in the literature tackling particular challenges [9][10][11]. In [9], the authors focused on the variety issue by proposing an architectural design of a schema-less big-data repository aiming at capturing all data types. To cope with velocity, the authors in [10] elaborated an approach for detecting concept drift by investing in ML techniques utilizing both real and artificial data. As for Yousfi et al., they proposed in [11] a framework combining different processing engines in order to handle velocity. These engines operate parallel to perform the relevant matching and deliver the most complete and accurate data insight. On the other side, with the emergence of a wide variety of big-data tools, we notice that many surveys have been conducted to discuss and even evaluate these tools [12]-[13], [18]-[19]. However, none of these comparisons have classified the tools according to their satisfaction with big-data features. In particular, in [12], the authors compared some popular big-data frameworks based on the employed-programming model, the types of data sources utilized, the programming languages supported, the fault-tolerance support, the scalability and whether or not iterative processing is supported. Additionally, in [13], the authors considered scalability, distributed architecture, parallel computation and fault tolerance as comparison criteria. Recently, in [20], the author addressed a theoretical study of the big-data value chain, without focusing on specific technological solutions or practical implementations. It explored the conceptual relationships between big-data characteristics and the different stages of the value chain, but did not provide operational details on implementation and interoperability constraints between technologies.

5.2 Big-data Warehousing and Analysis: Approaches to Tool Selection

In [14], the authors defined criteria for choosing big-data tools and proposed a customer data analytics architecture. Despite the originality of their work, the approach has been limited to a restricted selection of tools, focusing only on smart-grid application. In [15], the authors proposed an approach utilizing key performance indicators (KPI), weightage and scores to help choose the best-ranked data warehousing tool for enterprises. In [16], the authors set forward a big-data analytical approach architecture. They classified the investigated approaches for analytical processing into NoSQL-based architecture, parallel relational database-based architecture and graph-based architecture. For each type of these architectures, they examined a set of tools according to these criteria: query language used, scalability, OLAP support, fault-tolerance support, cloud support, programming model and ML support. Moreover, the work in [17] is relevant to the scope of our research. The authors aimed to incorporate an iterative methodology for defining big-data analytics architectures. With its various phases, this methodology covers all the modeling tasks that a designer should perform to define a big-data pipeline. By considering the phase requirements regarding big-data characteristics, the authors introduced some technologies that can be deployed to meet these needs. Despite the importance of the proposed methodology, we note that they did not propose an automatic and interactive solution to guide the users in their choice of tools for each phase of the pipeline. In [20], by examining 110 significant and recently published articles, the authors conducted a comprehensive and systematic literature review on big-data management (BDM) techniques in the Internet of Things (IoT). They categorized the investigated mechanisms into four groups: BDM processes, BDM architectures/frameworks, quality attributes and types of big-data analysis. A detailed comparative analysis was provided for each category. Moreover, the authors presented a holistic BDM framework for IoT, including the following steps: data collection, communication, data ingestion, storage, processing and analysis and post-processing. The reviewed articles were classified according to these framework steps. Additionally, the study evaluated and compared various tools, platforms and frameworks used in the IoT domain based on qualitative criteria, such as performance, efficiency, accuracy and scalability. Finally, despite the comprehensive study presented in this paper and the advice derived from the authors' and other researchers' experiences, it's important to note that it exclusively focuses on techniques deployed in IoT.

Despite the community's awareness of the technological revolution associated with big data and the numerous efforts enacted to compare tools and propose approaches for designing big-data pipelines, we note the following shortages: 1- None of these works has proposed a generic architecture from which we can instantiate multiple specific pipeline models dedicated to different use cases. 2- The proposed approaches are limited to specific case studies. 3- The proposed approaches handle a limited

number of big-data tools and often do not focus on checking the interoperability between the proposed tools and the overall consistency of the proposed pipeline. 4- None of the studies proposed a method for creating a tool database classified according to satisfaction with big-data characteristics. 5- None of these works erected an automatic framework based on interaction with the experts to deduce from an exchange form all the particular needs, data specificities and technical constraints. 6- None of the proposed solutions relies on an ML model to analyze the experts' specific needs and identify the most appropriate tools. 7- None of the suggested approaches provide step-by-step assistance to experts composing their end-to-end big-data pipeline model. To address all these issues, we have proposed this ML-based interactive framework driven by a generic architecture to assist experts step-by-step in designing a big-data pipeline customized to their specific needs.

6. CONCLUSION

This research paper proposes an architecture to support a big data pipeline modeling interactive framework based on ML (ArchiTectAI). This architecture is based on three main modules. An *ad-hoc* method for generating tool databases has been developed for the first module. This method, from the list of tools for each Tools Independent Architecture phase and task and the different categorical variables characterizing big-data challenges, generates tool bases categorized according to their characteristics for each task in the big-data pipeline. The second module generates ML models. This module has implemented and evaluated several ML methods to choose the best one. The third module relies on these ML models to predict the best tool for each task and pipeline phase for the experts while respecting the constraints and specificities of the data. At the completion of this research, we evaluated the satisfaction of our interactive framework based on the ISO/IEC 25022 standard and validated it on two use cases. The consistency of the resulting pipeline proves the framework's effectiveness in its choice for suggesting tool choices. As a final note, this research work is extremely valuable and promising, as it opens further fruitful lines of investigation and offers promising future research directions. Indeed, our framework has been developed to be generic and scalable. Its adaptability allows for future updates with changes to existing tools or the addition of new ones, facilitated by the flexible underlying method for generating the tool database on which it depends. Furthermore, it can also be enriched with additional forms to address more constraints and expert-specific requirements. We also intend to enrich the framework with comprehensive guidelines for deploying the proposed architecture, specifying the connectors between tools. In addition, the tools and platforms for big-data governance and security are beyond this research's scope. In this respect, the elaborated work can be extended by incorporating this type of tools.

REFERENCES

[1] T. P. Raptis and A. Passarella, "A Survey on Networked Data Streaming with Apache Kafka," IEEE Access, vol. 11, pp. 85333-85350, 2023.

[2] S. Mishra and A. Misra, "Structured and Unstructured Big Data Analytics," Proc. of the 2017 IEEE Int. Conf. on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 740-746, Mysore, India, 2017.

[3] A. Davoudian and M. Liu, "Big Data Systems: A Software Engineering Perspective," ACM Computing Surveys (CSUR), vol. 53, no. 5, pp. 1-39, 2020.

[4] K. Rahul, R. K. Banyal and N. Arora, "A Systematic Review on Big Data Applications and Scope for Industrial Processing and Healthcare Sectors," Journal of Big Data, vol. 10, Article no. 133, 2023.

[5] A. Dhaouadi, W. Paccoud, K. Bousselmi, S. Monnet, M. M. Gammoudi and S. Hammoudi, "Big Data Tools: Interoperability Study and Performance Testing," Proc. of the IEEE Int. Conf. on Big Data, MIDP Workshop (MIDP-2023), pp. 2386-2395, 2023.

[6] ISO/IEC, "ISO/IEC 25022:2016 - Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — Measurement of Quality in Use," ISO/IEC 25022:2016, [Online], Available: https://www.iso.org/standard/35746.html, 2016.

[7] J. Sulla-Torres, A. Gutierrez-Quintanilla, H. Pinto-Rodriguez, R. Gómez-Campos and M. A. Cossio-Bolaños, "Quality in Use of an Android-based Mobile Application for Calculation of Bone Mineral Density with the Standard ISO/IEC 25022," IJACSA, DOI: 10.14569/IJACSA.2020.0110821, 2020.

[8] A. Dhaouadi, K. Bousselmi, S. Monnet, M. M. Gammoudi and S. Hammoudi, "A Multi-layer Modeling for the Generation of New Architectures for Big Data Warehousing," Proc. of the 36th Int. Conf. on Advanced Information Networking and Applications (AINA- 2022), vol. 2, pp. 204–218, 2022.

[9] A. M. Olawoyin, C. K. Leung, C. CJ. Hryhoruk and A. Cuzzocrea, "Big Data Management for Machine

Learning from Big Data," Proc. of the 37th Int. Conf. on Advanced Information Networking and Applications (AINA-2023), vol. 1, pp. 393–405, 2023.
[10] A. Abbasi, A. R. Javed, C. Chakraborty, J. Nebhen, W. Zehra and Z. Jalil, "ElStream: An Ensemble Learning Approach for Concept Drift Detection in Dynamic Social Big Data Stream Learning," IEEE Access, vol. 9, pp. 66408–66419, 2021.
[11] S. Yousfi, D. Chiadmi and M. Rhanoui, "Smart Big Data Framework for Insight Discovery," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 10, pp. 9777–9792, 2022.
[12] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri and E. M. Nguifo, "An Experimental Survey on Big Data Frameworks," Future Generation Computer Systems, vol. 86, pp. 546–564, 2018.
[13] S. Riaz, M. U. Ashraf and A. Siddiq, "A Comparative Study of Big Data Tools and Deployment Platforms," Proc. of the IEEE Int. Conf. on Engineering and Emerging Technologies (ICEET), pp. 1–6, Lahore, Pakistan, 2020.
[14] H. Daki, A. El Hannani, A. Aqqal, A. Haidine and A. Dahbi, "Big Data Management in Smart Grid: Concepts, Requirements and Implementation," Journal of Big Data, vol. 4, no. 1, pp. 1–19, 2017.
[15] M. R. Sureddy and P. Yallamula, "Approach to Help Choose Right Data Warehousing Tool for an Enterprise", Int. J. of Advance Research, Ideas and Innovat. in Technol., vol. 6, no. 4, pp. 579-583, 2020.
[16] Y. Cardinale, S. Guehis and M. Rukoz, "Classifying Big Data Analytic Approaches: A Generic Architecture," Proc. of the 12th Int. Joint Conf. on Software Technologies (ICSOFT), Part of the Book Series: Communi. in Computer and Information Science, vol. 868, pp. 268-295, Madrid, Spain, 2018.
[17] R. Tardio, A. Mate and J. Trujillo, "An Iterative Methodology for Defining Big Data Analytics Architectures," IEEE Access, vol. 8, pp. 210597–210616, 2020.
[18] S. Alkatheri, S. A. Abbas and M. A. Siddiqui, "A Comparative Study of Big Data Frameworks," Int. J. of Computer Science and Information Security (IJCSIS), vol. 17, no. 1, pp. 66-73, 2019.
[19] M. Khalid and M. Murtaza Yousaf, "A Comparative Analysis of Big Data Frameworks: An Adoption Perspective," Applied Sciences, vol. 11, no. 22, p. 11033, 2021.
[20] A. A. Aydin, "A Comparative Perspective on Technologies of Big Data Value Chain," IEEE Access, vol. 11, pp. 112133 – 112146, 2023.
[21] A. Naghib, N. J. Navimipour, M. Hosseinzadeh and A. Sharifi, "A Comprehensive and Systematic Literature Review on the Big Data Management Techniques in the Internet of Things," Wireless Networks, vol. 29, no. 3, pp. 1085-1144, 2023.

ملخص البحث:

إنّ عمليـــة اســـتيداع البيانـــات تتطلّـــب ثـــورةً بِنيويـــةً لتســـوية تحـــدّيات البيانـــات الضّـــخمة والتّعامـــل مـــع مصـــادر البيانـــات الجديـــدة، مثـــل شـــبكات التّواصُـــل الاجتمـــاعي، وأنظمـــة التّوصـــية، والمـــدن الذّكيـــة، وشـــبكة الويـــب لاســـتخلاص قيمـــةٍ مـــن البيانـــات المتبادَلـــة. وفـــي هـــذا الإطـــار، فـــإنّ مجتمـــع نمذجـــة خطـــوط البيانـــات مـــن أجـــل اكتســـاب البيانـــات وتخزينهـــا ومعالجتهـــا بهـــدف تحليلهـــا يوظّـــف طيفـــاً واســـعاً مـــن الحلـــول التّكنولوجيـــة الّتـــي تنطـــوي بـــدورها علـــى تحـــدّياتٍ مهمّـــة وصـــعوباتٍ جمّـــة. وبشـــكلٍ أكثـــر تحديـــداً، فـــإنّ اختيـــار الأداة المثلـــى الملائمـــة لاحتياجـــات العمـــل الخاصّـــة بالمســـتخدم وموضـــوع تبادُليـــة التّشـــغيل بـــين الأدوات المختلفة أصبح من بين التّحديات الأساسية.

مـــن هـــذا المنطلـــق، نقتـــرح فـــي هـــذه الدراســـة إطـــار عمـــلٍ تفاعليـــاً جديـــداً مبنيـــاً علـــى تقنيـــات تعلُّـــم الآلـــة لمســـاعدة الخبـــراء فـــي نمذجـــة خـــطّ بيانـــاتٍ مـــن أجـــل اســـتيداع البيانـــات. وعلـــى نحوٍ أدّقّ، فإننا نعمل على:

أ) تحليل متطلّبات الخبراء وخصائص البيانات المطلوب معالجتها.
ب) اقتراح البِنية الملائمة لتلك المتطلّبات من بين مجموعةٍ من البِنى.
ج) تحقيـــق ذلـــك مـــن خـــلال عـــددٍ مـــن طـــرق تعلّـــم الآلـــة لتوقُّـــع أكثـــر الأدوات المناســـبة لكلّ مرحلة ولكلّ مهمّة داخل البِنية.

بالإضـــافة إلـــى ذلـــك، جـــرى التّحقُّـــق مـــن فاعليـــة إطـــار العمـــل المقتـــرح باســـتخدام حـــالتَي استخدامٍ من العالم الحقيقي، إلى جانب التّغذية الرّاجعة من المستخدمين.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

PARALLEL BUCKET-SORT ALGORITHM ON OPTICAL CHAINED-CUBIC TREE INTERCONNECTION NETWORK

Basel A. Mahafzah

(Received: 15-Mar.-2024, Revised: 12-May-2024 and 19-Jun.-2024, Accepted: 23-Jun.-2024)

ABSTRACT

The performance of sorting algorithms has a great impact on many computationally intensive applications. Researchers worked on parallelizing many sorting algorithms on various interconnection networks to improve their sequential counterpart performance. One of these interconnection networks is the optical chained-cubic tree (OCCT). In this paper, a parallel bucket sort (PBS) algorithm is presented and applied to the OCCT interconnection network. This PBS algorithm is evaluated analytically and by simulation in terms of various performance metrics including parallel runtime, computation time, communication time, concatenation time, speedup and efficiency, for a different number of processors, dataset sizes and data distributions including random and descending distributions. Simulation results show that the highest obtained speedup is approximately 912x on OCCT using 1020 processors, which means that the parallel runtime of the PBS on 1020 processors is 912 times faster than the sequential runtime of BS on a single processor.

KEYWORDS

Bucket sort, Parallel sorting algorithm, Interconnection network, Opto-electronic architecture.

1. INTRODUCTION

Many researchers concentrate their efforts on minimizing the run time needed to perform sorting algorithms efficiently on various architectures [1]-[11]. Also, several comparative sorting algorithms have been presented and analyzed in detail to show their advantages and disadvantages [12]-[17]. In general, sorting algorithms are among the most studied algorithms and are important in the computer science field, since sorting is one of the most essential operations used in many problems and applications, such as integer problems, databases, search engines, text data, image processing and information retrieval [18]-[24].

One of the well-known sorting algorithms is the bucket sort (BS) [14][19][25], which is a good choice for sorting elements with values uniformly distributed over an interval. In the BS algorithm, the interval is divided into consecutive non-overlapping sub-intervals called buckets to sort the input, where each element is placed in an appropriate bucket based on the element's value and each bucket is sorted using any sorting algorithm, such as quicksort, merge sort, count sort, insertion sort, …etc. Then, buckets are concatenated to form the sorted list [19], [25]-[26].

Practically, sorting a large number of elements using a sequential bucket-sort algorithm requires a high runtime. So, one way to improve the runtime of the bucket-sort algorithm is to run it on parallel or distributed architectures [27]-[30]. Examples of these architectures are optical chained-cubic tree (OCCT) [31] and optical transpose interconnection system (OTIS) and its variants, such as OTIS-Mesh, OTIS-Hypercube and OTIS Hyper Hexa-Cell (OHHC) [32]-[34].

The OCCT interconnection network is based on the chained-cubic tree (CCT) which is constructed from a tree and hypercubes in addition to electronic and optical links [31][35]. The electronic links connect processors within tree levels and hypercubes, whereas optical links are added on a certain level of the tree to reduce the distance between processors. In general, optical links can carry data with less power consumption and a high data rate compared to electronic links [36]-[37]. OCCT shows efficient topological properties including low diameter, high maximum node degree and high bisection width [31]-[32]. Also, the CCT was evaluated by implementing a parallel bitonic sort algorithm on this interconnection network, where it showed a great performance [3]. The efficient properties of

B. A. Mahafzah is with Department of Computer Science, King Abdullah II School of Information Technology, The University of Jordan and with Department of Computer Science, King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan. Emails: b.mahafzah@ju.edu.jo and b.mahafzah@psut.edu.jo

OCCT and the previous work on CCT motivate us to implement a parallel bucket-sort (PBS) algorithm on OCCT taking advantage of the OCCT-structure properties to get an efficient parallel sorting algorithm.

The main contribution of this paper is implementing an efficient PBS algorithm on the OCCT interconnection network and evaluating the PBS algorithm analytically and by simulation in terms of parallel runtime, computation time, communication time, concatenation time, speedup and efficiency, for different numbers of processors and dataset sizes and two types of data distributions; namely, random and descending distributions.

2. RELATED WORK

Several research works have been conducted on the parallel bucket-sort algorithm using various architectures and platforms. For example, in [27], the author showed how to convert a sequential bucket-sort algorithm into a parallel algorithm, which has been implemented and executed using OpenMP API. Experimental results showed that this parallel version is a scalable algorithm, where its performance can be improved as the number of cores is increased. Also, in [28], the authors implemented a parallel bucket-sort algorithm for a many-core architecture of graphics processing units (GPUs) based on convex optimization. Moreover, in [29], the author used threads and GPU programming to optimize the bucket-sort algorithm. Experimental results showed that for a small number of elements, it is better to carry out the sorting in a single thread. Also, using the bucket-sort algorithm, the bottleneck of GPU and CPU is shown in this research work clearly.

Additionally, several research works have been conducted on opto-electronic architectures. In [32], the authors presented a detailed review of nine optoelectronic architectures in terms of their topological structure and topological properties including the OCCT. These opto-electronic architectures are interconnection networks that use electronic and optical links to connect processors. All these architectures except the OCCT are based on OTIS. These architectures are evaluated in terms of various topological properties; namely, size, diameter, cost, bisection width, maximum node degree and minimum node degree and Hamiltonian path and cycle. Among these architectures, the OCCT showed great performance in terms of diameter, maximum node degree and bisection width [31]-[32]. However, up to this time and up to our knowledge, none of the parallel bucket-sorting algorithms has been applied to opto-electronic architectures, which motivates us to implement an efficient parallel bucket-sort algorithm on the OCCT opto-electronic architecture and evaluate it analytically and by simulation in terms of various performance metrics.

3. OCCT INTERCONNECTION NETWORK

The structure of the OCCT interconnection network [31] is based on CCT [35], where the CCT interconnection network is based on a binary tree and hypercubes. The height h of OCCT is floor(log G) and each hypercube in OCCT is a group G of 2^d processors of dimension d, in addition to a specific level lv that is chosen according to the height of the tree where the optical links are added in a cascading manner between distant hypercubes at that level. Thus, OCCT is referred to as OCCT (h, d, lv). An OCCT can be a full or complete binary tree network based on the status of its last level. Figure 1 shows a full OCCT(3, 2, 2) [31], where 3 is the height of the tree, 2 is the dimension of each hypercube group and 2 is the lv level number wherein at that level, the optical links are added (thick black lines). Figure 2 shows the lv level where lv = 2 in details of the OCCT(3, 2, 2) [31]. Also, as shown in Figure 2, the label of each processor is unique and contains a pair of numbers (G_i, p_j). For example, processor (3, 2) means processor number 2 in group number 3. However, more details regarding the labeling of groups and processors in OCCT can be found in [31].

The lv value depends on two factors; the type of binary tree whether it is full or complete. If the tree is a full binary tree, then the level lv = ceiling(h/2) and if the tree is a complete binary tree, then the level lv depends on the tree height type; whether odd or even and the number of groups in the last level. Thus, there are three cases; the first case is if the tree height h is even, then lv= h/2. In the second case, if the tree height h is odd and the number of groups in the last level is less than $(2^{(h-1)/2})\times3+1$, then lv = (h-1)/2. In the third case, if the tree height h is odd and the number of groups in the last level is greater than or equal to $(2^{(h-1)/2})\times3+1$, then lv = (h+1)/2 [31]. However, more details regarding implementing

the structures of OCCT and CCT can be found in [31][35].

The size is the number of processors in the OCCT interconnection network. The size of the OCCT(h, d, lv) is $G \times 2^d$, where G is the number of hypercube groups in the tree and 2^d is the number of processors in each hypercube of dimension d [31]-[32].

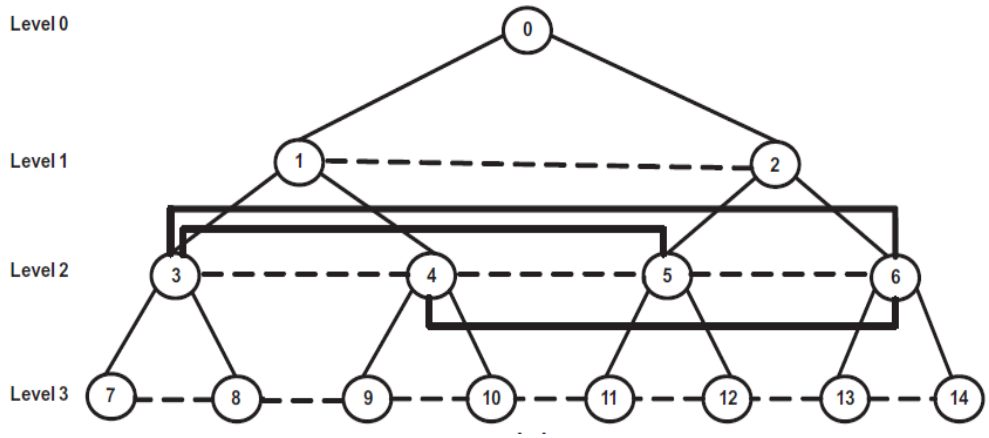


Figure 1. An OCCT(3, 2, 2).

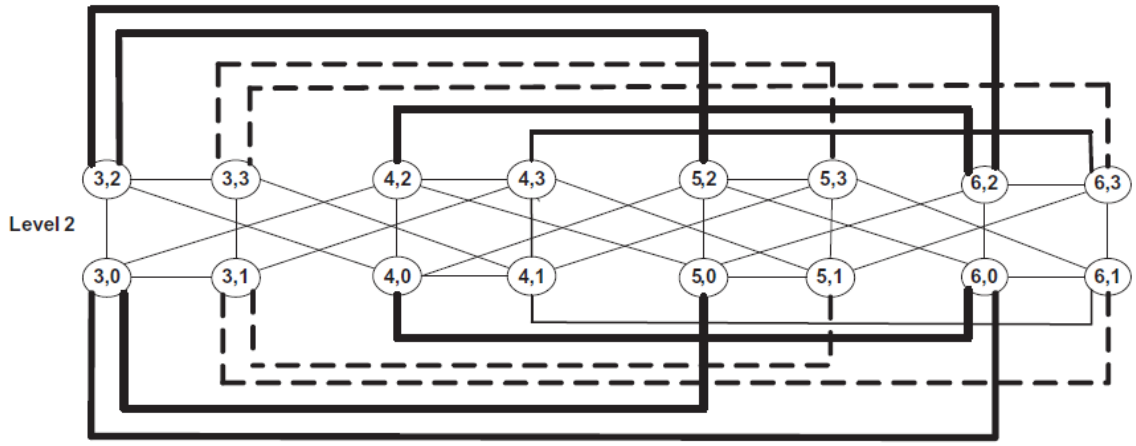


Figure 2. Level two of OCCT(3, 2, 2).

4. SEQUENTIAL BUCKET SORT ALGORITHM

The sequential BS algorithm is a well-known sorting algorithm. It sorts n elements the values of which are uniformly distributed over an interval [1, n], where this interval is divided into b equal-sized sub-intervals called buckets. That is, the BS algorithm uses buckets of the same size and each element is placed in the appropriate bucket according to its value. As a result, each bucket will have almost the same number of elements which is approximately n/b. Then, the BS algorithm uses an efficient and easy-to-implement sorting algorithm, such as quicksort [14][38], to sort the elements in each bucket. Finally, these sorted buckets are concatenated in the appropriate order to form the final sorted list. The run time of this sequential BS algorithm is $\Theta(n \log (n/b))$, where this low time complexity is due to the assumption that the n elements to be sorted are uniformly distributed over the interval [1, n].

The sequential BS algorithm's steps are shown in Algorithm 1, where the input parameters are defined in Table 1 and the size of a bucket (s) and its sub-interval are computed using Equations (1)-(3).

$$s = (max - min + 1) / b \tag{1}$$

$$Bucket_{start}(B) = min + B \times s, \qquad \text{where } B = 0, 1, 2, \ldots, b-1 \tag{2}$$

$$Bucket_{end}(B) = Bucket_{start}(B) + s - 1 \tag{3}$$

Algorithm 1	Sequential BS algorithm's steps.
Input	Initially, a single processor has the input of n elements, which are distributed uniformly over the interval [1, n].
Output	Sorted n elements in ascending order.
1	Sizes and sub-intervals of the buckets are computed by a single processor using Equations (1)-(3).
2	Using the created buckets in Step 1, each element is placed in its appropriate bucket according to its value and the bucket's sub-interval. That is the input array is scanned from left to right and each element is moved to its appropriate bucket.
3	After all elements are placed in their buckets, a call to a quicksort algorithm is executed to sort each bucket.
4	After each bucket is sorted, buckets are concatenated in the required order to produce the final sorted list.

Table 1. Input parameters of sequential BS algorithm.

Parameters	**Description**
n	Number of elements to be sorted (input size)
b	Number of buckets
B	Bucket number
max	Maximum element value in the input (last index of input array equals n)
min	Minimum element value in the input (first index of input array equals 1)
s	Size of the bucket, which is the number of elements in the bucket, where all buckets have almost the same size
$Bucket_{start}(B)$	First element in the sub-interval of bucket number B
$Bucket_{end}(B)$	Last element in the sub-interval of bucket number B

5. PARALLEL BUCKET SORT ALGORITHM ON OCCT

In this section, we introduce the PBS algorithm on the OCCT interconnection network, as shown in Algorithm 2. In the proposed PBS algorithm, we assume a list of n elements uniformly distributed over the interval [1, n] to be sorted using a number of buckets (b), where these b buckets are almost of the same size. Each bucket is assigned to a single processor in OCCT; that is, $p = b$, where p is the number of processors. Also, we assume that the bucket size is s, where $s = n/b$. Additionally, we assume initially that each processor has a complete copy of the input n elements.

The PBS algorithm consists of two phases: the computation phase and the communication and concatenation phase, where, in Algorithm 2, steps 1–4 present the computation phase and steps 5 and 6 present the communication and concatenation phase.

Algorithm 2 works as follows: In step 1, in parallel, each processor in OCCT calculates the size of the bucket and the sub-intervals of the buckets using Equations (1)-(3).

In step 2, each processor is assigned a bucket according to a global group sequential ordering, where for example every four buckets are assigned to one group (i.e., one hypercube of four processors) by sequential order. For example, the first four buckets numbered 0 to 3 are assigned to group 0, while the second four buckets which are numbered 4 to 7 are assigned to group 1, …and so on.

In step 3, in parallel, each processor scans the entire n input elements and determines the elements that belong to its bucket according to both its bucket sub-interval and the elements' values which must be within the bucket's sub-interval. As a result, each bucket will have almost the same number of elements, which is approximately n/b.

In step 4, in parallel, each processor applies the sequential quicksort algorithm to sort the elements of its bucket which are approximately n/b elements. The quicksort algorithm is an in-place sorting

algorithm that does not need additional memory space to sort the n/b elements and it is easy to implement using the divide-and-conquer approach. Also, it is efficient in terms of time complexity, which is $O(n/b \log_2 n/b)$, since it sorts n/b elements in each processor in parallel and independently.

In step 5, the processors communicate with each other to combine their sorted buckets at a single processor located at the left-most hypercube group of the *lv* level in OCCT, where the optical links exist to take advantage of these links. Thus, the communication pattern of the proposed PBS algorithm takes place in three major stages as follows:

1. Upper and lower tree communication stage: In this stage, processors at the *lv* level gather results from processors at upper and lower levels in the tree at the same time in parallel.
2. Hypercube-communication stage: In this stage, each group of processors from a hypercube at the *lv* level gathers their results at processor number 0 of that hypercube.
3. Optical-communication stage: In this stage, processor number 0 of the left-most hypercube at the *lv* level gathers results from other processors number 0 of other hypercubes of the same *lv* level.

Finally, in step 6, at this left-most group, the single processor that received all sorted buckets concatenates them as one list of elements according to the buckets' number from lowest to highest which presents the buckets' sub-intervals from lowest to highest to have the n elements sorted in ascending order.

Algorithm 2	PBS algorithm's steps on OCCT.
Input	Initially, each processor p_i has a complete copy of the input n elements which are uniformly distributed over the interval [1, n].
Output	Sorted n elements in ascending order.
1	In parallel, the size of the bucket and the sub-intervals of buckets are computed by each processor in OCCT using Equations (1)-(3).
2	Each processor p_i is assigned a bucket b_i according to a global ordering, where for example every four buckets are assigned to one group (i.e., one hypercube of four processors) by sequential order.
3	In parallel, each processor p_i scans the input n elements and determines the elements that belong to its bucket b_i according to its sub-interval and the elements' values.
4	In parallel, each processor p_i applies the sequential quicksort algorithm to sort the elements of its bucket b_i.
5	Processors communicate with each other to combine and concatenate their results at a single processor located at the left-most group of the *lv* level in OCCT.
6	At this single processor, the buckets are concatenated according to their number and sub-intervals from lowest to highest to have the n elements sorted.

The proposed PBS algorithm is modified slightly and customized to be applied to OCCT architecture efficiently. The modification is made in the computation phase by having the buckets almost equal in size and $p=b$ to distribute buckets on processors evenly (i.e., load-balanced) to have all processors finish approximately at the same time, which leads to better performance. The customization is made in the communication and concatenation phase, where the buckets are distributed and gathered in less communication time using the electronic and optical links; that is reaching all processors using the shortest path (the diameter of OCCT).

An example of the PBS algorithm's steps is shown in Figure 3. In this example, we do not show the communication phase in detail, for simplicity. Also, in this example, we assume that $n = 16$ and $b = p = 4$. Therefore, the size of each bucket is 4. Initially, each processor has a copy of the input list which contains uniformly distributed 16 elements over the interval [1, 16], as shown in Figure 3(a). Then the bucket's size and sub-interval of each bucket are computed using Eqs. (1–3) and each processor determines its elements according to its bucket's sub-interval, as shown in Figure 3(b). Then, each processor sorts its bucket, as shown in Figure 3(c) and finally, the processors communicate to gather

the buckets at processor(0), where it concatenates the buckets sequentially according to their number and sub-intervals from lowest to highest to have a sorted list in ascending order, as shown in Figure 3(d).

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(0)	5	3	1	2	6	8	10	11	15	16	14	13	12	7	9	4

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(1)	5	3	1	2	6	8	10	11	15	16	14	13	12	7	9	4

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(2)	5	3	1	2	6	8	10	11	15	16	14	13	12	7	9	4

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(3)	5	3	1	2	6	8	10	11	15	16	14	13	12	7	9	4

(a) Initially, each processor has a copy of the input list which contains uniformly distributed 16 elements over the interval [1, 16].

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(0) Subinterval [1, 4]	5	**3**	**1**	**2**	6	8	10	11	15	16	14	13	12	7	9	**4**

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(1) Subinterval [5, 8]	**5**	3	1	2	**6**	**8**	10	11	15	16	14	13	12	**7**	9	4

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(2) Subinterval [9, 12]	5	3	1	2	6	8	**10**	**11**	15	16	14	13	**12**	7	**9**	4

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Processor(3) Subinterval [13, 16]	5	3	1	2	6	8	10	11	**15**	**16**	**14**	**13**	12	7	9	4

(b) Bucket sizes and the sub-intervals of buckets are computed and each bucket is assigned to a processor.

Index	1	2	3	4		Index	1	2	3	4
Processor(0) Unsorted Bucket(0)	3	1	2	4	➔	**Processor(0) Sorted Bucket(0)**	1	2	3	4

Index	1	2	3	4		Index	1	2	3	4
Processor(1) Unsorted Bucket(1)	5	6	8	7	➔	**Processor(1) Sorted Bucket(1)**	5	6	7	8

Index	1	2	3	4		Index	1	2	3	4
Processor(2) Unsorted Bucket(2)	10	11	12	9	➔	**Processor(2) Sorted Bucket(2)**	9	10	11	12

Index	1	2	3	4		Index	1	2	3	4
Processor(3) Unsorted Bucket(3)	15	16	14	13	➔	**Processor(3) Sorted Bucket(3)**	13	14	15	16

(c) Each processor sorts its unsorted bucket.

Buckets subintervals	**1**	2	3	**4**	**5**	6	7	**8**	**9**	10	11	**12**	**13**	14	15	**16**
Concatenated sorted buckets at Processor(0)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Bucket(0)				**Bucket(1)**				**Bucket(2)**				**Bucket(3)**			

(d) Processor(0) concatenates all received buckets to form a sorted list.

Figure 3. Example of the PBS algorithm's steps.

6. ANALYTICAL EVALUATION

In this section, the proposed PBS algorithm is evaluated analytically in terms of several performance metrics, including parallel runtime complexity, speedup and efficiency. The parallel runtime is the time that passes from the moment that a parallel computation starts to the moment at which the last processor finishes execution, where it includes the time that a parallel program spends in computation and communication [13]. Thus, the analytical evaluation of the PBS algorithm is presented for its computation phase first in sub-section 6.1 followed by its communication and concatenation phase in sub-section 6.2 and lastly, in sub-section 6.3, the mentioned performance metrics are presented.

6.1 Computation-phase Analysis

In this sub-section, the computation analysis of the proposed PBS algorithm (Algorithm 2) is presented. In the computation phase, according to Algorithm 2, four steps are shown as follows, where each step is followed by its expected sequential and parallel runtime complexity:

1. *Finding bucket size and sub-intervals of all buckets*: Sequential and parallel runtime complexity for finding bucket size is constant O(1) and for finding the sub-intervals of all buckets is O(b), since in BS and PBS algorithms, finding the size and the sub-intervals of buckets can be carried out using Equations (1)-(3).
2. *Assigning buckets*: Sequential and parallel runtime complexity is O(1), since, in the BS algorithm, all buckets are assigned to a single processor and in the PBS algorithm, each processor is assigned one bucket according to a global sequential ordering; specifically, buckets are assigned according to group number and processor number.
3. *Putting each element in its proper bucket*: Sequential and parallel runtime complexity is O(n), since, in BS and PBS algorithms, each processor passes over the n elements to find its bucket elements.
4. *Sorting buckets*: Sequential runtime is O($b \times s \log s$) as best and average cases, since we have b buckets to sort each of size s, assuming that each bucket has the same number of elements and using quicksort to sort each bucket sequentially. The quicksort best and average-case time complexity is O($n \log n$) to sort n elements, where these cases occur when the elements are random [14][38]. Since $n = b \times s$, then the sequential runtime is O($n \log s$). In the worst case, the time complexity of quicksort is O(n^2) to sort n elements, where this case occurs when the elements are already ascendingly or descendingly sorted [14]. Thus, the bucket sort worst-case time complexity would be O($b \times s^2$) = O($n \times s$). Whereas, the parallel runtime complexity is O($s \log s$), since in parallel each processor uses a quicksort algorithm to sort its s elements in the best and average cases and in the worst case, it would be O(s^2). Note that $s << n$, since $s = n / b$.

The sequential and parallel computation runtimes of the PBS algorithm differ only in the fourth step, which is the step of sorting buckets. Thus, based on the four computational steps, the sequential runtime (T_{seq}) of the BS algorithm as best and average cases and worst case is shown in Equations (4)-(5), respectively. The parallel computation runtime (T_{comp}) of the PBS algorithm as best and average cases and worst case is shown in Equations (6)-(7), respectively.

$$T_{seq} = O(1) + O(b) + O(1) + O(n) + O(n \log s) \approx O(n \log s) \quad (\textit{best \& average cases}) \quad (4)$$

$$T_{seq} = O(1) + O(b) + O(1) + O(n) + O(n \times s) \approx O(n \times s) \quad (\textit{worst case}) \quad (5)$$

$$T_{comp} = O(1) + O(b) + O(1) + O(n) + O(s \log s) \approx O(s \log s) \quad (\textit{best \& average cases}) \quad (6)$$

$$T_{comp} = O(1) + O(b) + O(1) + O(n) + O(s^2) \approx O(s^2) \quad (\textit{worst case}) \quad (7)$$

6.2 Communication and Concatenation Phase Analysis

In this sub-section, the communication and concatenation phase of the proposed PBS algorithm (Algorithm 2) is presented. The communication pattern of the proposed PBS algorithm has the following three stages: The upper and lower tree communication stage, the hypercube-communication stage and the optical-communication stage.

In general, the communication time equals the number of required steps multiplied by the message size, where a message may contain one sorted bucket or two concatenated sorted buckets or more, multiplied by the time to transmit a word of data on an electronic (T_{we}) or an optical link (T_{wo}).

For the first stage, the upper tree communication time is presented in Equation (8). The number of required steps for the upper tree communication equals the value of *lv*, because we have to pass *lv* levels from the root of the tree (level 0) to the optical links level (*lv*). In each step, we pass a bucket of size $s = (n / b)$. Also, in this upper communication, buckets are transmitted using only electronic links. So, the total communication time required for the upper tree is T_{up}, as shown in Equation (8).

$$T_{up} = lv \times s \times T_{we} \tag{8}$$

The tree height from the last level of the tree to the *lv* level is $(h - lv)$, which is equal to *lv* or $(lv–1)$ depending on the place of the optical links, as discussed in Section 3. Thus, the number of communication steps is the maximum between *lv* and $(lv - 1)$, which is *lv*. In the first step, the size of the transferred message is *s*, which is the size of a single bucket. In the second step, the size becomes $(2 \times s)$, while in the third step, the size is $(4 \times s)$, …and so on. At maximum, the size of the transferred message is $(2^{lv-1} \times s)$; that is at every step, the size of the transferred message doubles, which means that the number of transferred buckets doubles. Also, only the electronic links are used to transfer the buckets to the *lv* level. So, the total communication time required for the lower tree is T_{low}, as shown in Equation (9), where the size of the transferred buckets is $(\sum_{i=1}^{lv} 2^{i-1} \times s)$. Equation 9 can be simplified, as shown in Equation (10). Note that, in our simulation runs, the values of *lv* vary according to the size of OCCT, specifically from 2 to 4, which is a small value.

$$T_{low} = lv \times (\sum_{i=1}^{lv} 2^{i-1} \times s) \times T_{we} \tag{9}$$

$$T_{low} = lv \times ((2^{lv} - 1) \times s) \times T_{we} \tag{10}$$

Since the upper and the lower tree communication are carried out in parallel, then the communication time of this stage is T_{up-low}, which is the maximum time between the upper and the lower tree communication times, as shown in Equation (11). Thus, since the lower tree communication time is larger than the upper tree communication time because a larger message size is transferred, the communication time of this stage is dominated by the lower-communication time.

$$T_{up-low} = \max((lv \times s \times T_{we}), (lv \times ((2^{lv} - 1) \times s) \times T_{we})) \tag{11}$$

The number of communication steps in the hypercube is *d*, which is the dimension of the hypercube, in our case $d = 2$. Specifically, in the first step, the size of the message equals the results (concatenated sorted buckets) gathered from stage 1 in addition to one bucket that each processor in the hypercube originally has. However, the hypercube-communication time in this stage depends on the number of received buckets from the previous stage. So, the total hypercube-communication time (T_Q) is shown in Equation (12), where the size of the received buckets from the upper tree levels is $(lv \times s)$ and the size of the received buckets from the lower tree levels including the bucket that each processor in the hypercube originally had is $\left(2 \times \left(2^{lv} - 1\right) \times s + s\right)$ where it can be simplified as $(\left(2^{lv+1} - 1\right) \times s)$. Also, the communication links used in the hypercubes are electronic links.

$$T_Q = d \times ((lv \times s) + (\left(2^{lv+1} - 1\right) \times s)) \times T_{we} \tag{12}$$

In the last stage of communication, the optical communication stage, processor number 0 at the left-most hypercube of level *lv* gathers all results (concatenated sorted buckets) from its counterpart processors numbered 0 of other hypercubes in the same level *lv* using the optical links. This required at most two optical steps, since there are no adjacent nodes and groups connected using optical links in the *lv* level [31]. The size of the transferred concatenated sorted buckets is the size of the received buckets from stage 2 (hypercube-communication stage) to processor 0, in addition to its bucket. Equation 13 presents the total optical communication time of this stage, which is T_{op}.

$$T_{op} = 2 \times (2^d \times (((lv - 1) \times s) + (\left(2^{lv+1} - 1\right) \times s))) \times T_{wo} \tag{13}$$

The total communication time (T_{comm}) of the PBS algorithm on OCCT is the summation of the upper and the lower-communication time, hypercube communication time and optical communication time, as shown in Equation (14), which are presented in Equations (11)-(13), respectively.

$$T_{comm} = T_{up-low} + T_Q + T_{op} \tag{14}$$

During each stage of communication, sorted buckets are concatenated once they are received by a processor according to the group and processor numbers. So, the parallel runtime complexity of the

concatenation of the buckets is T_{pc}, as shown in Equation (15), which is the number of communication steps in each stage, where lv is the number of communication steps in the upper and the lower communication stage, d is the number of communication steps in the hypercube-communication stage and 2 is the number of communication steps in the optical-communication stage. Also, lv and d are small values, where in our case d = 2 and lv varies between 2 and 4 according to the OCCT size. However, the sequential runtime complexity of the concatenation of buckets is O(b), where b is the number of buckets.

$$T_{pc} = O(lv + d + 2) \tag{15}$$

6.3 Performance Metrics

In this sub-section, the performance metrics of the proposed PBS algorithm are presented, including the parallel runtime complexity, speedup and efficiency. The parallel runtime complexity (T_p) of the PBS algorithm on OCCT is the summation of the computation, communication and concatenation times, as shown in Equation (16), which are presented in Equations (6) (14) (15), respectively, as shown in Equation (17). However, the parallel runtime complexity of the PBS algorithm, which is presented in Equation (17) as the best and average cases, is dominated by the computation time for large n, which is the common case. The speedup (S_p) is defined as the sequential runtime divided by the parallel runtime of solving the same problem, as shown in Equation (18). Thus, the speedup of the PBS algorithm on OCCT is shown in Equation (19), where the sequential runtime complexity (T_{seq}) of the BS algorithm is shown as the best and average cases. Accordingly, Equation (19) shows the speedup as the best and average cases. The efficiency (E_f) is defined as the speedup divided by the number of used processors, as shown in Equation (20). Thus, the efficiency of the PBS algorithm on OCCT is shown in Equation (21) as the best and average cases.

$$T_p = T_{comp} + T_{comm} + T_{pc} \tag{16}$$

$$T_p = O(s \log s) + (T_{up-low} + T_Q + T_{op}) + O(lv + d + 2) \tag{17}$$

$$S_p = \frac{T_{seq}}{T_p} \tag{18}$$

$$S_p = \frac{O(n \log s)}{O(s \log s) + (T_{up-low} + T_Q + T_{op}) + O(lv + d + 2)} \tag{19}$$

$$E_f = \frac{S_p}{p} \tag{20}$$

$$E_f = \frac{O(n \log s)}{p \times (O(s \log s) + (T_{up-low} + T_Q + T_{op}) + O(lv + d + 2))} \tag{21}$$

7. SIMULATION ENVIRONMENT AND RESULTS

In this section, the simulation environment and results are presented and discussed. The simulation results are evaluated in terms of two performance metrics; namely, speedup and efficiency.

7.1 Simulation Environment

The OCCT interconnection network does not exist as a real-machine or real-computing environment. Therefore, the OCCT interconnection network and the algorithms are implemented using Java Virtual Threads, simulated as a distributed memory model. However, the simulation runs under a shared memory multi-core computer machine.

In this sub-section, the simulation environment is presented, including software and hardware specifications and input-data distributions. The simulation implementation is programmed using Java Virtual Threads, which offer lightweight and efficient concurrency management within the Java Virtual Machine, on a multi-core computer machine with the specifications provided in Table 2. The parameter settings of the PBS algorithm's conducted simulation runs are shown in Table 3. Also, Table 4 presents the required parameter settings to implement the OCCT interconnection network which are the type of OCCT whether it is full or complete, the height h of the tree, the number of processors (p), the number of groups (G) in OCCT, the number of groups at the last level of OCCT (G_L) and the values of the level lv, where these values are calculated according to the equations presented in Section 3.

Table 2. Hardware specifications of the computer machine used for simulation runs.

Processor	Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz (4 Cores and 8 Threads)
RAM	16 GB
Operating system	64-bit Windows 10 Pro

Table 3. Parameter settings of the PBS algorithm's conducted simulation runs.

Type of sorting	Ascending
Type of input elements	4-byte integer number
Input size (MB)	0.4, 1.2, 2, 2.8, 4, 12, 20, 40, 60, 100, 150, 200, 400
Type of OCCT	Full OCCT(*h*, 2, *lv*), two-dimensional hypercube Complete OCCT(*h*, 2, *lv*), two-dimensional hypercube
Number of processors	Full OCCT: 60, 124, 252, 508, 1020 Complete OCCT: 92, 188, 380, 764
Input-data distribution	R: Random D: Descending and continuous (reverse ordered)

Table 4. Parameter settings of the OCCT interconnection network.

OCCT Type	Height (*h*)	Number of Processors (*p*)	Number of Groups (*G*)	Number of Groups at Last Level (G_L)	Level (*lv*)
Full	3	60	15	8	2
Complete	4	92	23	8	2
Full	4	124	31	16	2
Complete	5	188	47	16	3
Full	5	252	63	32	3
Complete	6	380	95	32	3
Full	6	508	127	64	3
Complete	7	764	191	64	4
Full	7	1020	255	128	4

Table 5. Communication-time parameters.

Parameters	Values
W_{size}	$4\ Bytes \times Byte\ Size = 4 \times 8 = 32\ bits$
L_{size}	$4\ Bytes \times Byte\ Size = 4 \times 8 = 32\ bits$
El_{speed}	$250\ Mbps$
Op_{speed}	$2.5\ Gbps$
T_{we}	$\frac{W_{size}}{El_{speed}} = \frac{32}{250 \times 1024 \times 1024} = 122.1\ nsec$
T_{wo}	$\frac{W_{size}}{Op_{speed}} = \frac{32}{2.5 \times 1024 \times 1024 \times 1024} = 12.2\ nsec$

Moreover, to compute and analyze the communication time, the values of word size (W_{size}) which is equal to the element size (L_{size}), electronic link speed (El_{speed}) [36]-[37], optical link speed (Op_{speed}) [36], time to transmit word of data on the electronic link (T_{we}) and time to transmit word of data on the optical link (T_{wo}), are shown in Table 5.

7.2 Simulation Results

In this sub-section, the simulation results are presented and evaluated in terms of speedup and efficiency using random and descending input-data distributions. Figures 4 and 5 show the speedup of

PBS using random and descending distributions of different sizes on a different number of processors on OCCT, respectively. In these figures, the speedup was highest when the input is 40 MB and lowest when the input is 0.4 MB. However, two main cases can be observed from Figures 4 and 5 as follows:

- For a certain size of the input data distribution, the speedup increases as the number of processors increases. This is because, as we increase the number of processors, the computation time on each processor decreases since the data on each processor is decreased in size.
- For a certain number of processors, the speedup increases as we increase the size of the input-data distribution. This is because the gap between parallel runtime and sequential runtime increases as the size of data is increased.

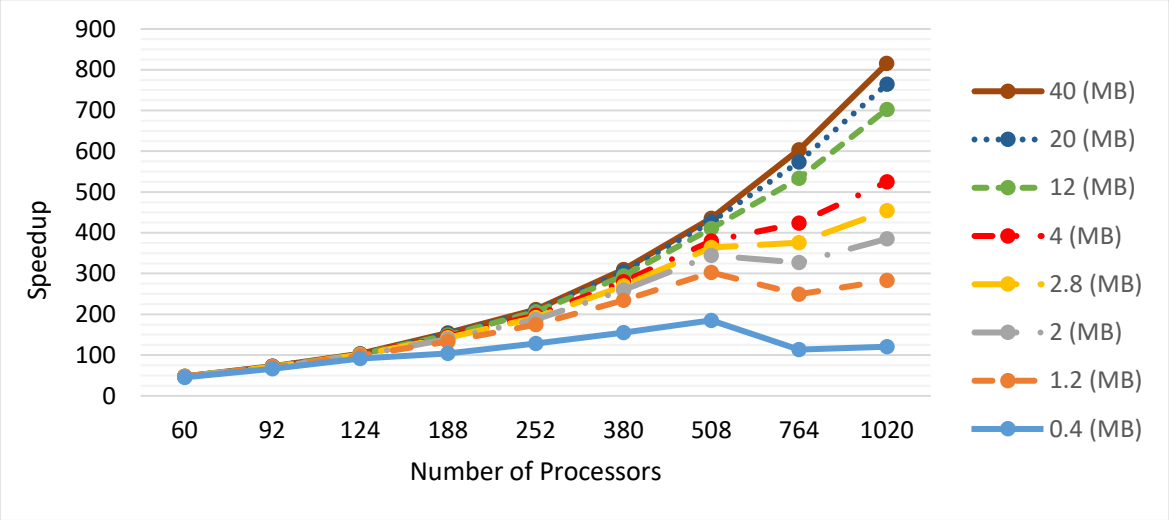


Figure 4. PBS speedup for various random data-distribution sizes on OCCT.

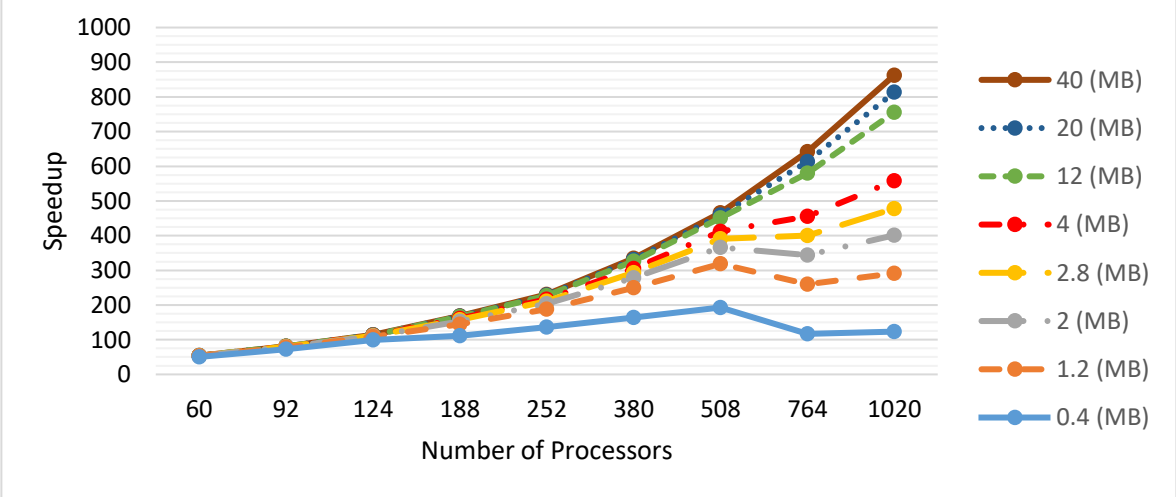


Figure 5. PBS speedup for various descending data-distribution sizes on OCCT.

Figures 6 and 7 show the efficiency of the PBS algorithm running on a different numbers of processors using random and descending distributions over OCCT, respectively. The highest efficiency is achieved, which is approximately 92%, when we used descending data distribution of size 40 MB on 124 processors, as shown in Figure 7. However, two main cases can be observed from Figures 6 and 7 as follows:

- For a certain small size of the input-data distribution, the efficiency decreases as the number of processors increases.
- For a certain number of processors, the efficiency decreases as we decrease the size of the input data distribution.

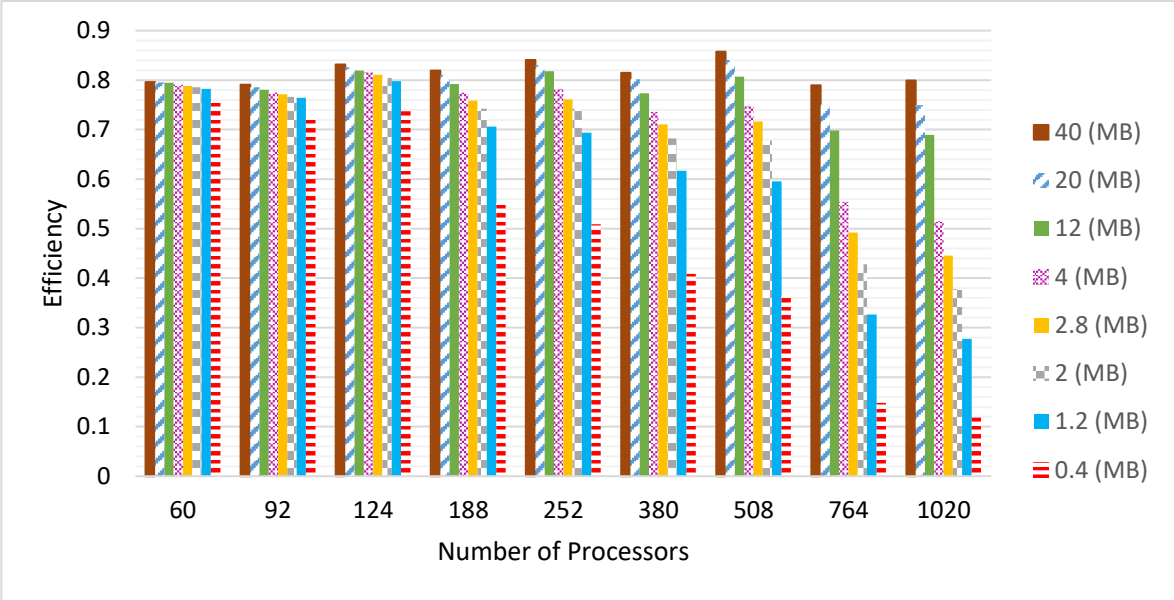


Figure 6. PBS efficiency for various random data-distribution sizes on OCCT.

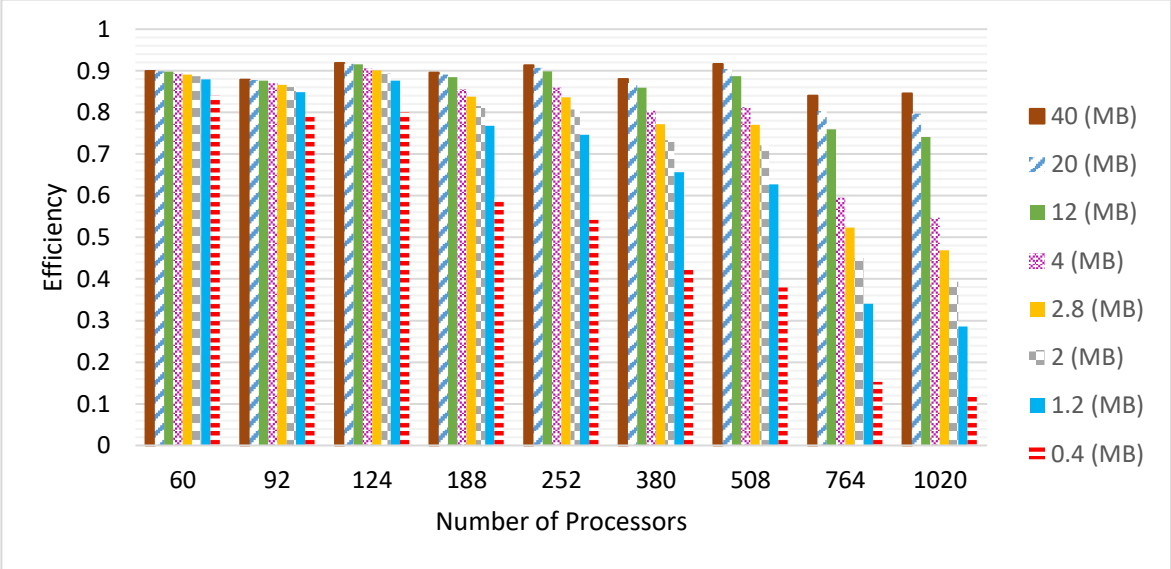


Figure 7. PBS efficiency for various descending data-distribution sizes on OCCT.

In general, data distribution affects the performance of the PBS algorithm. Specifically, sorting random data distribution on OCCT using the sequential quicksort on each processor, as mentioned in Algorithm 2, will lead to the best and average case scenarios, whereas sorting descending data distribution will lead to the worst-case scenario.

Figure 8 shows the scalability of the proposed PBS algorithm in terms of different large data sizes ranging from 60 MB to 400 MB presented as random data distributions on 1020 OCCT processors. Specifically, as shown in Figure 8, as the data gets larger, the speedup gets higher; that is for 60, 100, 150, 200 and 400 MB, the speedup is approximately 832, 846, 857, 861 and 871, respectively. Additionally, the proposed PBS algorithm is compared with the parallel quicksort (PQS) algorithm in terms of speedup, as shown in Figure 8. In this comparison, the PBS algorithm is applied on the OCCT interconnection network using 1020 processors, whereas the PQS algorithm is applied on the OHHC interconnection network using 1152 processors. The difference in the number of processors is due to the structures of the OCCT and OHHC interconnection networks, as shown in [31][34]. As shown in Figure 8, the PBS algorithm outperforms the PQS algorithm for all ranges of data sizes. However, for 400 MB, the PBS algorithm outperforms the PQS algorithm slightly; specifically the PBS algorithm achieved a speedup of 871, whereas the PQS algorithm achieved a speedup of 867. This is due to the number of processors in OHHC which has more processors than OCCT by 132.

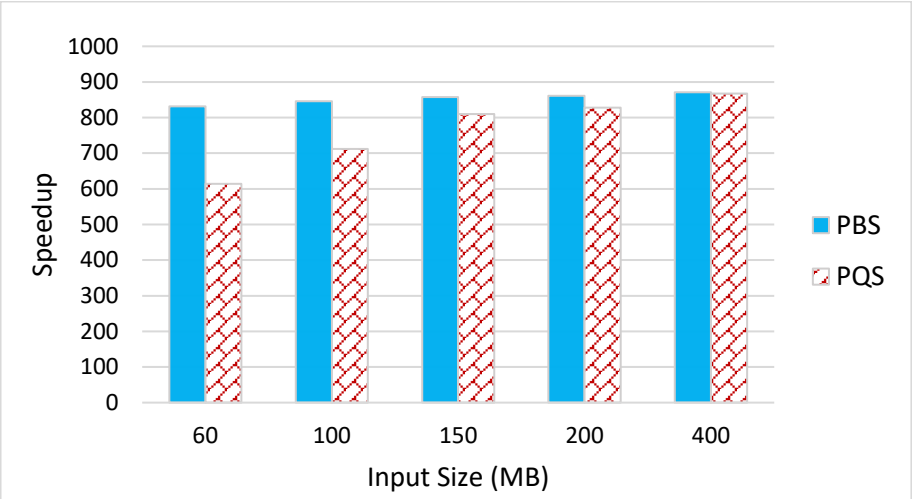


Figure 8. Speedup comparison between the PBS algorithm on 1020 OCCT processors and the PQS algorithm on 1152 OHHC processors for various random data-distribution sizes.

Additionally, as shown in Table 6, the PBS on OCCT is compared with the parallel bucket-sort algorithm using various techniques and architectures, where random data distribution of size 40 MB is used. In this table, the results show that PBS on OCCT outperforms these techniques and architectures since the parallel bucket-sort algorithms are implemented on shared-memory architectures with limited resources. Specifically, as the number of threads increases, the performance of these algorithms decreases; that is, creating more threads than cores degrades the performance of these algorithms, as mentioned in [27][29].

Table 6. Speedup comparisons between the PBS algorithm on OCCT and other parallel bucket-sort algorithms using different techniques and architectures.

Technique/Architecture	Speedup	Threads/Nodes	Number of Buckets
Multi-threaded [29]	1.88	8	100
OpenMP [29]	3.13	8	100
GPGPU using CUDA [29]	1.25	8	100
OpenMP API [27]	2.2	8	100
PBS on OCCT	**6.54**	8	100

Table 7 shows analytical *versus* simulation speedup for the PBS algorithm using 400 MB descending data distribution on OCCT for different numbers of processors. It can be seen from the table that the difference between the analytical and simulation speedup is small, ranging from 7% up to 12%, which validates the correctness of the obtained simulation results.

Table 7. Analytical versus simulation speedup results of PBS algorithm over various numbers of processors on OCCT using descending data distribution of 400 MB input size.

	Number of Processors								
	60	**92**	**124**	**188**	**252**	**380**	**508**	**764**	**1020**
Analytical	59.995	91.989	123.981	187.902	251.824	379.594	507.2651066	760.377	1013.471
Simulation	53.996	80.951	114.063	169.112	231.678	337.839	471.7565491	669.132	912.124
Difference	**10%**	**12%**	**8%**	**10%**	**8%**	**11%**	**7%**	**12%**	**10%**

8. CONCLUSIONS

In this paper, an efficient PBS is implemented on OCCT using up to 1020 processors, up to 400 MB of input-data size and two data distributions; namely, random and descending. The performance of the PBS algorithm on OCCT is evaluated analytically in terms of parallel runtime, which includes computation, communication and concatenation, in addition to speedup and efficiency. Also, the PBS algorithm is evaluated by simulation in terms of speedup and efficiency. Moreover, a comparison is

presented in terms of speedup between the PBS algorithm on 1020 processors of the OCCT and the PQS algorithm on 1152 processors of the OHHC for random data distribution ranges from 60 MB to 400 MB.

As simulation results, the PBS algorithm on OCCT outperforms the PQS algorithm on OHHC in terms of speedup for various random data sizes ranging from 60 MB to 400 MB. A maximum speedup of approximately 912x is obtained on OCCT using 1020 processors and descending input-data distribution of size 400 MB. Also, a maximum efficiency of approximately 92% is obtained on OCCT using 124 processors and descending input-data distribution of size 40 MB, which means that the utilization of the OCCT processors reaches 92%.

In general, the PBS algorithm has some limitations, where its performance can be affected by the type of data distribution. For example, when the data distribution is random, then the best and average cases are obtained, since we used the sequential quicksort to sort the data locally at each processor. Whereas the worst-case is obtained when we used the descending data distribution for the same mentioned reason.

Moreover, in general, bucket-sort performance is sensitive to the distribution of the input values; so, if you have tightly clustered values, it is not recommended. Also, the performance of bucket sort depends on the number of buckets chosen, which might require some extra performance tuning compared to other algorithms. However, these limitations need to be considered when the bucket-sort algorithm is applied to various architectures.

As potential future research directions to this work, the PBS algorithm can be applied to other opto-electronic interconnection networks, such as the OTIS and its variants, to show the performance of such opto-electronic interconnection networks [31] [33]. Moreover, the PBS algorithm can be applied to other well-known architectures, such as multi-threaded architectures, shared-memory multi-core architectures and mesh-connected multi-processors to evaluate their performance [1], [39]-[40].

ACKNOWLEDGMENT

The author would like to express his deep gratitude to the anonymous referees for their valuable comments and helpful suggestions, which enhanced this research manuscript. This research work was conducted during the sabbatical leave from the University of Jordan for the academic year 2022/2023, where this research work was accomplished at the Department of Computer Science, King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan.

REFERENCES

[1] L. Rashid, W. M. Hassanein and M. A. Hammad, "Analyzing and Enhancing the Parallel Sort Operation on Multithreaded Architectures," Journal of Supercomputing, vol. 53, pp. 293–312, 2010.

[2] N. R. Nitin, "Analysis of Multi-sort Algorithm on Multi-mesh of Trees (MMT) Architecture," Journal of Supercomputing, vol. 57, pp. 276–313, 2011.

[3] S. Al-Haj Baddar and B. Mahafzah, "Bitonic Sort on a Chained-cubic Tree Interconnection Network," Journal of Parallel and Distributed Computing, vol. 74, pp. 1744–1761, 2014.

[4] F. Dehne and H. Zaboli, "Parallel Sorting for GPUs," In: Adamatzky, A., editor. Emergent Computation. Emergence, Complexity and Computation (ECC), vol. 24, DOI: 10.1007/978-3-319-46376-6_12, Springer, Cham., 2017.

[5] A. Al-Adwan, R. Zaghloul, B. Mahafzah and A. Sharieh, "Parallel Quicksort Algorithm on OTIS Hyper Hexa-Cell Optoelectronic Architecture," Journal of Parallel and Distributed Computing, vol. 141, pp. 61–73, DOI: 10.1016/j.jpdc.2020.03.015, 2020.

[6] M. K. I. Rahmani, "Smart Bubble Sort: A Novel and Dynamic Variant of Bubble Sort Algorithm," Computers, Materials & Continua, vol. 71, no. 3, pp. 4895–4913, 2022.

[7] P. Preethi, K. G. Mohan, S. Kumar and K. K. Mahapatra, "Sorter Design with Structured Low Power Techniques," SN COMPUT. SCI., vol. 4, no. 129, DOI: 10.1007/s42979-022-01546-7, 2023.

[8] Y. Han and X. He, "More Efficient Parallel Integer Sorting," International Journal of Foundations of Computer Science, vol. 33, no. 5, pp. 411–427, DOI: 10.1142/S0129054122500071, 2022.

[9] B. Bramas, "A Fast Vectorized Sorting Implementation Based on the ARM Scalable Vector Extension (SVE)," PeerJ Computer Science, vol. 7, p. e769, DOI: 10.7717/peerj-cs.769, 2021.

[10] O. Obeya, E. Kahssay, E. M. Fan and J. Shun, "Theoretically Efficient and Practical Parallel In-place Radix Sorting," Proc. of the 31st ACM Symposium on Parallelism in Algorithms and Architectures,

DOI: 10.1145/3323165.3323198, 2019.

[11] T. Tokuue and T. Ishiyama, "Performance Evaluation of Parallel Sortings on the Supercomputer Fugaku," Journal of Information Processing, vol. 31, pp. 452–458, 2023.

[12] S. K. Gill, V. P. Singh, P. Sharma and D. Kumar, "A Comparative Study of Various Sorting Algorithms," Int. Journal of Advanced Studies of Scientific Research, vol. 4, pp. 367–372, 2019.

[13] A. Grama, A. Gupta, G. Karypis and V. Kumar, Introduction to Parallel Computing, 2ed Edition, Reading: Addison-Wesley, 2003.

[14] T. Cormen, C. Leiserson, R. Rivest and C. Stein, Introduction to Algorithms, 4th Edition, Massachusetts: The MIT Press, 2022.

[15] H. Wang et al., "PMS-Sorting: A New Sorting Algorithm Based on Similarity," Computers, Materials & Continua, vol. 59, no. 1, pp. 229–237, DOI: 10.32604/cmc.2019.04628, 2019.

[16] M. Nowicki, "Comparison of Sort Algorithms in Hadoop and PCJ," Journal of Big Data, vol. 7, no. 101, DOI: 10.1186/s40537-020-00376-9, 2020.

[17] M. Garland, "Chapter 13 – Sorting," in the Book: Programming Massively Parallel Processors: A Hands-on Approach, 4th Edition, Morgan Kaufmann Publisher, pp. 293–310, DOI: 10.1016/B978-0-323-91231-0.00019-7, 2023.

[18] W. X. Zhang and Z. Wen, "Efficient Parallel Algorithms for Some Integer Problems," Proc. of the 19th Annual Conference on Computer Science (CSC '91), pp. 11–20, San Antonio, USA, DOI: 10.1145/327164.327169, 1991.

[19] T. Rożen, K. Boryczko and W. Alda, "GPU Bucket Sort Algorithm with Applications to Nearest-Neighbour Search," Journal of WSCG, vol. 16, pp. 161–167, 2008.

[20] M. Amirul et al., "Sorting Very Large Text Data in Multi GPUs," Proc. of the 2012 IEEE Int. Conf. on Control System, Computing and Engineering, pp. 160–165, Penang, Malaysia, DOI: 10.1109/ICCSCE.2012.6487134, 2012.

[21] M. Asiatici, D. Maiorano and P. Ienne, "How Many CPU Cores Is an FPGA Worth? Lessons Learned from Accelerating String Sorting on a CPU-FPGA System," Journal of Signal Processing Systems, vol. 93, pp. 1405–1417, DOI: 10.1007/s11265-021-01686-8, 2021.

[22] H. Chen, S. Madaminov, M. Ferdman and P. Milder, "Sorting Large Datasets with FPGA-accelerated Sample Sort," Proc. of 27th IEEE Int. Symposium on Field-Programmable Custom Computing Machines (FCCM 2019), Art. no. 8735541, p. 326, DOI: 10.1109/FCCM.2019.00067, 2019.

[23] M. Kaur and V. Kumar, "Parallel Non-dominated Sorting Genetic Algorithm-II-based Image Encryption Technique," The Imaging Science Journal, vol. 66, no. 8, pp. 453–462, 2018.

[24] J. Xie, Z. Li, H. Wu, L. Li, B. Pan, P. Guo and G. Sun, "Application of Quicksort Algorithm in Information Retrieval," Journal on Big Data, vol. 3, no. 4, pp. 135–145, 2021.

[25] N. Faujdar and S. Saraswat, "The Detailed Experimental Analysis of Bucket Sort," Proc. of the 7th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence), pp. 1–6, Noida, India, DOI: 10.1109/confluence.2017.7943114, 2017.

[26] M. Khurana, N. Faujdar and S. Saraswat, "Hybrid Bucket Sort Switching Internal Sorting Based on the Data Inside the Bucket," Proc. of the 6th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 476–482, DOI: 10.1109/icrito.2017.8342474, Noida, India, 2017.

[27] H. Hong, "Parallel Bucket Sorting Algorithm Hiep Hong," [Online], Available: https://api.semanticscholar.org/CorpusID:33533373, 2014.

[28] G. Beliakov, G. Li and S. Liu, "Parallel Bucket Sorting on Graphics Processing Units Based on Convex Optimization," Optimization, vol. 64, pp. 1033–1055, 2015.

[29] H. I. S. Wijayabandara, Performance Analysis of Parallel Bucket Sort, Thesis for Master's Degree in Computer Science, University of Colombo, School of Computing, 2018.

[30] K. Chen, H. Chen and C. Wang, "Bucket MapReduce: Relieving the Disk I/O Intensity of Data-intensive Applications in MapReduce Frameworks," Proc. of the 29th Euromicro Int. Conf. on Parallel, Distributed and Network-based Processing (PDP), DOI: 10.1109/pdp52278.2021.00013, 2021.

[31] B. Mahafzah, M. Alshraideh, T. Abu-Kabeer, E. Ahmad and N. Hamad, "The Optical Chained-cubic Tree Interconnection Network: Topological Structure and Properties," Computers & Electrical Engineering, vol. 38, pp. 330–345, DOI: 10.1016/j.compeleceng.2011.11.023, 2012.

[32] B. Mahafzah, A. Al-Adwan and R. Zaghloul, "Topological Properties Assessment of Opto-electronic Architectures," Telecomm. Systems, vol. 80, pp. 599–627, DOI: 10.1007/s11235-022-00910-5, 2022.

[33] G. C. Marsden, P. J. Marchand, P. Harvey and S. C. Esener, "Optical Transpose Interconnection System Architectures," Optics Letters, vol. 18, pp. 1083–1085, 1993.

[34] B. A. Mahafzah, A. Sleit, N. A. Hamad, E. F. Ahmad and T. M. Abu-Kabeer, "The OTIS Hyper Hexa-Cell Optoelectronic Architecture," Computing, vol. 94, pp. 411–432, 2012.

[35] M. Abdullah, E. Abuelrub and B. Mahafzah, "The Chained-cubic Tree Interconnection Network," Int. Arab Journal of Information Technology, vol. 8, pp. 334–343, 2011.

[36] O. Kibar, P. J. Marchand and S. C. Esener, "High Speed CMOS Switch Designs for Free-space Opto-

electronic MINs," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 6, pp. 372–386, DOI: 10.1109/92.711309, 1998.

[37] I. Kaminow, T. Li and A. Willner, "Optical Fiber Telecommunications VB: Systems and Networks," 5th Edition, Academic Press, 2008.

[38] D. K. J. Lin and J. Chen, "Adaptive Order-of-Addition Experiments *via* the Quick-sort Algorithm," Technometrics, vol. 65, no. 3, pp. 396–405, DOI: 10.1080/00401706.2023.2174601, 2023.

[39] M. S. Mohammed and G. A. Abandah, "Characterization of Shared-memory Multi-core Applications," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 2, no. 1, pp. 37–54, DOI: 10.5455/jjcit.71-1448574289, 2016.

[40] J. Al-Azzeh, "Distributed Mutual Inter-unit Test Method for D-Dimensional Mesh-connected Multiprocessors with Round-Robin Collision Resolution," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 5, no. 1, pp. 1–16, DOI: 10.5455/jjcit.71-1539688899, 2019.

ملخص البحث:

إنّ لأداء خوارزميات التّصنيف أثراً كبيراً على العديد من التّطبيقات التي تنطوي على حساباتٍ مكثّفة. وقد عمل الباحثون على موازاة الكثير من خوارزميات التّصنيف في شبكات اتّصال متنوّعة بغية تحسين أدائها المقابل التّتابعي. ومن بين تلك الشّبكات ما يُعرف بالشّجرة المكّعبة المسَلْسَة ضوئياً.

في هذه الورقة، نقدّم خوارزمية تصنيف دَلْوِيّة متوازية، ونقوم بتطبيقها على شبكة اتّصال مكّعبة مُسَلْسَة ضوئياً. كذلك نعمل على تقييم الأداء التّصنيفي لتلك الخوارزمية عن طريق كلٍّ من التّحليل والمحاكاة من حيث مجموعة متنوّعة من مقاييس الأداء بما فيها زمن التّشغيل المتوازي، وزمن الحساب، وزمن الاتّصال، وزمن التّجميع، والسّرعة، والفعالية؛ وذلك لأعداد مختلفة من المعالِجات، ولأحجام مختلفة من مجموعات البيانات، ولأنواع مختلفة من توزيع البيانات (التّوزيع العشوائي، والتّوزيع التّنازلي).

وبيّنت نتائج المحاكاة أنّ أقصى سرعة تمّ الحصول عليها كانت لشبكة اتّصال فيها (1020) معالِجاً، حيث بلغ زمن التّشغيل المتوازي باستخدام (1020) مُعالِجاً (912) ضِعْفاً مقارنة بزمن التّشغيل التّتابعي للتّصنيف الدَّلْوِي باستخدام مُعالِجٍ واحد.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 03, September 2024.

THE DEEP LEARNING MODEL FOR DECAYED-MISSING-FILLED TEETH DETECTION: A COMPARISON BETWEEN YOLOV5 AND YOLOV8

Maya Fitria[1], Yasmina Elma[1], Maulisa Oktiana[1*], Khairun Saddami[1], Rizki Novita[2], Rizkika Putri[3], Handika Rahayu[1], Hafidh Habibie[1] and Subhan Janura[1]

(Received: 19-Mar.-2024, Revised: 17-May-2024, Accepted: 6-Jul.-2024)

ABSTRACT

Tooth decay is a dental condition characterized by the deterioration of tooth tissue originating from the outer surface and progressing to the pulp. Severe tooth decay, evolving into cavities, necessitates timely intervention to avert more serious dental-health issues. Common treatment procedures include filling and extraction of affected teeth. Presently, dentists conduct examinations for tooth decay by manually tallying affected, missing and filled teeth using an odontogram—a human tooth code diagram. This data is then recorded in patients' dental medical records. Recognizing the need for automation in assessing patients' experiences of tooth decay, this research endeavors to develop a model capable of detecting decayed, missing and filled teeth using variations of the YOLOv5 and YOLOv8 model architectures. The results of the training evaluation demonstrate the efficacy of YOLOv5l with a learning rate of 10^{-2}, exhibiting a high precision value of 0.97, a recall of 0.858 and a mean average precision (mAP) of 0.904 within 1 hour and 18 minutes. According to the curves obtained in the training process, YOLOv5l shows great performance on the dental caries dataset, but precautions like early stopping are needed for a reliable and generalizable model. In contrast, YOLOv8 offers better training stability and larger variants perform better on the dental caries dataset, improving detection capabilities with continued training epochs.

KEYWORDS

Caries detection, Detection model, Deep learning, DMF-T, Tooth decay.

1. INTRODUCTION

Although teeth are often known as the strongest part of the human body, they possess a vulnerable inner layer called the dental pulp tissues [1], [2], [3]. This tissue is vulnerable to bacteria and traumas that can lead to several tooth diseases [2]-[3]. One of the common chronic dental diseases is dental caries [4]. It is a complex infectious oral disease that progressively and accumulatively infects hard dental tissue, resulting in teeth loss [5]-[6]. The untreated caries teeth can cause pain and over an extended period, can cause inflammation to develop leading to subsequent swelling [7]. Numerous epidemiological and clinical studies additionally have suggested that tooth loss, particularly due to dental caries, could potentially be a risk indicator for cardiovascular and cognitive disorders [8]-[9]. Dental caries commonly occurs in children and almost 100% of adults [10]. Based on the Basic Health Research of Indonesia (RISKESDAS), 45,3% of the Indonesian population experience dental and oral health problems, with notable prevalence associated with cavities and damaged teeth and dental caries accounts for 88% of the severity prevalence [11]-[12].

Commonly, dentists employ manual examination methods to evaluate dental caries in each patient in the dental clinic or hospital, involving the inspection of cavity count (for teeth decay), the number of missing teeth and the count of filled teeth. Subsequently, dentists manually record the position of the teeth infected with caries in a form called the odontogram, an instrument to record the dental status of a person recorded in visual format [13]. This examination of caries status within the population typically requires the computation of the Decayed, Missing and Filled Teeth, known as the DMF-T index, to serve the preventive, curative and rehabilitative care, as well as for determination of dental-

1. M. Fitria, Y. Elma, M. Oktiana, K. Saddami, H. Rahayu, H. Habibie, and S. Janura are with Department of Electrical and Computer Eng., Universitas Syiah Kuala, Banda Aceh, Indonesia. Emails: mayafitria@usk.ac.id, yasmin06@mhs.usk.ac.id, maulisaoktiana@usk.ac.id, khairun.saddami@usk.ac.id, handika@mhs.usk.ac.id, habibie19@mhs.usk.ac.id, subhan.j@mhs.usk.ac.id
2. R. Novita is with Faculty of Dentistry, Universitas Syiah Kuala, Banda Aceh, Indonesia. Email: drg_rizkinovita@usk.ac.id
3. R. Putri is with Polyclinic of Dental and Oral Medicine, Regional General Hospital dr. Zainoel Abidin (RSUZA), Banda Aceh, Indonesia. Email: rizkikaputri@gmail.com

health status in a community [14]-[15]. As reported in interviews with dentists from Regional Hospital Zainoel Abidin, Aceh, Indonesia, conducting the dental caries assessment through DMF-T in a population is a time- consuming process. This is due to the assessment that requires a meticulous inspection of each decayed, missing and restored tooth individually, followed by the record count and manual calculation of the DMF-T index based on the gathered information. In epidemiological research, dental status such as DMF-T status, is used to describe caries prevalence in a certain population [16]. These records are utilized in forensics science, as intra-oral information is important to determine characteristics of individuals or corpses through criminal investigation or other civil cases [13], [16]. However, the odontogram of the patient is not completely filled by the dentists due to being preoccupied with treating other patients or the odontogram forms being run out [17]. Thus, an automated dental caries detection system is preferable to tackle this problem.

Several scientific studies have been undertaken to identify and detect dental caries, particularly employing the Convolutional Neural Network algorithm. A study conducted by Baydakar et al. utilized the U-Net and VGG-16 techniques to detect the cavities in radiographic bitewing images, resulting in 48% detection accuracy [18]. A similar study on radiographic bitewing dental data was also carried out by Kumari et al. employing an image enhancement process using CLAHE and FOC-KCC and a training process using M-ResNet-RNN. However, assessing dental caries in a population is not feasible and the utilization of radiographic images for this purpose is discouraged due to the potential risks associated with X-ray exposure [19]. To minimize the use of X-rays in the automated detection of dental caries, Fitria et al. undertook a study utilizing dental clinical images for the detection process using CNN architecture [20]. The work employed five sides of dental clinical dataset images; namely, anterior, left buccal, right buccal, upper occlusal and lower occlusal. The development model was conducted on 1400 augmented images implementing ResNet-50 architecture. However, the performance of this model is considered sub-optimal, as the datasets exhibit significant variability and the missing and filled teeth also need to be taken into account in addition to the classification of caries and non-caries cases [21]. In addition, the inclusion of other parts in the dataset images, such as gums, normal teeth, lips and various anatomical features creates a challenge for the system in accurately identifying the cavity areas, resulting in a relatively lower accuracy.

This research proposed a baseline work to address the limitation of manual caries inspection conducted by dentists. Moreover, this research aims to overcome the disadvantages of the CNN model developed in [20] by developing a deep-learning model to identify the caries experience based on Decayed, Missing, Filled Teeth (DMF-T) using a popular object-detection model, You Only Look Once (YOLO) [22]. This object-detection model is considered capable of detecting multiple objects in an image by using the bounding box technique for the object [23]. Decayed, missing and filled teeth are the objects that are detected in the work. Two different versions of YOLO models were implemented in this work, namely YOLOv5 and YOLOv8, as both of them tend to provide higher accuracy than other versions [22], [24]-[25]. Moreover, YOLOv5 is chosen to be adopted, as the models usually yield a significant accuracy with unaffected model's real-time performance [26]. Conversely, as the latest version of YOLO models, YOLOv8 is selected for its advancement, manifesting in a new neural-network architecture succeeding YOLOv5 [27]. The model features an anchor-free detection head that simplifies the detection process and improves accuracy. The clinical dataset used in this work is the dental clinical images obtained by Fitria et al [20]. The results of this work are expected to be a basis for future research to contribute to the practical development of dental diagnostic tools and telemedicine applications.

2. METHOD

Figure 1 shows the procedure conducted in this work, which involves four stages; 1) Problem analysis; 2) Dataset pre-processing; 3) Model development and verification; and 4) Result analysis. The procedure will be discussed in detail in the sub-sections below.

2.1 Dataset and Pre-processing

The datasets employed in this research were sourced from the dataset utilized by Fitria et al. in [20], gathered from the Dental and Oral Polyclinic of Regional Hospital Zainoel Abidin Banda Aceh, Indonesia. The dataset consists of 350 images identified in the caries class. However, only 294 of the caries images were selected, based on the considerations of light intensity, object clarity and image

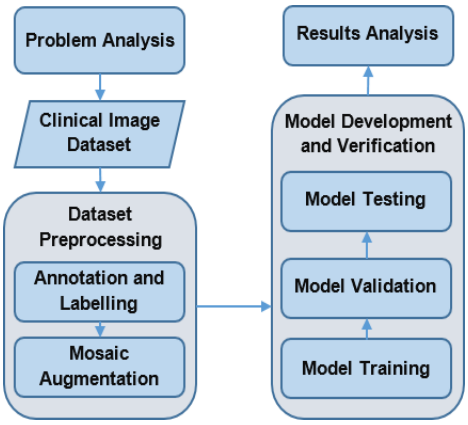


Figure 1. Research methodology.

sharpness, aiming to provide a strong feature representation. The removed dataset is considered weak in clarity, as some of the teeth are not exposed enough in the pictures and there is less light intensity and less sharpness, as the pictures contain more shadows, especially in the hindmost teeth. The dataset taken into account is discussed and chosen involving the dentists.

Figure 2 exhibits the sample of five sides of teeth used as datasets in this work; namely, labial (a), right buccal (b), left buccal (c), lower occlusal (d) and upper occlusal (e). Out of 294 images taken into consideration, there are 70 labial images, 58 right buccal images, 58 left buccal images, 52 lower occlusal images and 56 upper occlusal images. Moreover, the filled teeth shown in the images show temporary fillings of the patients.

(a) (b) (c) (d) (e)

Figure 2. A sample of five sides of teeth used as the dataset.

The pre-processing stage was conducted in this work to obtain optimized images for model training. The processes involved annotation, labeling and the mosaic augmentation process. The annotation and labeling process was carried out using the Roboflow image-labeling framework [28]. Dental image datasets were uploaded into Roboflow, where the annotation process involved placing bounding boxes on the images. The annotation process involved all authors of this work in collaboration with three dental experts and the annotated images were calibrated within the annotators. Subsequently, the annotated images were labeled according to their respective classes; decayed teeth were labeled as a red box, missing teeth were labeled as a yellow box and filled teeth were labeled as a blue box. For instance, there are two labels in the left image of Figure 3, decayed and missing teeth box, respectively and two boxes for decayed and filled teeth appear in the right image of Figure 3.

To enhance the diversity of the datasets and to make the model more robust, mosaic augmentation was established in the datasets. This type of augmentation involves the replacement and transformation of a certain area in the image with mosaic, such as breaking the images into small blocks [29]-[30]. This augmentation method is employed to assess the detection precision across various image dimensions that resemble the original image with a lower objective lens [31]. The reason for employing mosaic augmentation is to assess the detection precision across various image dimensions resembling those captured with a less powerful lens. Figure 4 illustrates the images that have already been augmented by implementing the mosaic augmentation technique. In total, there are 708 augmented images in the datasets, distributed among training, validation and testing data with a ratio of 88:8:4, respectively. Table 1 represents the distribution of data before and after augmentation and the allocation of training,

validation and testing data.

Figure 3. Bounding-box technique and image annotation in the datasets.

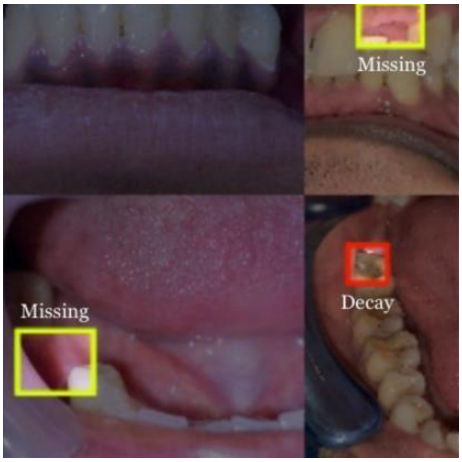


Figure 4. Mosaic augmentation.

Table 1. Distribution of datasets.

Dataset	D	M	F	Total	Augmented Dataset
Training	71	120	103	207	621
Validation	25	33	22	58	58
Testing	10	15	13	29	29
Total	106	168	148	294	708

2.2 Model Development and Verification

The model was developed to detect the caries condition; decayed, missing and filled teeth. The training process was carried out by implementing different variants of YOLOv5 and YOLOv8. The YOLOv5 network is divided into three parts; the backbone for the feature extraction on an input image using Cross-Stage Partial Network (CSPNet); the neck component for refining extracted features from the backbone; and the detection head for object detection using Feature Pyramid Network (FPN) [32]-[33]. YOLOv5 provides five different versions; namely, YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x, where every version provides different trade-offs in terms of calculation speed, average precision and the depth of channel [33]-[34]. Figure 5 shows the scales of corresponding versions of YOLOv5. The higher the versions, the larger and more accurate the models become, yet the slower the calculation speed [33]. Thus, this work selected the first versions only, which are YOLOv5n, YOLOv5s, YOLOv5m and YOLOv5l to avoid a long running time during the training process. Similar to YOLOv5, YOLOv8 also consists of a backbone network, a neck segment and a detection head [35]. However, YOLOv8 employs FPN for feature extraction in the backbone part and Cross-Layer Connection (CLC) in the neck [35]. The YOLOv8 versions specifically chosen in this experiment were YOLOv8n, YOLOv8s, YOLOv8m and YOLOv8l to prevent prolonged computation

time. YOLOv8 incorporates features that make it a highly precise object detector, particularly through the use of an anchor-free detection head [36]-[37]. This approach simplifies the architecture of the model and improves its accuracy in predicting object locations. This enhancement is especially advantageous for datasets containing objects of various shapes and sizes.

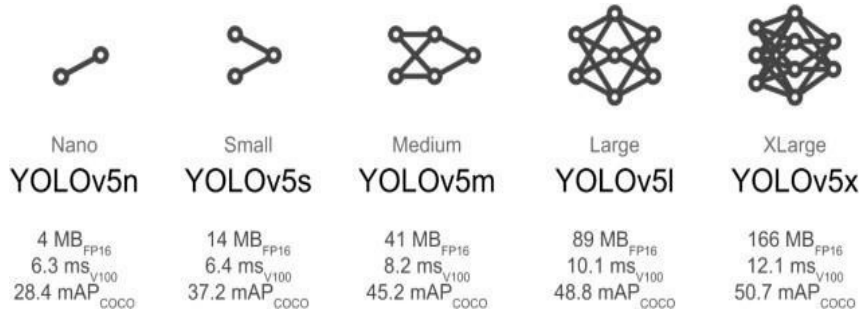


Figure 5. Variants of YOLOv5 [31].

The shift to an anchor-free detection head, away from the anchor box method used in previous YOLO versions, streamlines the detection process and boosts accuracy [36]. The variants of YOLOv8 can be seen in Figure 6. Considering the time constraints during the experiment, the hyper-parameters reported in this work, such as epoch, batch size, optimizer, momentum and learning rates, were set as shown in Table 2.

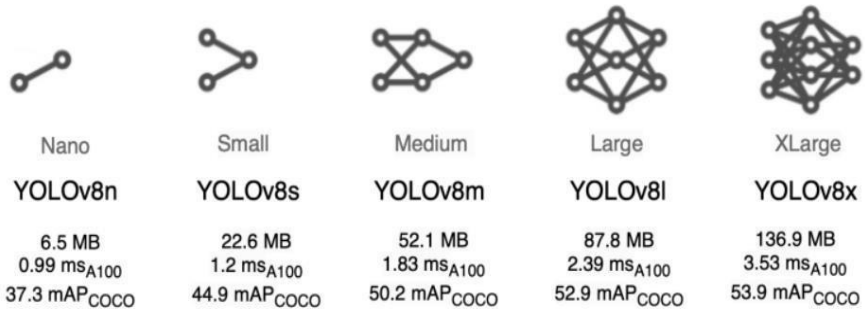


Figure 6. Variants of YOLOv8.

Table 2. Hyper-parameters for model development.

Hyper-parameters of YOLOv5 and YOLOv8		
Hyper-parameter	YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l	YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l
Epoch	400	100
Batch Size	16	16
Optimizer	SGD	SGD
Momentum	0.9	0.9
Learning Rate	10^{-2}, 10^{-3}, 10^{-4}	10^{-2}, 10^{-3}, 10^{-4}

The model-performance evaluation involves the analysis of the precision, recall and mean average precision (mAP) obtained by the models. Precision and recall serve as common metrics used for the evaluation of detection and classification models. Precision measures the accuracy of the model in identifying the positive class, while recall assesses how successfully the model identifies images of the positive class. In this work, precision is used to examine the accuracy of the model in identifying a dental caries image containing decayed, missing or filled teeth. Additionally, recall assesses the number of images containing caries that were correctly identified by the model.

The precision is computed by taking the ratio of true positive (TP) to the total predictions belonging to a positive class, as per Formula 1. On the other hand, recall is defined as the ratio of true positive (TP) to all predicted results, following Formula 2. True positive (TP) represents the number of accurately classified data as a positive class, while false positive (FP) denotes the number of incorrectly classified data as a positive class. In addition, true negative (TN) corresponds to the number of correctly classified data as a negative class. In contrast, false negative (FN) is the count of incorrectly classified

data as a negative class. The precision and recall values increase with a higher TP count as well as with lower FP or FN values.

$$recall = \frac{TP}{TP+FN} \quad (1)$$

$$precision = \frac{TP}{TP+FP} \quad (2)$$

In the object-detection model, model output is not solely confined to the object class; it also includes additional outputs, such as bounding-box annotation for the detected object. Consequently, Mean Average Precision (mAP) is employed in this study, evaluating the average precision for decision values ranging from 0 to 1. The calculation of mAP, as depicted in Formula 3, involves N which represents the number of average precision (AP).

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (3)$$

Intersection over Union (IoU) assesses the overlap between the predicted bounding box and the ground truth bounding box. If the IoU exceeds 0.5, the value is considered True Positive; conversely, it is considered False Positive if the IoU is less than 0.5. To yield the Inter-precision for APi, recall maximum needs to be taken into account by using Formula 4.

$$AP_i = maxAP_i \quad (4)$$

Suppose that the automated system for detecting decayed, missing and filled teeth (DMFT) in dental clinical images is tested on a dataset containing 100 images and assume that there are 50 images with actual DMFT cases and 50 images without any DMF-T. Now, we consider the following hypothetical results:

- True positive (TP): The system correctly identifies 40 images with DMF-T.
- False positive (FP): The system incorrectly identifies 5 images without DMF-T as having DMF- T.
- False negative (FN): The system misses 10 images with actual DMF-T cases.

Using these results, now Precision, Recall and mAP are calculated as below:

- Precision = TP / (TP + FP) = 40 / (40 + 5) = 0.889.
- Recall = TP / (TP + FN) = 40 / (40 + 10) = 0.8.

To calculate the mAP, the average precision for each image in the dataset needs to be computed, which involves ranking the predicted DMF-T cases by their confidence scores. For the sake of brevity, it is assumed that there is an average precision of 0.85 for this example.

- Mean Average Precision (mAP) = Average precision across all images = 0.85.

In this example, our system achieved a precision of 0.889, indicating that 88.9% of the positive predictions were correct. The recall of 0.8 demonstrates that the system correctly identified 80% of the actual DMF-T cases. The mAP of 0.85 suggests that, on average, our model's predictions are highly accurate across all images in the dataset.

3. RESULTS AND DISCUSSION

3.1 Results

Tables 3 and 4 and Figures 7 and 8 show the training results of different versions of YOLOv5 and YOLOv8, respectively. It can be seen in Table 3 and Figure 7 that the smaller the learning rate tuned, the smaller the precision and recall obtained, resulting in a smaller mAP value of the model. Table 3 also indicates the increasing mAP value in every newer version of YOLOv5. However, the computation speed exhibited in Table 3 is inverse to the mAP value. The newer version of YOLOv5 employed, the longer the training time consumed. The mAP value yielded by the YOLOv5l version set with the learning rate of 10^{-2} is highlighted as the highest training result, outperforming the other versions with a mAP value of 90.4%, followed by YOLOv5s with a slightly different mAP value of 90.2%. Nevertheless, the calculation time of YOLOv5l tends to be longer than that of YOLOv5s, consuming one hour, 18 minutes and 42 seconds of training time, while YOLOv5s takes only 37 minutes and 28 seconds, which is the fastest running time amongst all models. A significant drop in mAP is also

obtained in YOLOv5n, YOLOv5s and YOLOv5m tuned with a learning rate of 10^{-4}, where the mAP values are 36.9%, 47.5% and 56%, respectively.

Table 3. Result of YOLOv5.

Architecture	Learning Rate	Precision	Recall	mAP	Time
YOLOv5n	10^{-2}	0.954	0.803	0.878	45.24 m, s
	10^{-3}	0.831	0.712	0.772	47.46 m, s
	10^{-4}	0.483	0.313	0.369	45.36 m, s
YOLOv5s	10^{-2}	0.97	0.861	0.902	37.28 m, s
	10^{-3}	0.931	0.729	0.811	47.58 m, s
	10^{-4}	0.621	0.435	0.475	48.04 m, s
YOLOv5m	10^{-2}	0.971	0.856	0.895	40.05 m, s
	10^{-3}	0.90	0.785	0.841	01.04.12 h, m, s
	10^{-4}	0.788	0.485	0.56	01.00.24 h, m, s
YOLOv5l	10^{-2}	0.97	0.858	0.904	01.18.42 h, m, s
	10^{-3}	0.941	0.888	0.835	01.34.02 h, m, s
	10^{-4}	0.793	0.547	0.641	01.45.04 h, m, s

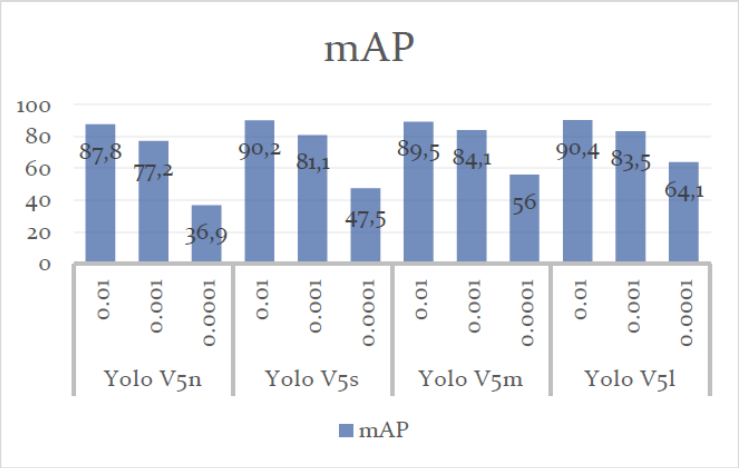


Figure 7. mAP value of different versions of YOLOv5 adjusted with different learning rates.

Similar to the YOLOv5, the four versions of YOLOv8 yielded a smaller mAP values as the learning rate decreases (Figure 8). Based on Tabel 4, the highest mAP value is received by YOLOv8m with a learning rate of 10^{-2} with a 90.6% mAP value, surpassing other models. This value is followed by YOLOv8l, YOLOv8n and YOLOv8s, yielding mAP values of 89.7%, 87.8% and 87.1%.

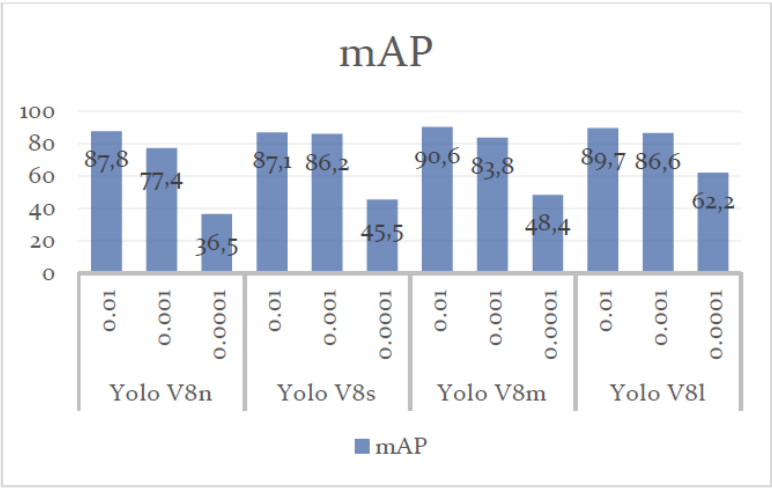


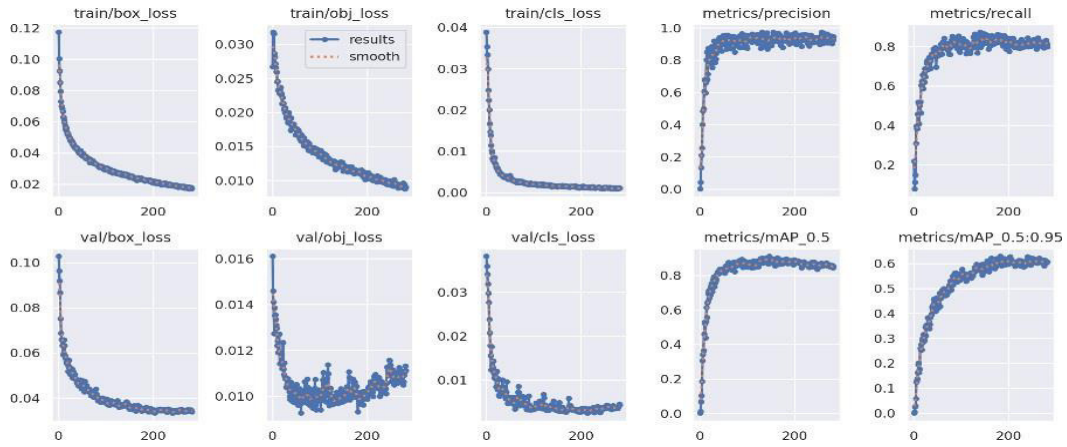
Figure 8. mAP value of different versions of YOLOv8 adjusted with different learning rates.

A notable decrease of mAP value also appears in every variant of the YOLOv8 set with a learning rate of 10^{-4}. Despite the similarity in the behavior in both YOLOv5 and YOLOv8, the training time of YOLOv8 tends to be longer than that of YOLOv5, with the fastest time of 4 hours, 15 minutes and 18 seconds and the slowest time of 5 hours, 32 minutes and 52 seconds, while the fastest computation time of YOLOv5 is 37 minutes and 28 seconds by YOLOv5s and the slowest time of one hour, 45 minutes and 4 seconds by YOLOv5l.

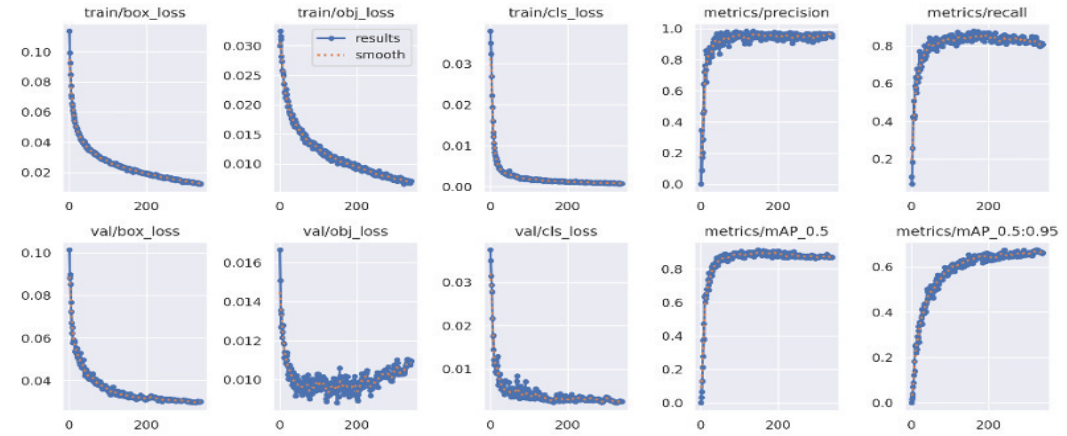
Table 4. Results of YOLOv8.

Architecture	Learning Rate	Precision	Recall	mAP	Time
YOLOv8n	10^{-2}	0.954	0.803	0.878	04.15.18 h, m, s
	10^{-3}	0.946	0.685	0.774	04.19.08 h, m, s
	10^{-4}	0.427	0.417	0.365	04.20.34 h, m, s
YOLOv8s	10^{-2}	0.955	0.811	0.871	04.55.06 h, m, s
	10^{-3}	0.956	0.778	0.862	04.22.32 h, m, s
	10^{-4}	0.573	0.454	0.454	04.25.28 h, m, s
YOLOv8m	10^{-2}	0.954	0.846	0.906	04.57.30 h, m, s
	10^{-3}	0.945	0.79	0.838	05.02.04 h, m, s
	10^{-4}	0.668	0.457	0.484	05.06.06 h, m, s
YOLOv8l	10^{-2}	0.953	0.839	0.897	05.20.16 h, m, s
	10^{-3}	0.986	0.806	0.866	05.21.04 h, m, s
	10^{-4}	0.751	0.518	0.622	05.32.52 h, m, s

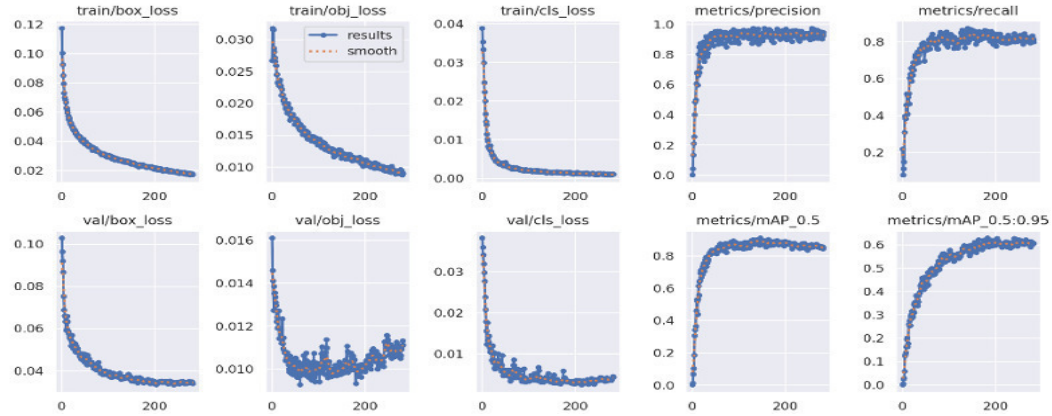
Figure 9 presents a comparison of training results across various versions of YOLOv5 on a dental caries dataset. It's evident from the graph that training with YOLOv5 produces the most effective model. The training curve highlights the outstanding performance of YOLOv5l. Based on Figure 10, YOLOv8 presents a different aspect compared to YOLOv5. YOLOv8 generates a more stable training graph than YOLOv5.



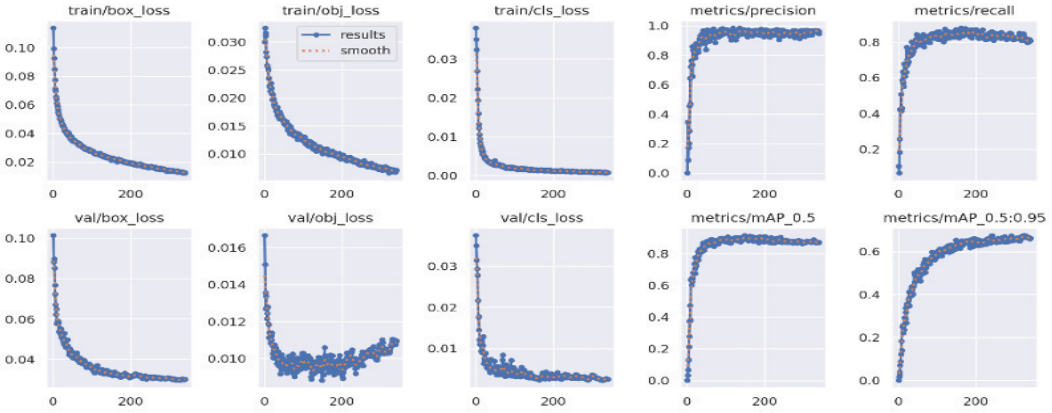
(a) YOLOv5n



(b) YOLOv5

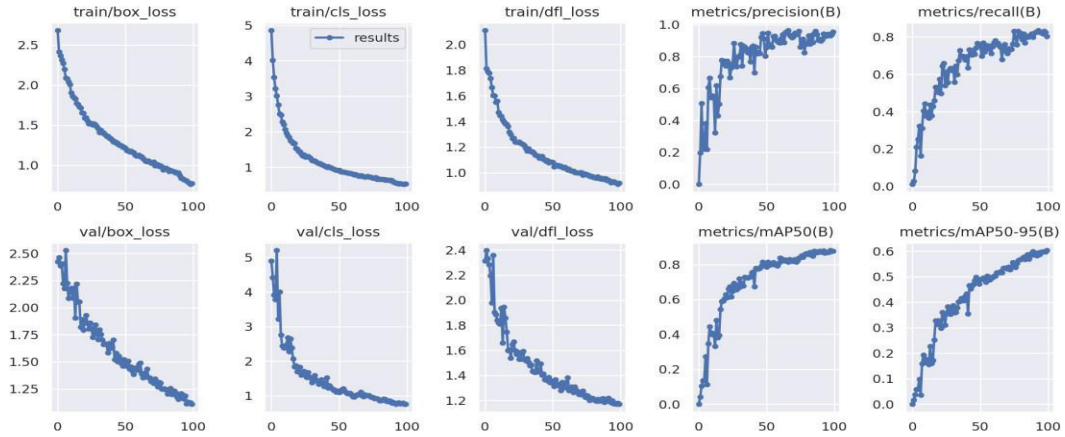


(c) YOLOv5m

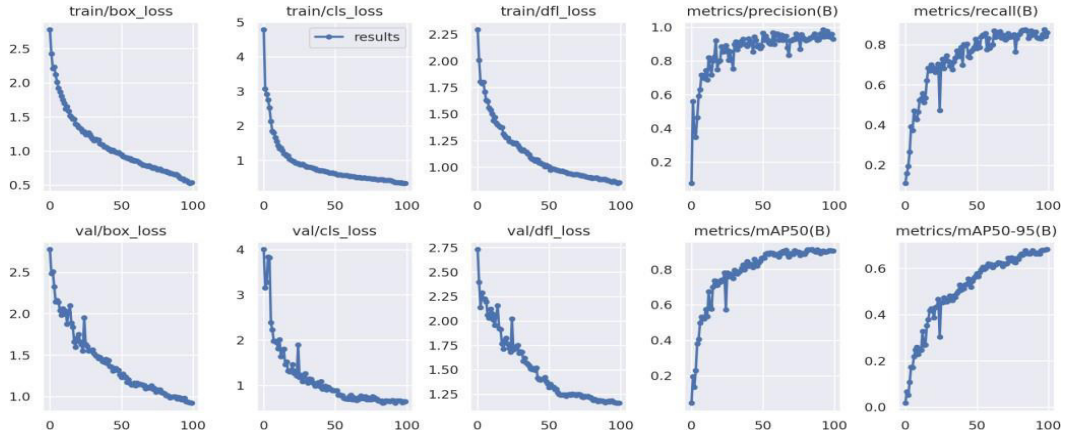


(d) YOLOv5l

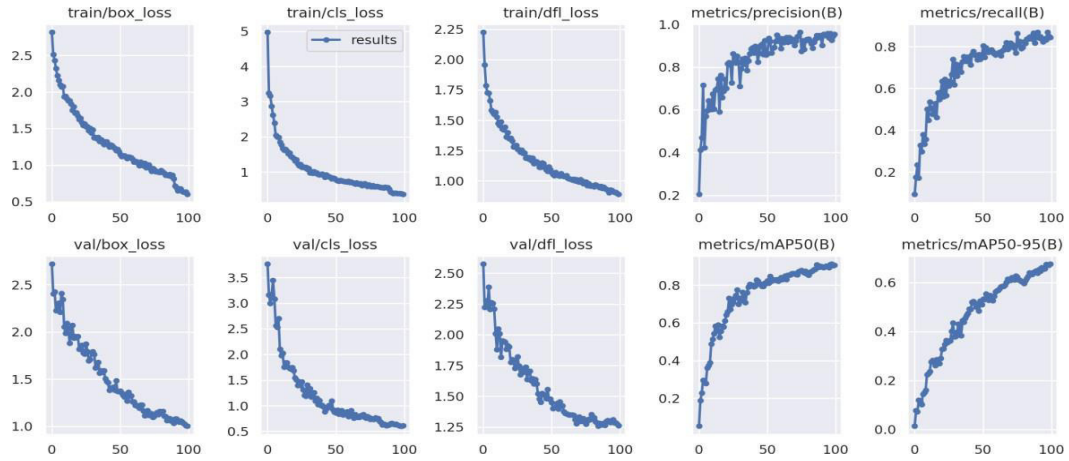
Figure 9. Training curves of YOLOv5.



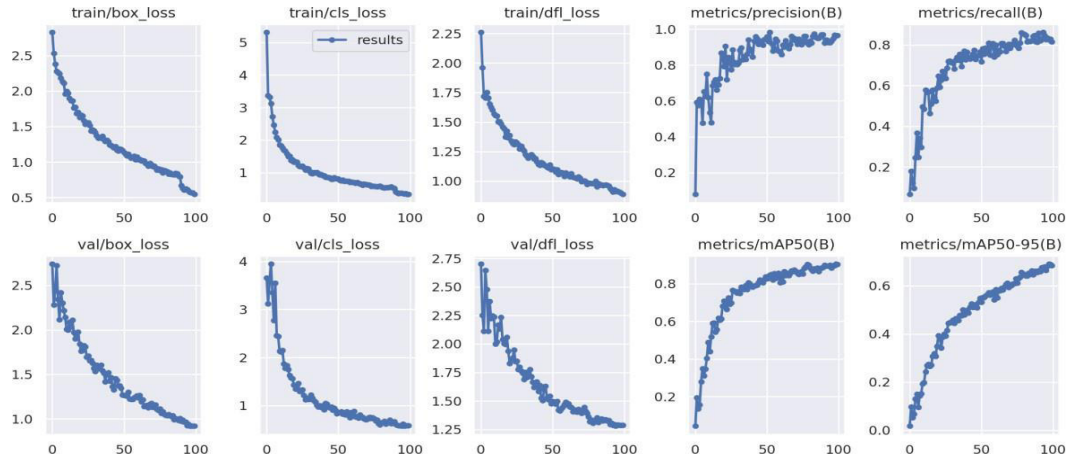
(a) YOLOv8n



(b) YOLOv8s



(c) YOLOv8m



(d) YOLOv8l

Figure 10. Training curves of YOLOv8.

Figures 11 and 12 depict the model detection of YOLOv5 and YOLOv8 variants on a sample dental clinical image, adjusted with a learning rate of 10^{-2}. It is shown that YOLOv5l detects the caries indications better than YOLOv8s with accuracies above 90% detected as decayed teeth (labeled as Karies) and filled teeth (labeled as Tambal). On the other hand, there is a filled tooth on the image detected by YOLO8s yielding an accuracy of 51%.

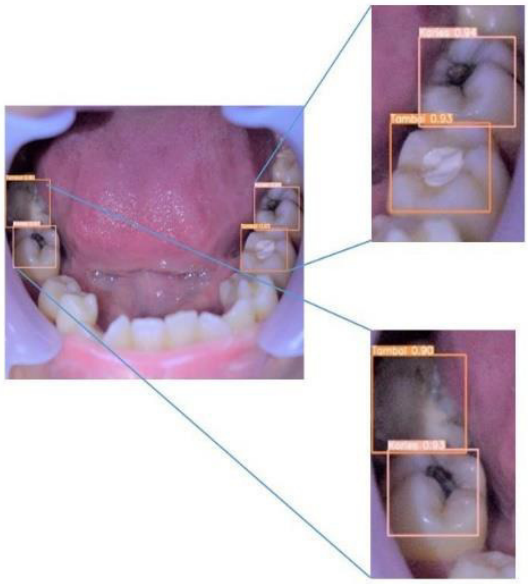


Figure 11. Model detection results of YOLOv5l with the learning rate of 10^{-2}.

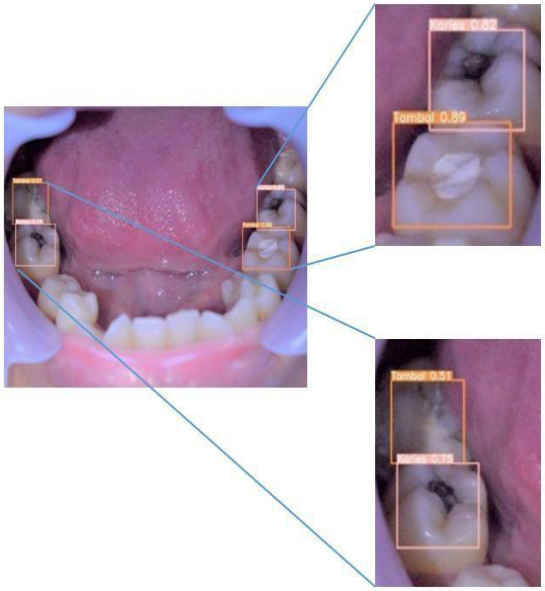


Figure 12. Model detection results of YOLOv8m with the learning rate of 10^{-2}.

Figure 13 visualizes the confusion matrix of the YOLOv5l model tested on 29 dental images. It can be seen that 87% of the missing teeth are correctly detected, while 77% of the decayed class were correctly detected and 93% of the filled teeth were correctly detected. This indicates that the model yielded robust performance and is effectively identifying the correct DMF-T in practical scenarios.

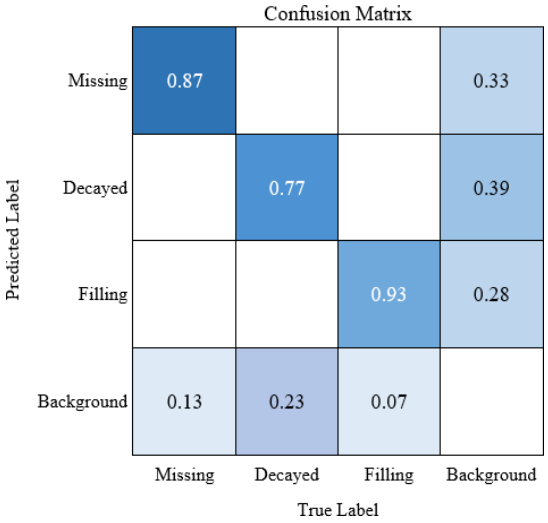


Figure 13. Confusion matrix of YOLOv5l with a learning rate of 10^{-2}.

3.2 Discussion

According to Tables 3 and 4, it is shown that the differences in mAP values and training times can indeed impact the practical applicability of the models in real-world scenarios. Higher mAP values generally indicate better model performance, which can lead to more accurate detection of decayed, missing and filled teeth (DMF-T) in dental clinical images. However, models with higher mAP values may also have longer training times and require more computational resources, which can be a concern in resource- limited settings or when deploying the model on edge devices. On the other hand, models with faster training times and lower computational requirements may have slightly lower mAP values, but may be more suitable for real-world applications, where efficiency and resource constraints are critical factors. In such cases, the trade-off between model accuracy and computation time should be carefully considered based on the specific requirements of the target environment and application. In this study, we have aimed to strike a balance between model accuracy and computation time by selecting the YOLOv5 and YOLOv8 versions that offer a reasonable compromise between these

factors. However, it is acknowledged that the optimal choice may vary depending on the specific use case and available resources.

Based on Figure 9, YOLOv5l yielded the most outstanding training curve. However, validation results for this model indicate overfitting occurring after surpassing 100 epochs, a trend also observed with the YOLOv5 medium model. Despite this, YOLOv5l still stands out as the superior option overall, although it necessitates the implementation of early stopping in order to prevent overfitting. Early stopping can prevent validation loss from increasing by halting the training process if the validation loss stops decreasing or starts to increase, which is a sign of overfitting. Overfitting, a common issue in machine learning, occurs when a model learns noise or irrelevant patterns from the training data, thereby hindering its ability to generalize well to unseen data. Early stopping serves as a regularization technique to mitigate overfitting by halting the training process when the model's performance on a validation dataset starts to decline. By doing so, it ensures that the model maintains its ability to generalize effectively.

Compared to YOLOv5, YOLOv8 in Figure 10 generated more stable training curves. The larger the size of YOLOv8, the better the model learns about the dental caries dataset. The same trend is also observed in the number of training iterations. The more epochs utilized, the closer the model gets to convergence. It's evident that YOLOv8 large yields the best-performing model. The stability of the training graph in YOLOv8 suggests improved training dynamics and possibly better handling of the dataset's complexities. Additionally, the correlation between model size and performance indicates that larger YOLOv8 variants are more adept at capturing intricate features within the dental caries dataset. Moreover, the convergence of the model with increased epochs implies a continuous improvement in learning, leading to enhanced model accuracy.

In summary, while YOLOv5l demonstrates exceptional performance on the dental caries dataset, precautions such as early stopping are necessary to ensure the model's reliability and generalization capabilities. This approach would lead to the development of a robust model capable of accurately detecting dental caries in real-world applications. However, YOLOv8 showcases superior stability in training, with larger variants demonstrating enhanced performance on the dental caries dataset. The convergence of the model with more training epochs signifies the ongoing refinement of the model's understanding, ultimately resulting in improved detection capabilities.YOLOv8 provided new features that improved its detection capabilities, increasing both accuracy and efficiency. This model particularly excels in segmentation tasks, offering precise segmentation and classification of various image parts, making it highly effective for diverse applications, like medical imaging and autonomous vehicle navigation.

4. CONCLUSION

This research proposed a comparison of deep learning-based models for dental caries detection, characterized by decayed, missing and filled teeth. Two different object detection architectures are implemented in this work; YOLOv5 and YOLOv8, including their variants. The models' performances are compared by calculating their precision, recall and mAP values. The results show that YOLOv5l and YOLOv8m outperformed other variants with the mAP values of 90.4% and 90.6%, respectively. However, the computational time required by YOLOv8m is considered extensive; namely, around four hours 57 minutes 30 seconds, while YOLOv5s takes only one hour, 18 minutes and 42 seconds. The YOLOv8 annotation format is unique and requires precise detailing of objects in images, usually using bounding boxes and labels. Preparing a dataset for YOLOv8 involves annotating numerous images, which can be a time-consuming task. The quality and accuracy of these annotations directly influence the model's ability to learn and make precise predictions. To further enhance performance, the YOLOv8 model should be integrated into the existing image-enhancement process.

ACKNOWLEDGMENTS

This research was supported by Lembaga Penelitian dan Pengabdian Masyarakat Universitas Syiah Kuala. In addition, we thank the dentists and nurses from Polyclinic of Dental and Oral Medicine, Regional General Hospital dr. Zainoel Abidin (RSUZA), who helped and assisted the authors throughout the process of dataset-gathering.

REFERENCES

[1] S. Prabhu, A. B. Acharya and M. V. Muddapur, "Are Teeth Useful in Estimating Stature?," Journal of Forensic and Legal Medicine, vol. 20, no. 5, pp. 460–464, DOI: 10.1016/j.jflm.2013.02.004, 2013.

[2] G. T. Huang, "Pulp and Dentin Tissue Engineering and Regeneration: Current Progress," Regenerative Medicine, vol. 4, no. 5, pp. 697–707, DOI: 10.2217/rme.09.45, 2009.

[3] X. Huang et al., "Microenvironment Influences Odontogenic Mesenchymal Stem Cells Mediated Dental Pulp Regeneration," Frontiers in Physiology, vol. 12, p. 656588, 2021.

[4] L. Cheng et al., "Expert Consensus on Dental Caries Management," Int. Journal of Oral Science, vol. 14, no. 1, p. 17, 2022.

[5] H. Nilsson, J. Sanmartin Berglund and S. Renvert, "Longitudinal Evaluation of Periodontitis and Tooth Loss among Older Adults," Journal of Clinical Periodontology, vol. 46, no. 10, pp. 1041–1049, 2019.

[6] M. J. Y. Yon, S. S. Gao, K. J. Chen, D. Duangthip, E. C. M. Lo and C. H. Chu, "Medical Model in Caries Management," Dentistry Journal, vol. 7, no. 2, p. 37, 2019.

[7] T. Kikuiri, K. Saito, A. Iida, Y. Yoshimura, Y. Yawaka and T. Shirakawa, "Occurrence of Subcutaneous Emphysema during a Caries Filling Procedure: A Case Report," Pediatric Dental Journal, vol. 32, no. 3, pp. 211–215, 2022.

[8] M. Zhou, J. Dong, L. Zha and Y. Liao, "Causal Association between Periodontal Diseases and Cardiovascular Diseases," Genes, vol. 13, no. 1, p. 13, 2021.

[9] J.-H. Lee, D.-H. Kim, S.-N. Jeong and S.-H. Choi, "Detection and Diagnosis of Dental Caries Using a Deep Learning-based Convolutional Neural Network Algorithm," Journal of Dentistry, vol. 77, pp. 106–111, 2018.

[10] M. Rathee and A. Sapra, "Dental Caries," [Online], Available: https://www.ncbi.nlm.nih.gov/books/NBK51699/, 2019.

[11] Balitbangkes, "National Report of Basic Health Research 2018 (in Indonesian)," Indonesian Ministry of Health, p. 206, 2019. [Online], Available: https://repository.badankebijakan.kemkes.go.id/id/eprint/3514/1/Laporan%20Riskesdas%202018%20Nasional.pdf, [Accessed: February 2024].

[12] N. D. Ardiyanti, R. Adhani and I. Hatta, "Correlation between DMF-T Caries Index, Consumption of Drinking Water and Tooth Brushing Behavior in Indonesian Communities (in Indonesian)," Dentin, vol. 6, no. 1, 2022.

[13] E. S. Wardhana, S. Suryono, A. Hernawan and L. E., Nugroho, "Evaluation of Format and Security of Dental Electronic Medical Record Systems in General Hospital Based on Legislation," Odonto: Dental Journal, vol. 9, Special Issue 1, pp. 80-89, 2022.

[14] G. Moradi, A. M. Bolbanabad, A. Moinafshar, H. Adabi, M. Sharafi and B. Zareie, "Evaluation of Oral Health Status Based on the Decayed, Missing and Filled Teeth (DMFT) Index," Iranian Journal of Public Health, vol. 48, no. 11, p. 2050, 2019.

[15] A. Alami, S. Erfanpoor, E. Lael-Monfared, A. Ramezani and A. Jafari, "Investigation of Dental Caries Prevalence, Decayed, Missing and Filled Teeth (DMF-T and DMF-T Indices) and the Associated Factors among 9-11 Years Old Children," Research Square, pp. 1-18, DOI: 10.21203/rs.2.21545/v1, 2020.

[16] J. A. Daza-Cardona, J. Vargas-Ramírez and M. A. Guapacha-Sánchez, "Doing Odontograms and Dentists in the Classroom. Materiality and Affect in Dental Education," Tapuya: Latin American Science, Technology and Society, vol. 4, no. 1, p. 1968635, 2021.

[17] E. D. Fadhillah et al., "Smart Odontogram: Dental Diagnosis of Patients Using Deep Learning," Proc. of 2021 IEEE Int. Electronics Symposium (IES), pp. 532-537, DOI: 10.1109/IES53407.2021.9594027, Surabaya, Indonesia, 2021.

[18] I. S. Bayrakdar et al., "Deep-learning Approach for Caries Detection and Segmentation on Dental Bitewing Radiographs," Oral Radiology, vol. 38, no. 4, pp. 468-479, pp. 1–12, 2021.

[19] Y. Al-Hadeethi, M. Sayyed, H. Mohammed and L. Rimondini, "X-ray Photons Attenuation Characteristics for Two Tellurite Based Glass Systems at Dental Diagnostic Energies," Ceramics International, vol. 46, no. 1, pp. 251–257, 2020.

[20] M. Fitria et al., "Development of Intraoral Clinical Image Dataset for Deep Learning Caries Detection," Proc. of the 2023 IEEE 2nd Int. Conf. on Computer System, Information Technology and Electrical Engineering (COSITE), pp. 194–198, DOI: 10.1109/COSITE60233.2023.10249428, Banda Aceh, Indonesia, 2023.

[21] L. Que et al., "Prevalence of Dental Caries in the First Permanent Molar and Associated Risk Factors among Sixth-grade Students in São Tomé Island," BMC Oral Health, vol. 21, no. 1, pp. 1–10, 2021.

[22] K. Kim, K. Kim and S. Jeong, "Application of YOLO v5 and v8 for Recognition of Safety Risk Factors at Construction Sites," Sustainability, vol. 15, no. 20, p. 15179, 2023.

[23] T. Diwan, G. Anirudh and J. V. Tembhurne, "Object Detection Using YOLO: Challenges, Architectural Successors, Datasets and Applications," Multimedia Tools and Applications, vol. 82, no. 6, pp. 9243–9275, 2023.

[24] V. D. Matta et al., "Single Use Plastic Bottle Recognition and Classification Using Yolo V5 and V8 Architectures," Proc. of the Int. Conf. on Cognitive Computing and Cyber Physical Systems, Part of the Book Series: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 537, pp. 99– 106, Springer, 2023.

[25] E. Casas, L. Ramos, E. Bendek and F. Rivas-Echeverría, "Assessing the Effectiveness of YOLO Architectures for Smoke and Wildfire Detection," IEEE Access, vol. 11, pp. 96554 – 96583, DOI: 10.1109/ACCESS.2023.3312217, 2023.

[26] J. Terven, D.-M. Córdova-Esparza and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-nas," Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1680–1716, 2023.

[27] I. P. Sary, S. Andromeda and E. U. Armin, "Performance Comparison of YOLOv5 and YOLOv8 Architectures in Human Detection using Aerial Images," Ultima Computing: Jurnal Sistem Komputer, vol. 15, no. 1, pp. 8–13, 2023.

[28] Q. Lin, G. Ye, J. Wang and H. Liu, "RoboFlow: A Data-centric Workflow Management System for Developing AI-enhanced Robots," Proc. of the 5th Conf. on Robot Learning, vol. 164, pp. 1789–1794, [Online], Available: https://proceedings.mlr.press/v164/lin22c.html, PMLR, 2022.

[29] Y. Lee, J. Choi and K. Jo, "VSNet: Vehicle State Classification for Drone Image with Mosaic Augmentation and Soft-label Assignment," Proc. of the Asian Conference on Intelligent Information and Database Systems, Part of the Book Series: Lecture Notes in Computer Science, vol. 13995, pp. 109–120, 2023.

[30] F. Dadboud, V. Patel, V. Mehta, M. Bolic and I. Mantegh, "Single-stage UAV Detection and Classification with YOLOv5: Mosaic Data Augmentation and PANet," Proc. of the 2021 17th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8, DOI: 10.1109/AVSS52988.2021.9663841, Washington, USA, 2021.

[31] T. B. Pun, A. Neupane, R. Koech and K. Walsh, "Detection and Counting of Root-knot Nematodes Using YOLO Models with Mosaic Augmentation," Biosensors and Bioelectronics: X, vol. 15, p.100407, DOI: 10.1016/j.biosx.2023.100407, 2023.

[32] Q. Song et al., "Object Detection Method for Grasping Robot Based on Improved YOLOv5," Micro-machines, vol. 12, no. 11, p. 1273, 2021.

[33] L. Wang et al., "Investigation into Recognition Algorithm of Helmet Violation Based on YOLOv5-CBAM- DCN," IEEE Access, vol. 10, pp. 60622–60632, 2022.

[34] W. Sheng et al., "Symmetry-based Fusion Algorithm for Bone Age Detection with YOLOv5 and ResNet34," Symmetry, vol. 15, no. 7, p. 1377, 2023.

[35] B. Selcuk and T. Serif, "A Comparison of YOLOv5 and YOLOv8 in the Context of Mobile UI Detection," Proc. of the Int. Conf. on Mobile Web and Intelligent Information Systems, Part of the Book Series: Lecture Notes in Computer Science, vol. 13977, pp. 161– 174, Springer, 2023.

[36] G. Wen, M. Li, Y. Luo, C. Shi and Y. Tan, "The Improved YOLOv8 Algorithm Based on EMSPConv and SPE-head Modules," Multimedia Tools and Applications, vol. 83, pp. 61007–61023, DOI: 10.1007/s11042-023-17957-4, 2024.

[37] H. Wang, C. Liu, Y. Cai, L. Chen and Y. Li, "YOLOv8-QSD: An Improved Small Object Detection Algorithm for Autonomous Vehicles Based on YOLOv8," IEEE Transactions on Instrumentation and Measurement, vol. 73, DOI: 10.1109/TIM.2024.3379090, 2024.

ملخص البحث:

تآكل الأسنان وضع صحّي يصيب الأسنان ويتميّز بتآكل نسيج الأسنان، حيث يبدأ من السّطح الخارجي للسّنّ ويتطوّر إلى أن يصل إلى لُبّ السّنّ. ويتطلّب التّآكل الشّديد للأسنان تدخّلاً في الوقت المناسب للحيلولة دون قضايا أشدّ خطراً. وتشمل إجراءات العلاج الشّائعة الحشو والاستئصال للأسنان المتضرّرة والمفقودة والمحشوّة باستخدام مخطّط رموز يفصّل الحالة السّنّية لكلّ مريض. ويتمّ تسجيل المعلومات ذات العلاقة في السّجلّات الطّبية للمرضى.

يسعى هذا البحث إلى تطوير نموذجٍ قادرٍ على الكشْف عن الأسنان المتآكلة والمفقودة والمحشوّة باستخدام البنى المعروفة باسم (يولو 5) و (يولو 8) بأشكالها المختلفة. وقد بيّنت النّتائج فعالية النّموذج (يولو 5)/البنية (l) بمعدل تعلّم مقداره (10^{-2})، محققاً دقّةً عاليةً في الكشْف بلغت 97%، ومعدّل استرجاعٍ قدرُهُ (0.858)، ومتوسط دقة كشفٍ قدره (0.904) في غضون زمنٍ مقداره ساعة واحدة و 18 دقيقة.

وبناءً على المنحنيات التي تمّ الحصول عليها في عملية التّدريب، حقّق النّموذج يولو 5/ البنية (l) أداءً عالياً عند تطبيقه على مجموعة بيانات تآكل الأسنان، لكن مع احتياطاتٍ مثل التّوقّف المبكّر ليكون النّموذج أكثر موثوقية وقابلية للتّعميم . بالمقابل، يوفّر نموذج يولو 8 استقرارية تدريب أفضل، وتعمل الأشكال الأكبر منه بشكل أفضل على مجموعة بيانات تآكل الأسنان.

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) مجلة علمية عالمية متخصصة محكمة تنشر الأوراق البحثية الأصيلة عالية المستوى في جميع الجوانب والتقنيات المتعلقة بمجالات تكنولوجيا وهندسة الحاسوب والاتصالات وتكنولوجيا المعلومات.

تحتضن وتنشر جامعة الأميرة سمية للتكنولوجيا (PSUT) المجلة الأردنية للحاسوب وتكنولوجيا المعلومات، وهي تصدر بدعم من صندوق دعم البحث العلمي في الأردن. وللباحثين الحق في قراءة كامل نصوص الأوراق البحثية المنشورة في المجلة وطباعتها وتوزيعها والبحث عنها وتنزيلها وتصويرها والوصول إليها. وتسمح المجلة بالنسخ من الأوراق المنشورة، لكن مع الإشارة إلى المصدر.

الأهداف والمجال

تهدف المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) إلى نشر آخر التطورات في شكل أوراق بحثية أصيلة وبحوث مراجعة في جميع المجالات المتعلقة بالاتصالات وهندسة الحاسوب وتكنولوجيا المعلومات وجعلها متاحة للباحثين في شتى أرجاء العالم. وتركز المجلة على موضوعات تشمل على سبيل المثال لا الحصر: هندسة الحاسوب وشبكات الاتصالات وعلوم الحاسوب ونظم المعلومات وتكنولوجيا المعلومات وتطبيقاتها.

الفهرسة

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات مفهرسة في كل من:

فريق دعم هيئة التحرير

عنوان المجلة

الموقع الإلكتروني: www.jjcit.org

البريد الإلكتروني: jjcit@psut.edu.jo

العنوان: جامعة الاميرة سمية للتكنولوجيا، شارع خليل السالم، الجبيهة، عمان، الأردن.

صندوق بريد: 1438 عمان 11941 الأردن

هاتف: +962-6-5359949

فاكس: +962-6-7295534

جامعة الأميرة سميّة للتكنولوجيا
Princess Sumaya University for Technology

صندوق دعم البحث العلمي والابتكار
Scientific Research and Innovation Support Fund

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات

ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

أيلول ٢٠٢٤ المجلد ١٠ العدد ٣

JJCIT

www.jjcit.org

jjcit@psut.edu.jo

مجلة علمية عالمية متخصصة تصدر
بدعم من صندوق دعم البحث العلمي والابتكار