IJCSIS Vol. 19 No. 5, May 2021
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

Indexing Service

IJCSIS has been indexed by several world class databases, for more information, please access the following links:

Global Impact Factor
http://globalimpactfactor.com/
Google Scholar
http://scholar.google.com/
CrossRef
http://www.crossref.org/
Microsoft Academic Search
http://academic.research.microsoft.com/
IndexCopernicus
http://journals.indexcopernicus.com/
IET Inspec
http://www.theiet.org/resources/inspec/
EBSCO
http://www.ebscohost.com/
JournalSeek
http://journalseek.net
Ulrich
http://ulrichsweb.serialssolutions.com/
WordCat
http://www.worldcat.org
Academic Journals Database
http://www.journaldatabase.org/
Stanford University Libraries
http://searchworks.stanford.edu/
Harvard Library
http://discovery.lib.harvard.edu/?itemid=|library/m/aleph|012618581
UniSA Library
http://www.library.unisa.edu.au/
ProQuest
http://www.proquest.co.uk
Zeitschriftendatenbank (ZDB)
http://dispatch.opac.d-nb.de/

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2021 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies, IoT
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Editorial Board

*It is our great pleasure to present the **May 2021 issue** (Volume 19 Number 5) of the **International Journal of Computer Science and Information Security** (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and technology. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over **19313 times** and this journal is experiencing steady and healthy growth. Google statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, LinkedIn, Academia.edu and EBSCO among others.*

A great journal cannot be made great without a dedicated editorial team of editors and reviewers. On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers for their outstanding efforts to review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status, making sure we deliver high-quality content to our readers in a timely fashion.

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." We would like to thank you, the authors and readers, the content providers and consumers, who have made this journal the best possible.

*For further questions or other suggestions please do not hesitate to contact us at **ijcsiseditor@gmail.com**.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 19, No. 5, May 2021 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	**Dr Riktesh Srivastava [Profile]** Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	**Dr. Jianguo Ding [Profile]** Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran	**Dr. Naseer Alquraishi [Profile]** University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	**Dr. Kai Cong [Profile]** Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	**Dr. Omar A. Alzubi [Profile]** Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	**Dr. Jorge A. Ruiz-Vanoye [Profile]** Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	**Prof. Ning Xu,** Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	**Dr . Bilal Alatas [Profile]** Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	**Dr. Ioannis V. Koskosas,** University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	**Dr Venu Kuthadi [Profile]** University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	**Dr. Zhihan lv [Profile]** Chinese Academy of Science, China
Dr. Vahid Esmaeelzadeh [Profile] Iran University of Science and Technology	**Prof. Ghulam Qasim [Profile]** University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	**Prof. Dr. Maqbool Uddin Shaikh [Profile]** Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	**Dr. Musa Peker [Profile]** Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	**Dr. Wencan Luo [Profile]** University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	**Dr. Ijaz Ali Shoukat [Profile]** King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	**Dr. Yilun Shang [Profile]** Tongji University, Shanghai, China
Dr. Sedat Akleylek [Profile] Ondokuz Mayis University, Turkey	**Dr. Sachin Kumar [Profile]** Indian Institute of Technology (IIT) Roorkee

ISSN 1947 5500

TABLE OF CONTENTS

1. PaperID 01052108: Stemming Algorithm Optimization Using Big Data Analytics Tools (pp. 1-26)

M. Bougar, Dr. El. Ziyati,
RITM LABORATORY EST/ENSEM, University Hassan II, Casablanca, Morocco

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

2. PaperID 01052109: Convolutional Neural Networks and Long Short Term Memory for Phishing Email Classification (pp. 27-35)

Regina Eckhardt, Department of Computer Science, University of West Florida, Pensacola, FL, USA
Sikha Bagui, Department of Computer Science, University of West Florida, Pensacola, FL, USA

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

3. PaperID 01052111: Security Assessment of Authentication Protocols in Mobile Adhoc Networks (pp. 36-40)

Megha Soni, Assistance Professor, SVCE, Indore India
Brijendra Kumar Joshi, Professor MCTE, MCTE, Mhow India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

4. PaperID 01052117: Prediction of Survival in Breast Cancer Patients using Random Forest Classifier and ReliefF Feature Selection Method (pp. 41-47)

*Diogo Albino de Queiroz (#*1), Gabriel Sousa Almeida Assunção (#2), Kamila Alves da Silva Ferreira (#3), Vilian Veloso de Moura Fé (#4), Vitória Paglione Balestero de Lima (#5), Fernanda Antunes Dias (#6), Túlio Couto Medeiros (#7), Karen Nayara de Souza Braz (#8), Rodrigo Augusto Rosa Siviero (#9), Pâmela Alegranci (#10), Eveline Aparecida Isquierdo Fonseca de Queiroz (#11)*
(#) Universidade Federal de Mato Grosso (UFMT), Av. Alexandre Ferronato, 1200, 78550-728 – Sinop, MT – Brasil
(*) Escola Técnica Estadual de Educação Profissional e Tecnológica, Av. das Sibipirunas, 1681, 78557-673 – Sinop, MT – Brasil

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

5. PaperID 01052123: Computer Aided Diagnostic System for Diabetic Retinopathy Detection using Image Processing and Artificial Intelligence (pp. 48-63)

Anitha T Nair, Department of CSE, FISAT, Ernakulam, India
Arun Kumar M N, Department of CSE, FISAT, Ernakulam, India
Anitha M L, Department of CSE, PES College of Engg., Mandya, India
Anil Kumar M N, Department of ECE, FISAT, Ernakulam, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

6. PaperID 01052124: Strategies for Correlating DB2 & CICS SMF Records to aid Problem Determination (pp. 64-67)

Dr. Latha Sadanandam, Senior Cloud Modernization Architect, Cloud Centre of Competency, IBM India Pvt Ltd., Bangalore, India.
Atul Misra, Executive IT Enterprise Architect, Cloud Center of Competency, IBM India Pvt Ltd., Bangalore, India.
James Roca, WW Technology Partner Architect, IBM Cloud & Cognitive Software, IBM Services, Austin, TX, United States.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

7. PaperID 01052126: Cloud-Based Enterprise Resource Planning for Sustainable Growth of SMEs in Third World Countries (pp. 68-84)

Anthony I. Otuonye, Department of Information Technology, Federal University of Technology Owerri, Nigeria.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

8. PaperID 01052131: PCA, SPCA & Krylov-based PCA for Image and Video Processing (pp. 85-91)

(1) Amanda Zeqiri, (2) Markela Muca, (3) Arben Malko
(1, 2) Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Albania Tirana, Albania
(3) Lev Tech, Software Development Company

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

9. PaperID 01052134: Backpropagation and fuzzy algorithm Modelling to Resolve Blood Supply Chain Issues in the Covid-19 Pandemic (pp. 92-96)

Aan Erlansari, Rusdi Efendi, Funny Farady C., Andang Wijanarko, Reza Herliansyah, Boko Susilo
Faculty of Engineering, University of Bengkulu, Indonesia.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

10. PaperID 01052135: Integrating Neurological Examination with Radiology Diagnosis through Ontology (pp. 97-105)

Suela Maxhelaku, Computer Science Department, Faculty of Natural Sciences, University of Tirana, Albania.
Alda Kika, Computer Science Department, Faculty of Natural Sciences, University of Tirana, Albania.
Ridvan Alihmehmeti, Departament of NeuroSciences, University of Medicine, Service of Neurosurgery, University Hospital Center Mother Teresa, Albania.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | DOI | Google Scholar]

Stemming Algorithm optimization using big data analytics tools

M. BOUGAR[1], Dr, EL. ZIYATI[2],
RITM LABORATORY EST/ENSEM, University Hassan II
Casablanca, Morocco
Marieme.bougar7@gmail.com[1], Ziyati@gmail.com[2]

Abstract: With the exploitation of BIG DATA this article focuses, from a technical point of view, on the optimization of algorithms for pre-processing and classification of Big Data. Indeed, the massive amount of data produced in the world reaches such large volumes that it is undeniably impossible to analyze them manually without the help of adequate technology and statistical tools. As a result, we have chosen to exploit the open-source softwares HADOOP and SPARK. This work is interested in the Arabic language because we wanted to give our language even more interest in the world of scientific research in the face of the universal English language. This work created the optimizations of powerful algorithms for the preprocessing and classification of data in Arabic, as well as a comparative study to choose the best Big Data technological tools according to one's need.

***Key words*: BIG DATA, Preprocessing, classification, Stemming,** HADOOP, SPARK

I. INTRODUCTION

Arabic is the mother tongue of more than 400 million people and ranks seventh among Internet users in 2010 [1]. Arabic content on the Web has increased in volume, especially after the advent of social media. There are more than 6.5 million Arab Internet users generating 10.8 million tweets per day [2]. As a result, information-seeking systems for Arabic languages are becoming increasingly sophisticated and their evaluation is an important guide in the field of scientific research.

The challenge remains that the Arabic language is perceived as difficult. Indeed, derivative forms are sometimes irregular and flexed, spelling is different for certain words, writing differs from certain combinations of characters, short and long vowels and the richness of affixes in words [3] [4]. In addition, it is also necessary to consider many specification points of Arabic words, such as unstructured forms, words that spell differently depending on the context and specification of diacritics [6], which have led to the emergence of many approaches to overcome its challenges and address the Arabic language [7] [8], and not missing to mention the different dialects.

In this work, we address the challenge of big data for analytics, which require an efficient system that can manage the elasticity and scalability of data, especially the speed and efficiency that the MapReduce model offers. Indeed, it is considered one of the wisest choices for the analysis of large data [5].

We have reproduced and optimized Khoja Stemmer with the Apache Hadoop Framework for word manipulation, storage and derivation. Yahoo developed Hadoop, an open source framework of the MapReduce system, to process several terabytes of data on as many as 10,000 cores. Hadoop is a distributed and scalable computing framework, as well as a powerful distributed storage and computer hardware tool for a much less expensive system [9].

The aim of this work is to analyze the results obtained using Spark on Hadoop, and to expose all possible synergies between the two Framework and their respective benefits. This completes our Stemmer Khoja optimization process and keeps the results provided by the essential algorithm in terms of the accuracy of the basic words output. This minimizes the running time and use of the hardware by implementing the clustering approach.

Towards the end of this two-part work, one dedicated to the comparison of the two Frameworks and another part for the implementation of Spark on Hadoop, a clear perception is designed to benefit from the characteristics and performance of both Frameworks. The results of our work also include renditions on the consumption aspect of CPU and memory, as well as the logical side of parallelism minimizing the execution time parameter.

The aim of this research is to propose an optimized algorithm based on the Khoja Stemmer, implementing existing storage principles and computational models to improve the performance and accuracy of classified outputs of the roots of Arabic words inputted into this mechanism. Several Stemmers have been selected to enter into this research work such as Khoja based on the roots of words, and the Light Stemmers.

Our algorithms can be applied to natural language processing areas and error detection for large volumes of data for specific analysis or classification purposes.

II. Stemming Algorithms:

Stemming's algorithm, or the "Stemmer," has three main objectives:

The first is to group words according to their subject. Many words are derivatives of the same root and we consider them to belong to the same concept. For example: investigation, investigation, investigator. These derivations are generated by affixes (prefixes, infixes, suffixes). In general, and more specifically in English, only suffixes are considered, because prefixes and infixes change the meaning of the word, which leads to errors of poor determination of the subject.

Some exceptions to this rule occur in very bending languages such as German and Dutch, or in documents belonging to specific themes, such as medicine or chemistry, where prefixes and suffixes retain the concept of the word. Of these suffixes, two types of diversions can be considered. In the first case, inflection derivations reflect grammatical information related to sex, number, or time. These derivations do not cause a change in the part of the speech of the original word, that is, the linguistic category of the word (name, verb or adjective), nor in its meaning. On the contrary, derivative suffixes create new words based on an existing word, with which it shares meaning or not (for example, words ending with -IZE, -ATION, -SHIP). By eliminating these suffixes from a derived word, we get its stem, which is almost its morphological root, and then we can identify thematically related words by matching their stems.

The second objective of a stem is directly related to the information-finding process, because having the words rods improves certain phases of this process. These include the possibility of indexing documents according to their themes, since their terms are grouped by rods (which are similar to concepts), or the widening of a query to obtain more and more accurate results.

Extending the query allows you to refine it by replacing the terms it contains with their related topics that are also present in the collection, or by adding these topics to the original query. This can be done automatically and transparently for users, or the system can offer users one or more improved query formulations, allowing them to decide whether one of them is more specific and better defines their needs. Even if the extension of the interactive query is better in principle, because the user has better feedback on what is happening, it generally can not be done directly with the result of the stem, because the stems are only incomprehensible by humans.

The confusion of words sharing the same rod results in a reduction of the dictionary to be taken into account in the process, because the entire vocabulary contained in the collection of untreated input documents can be reduced to a set of subjects or rods. This leads to a reduction in the space needed to store the structures used by an information search system (such as the document terms index). This also reduces the system's computational load.

III. Why is Arabic a challenge?

Arabic is a Semitic language, a linguistic family that also includes Hebrew, Aramaic and Amharic. It is estimated that there are about four hundred million first-language speakers of Arabic [11] [12]. It is their mother tongue or their second language. As this is the language of the religious teaching of Islam, many other speakers from different nations have a passive knowledge of the language. Arabic is also one of the six official languages of the United Nations and the fifth most used language in the world [10] [36].

Phrases in Arabic are delineated by dots, dashes and commas, while words are separated by spaces and other punctuation marks. Arabic writing is written from right to left, while Arabic numerals are written and read from left to right. Arabic writing consists of two types of symbols [11] [14]: letters and diacritics (also known as short vowels), which are certain spelling symbols, usually added to disambiguate Arabic words. Al-Salamah [10] stated that the Arabic alphabet has 28 letters and that, unlike English, there are no capital letters and lower case for Arabic letters. An additional character, which is the HAMZA(ء),has also been added, but it is generally not classified as the 29th letter.

Arabic words are divided into three parts: names including adjectives and adverbs, then verbs, and particles. In Arabic, particles are attached to verbs and names. The words in Arabic are either masculine or feminine. The feminine is often formed differently from the masculine, for example ʿand l, which means respectively: unique (female) mistress, unique (male) teacher.

The same characteristic also appears in the names and verbs in literary Arabic to indicate the number (singular, double to describe two entities, and the plural) as in معلمة et معلم.

Arabic has a complex morphology. Its diversion system is based on 10,000 independent roots[38]. The roots in Arabic are usually built from 3 consonants (triliterals) and it is possible that 4 consonants (quad-literal) or 5 consonants (pent-literal) are used. Of the 10,000 roots, only about 1,200 are still used in modern Arabic vocabulary [15].

Words are formed by widening the root with affixes using well-known morphological patterns (sometimes called measurements) [10] [16].

Words and morphological variations are derived from roots using patterns. Grammatically, the main motif, which corresponds to the triliteral root, is the pattern. More regular patterns, adhering to well-known morphological rules, may be derived from the main motif.

Different types of affixes can be added to derived pattern words to build a more complex structure. Defined articles, such as conjunctions, particles and other prefixes can be added at the beginning of the word, while suffixes can be added at the end. For example, the word لنعلمهم - which means: we will teach them, can be broken down as follows: (antefix:ل , prefix: ن , root:علم , and postfix: هم). Kadri and Nie [17] have identified an important research that amply explains the affixes used in the Arabic language.

Unconstitutionals, whether separated or not, are usually prepositions added at the beginning of words before prefixes. Prefixes are attached to illustrate forms of verbs in future time and imperative, and usually consist of one, two or three letters. The suffixes are added to indicate sex and number, for example in the female double and the male plural.

Postfixes are used to indicate pronouns and to represent the object. Usually, this morphology is used to create verbal and nominal sentences.

Arabic affixes may also include clitics, which have been used in the proposed stems and can be proclitic or enclitic depending on their location in the words. These are morphemas that have the syntactic characteristics of one word but are morphologically related to others [18]. Thus, the clicks are attached to the beginning or end of the words. These clitics include certain prepositions, defined articles, conjunctions, possessive pronouns, particles and pronouns. Examples of clitics are the letters pronounced as (KAF) and pronounced as (FAA), which mean respectively as and then.

Arabic adjectives are considered names. Thus, the different forms that can be derived from the adjective (مزارع) meaning"farmer" according to their two grammatical forms may include words such as: (مزارعة) for the singular feminine in nominative cases, (مزارعان)for the double male in the nominative case, (مزارعين) for the male, (مزارعتان)for the female double in the nominative case, (مزارعتين) for the female double etc.…

Morphology adds a degree of ambiguity and difficulty, making the exact keyword matching mechanism insufficient to extract words. Morphological ambiguity may appear in several cases. For example, clitics may accidentally produce homogeneous form (the same word with two or more different meanings) with another complete word [10] [11] [19][29]. For example, the word (علم) can be attached to the clitic ي (to construct the word علمي) which is homographic with the word "scientific."

In addition, Arabic grammar contributes to morphological ambiguity. According to some arabic grammar rules, vowels may be removed from the roots. The letter of the vowels in Arabic consists of three letters: ALIF, YAA and WAW (أ، ي ، و). These letters have different rules, which do not follow the system derived from Arabic, and make

them very variable. For example, the last letter YAA is deleted in a word such as (امشي) meaning "goes," which gives(امش), if that verb appears in an imperative form.

In addition to its complex morphology, Arabic also has an overly complex type of plural called "broken plural." Plurals in Arabic do not obey morphological rules.

Broken plurals make up 10% of Arabic texts and 41% of plurals [10] [20]. The plural in Arabic indicates any number greater than two; for two, it's double.

The term "broken" means that the plural form does not resemble the original singular form. For example, the plural of the word(نهر)meaning "river" is(أنهار)"rivers." In simple cases of broken plurals, the new inflected plural contains some letters common to the singular form, as in the previous example. But in many cases, the plural is totally different from the original word, for example, the plural of the word (إمراة) is (نساء), keeping no letter from the root.

Another challenge, which are simply caused by the Arabic letters ALIF with its various forms (آ , إ , أ) . In most cases, one of this letter is modified or abandoned [21].

A. Focus on the Khoja Stemmer algorithm

Khoja Stemmer has been used by numerous researches to analyze the Arabic text, at the pre-processing and filtering stage of the text, in particular to remove a stop word. Khoja Stemmer eliminates diacritics and the longest prefix. The Stemmer manages punctuation characters, diacritic characters, a large number of stop words and a defined article. The Khoja Stemmer proceeds as shown in Figure 1.

Khoja's Stemming algorithm removes the longest suffix and prefix. Then, the remaining word is compared to verbal and nominative models. To extract the root, it is based on a list of all diacritic signs, punctuation marks, defined articles and keywords.

KHOJA algorithm

1. remove diacritic signs
2. Delete: empty word, punctuation and numbers.
3. Delete the article defined by
4. Remove the separable conjunction
5. remove suffixes
6. remove the prefix
7. Match the results with a list of patterns: if the match is found, extract the characters from the pattern that the root represents. Compare the extracted root with a list of known valid roots.
8. Replace the letters of the week with.
9. Replace all hamza instances with a.
10. The roots of two letters are checked to see if they should contain a double sign. If this is the case, the sign is added to the root

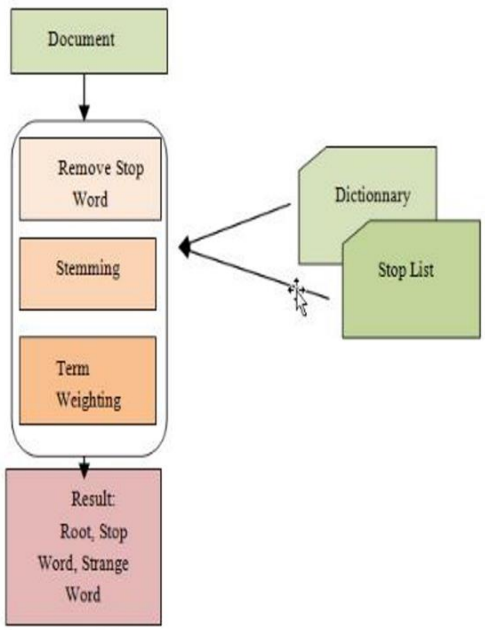


Figure 1 KHOJA algorithm

B. Conclusion

While not all researchers agree on the pros and cons of using Stemming as a process for seeking information in general terms, many agree on its benefits in specific contexts, when language is highly inflected, when documents are short or when data storage space is limited. Some researchers also argue that the nature of the documents can influence the performance and accuracy of the Stemmer.

Despite the fact that many researchers have studied this field for many years, a few questions remain open: how to evaluate a Stemming independently of the information-seeking process; how a Stemming improves the search for information in terms of speed. For these reasons, we have decided to try to answer these unknowns.

IV. LES BIG DATA:

A. Big data analytics tools

In the following, we have two ways of processing data. The first is to use the Mapreduce Framework Hadoop batch mode, the second is the streaming mode, it is a real-time processing, using Spark for data processing.

1) Hadoop

Apache HADOOP is a framework used to develop data processing applications that are run in a distributed computing environment. Similar to data residing in a local file system of a personal computer system, in Hadoop, the data resides in a distributed file system that is called Hadoop distributed file system. The processing model is based on the concept of "Data Locality," in which the calculation logic is sent to a group of nodes (servers) containing data.

This logic of calculation is nothing more than a compiled version of a program written in a high-level language, such as Java. Such a program processes data stored in Hadoop HDFS. Applications built with HADOOP are run on large datasets in widely available core computer groups. They are mainly useful for getting greater computing power at low cost. The computational cluster consists of a set of multiple processing units (storage disk - processor) that are connected to each other and act as a single system.

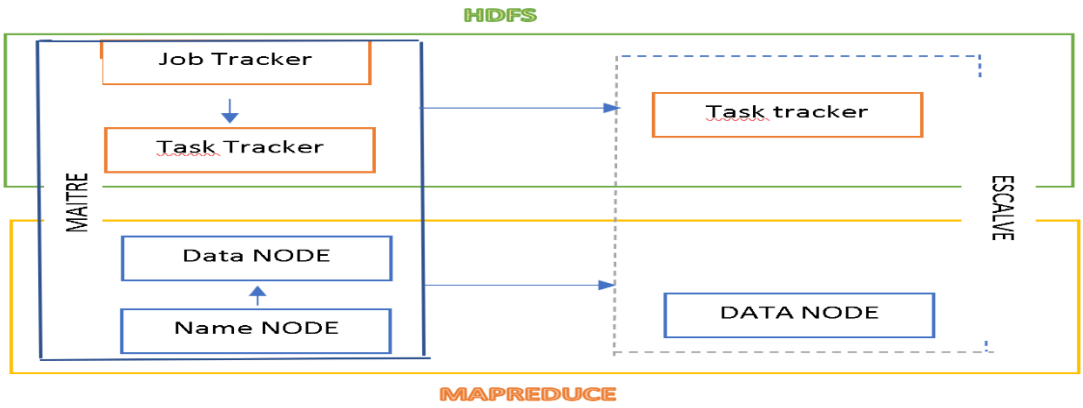


Figure 2: Clusters HADOOP2

Apache Hadoop consists of two sub-projects [23]:

1. Hadoop MapReduce: MapReduce (Figure 2) is a computational model and writing software framework that is performed on Hadoop. These MapReduce programs are capable of processing massive data in parallel on large clusters of computational nodes.

2. HDFS (Hadoop Distributed File System): HDFS(Figure 2) handles the storage part of Hadoop applications. MapReduce applications consume HDFS data. HDFS creates multiple replicas of data blocks and distributes them on cluster computing nodes [30]. This distribution allows for a reliable and extremely fast data management system in calculations.

HADOOP installation

To install HADOOP [80] we used the version (2.6) in "SINGLE NODE INSTALLATION" mode on UBUNTU 14.04 LTS (32 bit).

Below we describe the various stages of installation and implementation of this architecture

Step 1: Open the terminal using the next keyboard key combination (ctrl-alt-t), then write the next command marie@localhost sudo apt-get update

Step 2: $marie@localhost sudo apt-get upgrade

Étape 3 : marie@localhost$ sudo apt-get install openssh-server

Step 4: First check whether the jdk (java set up) is installed on the machine or not. To open the terminal we use (ctrl at-t) and then write the following command: marie@localhost$java-version (If Java is already installed do not run step 5).

Step 5: to install java you must write the following command: marie@localhost$sudo-get install openjdk-7-jdk marie@localhost$ java -version NOTER THAT: you will have a message that JAVA 1.7 VERSION is successfully installed, the jdk will be installed in the file /usr/lib/jvm/java-7-openjdk-i386

Step 6: Download Hadoop with the following command marie@localhost$wget http://mirrors.sonic.net/apache/Hadoop/common/Hadoop-2.6.0/Hadoop-2.6.0.tar.gz (then put the downloaded folder in the following folder:/usr/local file (this means that Hadoop is installed in the directory /usr/local) NOTE: you can install Hadoop wherever you want. marie@localhost$sudo tar -zxvf /Downloads/Hadoop-2.6.0.tar.gz -C /usr/local;

Note: to extract Hadoop Hadoop-2.6.0.tar.gz from the repertoire: /usr/local;marie@localhost$sudo mv /usr/local/Hadoop-2.6.0 /usr/local/Hadoop; Note: to rename the f file: marie@localhost$sudo chown -R marie /usr/local/Hadoop;

Step 7: Add a group and a Hadoop user: $marie@localhost$sudo addgroup Hadoop; marie@localhost$sudo adduser --ingroup Hadoop marie;(marie is our user name Hadoop) marie@localhost$sudo adduser marie sudo;

Step 8: Run the SSH certificate and generate your keys: marie@localhost$ssh-keygen -t rsa -P" marie@localhost$cat .ssh/id_rsa.pub .ssh/authorized_keys

Step 9: Get your IP address marie@localhost ifconfig marie@localhost$ sudo gedit /etc/hosts; marie@localhost$ sudo gedit /etc/hostname; marie@localhost$ssh localhost; Note: restart computer ($marie@localhost sudo reboot)

Step 10: Open the bashrc file and write the following code for Hadoop marie@localhost$sudo nano -/.bashrc marie@localhost$ source -..bashrc:

```
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-i386 export HADOOP_INSTALL=/usr/local/Hadoop export
PATH=marie@localhost$PATH:marie@localhost$HADOOP_INSTALL/bin export
PATH=marie@localhost$PATH:marie@localhost$HADOOP_INSTALL/sbin export
HADOOP_MAPRED_HOME=marie@localhost$HADOOP_INSTALL export
HADOOP_COMMON_HOME=marie@localhost$HADOOP_INSTALL export
HADOOP_HDFS_HOME=marie@localhost$HADOOP_INSTALL export
YARN_HOME=marie@localhost$HADOOP_INSTALL export
HADOOP_COMMON_LIB_NATIVE_DIR=marie@localhost$HADOOP_HOME/lib/native export
JAVA_LIBRARY_PATH=marie@localhost$HADOOP_HOME/lib/native export HADOOP_OPTS="-
Djava.library.path"=marie@localhost$HADOOP_INSTALL/lib
```

(Finally note that JAVA_HOME contains the path of my jdk and /usr/lib/jvm/java-7-openjdk-i386 HADOOP_INSTALL is the path where Hadoop is installed on my computer, Hadoop is installed in /usr/loca/Hadoop)

Step 10: set up the following xml files: marie@localhost$cd /usr/local/Hadoop/etc/Hadoop Edit core-site.xml: marie@localhost$ sudo gedit core-site.xml write: fs.default.name hdfs://localhost:9000

Step 10.1: set up hdfs-site.xml marie@localhost sudo gedit hdfs-site.xml write on the following lines: dfs.replication 1 dfs.permissions false

Step 10.2: set up mapred-site.xml NOTE: currently in the folder: /usr/local/Hadoop/etc/Hadoop , the mapred-site file.xml is not present. On the other hand the mapred site.templete file is available, so you have to convert this file from mapred-site.templet to mapred-site.xml using the command CONVERT mapred-site.templete at mapred-site.xml marie@localhost$ mv mapred-site.xml.mapred-mapred-site.xml

Étape 10.3 : change mapred-site.xml file marie@localhost$sudo nano mapred-site.xml write down following lines mapred.job.tracker localhost:9001

Step 10.4: set up the file Hadoop-env.sh marie@localhost$sudo nano Hadoop-env.sh /usr/lib/jvm/java-7-openjdk-i386 (just write this path in JAVA_HOME) my jdk is installed in /usr/lib/jvm/java-7-openjdk-i386)

Step 11: FORMATER NAMENODE using the following command on the terminal: cd /usr/local/Hadoop/bin then type the command: marie@localhost$Hadoop namenode -format

Step 12: Restart all DEMONS $marie@localhost start-all.sh

Step 13: Check that all DEMONS are in RUNNING status to write the following command: marie@localhost$ jps (if all the demons are present you have successfully installed your Hadoop)

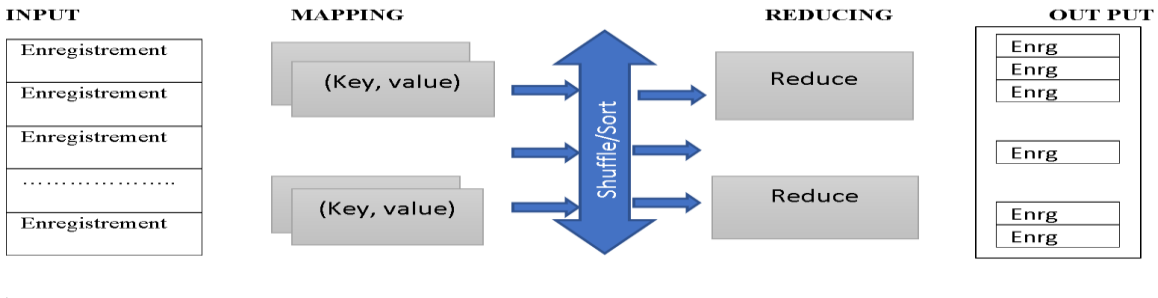


Figure 3: The MAPREDUCE Model3

Inputs and outputs

The MapReduce model (Figure 3) works exclusively on key, value pairs, i.e. they are seen at the entrance to the Job as a set of 'key', value-for-money' pairs and produces a set of key, valueable pairs like job exit, possibly of different types.

Key and value classes must be seable by the model and must therefore implement an inscriptible interface. In addition, key classes must implement the WritableComparable interface to make it easier for the model to sort.

Types of entry and exit of a MapReduce job:

(Entry) - v1, v1 - map - 'lt'k2, v2 - combine - -k2, v2 - reduce --t'3, v3

execution

To run the java code:

- Copy the code into a java file. Execute the following commands.
- Configure environment variables: export JAVA_HOME=/usr/java/default export PATH=marie@localhost${JAVA_HOME}/bin:marie@localhost${PATH} export HADOOP_CLASSPATH=marie@localhost${JAVA_HOME}/lib/tools.jar
- Compile Name.java and create jar: $marie@localhost bin/Hadoop com.sun.tools.javac.Main Name.java marie@localhost$ jar cf wc.jar Nom-.class

Make sure that:

- /user/marie/Name/input - the input path exists in HDFS
- /user/marie/wordcount/output - output path exists in HDFS
- To run any code or test: marie@localhost$ bin/Hadoop jar file.jar /user/marie/contact/user/user/marie/name/output

The MapReduce programming model includes two functions, map () and reduce (). Users can implement their own processing logic by providing a custom map () and reduce () function. The map function takes a key/entry value pair and creates a list of key/intermediate pairs. MapReduce's execution system groups all intermediate pairs according to intermediate keys and transmits them to the reduce function () to obtain the final results [86].

Map Reduce can be divided into two steps:

a) The mapped key/entry value pairs to a set of key/intermediate-value pairs. Maps are the individual tasks that turn input records into intermediate records. Processed intermediate records do not need to be the same type as input records. A given input pair can map on zero or several output pairs.

b) Reducator Steps: The gearbox has three main steps: mix, sort and shrink.

   1. Random play: Random playback is a step on intermediate data, used to combine all values in the key associated set. After that, there will be no more duplicate keys in the intermediate data.

   2. Sorting: All intermediate keys on a single node will be automatically sorted before being presented to the gearbox. The sorting is done thanks to the Box class. Random reading and sorting phase at the same time; When extracting the exit of the card, they must be merged.

   3. Streamline: Provide shuffle output data and sort the maapper. Reduce at this stage for each pair to value list as long as a group entry.

2) Apache Spark Perspective

Apache Spark is a general-purpose cluster computing engine, very fast and reliable. This system provides application programming interfaces in various programming languages such as Java, Python, Scala.

Spark is an Apache cluster computing system with incubator status. This tool specializes in making data analysis faster, both for running programs and for writing data. Spark supports memory computing, allowing it to query data much faster than disk-based engines like Hadoop. It also offers a general execution model that can optimize an arbitrary operator graph. Initially, the system was developed at Berkeley University as a research project and acquired incubator status in Apache in June 2013 [5].

In general, Spark is an advanced and high-performance Upgrade from Hadoop to improve Hadoop's advanced analytics capability. The functions of the Spark engine are very advanced and different from those of Hadoop. The Spark engine is developed for memory processing as well as disk processing. This memory processing capability makes it much faster than any traditional data processing engine. For example, the project's sensors report that logistic regression in Spark is 100 times faster than in Hadoop MapReduce [25].

This system also provides many impressive high-level tools such as the MLib machine learning tool, structured data processing, Spark SQL, graphics processing taken GraphX, the flow processing engine called Spark Streaming, and Shark for the quick interactive question device [26]. This is shown in Figure 8.

Spark Components

The following illustration shows the different components of Spark.

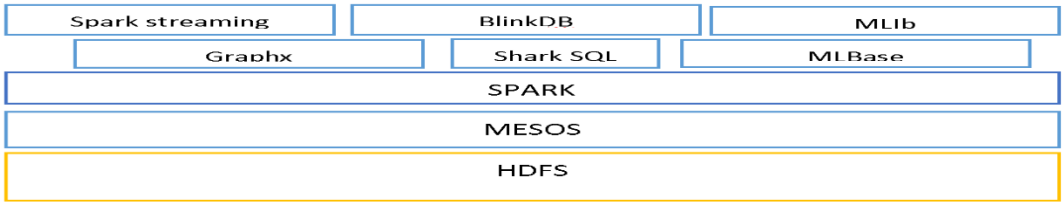


Figure 4: SPARK Framework4

- Spark Core is the underlying general performance of the engine for the Sparks platform that all other features benefit from it It provides in memory the calculation and referencing of data sets in storage systems.
- Spark SQL is a component that is added to Spark Core. It introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.
- Spark Streaming leverages Spark Core's speed and programming capability to stream analysis. It ingests data in mini batches and performs RDD (Resilient Distributed Datasets) transformations on these mini-batches of data.

- MLlib is a machine learning bookstore distributed on Spark because of spark's distributed memory-based architecture. According to benchmarks, between MLlib developments with ALS (ALS: Alternating Least Squares) implementations,Spark MLlib is nine times faster than Hadoop's Apache Mahout-based version.

  GraphX

- GraphX is a graphics processing distributed above Spark. It provides an API expressing graphic calculations to model user-defined graphs using Pregel abstraction. It also provides an optimized execution time for this abstraction.

3) Spark built on Hadoop

There are three ways to deploy Spark [32], explained below and illustrated in Figure 13.

- Standalone: Spark's stand-alone deployment means that Spark occupies HDFS [9] (Hadoop Distributed File System) and space is allocated for HDFS, explicitly. Here, Spark and MapReduce work side by side to cover all of Spark's jobs on the cluster. This mode is used when resources are allocated statically on the cluster or on a subset of machines in a Hadoop cluster. Spark can be run in parallel with Hadoop MapReduce, so the user can perform arbitrary tasks on his HDFS data. This deployment makes it simple for many Hadoop 1.x users.

- Hadoop Yarn: The deployment of Hadoop Yarn [111] [28] means that Spark operates on Yarn without pre-installation or access to the required root. This allows Spark to be integrated into the Hadoop or Hadoop Battery ecosystem. This mode makes it easy for Hadoop users to integrate Spark. It allows you to take full advantage of Spark, as well as other components running on Spark. Hadoop users have the option to run Sparks on their Hadoop wire. Even users who already use the Hadoop wire easily integrate Spark. There are no requirements such as pre-installation or administrator access.

- Spark in MapReduce (SIMR): Spark in MapReduce is used to launch Spark work on stand-alone deployment. With SIMR, the user can launch Spark and use his shell without any administrative access. In addition to stand-alone deployment, Hadoop users who do not yet use YARN can switch to SIMR. We can use this mode to start work in MapReduce.

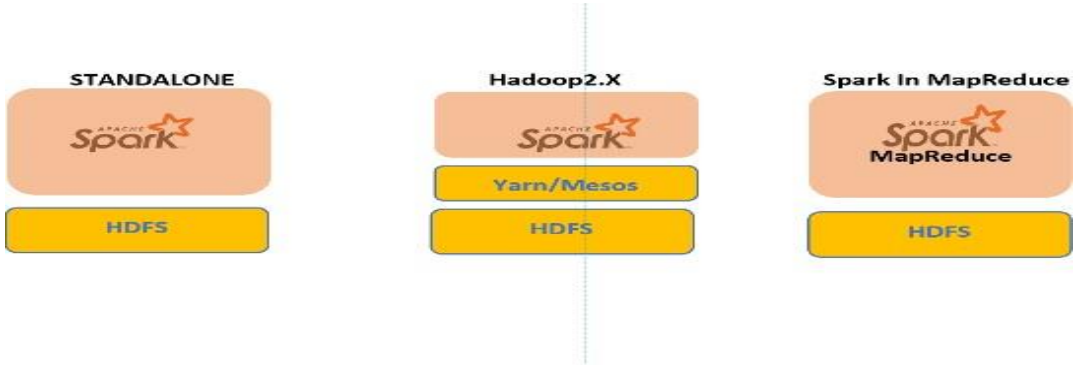


Figure 5: Sparks cluster architecture.5

4) Comparison of Mapreduce and Spark

Below is a summary of comparative studies based on the parameters of Big Data's analytics tools [24] [27] [31]. We try to understand and reproduce this comparison in order to construct a succinct comparison and allow the reader to make a first decision on the choice of tool, and for this proposed topic, we facilitate applications for the algorithm of Stemming in parallel.

Table 1: Comparison Spark vs Hadoop1

Mapreduce	Spark
MapReduce is ineffective for multi-pass applications Need to share over multiple low-latency data Parallel operation.	Spark allows large amounts of data to be processed in input. Use for online machines and enable real-time analysis
Intermediate data/results are stored more slowly on the hard drive	Compared to Hadoop, the speed can be increased up to 100 times for iterative operations, as the data/intermediate results are preserved in the memory.
Mainly a batch processing engine	Spark as a batch processing engine also includes Spark Streaming for streaming data processing, MLLIb GraphX for machine learning.
Less memory needs	Memory requirements are higher. This degrades performance when data doesn't fit in memory
Each mapping task generates data in a Key, value union. The output is stored in the Buffer Cache instead of writing on the disk.	The exit from the mapping side was written in the buffer cache. It is up to the operating system to decide whether the data can be stored in the buffer cache or should be stored.

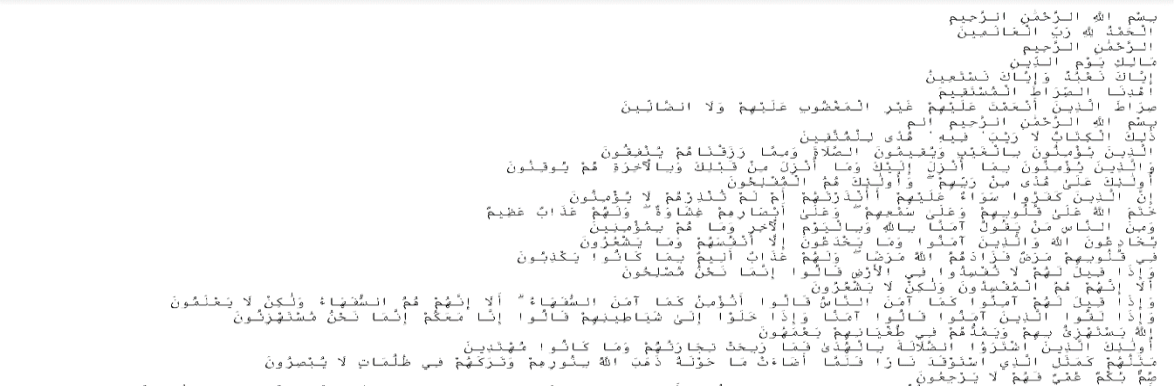


Figure 6: Extract of the data (coran) input from our algorithms.6

V. OUR ALGORITHMS: RESULTS, AND MODEL

A. Study of the existing algorithm

1) KHOJA Algorithm

The KHOJA Stemmer algorithm is developed in javaand available online:[http://zeus.cs.pacificu.edu/shereen/research.htm],as well as other versions taken up by researchers.

We had in our disposal:

- The MAIN class ARABICStemmer in java (Main Class).
- The Stemmer Class in java: containing the feature that treats the entire Khoja Stemmer process
- Gui Classes, containing the development of the graphical interface (Figure 7)
- The folder Stemmer Files building the data dictionary necessary for the conversion of characters and words, Khoja is based on the dictionary approach.

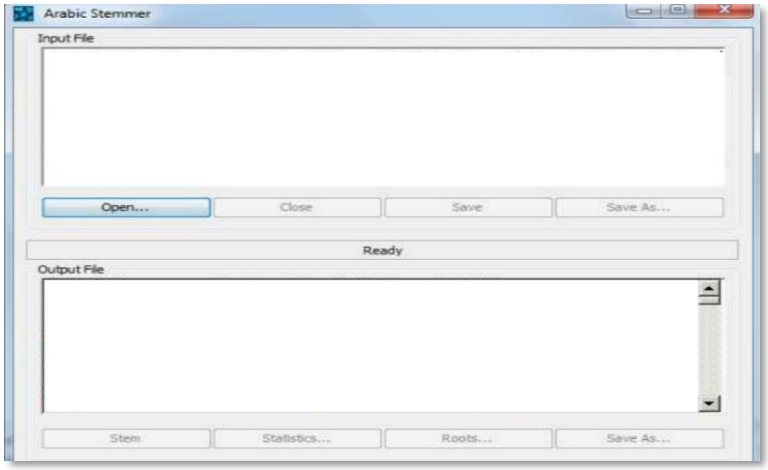


Figure 7 : Graphic interface of Khoja Stemmer

This interface contains several functions, mainly cited:

- Open: Button to load the desired document.
- Stem: Button to start Stemming treatment,
- Statistics: Button to view results measuring the accuracy of the roots extracted from the words provided on the basis
- Save as: Button to download and save the results of the stemming.

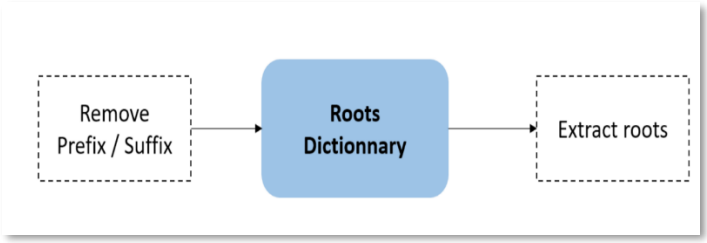


Figure 8: Khoja Stemmer's Approach7

2) Preliminary results

Examples of results are presented in Table 2, which presents different information: word, radical, type, prefix, affix and suffix. Its elements are important for identifying words, here is some type of information provided by the algorithm.

Table 2: Results provided by the Stemmer Khoja Algorithm.2

Word	Stem	Type	Pattern	Préfix	affix	suffix
from	from	Stop WORD	Null	Null	Null	Null
U,000	Uterus	Root	make	the	Null	nun
owner	possess	Root	make	Null	alif	Null

For a unit test performed on a word set that reaches 82624 words, the result displayed is 98.87%.

3) Lucene light Stemmer

In Figure 8.1 we present Lucene's algorithm [22] which is an Apache Arab Stemmer.

1. **Normaliser le mot :**
   - **1.1** Supprimer les diacritiques.
   - **1.2** Remplacer آ أ إ avec ا.
   - **1.3** Remplacer ة avec ه
   - **1.4** Remplacer ى avec ي
   - **1.5** Finir un re-process d'élimination des diacritiques
2. **Stemming :**
   - **2.1** Eliminer les préfixes : **ال، وال ، فال ، كال ،بال ،لل، و**
   - **2.2** Eliminer les suffixes : **ها، ان ، ات، ون ، ين، ية، ه،ي**

Figure 8.1: Light Stemmer Lucene algorithm8

4) Advantage of a Light Stemmer's process

The main idea of using light Stemmer is that many word variants do not have similar meanings or semantics. However, these word variants are generated from the same root.

Thus, root-based Stemming algorithms affect the meaning of words. The Light Stemming aims to improve classification performance while maintaining the true meaning of words. It certainly removes the prefixes and suffixes

of the word instead of seeking to extract the original root, for example the Arabic words (المكتبة الكاتب الكتاب)which mean (the library, the book, the writer) respectively, belong to the same root (كتب) although they have not the same meaning. Thus, the Lucene's approach is to keep the original meaning of the word. As a result of Lucerne, the new results will match the word(الكتاب) which means (the book) to (كتاب)(which means (the book).

5) The Mapreduce Computing Model

Map() is the first step in the MapReduce algorithm. It takes input tasks and divides them into smaller sub-tasks. The release of this Mapping function is a set of pairs of keys and values (key, value).

The reduction step takes a list of outputs from the Mapping function and performs these two sub-steps on each key-value pair.

Hadoop's Framework MapReduce uses a distributed file system to read and write its data. Hadoop MapReduce uses the Hadoop Distributed File System (HDFS), as well as file system. Therefore, the entry/exit performance of a Hadoop MapReduce job is closely dependent on HDFS.

The idea of our PSA algorithm and to use this model properly and admit that the key and the word is referred to in the same way and this in order to keep track of the original word to be milled in the Stemmer that will be highlighted in the output. Thus a stage pair of Mapping is defined for example as follows (الرحمٰن ,الرحمٰن), then at the intermediate phase called shuffling() sets of pairs will be assembled by common key finally in the stage of the reduce function() or the algorithm Khoja will be applied to get out the roots of the corresponding words out a file of pairs (original word, word after stemming) is built.

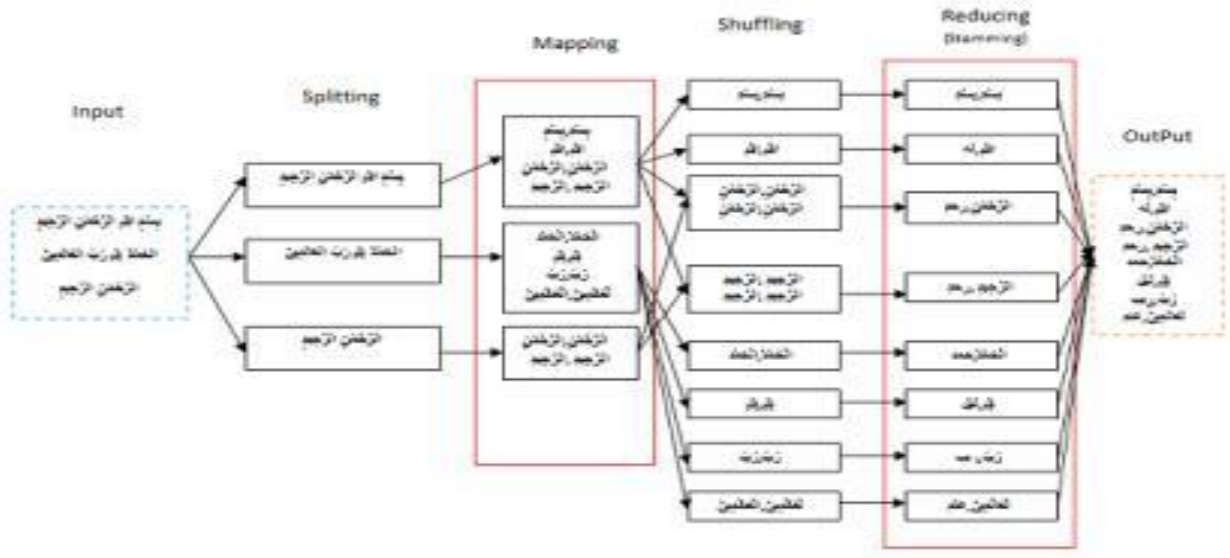


Figure 9:9 MapReduce model applied to Khoja algorithm.

6) PSA algorithm

Algorithm 1: Distributed Stemmer Using MapReduce

1 dStemming function ;
Input : Arabic Text such as the Holy Quran
Output: Stemmed Words
2 **Splitting text :** *each line is treated separately in parallel*;
3 **Mapping :** *each word is associated to itself as value*;
4 **Reducing:** *apply stemming function to the using Khoja stemmer*;

Figure10: Parallel Stemming Algorithm (PSA)

7) Result and discussion

This program takes as an entry a heavy text document of the Koran. The output is the original word, its root with the type of belonging. All results, as shown in the figures and tables below, are directed to a web page. Then we calculate the run time with parallel processing by implementing our method and without parallel treatment, the comparison with other Root-based Stemmer or light Stemmer shows an improvement in workload.

The application allows the user to view the algorithm classification from the fastest, the dashboard shows the size of the derivative document and gives permission to download the Word derived document.

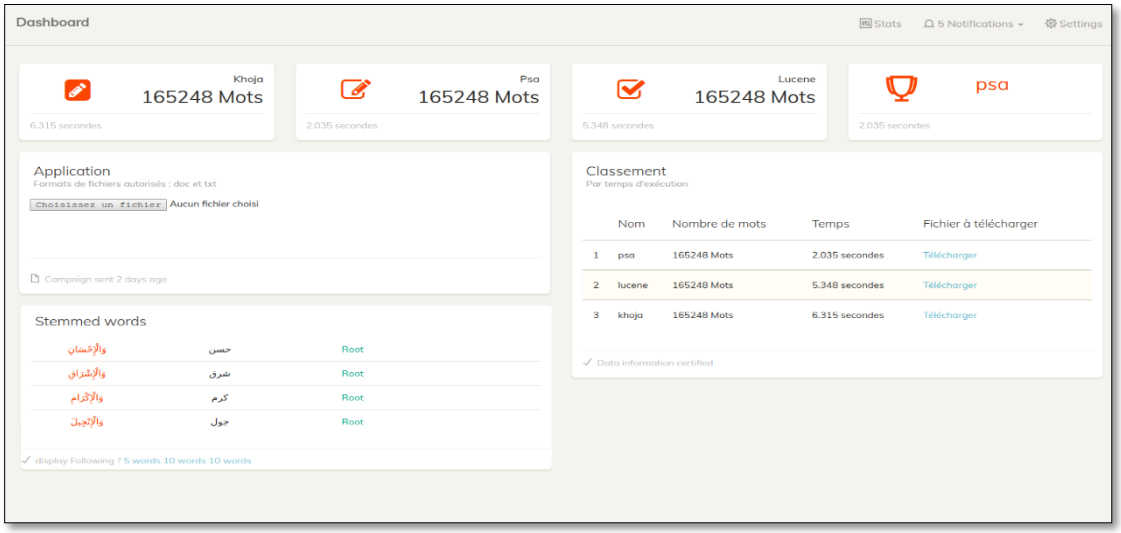


Figure11: DASHBOARD web comparative PSA, Lucene, Khoja.

Figure 12 shows that PSA's run time increases linearly in volume, but for other algorithms, the graph grows exponentially, like Khoja's algorithm.

In conclusion, the new version of Stemmer's algorithm based on distributed treatment MapReduce, shows a remarkable improvement, especially when the volume increases, the PSA acts linearly instead of other algorithms that are parabolic or even exponential.

8) Implementation of Spark solutions

Clustering en Spark

It's always a misrepresentation that Spark replaces Hadoop, but it affects Hadoop's functionality. From the beginning, Spark reads data from HDFS (Hadoop Distributed File System) and writes data. Therefore, apache Spark is a Data Processing Engine based on Hadoop. It can support batch processing and data streaming. Therefore, running Spark on Hadoop can provide improved features and more features.

Apache Spark is a framework for a computation distributed in memory, flexible in terms of writing the Spark application and deploying it to clusters. Writing applications on Spark allows calculations to evolve by adding machines and running them in cluster mode (Figure 13). You can then expand your app locally and then deploy it to

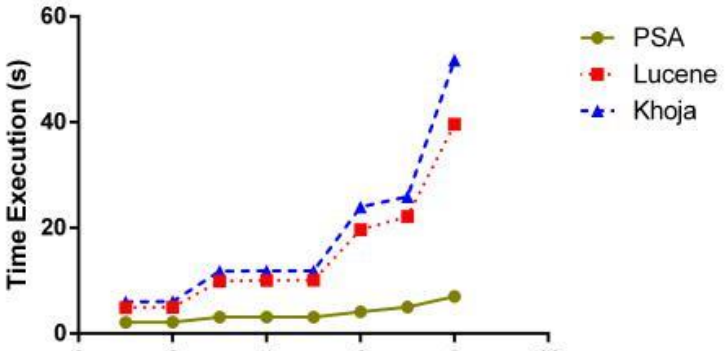


Figure 12:10 Evolution PSA run time versus another Stemming algorithm.

multiple clusters without changing the lines of code.

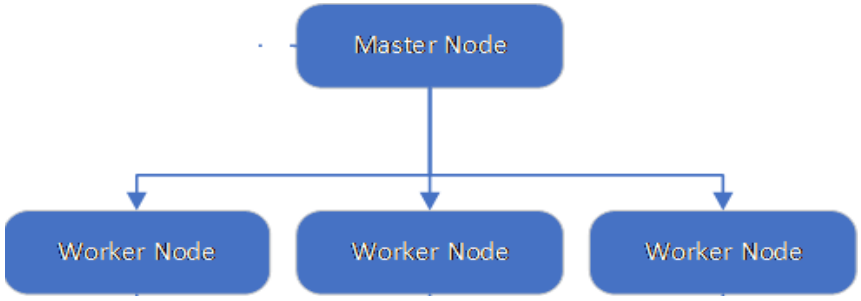


Figure 13: STANDLONE mode: Setting up Master/Slave architecture

Resilient Distributed Data Set (RDD)

Spark RDD (API-RDD) stands for Elastic Distributed Data Set. It is a collection of recording scores in playback alone. RDD is Spark's core data structure. It allows programmers to perform memory calculations on large clusters in a way that is tolerant of failures. As a result, the task is accelerated.

Spark made RDD available to users via an API developed in Scala (basic language), Java or Python. The datasets in the RDDs are represented as objects (class instances) and the transformations that are called to support the methods of these objects.

RDDs are collections of objects - basic elements of Spark that allow a calculation tolerant to failures and which are characterized by two main operations: transformations (map, reduce, filter ...) and actions (collect, count....) using filter and reduceByKey on RDD.

A RDD collection uses the cache to store RAM data for reuse, thus avoiding the disk data replication needed in Hadoop to ensure cluster availability. It is through this mechanism that you can provide high availability and tolerance to cluster failures.

In our implementation of the PSA algorithm with RDD (Figure 14), each line of the file (Holly Quran/Arabic Text) is read as an entire chain in RDD. Then the current line is divided into words that are put into a table generated by the "split" command. At the end of processing all lines. The order flatMap produces a table that includes all the tables generated in the division result. The final table is filtered to get only separate words. Each word is matched with a

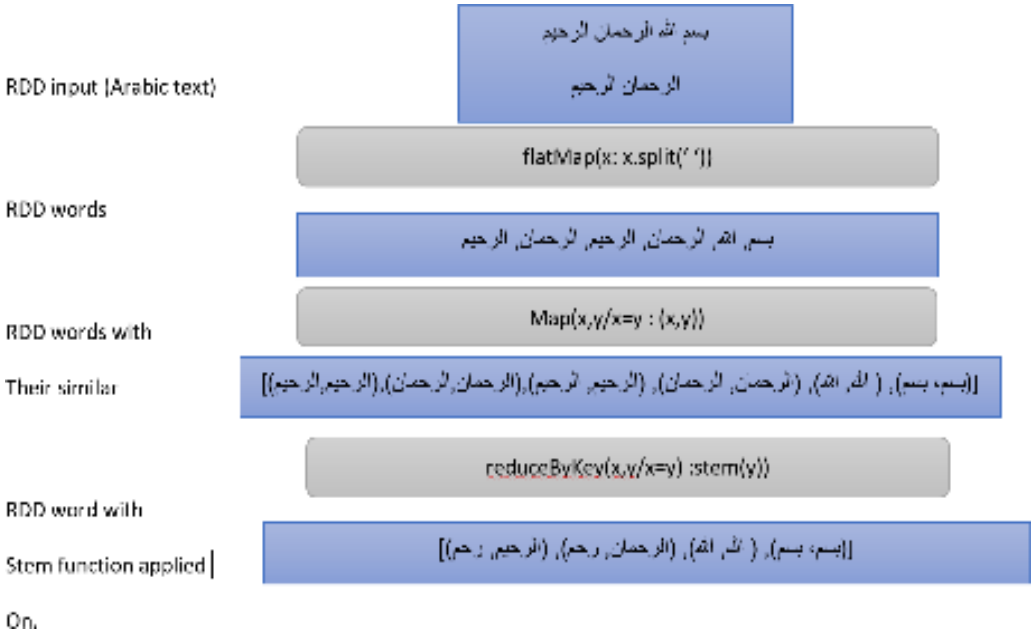


Figure 14:11 PSA algorithm improved in Spark RDD context

key/value pair. Finally, the operation reduceByKey is used on the key/value pair to apply the function Stemming on the value of the words in order to generate the root out.

Not to mention that we specify in our Spark level code the URL of our Master Node implementation that we get via the address localhost:8080, in that same page also we can recover several information including the number of workers or Slavs we created and the number of Jobs executed.

9) Results and discussion

In this section we set out our results in terms of the accuracy of the root words, the time it takes to complete the various implementations chosen during the study phase, the consumption in terms of memory and finally the use of theCPU.

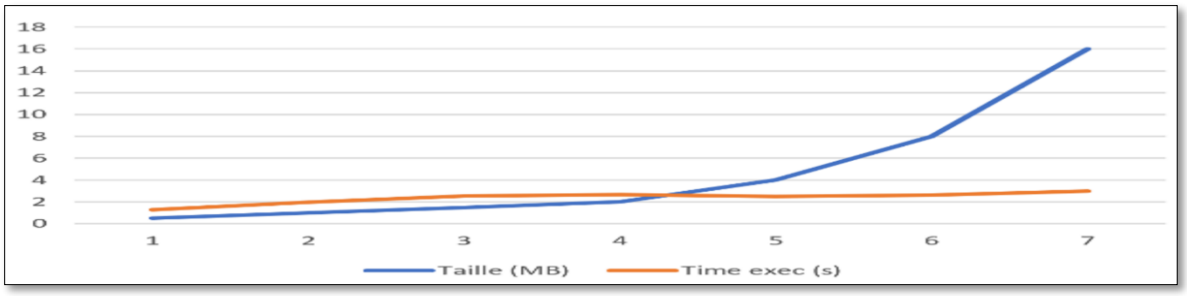


Figure 15: Evolution in time execution by file size12

As figure 15 shows, our algorithm using Spark makes our temporal function virtually nil.

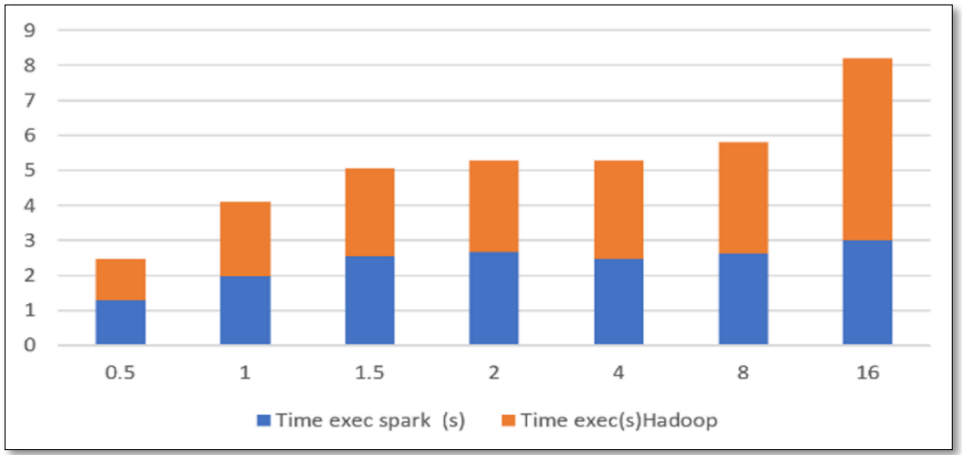


Figure 16: Time run by 13HADOOP vs. SPARKfile.

The results obtained for the Standalone implementation mode (HDFS for file system) where our optimization of the PSA algorithm with the RDD design, and are represented by tables later.

The stick graph in (Figure 16) allows you to visualize the execution time achieved by deploying the two frames separately, so our architecture (Standlone mode) used for Spark gives important results in terms of reducing execution time.

Both Hadoop and Spark require high memory consumption (Figure 17), however Spark is performed to save the use of CPU (Figure 17), these results are explained by the fact that Hadoop rewrites the value of the key on the disk, while Spark's storage is in memory.

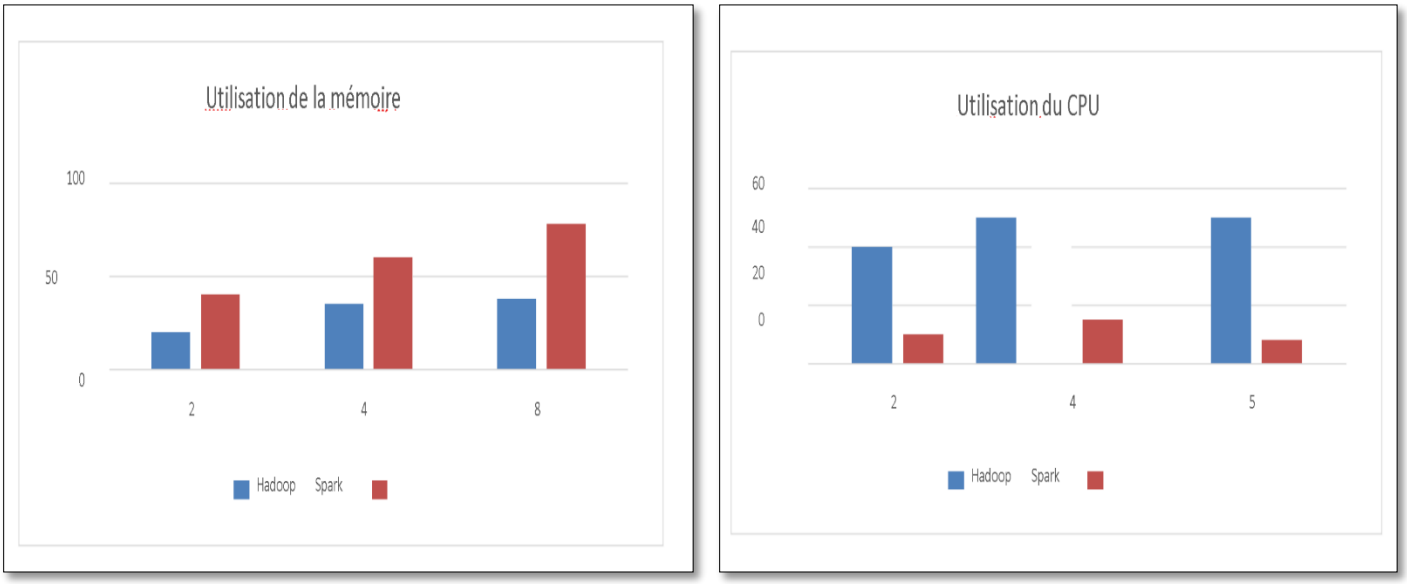
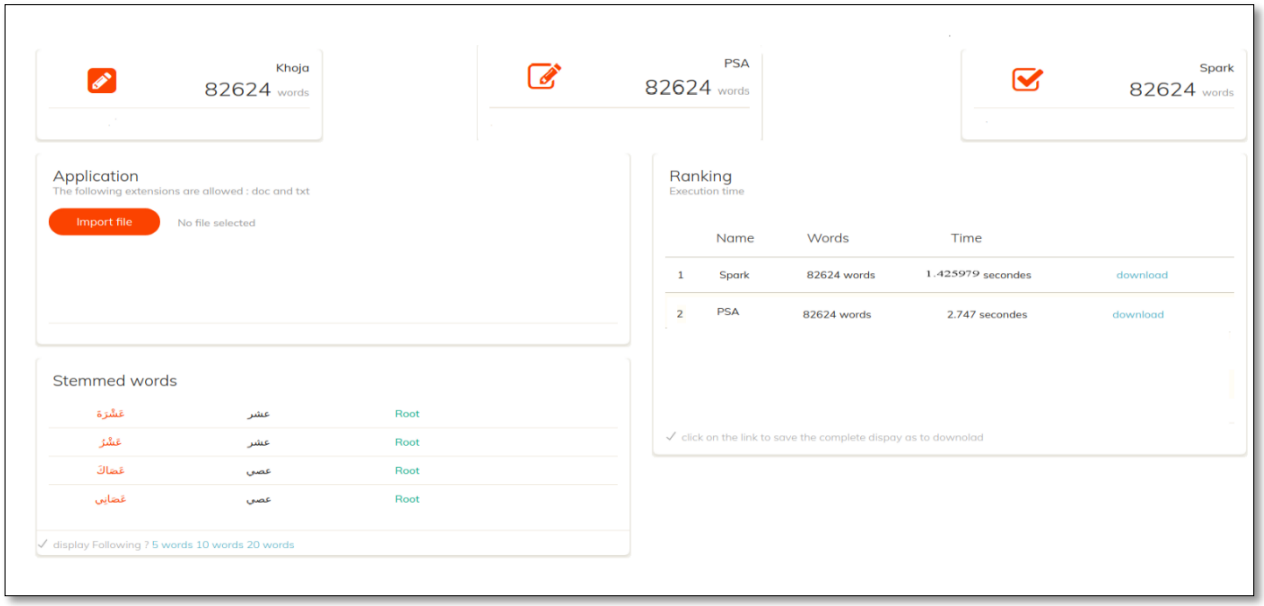


Figure 17:14Consumption of SPARK vs. HADOOP Memory and CPU

The consumption rates of resources memory are carefully recovered using the usual commands: free -m for globality, ps for the detail of processes.



In addition, the use of the CPU is supervised thanks to the "top" surveillance command, standard for most systems, and which was of great use to us. Normally, the "top" command is used dynamically. At the start, a table appears with values that change dynamically depending on the activities of each process, we also trace the time of the release of the Stemming process allowed us to align the results at a specific moment T.

We chose to track our results in terms of execution time and stemming results through a web application (Figure 18) for form and facilitate analysis as we tackle a lot of tests, hence the mastery of managing the history in visual result capture and output file loading for consultation.

Table 4: Result 1- Time Performance Spark vs Hadoop3

Framework	File size	Number of words	Results in seconds
Hadoop	112 bytes	6	1.156
Spark	112 bytes	6	0,262955

Figure 18:15Return of PSA exit results (Hadoop/Spark)

Table 5: Result 2 -Time Execution Spark vs Hadoop4

Framework	File size	Number of words	Results in seconds
Hadoop	1.3 Mo	82624	2.747 sec
Spark	1.3 Mo	82624	1.425979 sec

Table 6: Result 3- Time Execution Spark vs Hadoop5

Framework	File size	Number of words	Results in seconds
Hadoop	3 Mo	165248	2.9 sec
Spark	3 Mo	165248	1.5 sec

Table 7: Result 4-time execution Spark vs Hadoop6

Framework	File size	Number of words	Results in seconds
Hadoop	6 Mo	415173	10 sec
Spark	6 Mo	415173	8 sec

The tables from 4 to 7, just consolidates the graph (figure 15) is shows in detail the number of words the size of the file and the number of seconds manifest at the end of the execution and the layout of the Output file. The running time is managed in the algorithms.

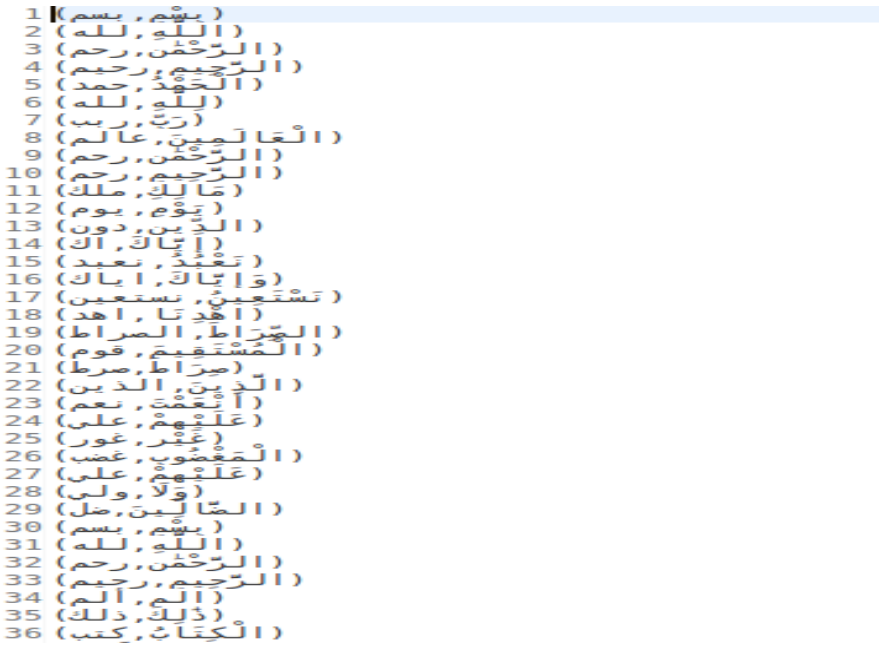


Figure 19: file example of algorithm ***OUTPUT***

16

VI. CONCLUSION

The research carried out during this thesis has optimized data pre-processing algorithms and classified them as classification techniques, applied to the classical Arabic language, as we found in the early parts of this manuscript. Several factors give a major concern of researchers for the classification of the Arabic text.

We have detailed throughout the chapters of this report what has been done in the literature to help improve the performance of data pre-processing and analysis and to fully exploit high-performing tools in the context of high volume data processing.

Indeed, in the first chapter, after defining the problem and the need to improve the algorithms of "Stemming" in the research, we synthesized the solutions already used and the algorithms applied to meet this criterion and defined the main pillars. Then we will focus on the KHOJA algorithm, and add our contributions in this regard in the sense, where the goal was to improve the algorithm on the one hand under the framework of Hadoop and Spark and make very fluid the use of these two Framework for their algorithms likely to use large volume of input data.

The first contribution is an algorithm based on Hadoop capable of processing very large data that is Arabic text taking into account its processing complexity. We proposed the Qur'an in Corpus of Data which represents a rich source of

any usual or specific morphological aspect. In a second step, we expanded the perimeter by working on specific domain data from Corpus Nada having several properties. This data is organized in a data warehouse.

The second part of our Contribution in this work and shed light on the usefulness of working with Spark and Hadoop not as much as competitor but as accompanying to have better results in result accuracy and material convenience in the air of Big Data, so we proposed an improvement based on both Framework to make further optimizations on the algorithm of root removal of words and classification of its latest for analysis and decision-making in several areas.

REFERENCES

[1] MAHDAOUY, ABDELKADER EL, ERIC GAUSSIER, ET SAID OUATIK EL ALAOUI. EXPLORING TERM PROXIMITY STATISTIC FOR ARABIC INFORMATION RETRIEVAL . IN INFORMATION SCIENCE AND TECHNOLOGY (CIST), 2014 THIRD IEEE INTERNATIONAL COLLOQUIUM IN, 272277. IEEE, 2014.

[2] "SOCIAL MEDIA USAGE IN MIDDLE EAST - STATISTICS AND TRENDS (INFOGRAPHICS)," GO-GULF, 4 JUN 2013. (ONLINE) . AVAILABLE: HTTPS://WWW.GOGULF.COM/BLOG/SOCIAL-MEDIA-MIDDLE-EAST (ACCESSED NOVEMBRE 2014)

[3] BUCKWALTER,T.QAMUS: ARABIC LEXICOGRAPHY HTTP://MEMBERS.AOL.COM/ARABICLEXICONS/

[4] KHOJA, S.AND GARSIDE, R. STEMMING ARABIC TEXT. COMPUTING DEPARTMENT, LANCASTER UNIVERSITY, LANCASTER. INTERNET HOME PAGE, 1-7 (1999). [AVAILABLE AT:HTTP://WWW.COMP.LANCS. AC.UK/COMPUTING/USERS/KHOJA/STEMER.PS]

[5] GROLINGER, KATARINA; HAYES, MICHAEL; HIGASHINO, WILSON A.; L'HEUREUX, ALEXANDRA; ALLISON, DAVID S.; AND CAPRETZ, MIRIAM A.M., "CHALLENGES FOR MAPREDUCE IN BIG DATA" (2014). ELECTRICAL AND COMPUTER ENGINEERING PUBLICATIONS. PAPER 44.

[6] BUDIMAN, R. (2013). UTILIZING SKYPE FOR PROVIDING LEARNING SUPPORT FOR INDONESIAN DISTANCE LEARNING STUDENTS: A LESSON LEARNT. PROCEDIA - SOCIAL AND BEHAVIORAL SCIENCES, 83: 5-10

[7] A. NEHAR, , ATTIA, DJELLOUL ZIADI, HADDA CHERROUN, AND YOUNES GUELLOUMA. AN EFFICIENT STEMMING FOR ARABIC TEXT CLASSIFICATION. IN INNOVATIONSIN INFORMATION TECHNOLOGY (IIT), 2012 INTERNATIONAL CONFERENCE ON,32832. IEEE, 2012.

[8] J. QIAN, D. MIAO, Z. ZHANG, AND X. YUE, "PARALLEL ATTRIBUTE REDUCTION ALGORITHMS USING MAPREDUCE," INF. SCI. (NY)., 2014.

[9] https://Hadoop.apache.org/docs/r1.2.1/hdfs_design.html, access, 2020

[10] MUSTAFA (2013) MIXED-LANGUAGE ARABIC-ENGLISH INFORMATION RETRIEVAL. PHD THESIS, UNIVERSITY OF CAPE TOWN, CAPE TOWN.

[11] DARWISH, K. AND MAGDY, W. (2014) ARABIC INFORMATION RETRIEVAL. FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL , 7, 239-342. HTTPS://DOI.ORG/10.1561/1500000031

[12] MIRKIN, B. (2010) POPULATION LEVELS, TRENDS AND POLICIES IN THE ARAB REGION: CHALLENGES AND OPPORTUNITIES. ARAB HUMAN DEVELOPMENT, REPORT PAPER 1.

[13] CHEUNG, W. (2008) WEB SEARCHING IN A MULTILINGUAL WORLD. COMMUNICATIONS OF THE ACM, 51, 32-40. HTTPS://DOI.ORG/10.1145/1342327.1342335

[14] HABASH, N. AND RAMBOW, O. (2007) ARABIC DIACRITIZATION THROUGH FULL MORPHOLOGICAL TAGGING. HUMAN LANGUAGE TECHNOLOGIES : THE CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, ROCHESTER, 22-27 APRIL 2007, 53-56.

[15] HEGAZI, N. AND EL-SHARKAWI, A. (1985) AN APPROACH TO A COMPUTERIZED LEXICAL ANALYZER FOR NATURAL ARABIC TEXT. PROCEEDINGS OF THE ARABIC LANGUAGE CONFERENCE, KUWAIT, 14-16 APRIL 1985.

[16] MUSTAFA, M. AND SULEMAN, H. (2011) BUILDING A MULTILINGUAL AND MIXED ARABIC- ENGLISH COLLECTION. 3RD ARABIC LANGUAGE TECHNOLOGY INTERNATIONAL CONFERENCE , ALEXANDRIA, 17-18 JULY 2011, 28-37.

[17] KADRI, Y. AND NIE, J.Y. (2006) EFFECTIVE STEMMING FOR ARABIC INFORMATION RETRIEVAL. PROCEEDINGS OF THE CHALLENGE OF ARABIC FOR NLP/MT CONFERENCE , LONDON, 23 OCTOBER 2006, 68-74.

[18] ATTIA, M.A. (2008) HANDLING ARABIC MORPHOLOGICAL AND SYNTACTIC AMBIGUITY WITHIN THE LFG FRAMEWORK WITH A VIEW TO MACHINE TRANSLATION. PHD THESIS, THE UNIVERSITY OF MANCHESTER, MANCHESTER.

[19] ATTIA, M.A. (2007) ARABIC TOKENIZATION SYSTEM. PROCEEDINGS OF THE 2007 WORKSHOP ON COMPUTATIONAL APPROACHES TO SEMITIC LANGUAGES : COMMON ISSUES AND RESOURCES , PRAGUE, 28 JUNE 2007, 65-72. HTTPS://DOI.ORG/10.3115/1654576.1654588

[20] GOWEDER, A., POESIO, M., DE ROECK, A. AND REYNOLDS, J. (2005) IDENTIFYING BROKEN PLURALS IN UNVOWELISED ARABIC TEXT. PROCEEDINGS OF HUMAN LANGUAGE TECHNOLOGY CONFERENCE AND CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, VANCOUVER, 6-8 OCTOBER 2005, 246-253.

[21] BUCKWALTER, T. (2004) ISSUES IN ARABIC ORTHOGRAPHY AND MORPHOLOGY ANALYSIS. PROCEEDINGS OF THE WORKSHOP ON COMPUTATIONAL APPROACHES TO ARABIC SCRIPT -BASED LANGUAGES , GENEVA, 28 AUGUST 2004, 31-34. HTTPS://DOI.ORG/10.3115/1621804.1621813

[22] ALJLAYL, M. AND FRIEDER, O. (2002) ON ARABIC SEARCH: IMPROVING THE RETRIEVAL EFFECTIVENESS VIA LIGHT STEMMING APPROACH. PROCEEDINGS OF THE 11TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT , ILLINOIS, 4-9 NOVEMBER 2002, 340-347. HTTPS://DOI.ORG/10.1145/584792.584848

[23] H S. BHOSALE1, PROF. D. P. GADEKAR2, A REVIEW PAPER ON BIG DATA AND HADOOP, INTERNATIONAL JOURNAL OF SCIENTIFIC AND RESEARCH PUBLICATIONS, 4(10),2014.

[24] LEKHA R.NAIR, DR. SUJALA,D.SHETTY, STREAMING TWITTER DATA ANALYSIS USING SPARK FOR EFFECTIVE JOB SEARCH, JOURNAL OF THEORETICAL AND APPLIED INFORMATION TECHNOLOGY ,. VOL.80. NO. 2 2005 – 2015.

[25] M. DHAVAPRIYA, N. YASODHA, BIG DATA ANALYTICS: CHALLENGES AND SOLUTIONS USING HADOOP, MAP REDUCE AND BIG TABLE, INTERNATIONAL JOURNAL OF COMPUTER SCIENCE TRENDS AND TECHNOLOGY (IJCST) – VOLUME 4 ISSUE 1, JAN - FEB 2016

[26] REYNOLD XIN, JOSHUA ROSEN, MATEI, ZAHARIA, MICHAEL J. FRANKLIN, SCOTT SHENKER, ION STOICA. SHARK: SQL AND RICH ANALYTICS AT SCALE. SIGMOD 2013. JUNE 2013.

[27] Y. SAMADI, M. ZBAKH AND C. TADONKI, "COMPARATIVE STUDY BETWEEN HADOOP AND SPARK BASED ON HIBENCH BENCHMARKS," 2016 2ND INTERNATIONAL CONFERENCE ON CLOUD COMPUTING TECHNOLOGIES AND APPLICATIONS (CLOUDTECH), MARRAKECH, 2016, PP. 267-275. DOI: 10.1109/CLOUDTECH.2016.7847709

[28] W. HUANG, L. MENG, D. ZHANG AND W. ZHANG, "IN-MEMORY PARALLEL PROCESSING OF MASSIVE REMOTELY SENSED DATA USING AN APACHE SPARK ON HADOOP YARN MODEL," IN IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 10, NO. 1, PP. 3-19, JAN. 2017. DOI: 10.1109/JSTARS.2016.2547020

[29] M. U. ÇAKIR AND S. GÜLDAMLASIOĞLU, "TEXT MINING ANALYSIS IN TURKISH LANGUAGE USING BIG DATA TOOLS," 2016 IEEE 40TH ANNUAL COMPUTER SOFTWARE AND APPLICATIONS CONFERENCE (COMPSAC), ATLANTA, GA, 2016, PP. 614-618. DOI: 10.1109/COMPSAC.2016.203

[30] SARD M, BUDIU M, YU Y, BIRRELL A. DYRAD: DISTRIBUTED DATA-PARALLEL PROGRAMS FROM SEQUENTIAL BUILDING BLOCKS. 2ND ACM SIGOPS/EUROSYS EUROPEAN CONFERENCE ON COMPUTER SYSTEMS 2007. 2007. PP. 59–72.

[31] A. VERMA, A. H. MANSURI AND N. JAIN, "BIG DATA MANAGEMENT PROCESSING WITH HADOOP MAPREDUCE AND SPARK TECHNOLOGY: A COMPARISON," 2016 SYMPOSIUM ON COLOSSAL DATA ANALYSIS AND NETWORKING (CDAN), INDORE, 2016, PP. 1-4. DOI: 10.1109/CDAN.2016.7570891

[32] https://www.infoq.com/articles/apache-Sparkintroduction

Convolutional Neural Networks and Long Short Term Memory for Phishing Email Classification

Regina Eckhardt
Department of Computer Science
University of West Florida
Pensacola, FL, USA
reginaeckhardt20@gmail.com

Sikha Bagui
Department of Computer Science
University of West Florida
Pensacola, FL, USA
bagui@uwf.edu

***Abstract*— The focus on this work is on classifying phishing emails using deep neural networks. Since phishing emails have no specific characteristic, they are difficult to detect and classify, and little research has been done on the detection of phishing emails. In this work, two deep neural networks, Long Short Term Memory (LSTM), a form of Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN), were compared and used for classification of phishing emails. RNN is the most used neural network for text classification. CNNs have also shown to be effective in text classification. In addition to tuning hyperparameters, different activation functions and optimizers are used for comparing the performance of CNN and LSTM on the basis of accuracy and the ROC-score. LSTM achieved a higher accuracy than CNN, and overall the Adam Optimizer performed better than the SGD optimizer. The best parameters for higher accuracy and ROC-score are also presented.**

Keywords: Phishing Email Classification; Convolutional Neural Networks; Long Short Term Memory; Hyperparameters; Recurrent Neural Networks; Deep Learning

I. INTRODUCTION

Phishing is a cleverly crafted social engineering attack that is characterized by an attacker imitating a trustworthy source to gain confidential and private information from a user for malicious purposes [1, 2]. Phishing attacks, primarily carried out via email [3] or other electronic communication channels [4], affect both businesses and private individuals [2], and since emails are widely used in both personal and professional contexts, phishing has become a rising threat [5]. In 2019 alone, more than 114,000 private individuals in the US lost in total more than $57.8 million through phishing [6]. In the same year, 90% of all organizations experienced targeted phishing attacks [7].

Phishing attackers also exploit any crisis, and the coronavirus outbreak has been no different [8]. Google's Threat Analysis Group reported that they blocked 18 million COVID-19 phishing emails per day in mid-April of 2020 [8]. Amidst the spike in remote work tied to the COVID-19 pandemic, phishing campaigns have even been targeting remote working software like Skype and Zoom [9]. Attackers sent phishing emails looking eerily similar to legitimate pending notifications coming from Skype or Zoom, with a sender's address that appears very legitimate at first and a very convincing landing page [9].

The focus on this work is on classifying phishing emails. Since phishing emails have no specific characteristic, they are difficult to detect and classify, and little research has been done on the detection of phishing emails. In this work, two different deep neural networks, Long Short Term Memory (LSTM), a form of Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN), were compared and used for classification of phishing emails. RNN is the most used neural network for text classification. CNNs have also shown to be effective in text classification. In this work, in addition to tuning hyperparameters, different activation functions and optimizers are used for comparing the performance of CNN and LSTM on the basis of accuracy and the ROC-score.

The rest of this paper is organized as follows. Section 2 presents the related works. Section 3 presents the data and preprocessing technique used. Section 4 presents the algorithms, CNN, RNN and LSTM. Section 5 presents the experimental design and parameters and section 6 presents the results and discussion. The last section, Section 7, is the conclusion.

II. RELATED WORKS

Several papers have shown that the CNN shows good performance for text classification. [10] applied CNN for sentence classification and achieved good results by using hyperparameter tuning. [10] included a dropout on the last layer and found the best hyperparameters by applying a grid search on one of the development data sets. In this work, six data sets were used to train and evaluate the performance of CNN for text classification. Several variations of the CNN were implemented and compared. The result was that even the simple CNN-model performed pretty well, but improvements can be made by fine-tuning (e.g. multi-channel for static CNN).

[11] performed an empirical exploration on the user of character-level CNN (ConvNets) for text classification, and showed how CNN can achieve competitive results. In this work, One-Hot Encoding was applied on characters and compared to traditional models such as bag of words, *n*-grams and their TFIDIF variants. This work compared CNN with LSTM.

Another work that made use of the One-Hot Encoding for text classification was [12]. This work analyzed the use of

discrete-time RNN and it's capability for predicting the next symbol in a sequence in order to implement a model. This study, however, focused more on online prediction.

[13] used RNN to classify textual data. In this work, on the basis of RNNs, three different architectures are implemented and compared. All showed good performance in classification of textual data.

[14] focused on extracting and analyzing features of emails, and using the most important features for a multi-classifier prediction model. J48, SVM and IB1 were used in the multi-classifier prediction model. The dataset used in this study was a publicly accessible dataset. These results showed that a single classifier would not be enough to classify phishing emails, and they were able to achieve almost perfect accuracy with their multi-classifier model (a low false positive rate).

[15] also covered the classification of phishing. Features that were used in this work were, age of the domain name, IP address, number of domains or key words.

III. THE DATA AND PREPROCESSING

A. The Data

The dataset was provided by AppRiver, a company headquartered in Pensacola, Florida, USA, that offers secure cloud-based cybersecurity solutions. The dataset contains 18,365 emails, of which 3,416 are phishing emails.

B. Preprocessing: One-Hot Encoding

Phishing data, rather, phishing email data, is mostly text data. One-Hot Encoding was applied to transform this textual data into numerical vectors. One-Hot Encoding works as follows: It takes a word and assigns a vector containing only zeros and ones to that word. This vector then represents the specific word. To avoid the assignment of different outcomes of the same word to different vectors, tokenizing and lemmatizing methods are applied. Tokenizing refers to splitting the text into tokens, in our case words, where each token will be preprocessed separately. Lemmatizing is the transformation of words back to their stem using morphological techniques [16].

After lemmatizing, the vocabulary size needs to be set for the one-hot encoding. The vocabulary size describes how many different words should be assumed in the data. There are different approximations possible based on the size of the data. In this work the vocabulary size was set to 100. After setting the vocabulary size, each word will be one-hot encoded until all unit vectors are assigned to words. The remaining words will be assigned to a PAD vector, a vector only containing zeros.

Also, the maximum document length and the vocabulary size needs to be specified. The vocabulary size was set to 100 and the maximum document length to 70.

IV. ALGORITHMS

A. Convolutional Neural Networks

Studies have shown that CNN has performed well with text classification [10, 11]. CNN consists of several layers of convolutions. Activation functions are applied to the results of the layers of convolutions. CNN tries to find features that best characterize the data by using filters throughout the convolutional layers. Each layer applies different filters. The filter and its values are learned by the network during the training phase. For each filter (each text) the output of the convolutional layer is a vector. This vector contains values that are the sum of the component wise multiplication when applying the filter on each region of the matrix containing the text [17].

In the pooling layer, the results of the convolutional layer will be pooled. There are different ways of pooling, e.g. max-pooling or average-pooling. Max-pooling takes the maximum value of each vector while average-pooling calculates the average of each vector. The last layer of CNN is the fully connected layer. This layer combines the results from the prior layers, i.e. the extracted features with the classification of the data. It finds the features that best characterize the data in order to classify the data correctly [17].

B. Recurrent Neural Networks

Studies have also shown that RNN has been successful in text classification [11, 13]. RNNs are a neural network in which each neuron receives input from a prior neuron, produces an output for the next neuron and sends an output back to itself. This means the input a neuron gets at the time step t is the input corresponding to this time step and its output from the previous time step. Each input is weighted. Taking into account not only the previous time step, but all the time steps before the output of a neuron, is a function of all the inputs from the previous time steps. There are different ways of implementing RNN. The RNN of interest for phishing email classification feeds the network with a sequence of inputs and ignores all outputs except for the last one, i.e. sequence-to-vector network. This work focuses on the recurrent neural network, Long Short Term Memory (LSTM).

1) Long Short Term Memory

LSTM has been successfully used in text classification by [11]. The LSTM cell can be considered a black box, like a basic cell but with better performance. The training converges faster and it detects long-term dependencies in the data. The state of the LSTM is split in two vectors: $h(t)$ and $c(t)$ where $h(t)$ describes the short-term state and $c(t)$ the long-term state. The long-term state $c(t–1)$ goes through the network, passes a forget gate (where some memories can be dropped) and the addition operation to add new memory. The current input vector $x(t)$ and the previous short-term state $h(t–1)$ are fed to four different fully connected layers: The main layer is the one that outputs $g(t)$. The other layers are gate controllers: the forget gate, the input gate and the output gate. This means that, in general, an LSTM cell can learn to recognize an important input, store it in the long-term state, learn to keep it and learn to extract it whenever it is needed [18].

C. Activation Functions

In order to process the data through the different layers, activation functions are used. These functions are used to transform the activation level of a unit into an output signal. This specifies how much information of a unit should be transported to the next layer. In this work, the Sigmoid and ReLU activation functions were used. Figure 1 illustrates the Sigmoid and ReLU activation functions respectively.

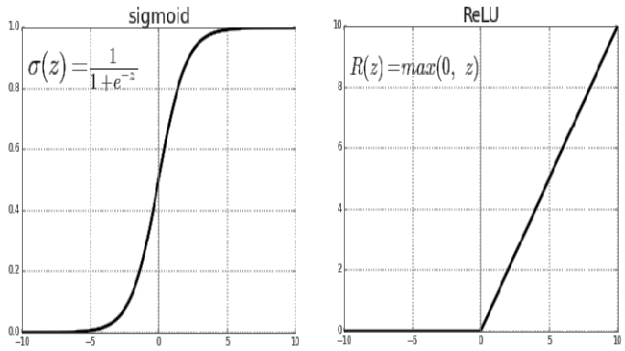


Figure 1. Sigmoid and ReLU Activation Functions [25]

1) The Sigmoid Activation Function

The sigmoid activation function assigns a positive number between 0 and 1 to every input. The sigmoid activation function, one of the most used is most activation functions [19], is useful for training data that is between 0 and 1.

2) The Rectified Linear Unit

ReLU, the rectified linear unit, is conventionally applied as an activation function for hidden layers in a deep neural network but it can also be used to learn the weight parameters of the ReLU classification layer through backpropagation. It is defined as a maximum of zero and a specific input. This means that it always assigns a value greater or equal to zero [20].

V. Experimentation

This section describes the parameters used for CNN and LSTM experimentations respectively.

A. Convolutional Neural Networks

First, the parameters that are used for CNN are described, and the next section describes how these parameters have been used.

1) Parameters Used for Convolutional Neural Networks

The experimental parameters that were used for CNN were: Number of channels, pooling method, dropout, Kernel size and the optimizer. Next the experimental parameters are explained.

Number of channels. The number of channels was kept at three.

Pooling method. Max-pooling takes the maximum value of each resulting vector of the convolutional layer while Average-pooling uses the average value of all entries of that vector [21]. Both Max-pooling and Average-pooling were used in this work.

Dropout. Dropouts are included in neural networks in order to avoid overfitting by dropping units when they exceed a specific value. Dropping a unit means excluding this unit and all connections of that unit in the neural network. A common value for the dropout rate is 0.5 [22]. The dropout rate was set to 0.5 in this work.

Kernel size. The kernel size is the size of the filters in the convolutional layer. Since it is necessary to use the same number of columns for the filter as the one-hot encoded matrix, a variation can be included in the number of rows of the filter. Commonly used numbers of rows, i.e. kernel sizes, are used in this work. The kernel size or the length of the 1d convolution window was set to 4, 6 and 8 (three different sizes for the three channels respectively).

Optimizer. Optimizer methods try to reduce the error of the model which can be measured by a loss function. Optimizing minimizes the loss function. The first optimizing method used was the stochastic gradient descent (SGD). This is a vector of partial derivatives. From this, the settings for minimal error can be calculated and the weights adjusted by using backpropagation [23]. The Adam Optimizer is generally better for working with high-dimensional parameter spaces. It combines the advantages of the AdaGrad (good performance with sparse gradients) and RMSProp (good performance for non-stationary settings) [26]. The Adam Optimizer and the Stochastic Gradient Descent were used in this work.

2) Experimental Design for CNN

Figure 2 presents the architecture for the design of experiments for CNN. For the filter sizes of 16, 32, 48 and 64, the following eight different settings were used. For Setting 1, the Sigmoid Activation Function, Max-pooling and Adam Optimizer were used. For setting 2, the Sigmoid Activation Function, Max-pooling and SGD optimizer were used, and so on.

Figure 2. Experimental Design for CNN

B. *Experimental Parameters for Long Short Term Memory*

The experimental parameters that were used for LSTM were, the optimizer, the maximum email length, the batch size and the number of epochs.

Optimizer. The Adam Optimizer and the Stochastic Gradient Descent were used.

Maximum Email Length. This was set to 500.

Batch size. Since the whole dataset should not be processed through the network at once, the dataset was split into batches, i.e. into smaller datasets. These were then fed to the neural network. The batch size was set to the commonly used batch size of 64.

The number of epochs. One epoch is when the whole dataset is passed forward and backward through the neural network only one time. Since optimizing a neural network is an iterative process, it is more efficient to not pass the whole dataset through the network at once. Low number for epochs can lead to underfitting (e.g. if the number of epochs is set to one), while a high number of epochs can lead to overfitting. Therefore, a balance has to be achieved to find the optimal balance between over- and underfitting [24]. In this work, the number of epochs was varied from 1 to 3.

1) *Experimental Design for LSTM*

Figure 3 presents the architecture for the design of experiments for LSTM. For Setting 1, the maximum email length was set to 500, batch size to 64, using one epoch, the Adam Optimizer, and the Sigmoid Activation Function. For setting 2, the maximum email length was set to 500, batch size to 64, using one epoch, the Adam Optimizer, and the ReLU Activation Function, and so on.

Figure 3. Experimental Design for LSTM

VI. Results and Discussion

In this section, first the measures of performance are presented, then the CNN and LSTM results are presented.

A. *Measures of Performance*

For a measure of performance, the accuracy, true/false positive rates and ROC-Score were used to measure the performance of the neural networks.

1) *Accuracy*

Accuracy is the ratio of correctly classified data.

2) *True / False Positives*

The false positive rate (FPR) is the ratio of non-phishing emails that were incorrectly classified as phishing emails. It can also be calculated by subtracting the ratio of correctly classified non-phishing emails from one. The true positive rate is the ratio of phishing emails that were correctly classified as phishing emails. The better a model, the higher the true positive rate, and the lower the false positive rate. The ROC curve plots TPR against FPR.

True and False Positive rates give a better impression of the fit of a model than accuracy, especially for data where the ratio of the classes is not equal. In this study, less than 20% of the total dataset were phishing emails. Hence, using only accuracy as a measure of the model fit can give an unusually high accuracy, which might not be correct.

3) The ROC-Curve

The ROC-curve shows the tradeoff between the true positive rates and the false positive rates (any increase in the true positive rates will be accompanied by a decrease in the false positive rates). The closer the curve follows the left border and the top border of the ROC space, the more accurate the test.

B. CNN Results

Figure 4 graphs each of the settings from Figure 2 by the number of filters and accuracy. Figure 4 shows that the number of filters did not have much of an effect on the accuracy of CNN. The settings, however, have an effect on the accuracy of CNN, hence this was further analyzed in Figures 5-8.

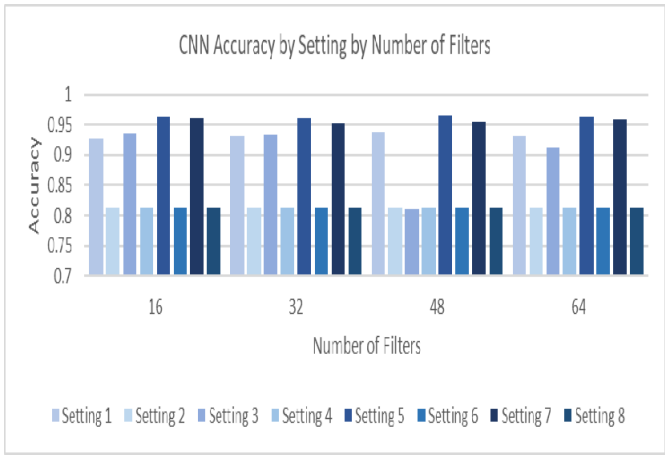


Figure 4. CNN Accuracy by Setting by Number of Filters

Figures 5-8 graph CNN accuracy by activation function for 16, 32, 48 and 64 filters respectively. In each of these figures, the first four bars of represent the usage of the sigmoid activation function and the second four bars represent the usage of the ReLU activation function.

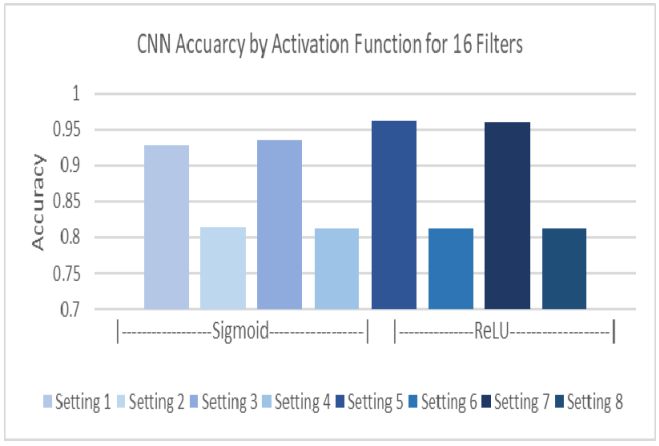


Figure 5. CNN Accuracy by Activation Function for 16 Filters

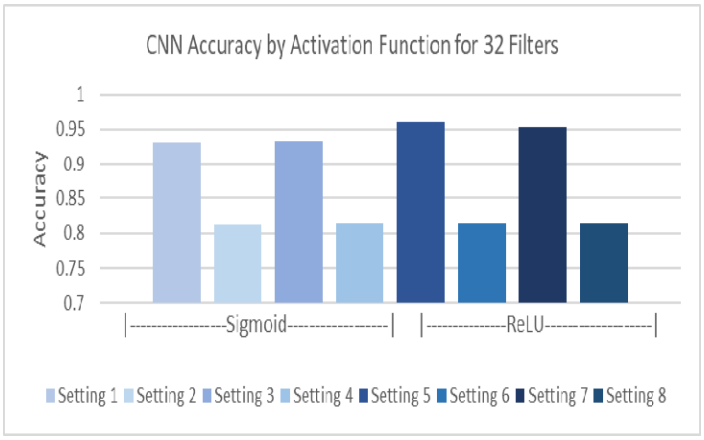


Figure 6. CNN Accuracy by Activation Function for 32 Filters

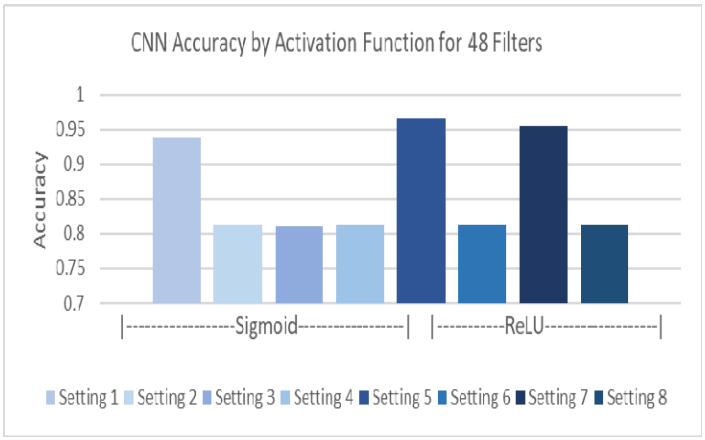


Figure 7. CNN Accuracy by Activation Function for 48 Filters

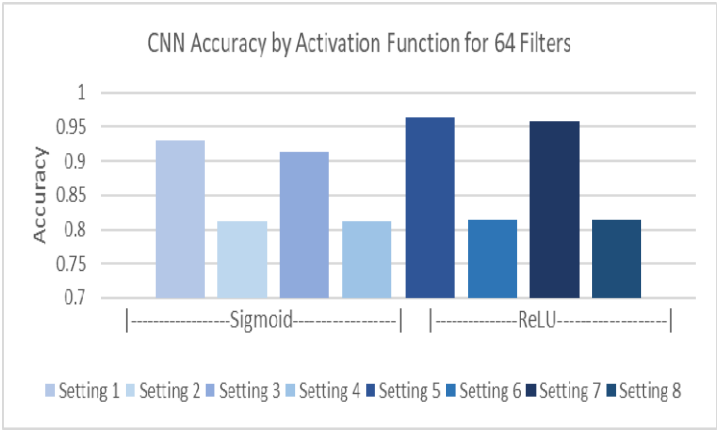


Figure 1. CNN Accuracy by Activation Function for 64 Filters

Figures 5 – 8 show that:

- Overall, the ReLU Activation function performed better than the Sigmoid Activation function.
- The Adam Optimizer performed better than the SGD Optimizer, both for Average-pooling as well as Max-pooling, for both the Sigmoid as well as the ReLU activation functions.

Detailed results for the various runs, as per settings presented in Figure 2, are presented in Tables 1-8. Table 1 presents the results of the runs for Setting 1, Table 2 are the results of the runs for Setting 2, and so forth.

Number of Filters	Accuracy	ROC-Score
16	92.82 %	95.15 %
32	93.15 %	95.67 %
48	93.84 %	94.11 %
64	93.07 %	93.92 %

Table 1. Setting 1: Kernel size 4,6,8, Sigmoid activation function, 3 channels, Max-pooling, Adam optimizer

Number of Filters	Accuracy	ROC-Score
16	81.36 %	50.00 %
32	81.30 %	50.00 %
48	81.23 %	50.00 %
64	81.31 %	50.00 %

Table 2. Setting 2: Kernel size 4,6,8, Sigmoid activation function, 3 channels, Max-pooling, SGD optimizer

Number of Filters	Accuracy	ROC-Score
16	93.48 %	95.55 %
32	93.30 %	95.08 %
48	81.01 %	50.00 %
64	91.25 %	94.89 %

Table 3. Setting 3: Kernel size 4,6,8, Sigmoid activation function, 3 channels, Average-pooling, Adam optimizer

Number of Filters	Accuracy	ROC-Score
16	81.28 %	50.00 %
32	81.35 %	50.00 %
48	81.28 %	50.00 %
64	81.29 %	50.00 %

Table 4. Setting 4: Kernel size 4,6,8, Sigmoid activation function, 3 channels, Average-pooling, SGD optimizer

Number of Filters	Accuracy	ROC-Score
16	96.25 %	94.10%
32	96.10 %	96.28%
48	96.52%	96.16%
64	96.36%	96.16%

Table 5. Setting 5: Kernel size 4,6,8, ReLU activation function, 3 channels, Max-pooling, Adam optimizer

Number of Filters	Accuracy	ROC-Score
16	81.25 %	50.00 %
32	81.36 %	50.00 %
48	81.29 %	50.00 %
64	81.34 %	51.59 %

Table 6. Setting 6: Kernel size 4,6,8, ReLU activation function, 3 channels, Max-pooling, SGD optimizer

Number of Filters	Accuracy	ROC-Score
16	96.04 %	96.05 %
32	95.27 %	94.95 %
48	95.57 %	96.15 %
64	95.86 %	95.91 %

Table 7. Setting 7: Kernel size 4,6,8, ReLU activation function, 3 channels, Average-pooling, Adam optimizer

Number of Filters	Accuracy	ROC-Score
16	81.20 %	50.00 %
32	81.35 %	50.00 %
48	81.24 %	50.00 %
64	81.32 %	50.00 %

Table 8. Setting 8: Kernel size 4,6,8, ReLU activation function, 3 channels, Average-pooling, SGD optimizer

From Tables 1-8, it can be observed that:

- The highest accuracy (96.52%) as well as highest ROC-Score (96.16%) was obtained with Setting 5: Kernel size 4, 6, 8, ReLU activation function, 3 channels, Max-pooling, Adam optimizer, with 48 filters. The other filter sizes for this Setting, Setting 5, also had higher accuracy and ROC-scores than any other settings.
- The second highest accuracy as well as second highest ROC-score group can be considered as Setting 7: Kernel size 4, 6, 8, ReLU activation function, 3 channels, Average-pooling, Adam optimizer, with 16

filters, though the performance of the other filters (for this setting) were close.

- The Adam optimizer performed better than the SGD optimizer.
- Both the Sigmoid and ReLU Activation functions performed better with the Adam Optimizer.
- Both Max-pooling as well as Average-Pooling performed better with the Adam Optimizer.
- For CNN accuracy, it is difficult to say if the Sigmoid Activation performed better or the ReLU Activation function performed better. For the Sigmoid Activation function, Settings 1 and 3 gave good results, and for the ReLU Activation function, Settings 5 and 7 gave good results. Hence, other factors have to be considered besides the Activation function.

1) ROC-Curves

Figures 9 and 10 show the ROC-Curves for the ReLU and Sigmoid Activation functions respectively, for the best performing settings. The closer the curve follows the left border and the top border of the ROC space, the more accurate the test. For both of the graphs, Figures 9 and 10, we can observe that both activation functions had high results in terms of correctly classified data.

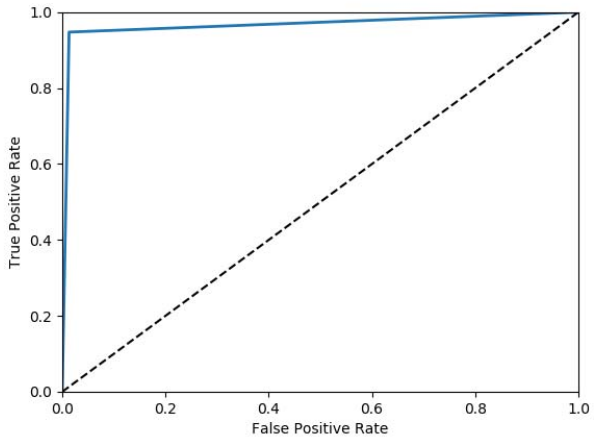


Figure 9. ROC Curve for CNN using the ReLU Activation function (the best performing settings)

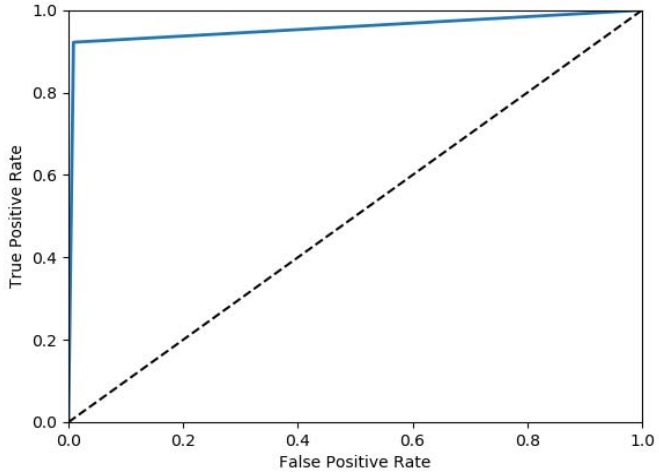


Figure 10. ROC Curve for CNN using the Sigmoid Activation function (the best performing settings)

C. Long Short Term Memory Results

Figure 11 presents the LSTM accuracy by Activation function by each setting as per Figure 3 (the color codings are matched up to the color coding used in Figure 3 for the different settings, hence the legends are omitted in the following Figures).

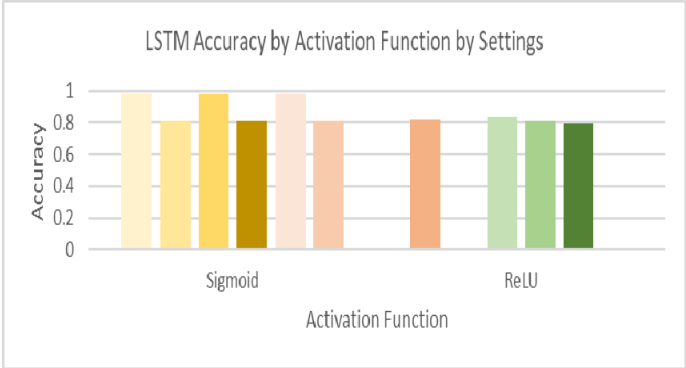


Figure 11. LSTM Accuracy by Activation Function by Settings

Comparing the performance of LSTM for the two activation functions, Sigmoid and ReLU, from Figure 11 it can be observed that the sigmoid activation function performed better than the ReLU activation function for the LSTM model. In some cases, for the ReLU activation function, the LSTM model even had an accuracy of 0%.

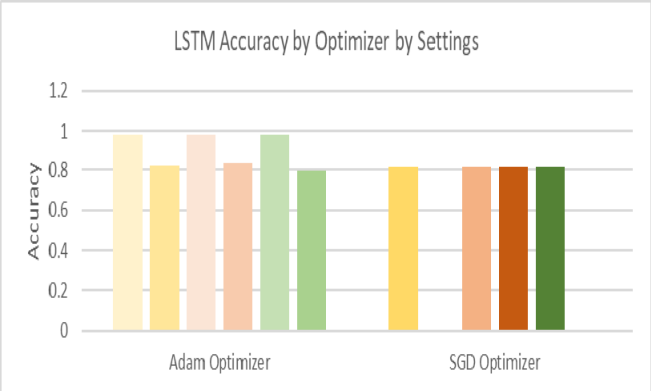


Figure 12. LSTM Accuracy by Optimizer by Settings

Figure 12 compares the performance of LSTM for the two optimization methods used, for the different settings. As per Figure 12, the Adam optimizer performed better than the SGD Optimizer. The SGD optimizer had 0% accuracy for some runs.

The accuracy and ROC-score for the various runs as per the settings in Figure 3 are presented in Tables 9 – 14. The highest accuracy was obtained by the Setting 9: the Adam optimizer, max email length= 500, batch size= 64, 3 epochs, with the Sigmoid Activation function. This was an accuracy of 98.32% and the ROC-Score was 96.57%. The second highest accuracy was obtained using Setting 5: Adam optimizer, max email length= 500, batch size= 64, 2 epochs, with the Sigmoid Activation function. The second highest accuracy was 98.02% and the ROC-Score was 95.71%. And, the third highest accuracy was obtained using Setting 1: Adam optimizer, max email length= 500, batch size= 64, 1 epoch, with the Sigmoid Activation Function. The third highest accuracy was 97.85% and ROC-Score was 95.83%. All three highest accuracy scores obtained by LSTM were higher than the CNN accuracy scores.

Activation Function	Accuracy	ROC-Score
Sigmoid	97.85 %	95.38 %
ReLU	81.88 %	94.95 %

Table 9. Setting 1 & 2: Adam optimizer, max email length= 500, batch size= 64, 1 epoch

Activation Function	Accuracy	ROC-Score
Sigmoid	81.28 %	50.00 %
ReLU	0.00 %	50.00 %

Table 9. Setting 3 & 4: SGD optimizer, max email length= 500, batch size= 64, 1 epoch

Activation Function	Accuracy	ROC-Score
Sigmoid	98.02 %	95.71 %
ReLU	83.56 %	95.18 %

Table 10. Setting 5 & 6: Adam optimizer, max email length= 500, batch size= 64, 2 epochs

Activation Function	Accuracy	ROC-Score
Sigmoid	81.28 %	50.00 %
ReLU	81.28 %	50.00 %

Table 11. Setting 7 & 8: SGD optimizer, max email length= 500, batch size= 64, 2 epochs

Activation Function	Accuracy	ROC-Score
Sigmoid	98.32 %	96.57 %
ReLU	79.79 %	57.28 %

Table 12. Setting 9 & 10: Adam optimizer, max email length= 500, batch size= 64, 3 epochs

Activation Function	Accuracy	ROC-Score
Sigmoid	81.28 %	50.00 %
ReLU	0.00 %	50.00 %

Table 13. – Setting 11 & 12: SGD optimizer, max email length= 500, batch size= 64, 3 epochs

VII. Conclusion

The highest accuracy was achieved with the LSTM model, with an accuracy of 98.32% and a ROC-Score of 96.57%. This conforms with previous studies that have shown that RNNs are a reasonable method of classifying textual data. Nevertheless, CNN's highest accuracy of 96.52% and ROC-Score of 96.16% was pretty close. The histograms will show, for both CNN and LSTM, that the Adam Optimizer always performs better than the SGD optimizer. These results are not surprising since the Adam Optimizer makes use of the characteristics of the SGD and also includes features of another optimizer. For the activation function, however, there is not a clear answer. The ReLU activation function performed better with CNN, but on the average, the Sigmoid activation function performed with LSTM.

Acknowledgments

This work was partially supported by the Askew Institute of the University of West Florida.

REFERENCES

[1] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. I. Hong, "Teaching johnny not to fall for phish," ACM Trans. Internet Techn., 10:7:1–7:31, 2010.

[2] D. Pienta, J. Thatcher, and A. Johnston, "A Taxonomy of Phishing: Attack Types Spanning Economic, Temporal, Breadth, and Target Boundaries," In Proceedings of the 13th Pre-ICIS Workshop on Information Security and Privacy, San Francisco, 2018.

[3] C. Nguyen, "Learning Not To Take the Bait: An Examination of Training Methods and Overlerarning on Phishing Susceptibility," PhD thesis, 2018.

[4] S. Abu-nimeh, D. Nappa, X. Wang, and S. Nair, "Distributed Phishing Detection by Applying Variable Selection using Bayesian Additive Regression Trees," 2009.

[5] B. B. Gupta, N. A. Arachchilage, and K. E. Psannis, "Defending against phishing attacks: taxonomy of methods, current issues and future directions," Telecommunication Systems, vol. 67(2), pp.247–267, 2018.

[6] A. Holmes, "Hackers are getting better at tricking people into handing over passwords - here's what to look out for, according to experts," July 2020. Hackers are getting better at tricking people into handing over passwords — here's what to look out for, according to experts (msn.com)

[7] G. Egan, "2019 State of the Phish Report: Attack Rates Rise, Account Compromise Soars," 2019. https://www.proofpoint.com/us/corporate-blog/post/2019-state-phish-report-attack-rates-rise-account-compromise-soars

[8] T. Kelly, Security, "How hackers are using COVID-19 to find new phishing victims," June 2020. https://www.securitymagazine.com/articles/92666-how-hackers-are-using-covid-19-to-find-new-phishing-victims

[9] J Davis, Health ITSecurity, "New COVID-19 Phishing Campaigns Target Zoom, Skype User Credentials," April 2020. https://healthitsecurity.com/news/new-covid-19-phishing-campaigns-target-zoom-skype-user-credentials

[10] Y. Kim, "Convolutional Neural Networks for Sentence Classification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1746-1751, 2014. DOI: 10.3115/v1/D14-1181

[11] X. Zhang, J. Zhau, Y. LeCun, "Character-level Convolutional Neural Networks for Text Classification," CoRR, vol. 1509.01626, 2016. arXiv:1509.01626v3

[12] J. A. Pérez-Ortiz, J. Calera-Rubio, and M. L. Forcada, "Online Text Prediction with Recurrent Neural Networks," Neural Processing Letters, vol. 14, pp. 127–140, 2001. DOI: https://doi.org/10.1023/A:1012491324276

[13] P. Liu, X. Qiu, and X. Huang, "Recurrent Neural Network for Text Classification with Multi-Task Learning," CoRR, vol. 1605.05101, 2016. arXiv:1605.05101v1

[14] S. Sarju, R. Thomas, and E. C. Shyni, "A Multi-Classifier Prediction Model for Phishing Detection," International Journal of Research in Engineering and Technology, vol. 3(3), pp. 31 – 34, 2014.

[15] R. Basnet, S. Mukkamala, and A. H. Sung, "Detection of Phishing Attacks: A Machine Learning Approach," In Prasad B. (eds). Soft Computing Applications in Industry. Studies in Fuzziness and Soft Computing, vol. 226, 373, 383, 2008. DOI: https://doi.org/10.1007/978-3-540-77465-5_19

[16] Stemming and lemmatization. Cambridge University Press, 2008. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html [Accessed 12 April 2021]

[17] H. Mouzannar, "CNN for Text Classification," 2019. https://www.researchgate.net/figure/CNN-for-Text-Classification-adapted-from-Kim-2014_fig1_325386740 [Accessed 20 February 2021]

[18] A. Gerón, "Hands-on Machine Learning with Scikit-Learn and Tensorflow," Concepts, Tools, and Techniques to build intelligent Systems, pp.320-321, 2017.

[19] P. Sibi, S. A. Jones, and P. Siddarth, "Analysis of different activation functions using back propagation neural networks," Journal of Theoretical and Applied Information Technology, vol. 47(3), pp. 1264-1268, 2013.

[20] A. Agarap, "Deep Learning using Rectified Linear Units (ReLU)", Cornell University. https://arxiv.org/abs/1803.08375

[21] J. Brownlee, "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks," 2019. https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salukhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, vol. 15(56), pp.1929-1958, 2014.

[23] A. S. Walia, "Types of Optimization Algorithms used in Neural Networks and Ways to Optimize Gradient Descent," 2017. https://towardsdatascience.com/types-of-optimization-algorithms-used-in-neural-networks-and-ways-to-optimize-gradient-95ae5d39529f

[24] S. Sharma, "Activation Functions in Neural Networks," 2017. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6 [Accessed 20 April 2021]

[25] Deep Autoencoders, 2015. https://skymind.ai/wiki/deep-autoencoder [Accessed 2 April 2021]

[26] D. Kingma, J. Lei Ba, and Adam, "A Method For Stochastic Optimization", Proceedings of ICLR 2015.

AUTHOR PROFILES

Regina Eckhardt was an Exchange Student at The University of West Florida. She completed has a Master's from University of Ulm in Germany. Her interests are in the area of data science.

Dr. Sikha Bagui is Professor and Askew Fellow in the Department of Computer Science, at The University West Florida, Pensacola, Florida. Dr. Bagui is active in publishing peer reviewed journal articles in the areas of database design, data mining, Big Data analytics, machine learning and AI. Dr. Bagui has worked on funded as well unfunded research projects and has over 80 peer reviewed publications, some in highly selected journals and conferences. She has also co-authored several books on database and SQL. Bagui also serves as Associate Editor and is on the editorial board of several journals.

Security Assessment of Authentication Protocols in Mobile Adhoc Networks

Megha Soni
Assistance Professor
SVCE, Indore India
meghasoni@svceindore.ac.in

Brijendra Kumar Joshi
Professor MCTE,
MCTE, Mhow India
brijendrajoshi@yahoo.com

***Abstract*—Mobile Ad-hoc Networks can be easily targeted by different attacks such as Denial of Service, Wormhole and Man-in-the-Middle. A packet authentication is required in wireless communication to combat such attacks from outsider nodes. We have found out the parameters by which security of different protocols like HEAP, LHAP, TESLA, and Lu and Pooch's can be assessed. Also vulnerabilities of these protocols for different attacks have been discussed. To grade the performance of these authentication protocols, different parameters such as throughput, latency and packet delivery ratio have been used.**

Index Terms—HEAP, Authentication, Security, MANETs

I. Introduction

Mobile Ad-hoc Networks (MANETs) are getting noteworthy attention from industry and research area due to its versatile applications and security challenges. For example, in military applications, for infrastructure-less networks, MANETs are able to exchange strategic information and perform with high mobility. MANETs are also suitable for mobile conferencing in a big group of people. It can be quickly installed in emergencies like disaster management situations [1].

MANETs have to offer different levels of security in various applications for their successful utilization. However, due to absence of central authority and wireless links among nodes, they have much greater security issues than wired networks. An attacker can easily join or leave and snoop a network, as physical link is not required. Their aim is to disrupt the network, drop packets or inject fake packets. As a result, it is easy to launch Denial of Service (DoS) attack, Man-in-the-Middle attack and Wormhole attacks or imitate another node.

To improve MANET's security, different schemes such as secure routing using symmetric and asymmetric cryptography for key establishment and distribution have been proposed [2]. But, all these protocols are able to authenticate only control packets. If these are used for authentication of data packets, network overhead would increase. On the other hand, unauthenticated data packets make protocols vulnerable for different routing attacks, as it is essential to authenticate control and data packets both to provide guard against different attacks. Many authentication protocol like Hop-by-Hop Efficient Authentication Protocol (HEAP), Lightweight Hop-by-Hop Authentication Protocol (LHAP), Timed Efficient Stream Loss-tolerant Authentication (TESLA) and Lu and Pooch's algorithm have been designed to authenticate both types of packets.

II. Authentication In Manets

In this process, an authentication protocol is used by authenticator to verify credentials presented by a supplicant. In this way supplicant's access privileges are established by the authenticator. Such an authentication protocol may also use a Trusted Third Party (TTP) during such a process. Here an supplicant is defined as an entity seeking access of protected resources through an authenticator, which is an entity that controls access to some resources. Further, it makes authentication decisions during the authentication process. A sequence of messages are exchanged between entities to identify each other. Here either supplicant or authenticator distribute secrets or allow secrets to be recognized. Further, an identifier that is used to authenticate a supplicant with high confidence is called a credential. Also, an entity trusted by both, supplicant and authenticator, is called TTP.

III. Need of Authentication Protocols

Communication links in MANETs, in contrast to fixed networks, are open shared medium. As a result, communication between neighboring nodes is more vulnerable to attacks. In MANETs, due to constrained resources; limited battery power, small computational capacity and rapid changes in topography; both data packet delivery and authentication protocol used for routing need to be scalable and light weight. In MANETs, techniques such as asymmetric cryptography, being very intensive, are prohibitively insufficient due to associated computational complexity and message overhead. Contrastingly, symmetric cryptographic algorithms are fast; yet

complex in key maintenance, thus it creates difficulty in authentication of multicast and broadcast communication.

Need for efficient and large-scale data dissemination is driving popularity of broadcast communication. Ability of broadcast networks to distribute packets to multitude of receivers also frequently facilitates malicious users to impersonate as a sender and inject packets in a broadcast network. This gives rise to need for authentication protocol which will enable receivers to verify that a given received packet was indeed sent by the claimed sender.

Appending a Message Authentication Code (MAC); generated by use of a shared secret key; as usually deployed in point-to-point authentication mechanism; is actually insufficient to provide secure broadcast authentication. This is because any receiver with a secret key can forge data and function as a sender. To prevent such attacks, asymmetric cryptographic protocol becomes a natural choice. While its action of signing each data packet indeed provides secured broadcast authentication, it is associated with high overhead, time required to sign and verify as well as the consequent use of bandwidth.

There are many techniques which gradually reduce this overhead by using single signature over several packets; yet none of those offer complete satisfaction about their bandwidth deployed, scalability and processing time in network. And as against this, serious vulnerabilities against attacks; like DoS, replay attack, Man-in-the-Middle attack and Wormhole attack are possible If data packets are unauthenticated, loss of robustness to packets loss are observed. Serious attacks like DoS are possible If an attacker floods the receiver with bogus packets supposedly containing a signature while authentication deploying schemes amortize a digital signature over multiple data packets. The receiver gets overwhelmed while verifying bogus signatures as the signature verification being computationally costly.

Researchers have recognized that to protect against such attacks, it is important to authenticate both data as well as control packets; and have accordingly designed requisite authentication protocols.

IV. Authentication protocols

In TESLA [3], packets are not authenticated at every Hop; instead it uses end-to-end authentication in which packets are authenticated by final receiver, that too after a delay of several seconds. As a result, TESLA's throughput for mobile nodes is mediocre and suffers from long latency. Moreover, it also requires loose time synchronization between sender and receiver.

LHAP is built on principles of TESLA and it makes an attempt to overcome its drawbacks. Actually, LHAP was designed for MANETs and makes use of Hop-by-Hop authentication [4].

It deploys twin techniques of lightweight packet authentication and lightweight trust management and is consequently more efficient compared to traditional authentication algorithms. It is characterized by reduced number of public key operations as it makes use of TESLA for trust bootstrapping and maintaining trust relationships among a set of nodes. LHAP employs one-way hash chains in packet authentication technique.

Lu and Pooch's algorithm is based on LHAP. Even this algorithm uses Hop-by-Hop authentication and is known to be efficient like LHAP[5][6]. Yet, it uses only one key at every node instead of two as used by LHAP. Lu and Pooch's algorithm makes use of 'delayed key disclosure' like TESLA; resulting in network latency.

HEAP is independent of routing protocol used. It is suited for most applications, whether unicast, multicast or broadcast. It is very efficient protocol as it uses two keys and is based on modified HMAC algorithm [7].

In HEAP; as and when due to mobility, a given node's neighborhood alters; the said node shares the *Ikey* and a new *Okey* with each new neighbor. Even TESLA, LHAP and Lu function on similar lines. These keys need to expire after a certain amount of time and new keys should be generated, e.g. after every few hours is another requirement. This is one way to guard against crypt-analysis attacks by an adversary.

V. Security attacks on authentication protocols

Authentication Protocols in MANETs are prone to different types of attacks on their mechanism. Such attacks are mainly divided into two groups [4]; Outsider attack and Insider attack.

A. Outsider Attacks

Outsider attacks are defined as attacks from nodes which do not hold a valid certificate. Such attacks are further classified into two types.

1) Single Outsider Attack: This type of attack is possible when a valid node moves out of the range of other nodes for a period of time; and this transmitting node is not aware of the absence of node. Further, transmitting node happens to have disclosed its TESLA and TRAFFIC keys for the said time period and a malicious node now obtains these keys and uses the same to pretend a node from the network.

2) Collaborative Outsider Attack : Wormhole attack [8] is a type of collaborative outsider attack. It is launched by multiple attackers. These attackers form a private tunnel by which they communicate directly. For example, node A wants to send packets to node E but node P1 eavesdropped key messages and traffic packets of node A and forwards to P2 via

a wormhole tunnel. Then P2 can alter and rebroadcast to node E.

B. Insider attack

These attacks are launched by one or more nodes which are compromised and hold legal certificates. Insider attacks can be classified in to two types.

1) Single / Multiple Insider Attack : To increase the traffic inside the networks a compromised node may try to flood many traffic packets into the network. These attacks can be initiated by many compromised nodes which were legitimate previously.

2) Insider Clone Attack : When a compromised legitimate node shares its identity or private key with an outsider attacker node and both holding the same identity. It is less likely to detect an outsider node. The cloned nodes are mostly spread in different network areas. The collaborative attack by these nodes is called clone attack.

A few important attacks that affect authentication protocols are: Man-In-The Middle attack and DoS attack. In case of Man-In-The Middle, the attacker impersonates both sender as well as the receiver with respect to each other and without their knowledge of attack. In this attack, which is also called Bucket Brigade attack, the attacker is actively eavesdropping upon the link between the verifier and the prover and thereby intercepting all authentication messages being exchanged by the sender and the receiver in the network under attack [9].

In DoS attack, an attacker causes unnecessary communication delay in data reaching its destination or network traffic is dropped altogether or even redirected to another destination.

TABLE I. ATTACKS ON PROTOCOLS

Name of Protocol	Attack		
	Man in Middle	***Wormhole***	***DoS***
Lu and Pooch's	No	No	Yes
TESLA	No	Yes	Yes
LHAP	Yes	Yes	Yes
HEAP	No	No	No

Source authentication schemes like TESLA having a upper bound limit on traffic rate but it can prevent some attacks on MANETs. In case of Hop-by-Hop authentication every node authenticates only neighborhood nodes in place of the sources of original traffic, compromised node is able to pretending itself as a valid forwarding node and transmits malicious traffic inside the network. It does not offer strong source authentication,

TESLA, LHAP and Lu and Pooch's algorithm refer Table I are vulnerable to DoS attack and requires that all the nodes should be synchronized in time [7]. To prevent TESLA, LHAP and Lu algorithm from certain attacks, it is required that when a node moves from the range, one or two keys should be exchanged with all new neighbor nodes. The keys should be valid for fix time and after that it should expire and a new key should be generated for same time period.

To prevent "Man in-the-Middle" attacks from the TESLA it uses delayed key disclosure. LHAP is vulnerable to Wormhole and Man-in-the-Middle attacks refer Table I because periodic delayed key disclosure is not used in these algorithms [4].

HEAP offers some level of protection against insiders who try to forge packets and impersonate other insiders in order to incriminate them. Any packet transmitted by an outsider node should be immediately dropped by the receiving insider node at the first hop with a very high probability.

HEAP successfully guards against many outsider attacks refer Table I , such as DoS attack that attempt to flood the network, Wormhole attack, Man-in-the-Middle attack, and Flooding etc. [7].

VI. SECURITY PARAMETERS

Security of any protocol can be assesed by following parameters-

- Protocol is based on source authentication or hop by hop authentication.
- Technique used for message broadcasting, weather protocol is multicast or unicast.
- Algorithm used for trust maintenance in authentication.
- No. of Keys used in trust bootstrapping.
- Condition of trust termination between two nodes.
- Use of symmetric cryptography (Digital Signature) in trust management.
- Delayed authentication with indexing in transmission.
- Delay time of key disclosure
- Use of varied delay in key disclosure

VII. PERFORMANCE PARAMETERS

The performance metrics employed to analyze different protocols are- Authentication Latency, Throughput, and Packet Delivery Ratio (PDR).

A. Authentication Latency

The latency in a packet authentication is due to MAC verification and delay key disclosure. The latency of MAC verification is less than one millisecond and it is due to computing one hash. The latency in packet authentication is mainly due to the key disclosure delay. The value of key disclosure delay is based on current traffic pattern and it is decided by the packet sender.

B. Throughput

Mean Throughput can be defined as the ratio of the number of packets successfully received by each node and total simulation time [10].

C. PDR

PDR is measured as the ratio of number of packets successfully received by each node and total number of packets sent [10].

Authors [7] have performed the simulation on GloMoSim v2.03. Following approximate values of Mean latency variation (Table II), PDR variation (Table III) and Throughput variation (Table IV) are found by the study of simulation results graph of TESLA, LHAP, Lu Pooch's algorithm and HEAP.

Mean latency of HEAP is very low as compared to other protocols.

In case of TESLA, LHAP and HEAP, the peak value of PDR (Table III) is reached approximately 85% at packet rate 20 packets/sec. Initially PDR remains constant up to 25 pkts/sec. As can be seen, at higher packet rates, PDR quickly reduces and tapers off to around 10%., as the throughput is a function of the product of the PDR and packet rate.

The peak value of throughput (Table IV) is reached at 25 packets/sec, because throughput is proportional to product of the PDR and packet rate. For high packet rate throughput falls sharply cutting with PDR but for packet rate more than 50 packets/sec. throughput is nearly constant and effect of low PDR is offset by the high packet rate.

The performance of Lu's algorithm is significantly poorer than all other algorithms in both PDR and throughput. It is due to caches of packets at first forwarding nodes. First forwarding node cached packets until it would not receive a key update packet.

TABLE I. VARIATION OF MEAN LATENCY

S.No.	Number of Hops	Protocol			
		LHAP	***HEAP***	***TESLA***	***Lu Pooch's***
1	1	1000	4	20000	2000
2	2	1000	10	20000	6000
3	3	1000	12	25000	9000
4	4	1000	14	30000	10000
5	5	1000	16	35000	11000

TABLE II. PDR (%) VARIATION

S.No.	Packet Rate (Pkts / sec)	Protocol			
		LHAP	***HEAP***	***TESLA***	**Lu Pooch's**
1	20	85	85	85	40
2	40	36	36	36	19
3	60	21	20	20	10
4	80	16	16	15	9
5	100	15	15	13	9

TABLE III. VARIATION OF THROUGHPUT(bytes/s)

S.No.	Packet Rate (Pkts /sec)	Protocol			
		LHAP	***HEAP***	***TESLA***	**Lu Pooch's**
1	20	8800	8800	8800	4000
2	40	7200	7200	7200	3700
3	60	6800	6000	6800	3500
4	80	6900	6600	6500	3900
5	100	6900	6600	6300	4700

VIII. CONCLUSION

We compared the performance of HEAP, TESLA, LHAP and Lu and Pooch's algorithms and it is observed that TESLA is vulnerable to DoS attack and requires secure time synchronization of all the nodes. It introduces very large latencies of several seconds making it unsuitable for real time or QoS applications. LHAP is vulnerable to Wormhole and Man-in-the Middle attack. It also has very large memory requirements at every node. Lu's scheme has overall poor performance. In this scheme, throughput and PDR significantly degrade. It has extremely low memory requirements. HEAP is resistant to several outsider attacks such DoS, Wormhole, Replay. HEAP is suitable for use in MANETs for unicast, multicast or broadcast applications.

IX. REFERENCES

[1]. Perkins, C. E. "Ad hoc Networking", Pearson Publication, India, 2008,

[2]. M. Soni, and B. K. Joshi, "Security Assessment of SAODV Protocols in Mobile Ad-hoc Networks", Proceeding of the International Symposium Data Science and Big Data Analytics (ISDB ACM-WIR 2018), Indore; 5-6 January 2018, pp. 347-355.

[3]. A. Perrig, R. Canetti, J. D. Tygar, and D. Song; "The TESLA Broadcast Authentication Protocol", Journal of Crypto Bytes, vol. 5, no. 2, Summer / Fall, 2002, pp. 2-13.

[4]. S. Zhu, S. Xu, S. Setia, and S. Jajodia, "LHAP: A Light-weight Hop-by-hop Authentication Protocol for Ad-Hoc Networks", Proceeding of 23rd IEEE International Conference on Distributed Computing Systems Workshops(ICDCSW03), Providence, USA, 19-22 May 2003, pp. 749.

[5]. Lu and U. W. Pooch, "A Lightweight Authentication Protocol for Mobile Ad Hoc Networks", Proceeding of the IEEE International Conference on Information Technology (ITCC'05), Las Vegas, USA, 4-6 April 2005, pp. 546-551.

[6]. Lu and U. W. Pooch, "A Lightweight Authentication Protocol for Mobile Ad Hoc Networks", International Journal of Information Technology, vol. 11, no. 2, 2005, pp. 119-135.

[7]. R. Akbani, T. Korkmaz and G.V.S. Raju; "HEAP: Hop-by-hop Efficient Authentication Protocol For Mobile Ad-hoc Networks", Proceeding of the Spring Simulation Multiconference, SpringSim 2007, Norfolk, Virginia, USA, 25-29 March 2007, pp. 157-165.

[8]. B. K. Joshi and M. Soni; "Security Assessment of AODV Protocol under Wormhole and DOS Attacks", Proceeding of the 2nd International IEEE Conference on Contemporary Computing and Informatics (IC3I2016), Noida , India, 14-17 December 2016, pp. 173-177.

[9]. N. Pari S, "Investigation of malicious nodes by security improvisation of routing in mobile ad hoc networks", PhD Dissertation, Anna University, Chennai 2014, pp. 111-112.

[10]. Megha Soni, and Brijendra Kumar Joshi; "Security Assessment of Routing Protocols in Mobile Ad-hoc Networks"; Proceeding of the International IEEE Conference on ICT in Business, Industry and Government (ICTIBIG2016) Indore; 18-19 November 2016, pp. 24.

[11]. M. Soni, and B. K. Joshi, "Security Assessment of DSDV Protocol in MANET", International Journal of Advance Computational Engineering and Networking (IJACEN), vol. 5, no. 9, September 2017, pp. 107-111.

AUTHORS PROFILE

Megha Soni doning PhD research work at Electronics and Telecommunication Department of MCTE, Mhow, DAVV University Indore, India. She has obtained BE in Electronics and Telecommunication Engineering from Govt Engg College, Sagar; M.E in Digital Communication from Davi Ahilya University Indore. She joined as an Assistant Professor in Electronics & Communication in Dec. 2005. Her research interest is in security assessment of routing and authentication protocols of Mobile Ad Hoc Networks.

Dr Brijendra Kumar Joshi is associated with as a Professor of Electronics &Telecommunication and Computer Engineering at Military College of Telecommunication Engineering, MHOW (MP), India. He has obtained BE in Electronics and Telecommunication Engineering from Govt Engg College, Jabalpur; ME in Computer Science and Engineering from IISc, Banglore, PhD in Electronics and Telecommunication Engineering from Rani Durgavati University, Jabalpur, and MTech in Digital Communication from MANIT, Bhopal. He has more than 34 years of teaching experience. His research interests are programming languages, compiler design, digital communications, mobile ad-hoc and wireless sensor networks, image processing, software engineering and formal methods. He has number of research publications to his credit. He has supervised six Ph D thesis and currently supervising nine research scholars. He has authored two books on Data Structures and Algorithms in C/C++ published by Tata McGraw-Hill, New Delhi.

Prediction of survival in breast cancer patients using Random Forest classifier and ReliefF feature selection method

Diogo Albino de Queiroz [#*1], Gabriel Sousa Almeida Assunção [#2], Kamila Alves da Silva Ferreira [#3], Vilian Veloso de Moura Fé [#4], Vitória Paglione Balestero de Lima [#5], Fernanda Antunes Dias [#6], Túlio Couto Medeiros [#7], Karen Nayara de Souza Braz [#8], Rodrigo Augusto Rosa Siviero [#9], Pâmela Alegranci [#10], Eveline Aparecida Isquierdo Fonseca de Queiroz [#11]

[#] *Universidade Federal de Mato Grosso (UFMT)*
Av. Alexandre Ferronato, 1200, 78550-728 – Sinop, MT – Brasil

[1] diogoqueiroz@gmail.com
[2] gabrielassuncao53@gmail.com
[3] kamila.alves.ferreira_@hotmail.com
[4] vilianmourafe@gmail.com
[5] vitoria.paglione@gmail.com
[6] fernandaantunesdias@gmail.com
[7] tuliocoutomed@gmail.com
[8] karennayara.ks@gmail.com
[9] rodsiviero92@gmail.com
[10] palegranci@gmail.com
[11] eveline.ufmt@gmail.com

[*] *Escola Técnica Estadual de Educação Profissional e Tecnológica*
Av. das Sibipirunas, 1681, 78557-673 – Sinop, MT – Brasil

***Abstract*— Studies have evaluated the use of machine learning to support clinical evaluation in cancer patients. Random Forest has demonstrated to be a relevant technique in predicting survival based on staging, treatment, prognosis, and patient characteristics. This study evaluated the performance of Random Forest classifier to predict the breast cancer patients' survival at starting the treatment phase in two different classes: up to or over 5 years; and evaluated performance by selecting the most relevant characteristics of the dataset using ReliefF feature selection method. The data were collected from the breast cancer patients' medical records. Altogether, 60 patient records were selected and the 10-folds cross-validation technique was used in the execution of Random Forest. Random Forest presented a significant result to predict the patients' survival, once its analysis demonstrated an accuracy of 91.67%, precision of 91.8%, F-Measure of 91.7%, sensitivity of 91.18% and specificity of 92.31%. Using the ReliefF method the accuracy was 93.33%, precision 93.3%, F-Measure 93.3%, sensitivity 94.12% and specificity 92.31%. It was possible to observe that the classifier presented excellent results of classification prediction. Therefore, this model could be recommended as a useful tool to predict the survival rate of breast cancer patients and to support medical decisions.**

Keywords- classification; prediction; breast cancer patients; Random Forest classifier; ReliefF method.

I. INTRODUCTION

Cancer is a chronic and multifactorial disease, characterized by uncontrolled growth of cells that can invade other tissues leading to metastasis and can promote severe metabolic, immunologic and endocrine alterations in the body, also can induce cachexia-anorexia syndrome, which can lead the patient to death [1], [2]. The patients can be treated by different types of anticancer therapies, such as surgery, chemotherapy, radiotherapy and hormone therapy, depending on the cancer stage and the patients' clinical condition, as well as, in accordance with the therapies available by the medical health system [3], [4].

Cancer incidence and prevalence have been increasing in an alarming rate worldwide and is considered the second leading cause of death [5]. Breast cancer is one of the main cancer types in Women [5]. In Brazil, breast cancer is the most common cancer and is one of the main causes of death in women [6]. In accordance with data from the National Cancer Institute (INCA), it is estimated that Brazil will have 625,000 new cancer cases in 2020, of which 66,280 will be new breast cancer cases [6].

Recent studies have evaluated the use of machine learning techniques to support the clinical assessment performed by the physician [7]. These tools assist in the diagnosis and can assist

in defining the most appropriate technique to be applied in the treatment of the patient. Vibha et al. [8] compared the Random Forest classifier and the algorithms Back Propagation and the Association Rule Classifier to confirm the diagnosis of tumor in breast cancer patients by mammography. Ghongade and Wakde [9] proposed to diagnose breast cancer patients using the machine learning method based on the Random Forest classifier using Fast Correlation-based feature Selection (FCBF) technique. Wang, Cheng, and Chiu [10] trained an Artificial Neural Network (ANN) to predict the patients' five-year survival rate. Montazeri et al. [11] evaluated seven models of machine learning methods for the prediction of breast cancer survival in Iranian adult people, and it was observed that the best model to predict survival was Random Forest.

In addition, Wang et al. [12] adopted a tree ensemble classification method that took imbalanced data into account to evaluate the 5-year survival rate of colorectal cancer patient. It was obtained the accuracy, sensitivity and specificity of 70.69%, 84.52% and 66%, respectively. Shouket et al. [13] evaluated six classifiers for the prediction of breast cancer survival. Concerning the 5-year survival rate, JRip and Random Forest classifiers had better accuracy, 96.25% and 95.83%, respectively.

Moreover, Singh et al. [14] evaluated a system and investigated the use of the AdaBoost feature selection method for the execution of the Random Forest to detect normal and abnormal mammography images. Ahmad and Yusoff [15] evaluated the classification of breast cancer lesions using the Random Forest classification method into two categories: benign and malignant cancer. Ganggayah et al. [3] researched machine learning techniques to build models for identifying the most important prognostic factors influencing the survival rate of breast cancer patients in the Asian setting, at the end, the Random Forest algorithm produced better performance using all data. Sun et al. [16] proposed a framework to detect cervical cancer patients based on Pap smear images using Random Forest classifier and ReliefF feature selection method.

Thus, this study aims to evaluate the performance of the Random Forest classifier to predict the breast cancer patients' survival rate at the beginning of the treatment phase into two different classes: up to 60 months (five years) or above 60 months (five years). The data were collected from medical records carried out by oncologist doctors of Santo Antônio's Hospital in Sinop, Mato Grosso, Brazil. Furthermore, the objective was also to evaluate performance by selecting the most relevant characteristics of the dataset using the ReliefF feature selection method in the Random Forest classifier.

II. Material and Methods

The architecture of the article is described in Figure 1. The first step was the collection of real data from the patients' medical records. During the data preparation phase, some instances were removed due to lack of pattern, noises, and when they did not address the purpose of this article. Then, the features were selected, and the classifier was executed.

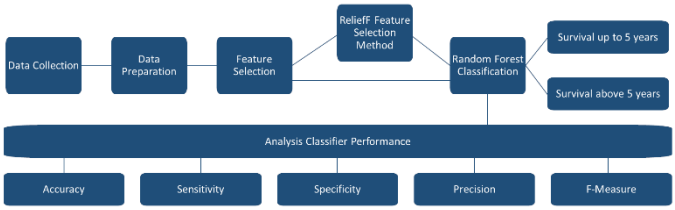


Figure 1. Classification architecture of data collected from the patients' medical records with breast cancer.

A. Data Collection

The data were collected from handwritten medical records carried out by oncologist doctors of Santo Antônio's Hospital in Sinop, Mato Grosso, Brazil. In total, 197 patients with breast cancer between 2013 and 2018 were analyzed. The research targets were female or male patients who were preliminarily diagnosed with breast cancer and received their entire treatment in the study hospital. The study and data collection were approved by the Hospital Ethics Committee and by the Ethics Committee for Human Study at Universidade Federal de Mato Grosso (protocol n. 2.414.600). In Figure 2 are listed the extracted data.

B. Data Preparation

Considering that data collection ended in December 2019 and this paper aims to predict the breast cancer patients' survival at the beginning of the treatment phase into two different classes: up to 60 months (five years) or above 60 months (five years), the data between 2015 and 2018 of alive patients were removed due to the impossibility to verify if these patients will survival up to or above five years.

The data between 2013 and 2014 of alive patients where on the day of data collection the patient had survived less than 5 years were removed, due to the impossibility to determine whether the survival of these patients was up to or above 5 years. The data between 2013 and 2018 of all the dead patients were considered. Some instances were removed due to lack of pattern, and noises. Thus, 60 patients' data were considered for prediction analysis.

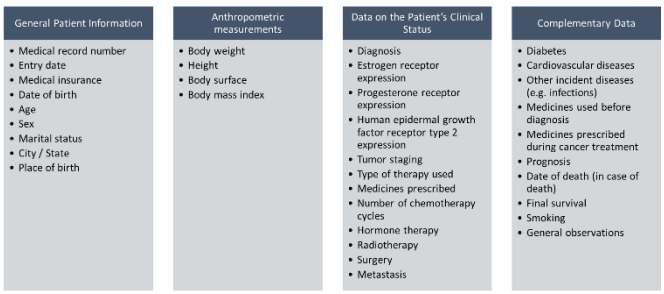


Figure 2. Data collected from the patients' medical records.

C. Feature Selection

The data were organized, and the feature selection was based on clinical availability and [10], [11] and [7] studies. In the present work, were selected twelve features to predict the five years survival rate of the patients: age, sex, body mass index, estrogen receptor expression (ER), progesterone receptor expression (PR), human epidermal growth factor receptor type 2 expression (HER-2), tumor staging, hormone therapy,

radiotherapy, surgery, metastasis and diabetes. The class target was final survival which if higher than or equal to five years, set as 0, otherwise, set as 1. In Table I, the attributes of the dataset were described in detail.

To compare the performance of the Random Forest classifier, the features were also ranked based on the ReliefF algorithm, as seen in Table II. The features with positive weights - sex, body mass index, ER, PR, tumor staging, hormone therapy, radiotherapy, surgery, metastasis and diabetes - were selected and submitted to the classifier. ReliefF algorithm is used to identify which attributes are most relevant to the set of instances and available data. A weight is linked to each attribute, where the most relevant attribute has the highest weight. It is not limited to two classes problems, it is a robust algorithm that can handle noisy and incomplete data [17]. It was executed by the Weka program.

Sun et al. [16] evaluated the ReliefF algorithm to select the features of the dataset of cervical cancer patients and obtained significant results. Sangaiah and Kumar [18] proposed to combine the ReliefF and entropy-based genetic algorithm to select the features. From a total of 24,482, it was reduced to 75 features for classifying normal breast cancer patients, obtaining results with better accuracy.

D. Random Forest Classification

The Weka program version 3.8.4 (Waikato Environment for Knowledge Analysis) was used in the simulations proposed in this article, which has several implementations contemplated that facilitates data analysis and simulations.

The use of machine learning techniques is an important methodology for tumor detection and assessment of the survival rate of breast cancer patients [19]. The Random Forest classifier proved to be one of the most efficient techniques in this process as can be seen in [11] and [9]. This classifier is defined in [20] and [21]. In this study, it was used the technique 10-folds cross-validation. In this technique, the data set is divided into 10-folds and 10 training iterations are performed, in each iteration 9-folds are used for training and one fold for validation, the calculated performance is the average performance of the 10 iterations, as shown in Figure 3 [16].

Figure 3. Interactions that occur in the technique 10-folds cross-validation.

E. Analysis Classifier Performance

A normal method for evaluating the classifier result is to compare the classifier results with the current condition of the patients. This paper used the indicators accuracy, sensitivity, specificity, precision, and F-Measure to evaluate the results classifier using the variables true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

TABLE I. BREAST CANCER DATASET DESCRIPTION OF ATTRIBUTES.

Feature number	Feature description	Value	Amount	Proportion (%)
1	Age	<=30	0	0.00%
		31-40	6	10.00%
		41-50	16	26.67%
		51-60	21	35.00%
		>60	17	28.33%
2	Sex	Female	59	98.33%
		Male	1	1.67%
3	Body mass index	Normal weight	17	35.42%
		Overweight	14	29.17%
		Obesity	17	35.42%
4	Estrogen receptor expression	Positive	48	84.21%
		Negative	9	15.79%
5	Progesterone receptor expression	Positive	45	78.95%
		Negative	12	21.05%
6	Human epidermal growth factor receptor type 2 expression	Positive	11	20.37%
		Negative	42	77.78%
		Inconclusive	1	1.85%
7	Tumor staging	1	16	27.12%
		2	16	27.12%
		3	17	28.81%
		4	10	16.95%
8	Hormone therapy	Yes	42	71.19%
		No	17	28.81%
9	Radiotherapy	Yes	37	61.67%
		No	23	38.33%
10	Surgery	Yes	51	85.00%
		No	9	15.00%
11	Metastasis in another organ	Yes	22	37.93%
		No	36	62.07%
12	Diabetes	Yes	5	9.09%
		No	50	90.91%
13	Class target (final survival)	0	34	56.67%
		1	26	43.33%

TABLE II. LIST OF ATTRIBUTES RANKED USING THE RELIEFF METHOD.

Ranking	Feature	Weight
1	Metastasis	0.66222
2	Tumor staging	0.2375
3	PR	0.11056
4	Hormone therapy	0.07167
5	Diabetes	0.05167
6	Surgery	0.02167
7	Body mass index	0.01958
8	Radiotherapy	0.01833
9	ER	0.01722
10	Sex	0.00667
11	HER2	-0.02208
12	Age	-0.04833

PR: progesterone receptor; ER: estrogen receptor; HER2: human epidermal growth factor receptor type 2.

- True positive: breast cancer patients' who lived above five years and were correctly classified as patients who lived above five years;
- False positive: breast cancer patients' who lived up to five years and were incorrectly classified as patients who lived above five years;
- True negative: breast cancer patients' who lived up to five years and were correctly classified as patients who lived up to five years;
- False negative: breast cancer patients' who lived above five years and were incorrectly classified as patients who lived up to five years.

Based on these variables, the accuracy is defined by the proportion of instances correctly classified concerning the total number of instances. It can be obtained from Equation 1. Precision is the proportion of instances correctly classified concerning instances classified as positive, as described in Equation 2. Finally, F-Measure is the combined metric of precision and recall. This analysis demonstrates how precise and robust is the classifier, being described by Equation 3.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$F - Measure = \frac{2 * Recall * Precision}{Precision + Recall} \quad (3)$$

Sensitivity or true positive rate (TPR) evaluates the positive instances correctly classified concerning the total number of positive instances, that is, the proportion of patients who lived above five years and were correctly classified. This demonstrates the capacity of the algorithm to classify correctly the positive cases. It can be obtained from Equation 4.

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

Specificity evaluates negative instances correctly classified concerning the total number of negative instances, that is, the proportion of people who lived up to five years and were correctly classified. This demonstrates the capacity of the algorithm to correctly classify the negative cases. It can be obtained from Equation 5. The false positive rate (TPR) is equal to 1-specificity.

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

Confusion Matrix		
Classified Positive	Classified Negative	
True Positive	False Negative	Actual Positive
False Positive	True Negative	Actual Negative

Figure 4. General confusion matrix for two classes.

Confusion matrix shows how many instances have been assigned to each class. In this study, we considered survival over 5 years as positive and survival up to 5 years as negative. Thus, the line "Actual Positive" in Figure 4 represents the number of patients who survived over 5 years, according to the data collected from medical records. And the "Actual Negative" line shows the number of patients who survived up to 5 years. However, the "Classified Positive" and "Classified Negative" columns represent the classification performed by the classifier.

III. RESULTS AND DISCUSSION

This study included data from 60 patients who started treatment between 2013 and 2018. In the classification based on the patients' survival status, 26 patients died within 5 years, which represented 43.33% of the total patients' population, and 34 patients survived above 5 years, 56.67% of the patients' population. In the classification based on cancer stage, the number of patients diagnosed as in Stage I to Stage IV were 16, 16, 17, and 10, respectively, accounting for 27.12%, 27.12%, 28.81%, and 16.95% of the data collected, respectively.

The 5-year survival rate predicted using the Random Forest classifier was compared to the original data as shown in Confusion Matrix in Figure 5. The "Survival above 5 years" line describes the number of patients who survived over 5 years, according to the medical record, which resulted in 34 patients. And the line "Survival up to 5 years" shows survivors up to 5 years old, which totaled 26 patients. However, the "Survival above 5 years" column presents the number of records classified by the Random Forest classifier as survivors over 5 years, and the "Survival up to 5 years" column describes the number of subjects classified as survivors up to 5 years. In Figure 5B, the

characteristics used in Random Forest were based on the selection of characteristics performed by the ReliefF algorithm. And in Figure 5A, we submit all the characteristics in the execution of the classifier.

A

Confusion Matrix		
Survival above 5 years	Survival up to 5 years	classified as
31	3	Survival above 5 years
2	24	Survival up to 5 years

B

Confusion Matrix		
Survival above 5 years	Survival up to 5 years	classified as
32	2	Survival above 5 years
2	24	Survival up to 5 years

Figure 5. Survival prediction results. A) Results using the Random Forest classifier. B) Results used the Random Forest classifier based on the selection of characteristics performed by the algorithm ReliefF.

Figure 6 presents the comparison between Random Forest classifier performance with or without the use of the ReliefF algorithm. It is possible to observe that the Random Forest classifier presented a significant result to predict the survival of these 60 patients, once its analysis demonstrated an accuracy of 91.67%, precision of 93.94%, F-Measure of 92.54%, sensitivity of 91.18% and specificity of 92.31%. Evaluating twelve features to predict the five years survival rate of the Brazilian patients by the Random Forest, it is possible to observe that the analysis was effective presenting excellent accuracy and sensitivity results.

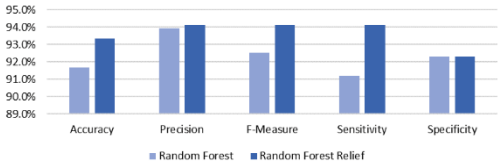


Figure 6. Comparison between Random Forest performance with or without the use of the ReliefF algorithm.

Similarly, Ganggayah et al. [3] also evaluated machine learning techniques to build models for identifying the important prognostic factors influencing the survival rate of breast cancer patients. They analyzed a large dataset with 8066 data with diagnosis information between 1993 and 2016 and selected 24 variables for the analysis, observing that the Random Forest algorithm produced better performance using all data: accuracy of 82.70%, sensitivity of 83%, specificity of 81% and precision of 93%. Also, the six most important variables identified were cancer stage classification, tumor size, number of total auxiliary lymph nodes removed, positive lymph nodes, primary treatment types, and methods of diagnosis.

In the present study, to compare the performance of the Random Forest classifier, the features were also ranked based on the ReliefF algorithm, and it was possible to observe that using the ReliefF significant results were obtained, once the data demonstrated an accuracy of 93.33%, a precision of 94.12%, an F-Measure of 94.12%, sensitivity of 94.12%, and specificity of 92.31%, as shown in Figure 6. The features used in this analysis were: sex, body mass index, ER, PR, tumor staging, hormone therapy, radiotherapy, surgery, metastasis and diabetes.

Tapak et al. [7] evaluating eight models of machine learning methods for the prediction of breast cancer survival and metastasis in Iranian women, analyzed 9 different risk factors to compare the models and presented that the Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) were the best models to predict survival, with the sensitivity of 73% and accuracy of 93%. However, LDA was the best model to predict metastasis, with an accuracy of 86% and precision of 61%. None of the methods had a sensitivity greater than 50% to predict breast cancer metastasis.

Wang, Cheng, and Chiu [10] also trained an Artificial Neural Network (ANN) to predict the patients' five-year survival rate. The authors selected five variables, including age, tumor size, condition of tumor metastasizing to lymph nodes, whether the tumors had metastasized to other organs and whether breast-conserving surgery was performed followed by radiotherapy. The ANN method had an accuracy of 85.1%, sensitivity of 94.83% and specificity of 52.86%. Sun et. al [16] evaluated the combination of the Random Forest classifier with the ReliefF method to generate a reliable classification framework for the cervical cell images and compared the results obtained with the performance of the Naive Bayes, C4.5 and Logistic Regression classifiers. The results based on Random Forest with the top 13 features selected by the ReliefF method achieved the best performance in all the experiments evaluated [16].

Moreover, Wang et al. [12] proposed a two-stage model for advanced-stage colorectal cancer survival prediction, where the first stage was to predict whether a patient can survive more than five years. In the first stage adopted a tree ensemble classification method that took imbalanced data into account. In total, 1568 records with survival over 60 months and 1503 records with survival less than 60 months, were retained for classification experiments. It was 10-fold cross-validation, and the accuracy, sensitivity and specificity were 70.69%, 84.52% and 66%, respectively.

In addition, Shouket et al. [13] evaluated six classifiers to predict breast cancer survival in Pakistani patients, based on a record of 200 breast cancer patients and a data set consisting of 10 features. Concerning the 5-year survival rate, JRip and Random Forest classifiers had better accuracy, 96.25% and 95.83%, respectively. Montazeri et al. [11] evaluated seven models of machine learning methods for the prediction of breast cancer survival in Iranian adult people, selecting 9 attributes that are associated with the survival of breast cancer patients. The dataset included 900 patient records in which 803 patients were alive and 97 patients were dead. They compared the models by 10-fold cross-validation strategy and was used Weka software, observing that the best models to predict survival was Tree Random Forest with accuracy, true positive rate and precision of 96% and false positive rate of 24.1%.

IV. Conclusion

The models presented significant results. At first, twelve characteristics were selected to be submitted to the Random Forest classifier, where it obtained accuracy of 91.67%, precision of 93.94%, F-Measure of 92.54%, sensitivity of 91.18% and specificity of 92.31%. Then, the features were ranked based on the ReliefF algorithm and submitted again in the Random Forest classifier, obtaining an accuracy of 93.33%, a precision of 94.12%, an F-Measure of 94.12%, sensitivity of 94.12% and specificity of 92.31%. In this sense, this model proved to be a useful tool to predict the survival rate of breast cancer patients and to support medical decisions. A limitation of

this work is the number of records available, as a suggestion it is possible to consider a longer period and perform the classification again to evaluate the accuracy of the results.

ACKNOWLEDGMENTS

The authors are grateful to UFMT and Hospital Santo Antônio for supporting and collaborating with the research. The authors are grateful to the oncologists Dr. Érico and Dr. Airton for their support and collaboration during this research and to Mari for technical assistance during the data collection in the Santo Antônio's Hospital in Sinop, Mato Grosso, Brazil. The authors are also grateful to the Brazilian agencies Mato Grosso State Research Support Foundation (FAPEMAT) (proc. nº. 0475822/2018 - Tulio Couto Medeiros and proc. nº. 0460139/2019 - Kamila Alves da Silva Ferreira) for financial support with their scholarship.

DECLARATION OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

[1] J. M. Argilés, B. Stemmler, F. J. López-Soriano, and S. Busquets, "Inter-tissue communication in cancer cachexia," *Nature Reviews Endocrinology*, vol. 15, no. 1. pp. 9–20, Jan. 21, 2018, doi: 10.1038/s41574-018-0123-0.

[2] W. H. O. WHO, "Cancer," 2020. https://www.who.int/cancer/en/ (accessed Apr. 11, 2020).

[3] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–17, 2019, doi: 10.1186/s12911-019-0801-4.

[4] B. L. Lee, P. E. R. Liedke, C. H. Barrios, S. D. Simon, D. M. Finkelstein, and P. E. Goss, "Breast cancer in Brazil: Present status and future goals," *Lancet Oncol.*, vol. 13, no. 3, pp. e95–e102, 2012, doi: 10.1016/S1470-2045(11)70323-0.

[5] W. H. O. WHO, "Global cancer observatory," 2020. http://gco.iarc.fr/ (accessed Apr. 11, 2020).

[6] INCA, "Estatísticas de câncer," 2020. https://www.inca.gov.br/numeros-de-cancer (accessed Apr. 11, 2020).

[7] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 3, pp. 293–299, 2019, doi: 10.1016/j.cegh.2018.10.003.

[8] L. Vibha, G. M. Harshavardhan, K. Pranaw, P. Deepa Shenoy, K. R. Venugopal, and L. M. Patnaik, "Statistical classification of mammograms using random forest classifier," *Proc. - 4th Int. Conf. Intell. Sens. Inf. Process. ICISIP 2006*, pp. 178–183, 2006, doi: 10.1109/ICISIP.2006.4286091.

[9] R. D. Ghongade and D. G. Wakde, "Computer-aided diagnosis system for breast cancer using RF classifier," *Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 1068–1072, 2018, doi: 10.1109/WiSPNET.2017.8299926.

[10] T. N. Wang, C. H. Cheng, and H. W. Chiu, "Predicting post-treatment survivability of patients with breast cancer using Artificial Neural Network methods," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 1290–1293, 2013, doi: 10.1109/EMBC.2013.6609744.

[11] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," *Technol. Heal. Care*, vol. 24, no. 1, pp. 31–42, 2016, doi: 10.3233/THC-151071.

[12] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin, and Y. Jin, "A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction," *Inf. Sci. (Ny).*, vol. 474, pp. 106–124, 2019, doi: 10.1016/j.ins.2018.09.046.

[13] T. Shouket, S. Mahmood, M. T. Hassan, and A. Iftikhar, "Overall and disease-free survival prediction of postoperative breast cancer patients using machine learning techniques," 2019, doi: 10.1109/INMIC48123.2019.9022756.

[14] V. P. Singh, A. Srivastava, D. Kulshreshtha, A. Chaudhary, and R. Srivastava, "Mammogram Classification Using Selected GLCM Features and Random Forest Classifier," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 6, pp. 82–87, 2016.

[15] F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," in *International Conference on Intelligent Systems Design and Applications, ISDA*, Dec. 2014, pp. 121–125, doi: 10.1109/ISDA.2013.6920720.

[16] G. Sun, S. Li, Y. Cao, and F. Lang, "Cervical cancer diagnosis based on random forest," *Int. J. Performability Eng.*, vol. 13, no. 4, pp. 446–457, 2017, doi: 10.23940/ijpe.17.04.p12.446457.

[17] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1–2, pp. 23–69, 2003, doi: 10.1023/A:1025667309714.

[18] I. Sangaiah and A. Vincent Antony Kumar, "Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer

prediction," *Cluster Comput.*, vol. 22, pp. 6899–6906, 2019, doi: 10.1007/s10586-018-1702-5.

[19] C. Beaulac, J. S. Rosenthal, Q. Pei, D. Friedman, S. Wolden, and D. Hodgson, "An evaluation of machine learning techniques to predict the outcome of children treated for Hodgkin-Lymphoma on the AHOD0031 trial," *Appl. Artif. Intell.*, vol. 34, no. 14, pp. 1100–1114, 2020, doi: 10.1080/08839514.2020.1815151.

[20] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[21] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," in *Ensemble Machine Learning*, Boston, MA: Springer, 2012, pp. 157–175.

AUTHORS PROFILE

Diogo Albino de Queiroz

Bachelor of Computer Science from the Universidade Estadual de Londrina (UEL). Master in Mathematics from Universidade do Estado de Mato Grosso (UNEMAT), Sinop-MT. Ph.D. in progress in Applied Computing by the Universidade do Vale do Rio dos Sinos (UNISINOS). Technician in Information Technology, Universidade Federal de Mato Grosso (UFMT), Sinop-MT. Professor of Informatics Area, Escola Técnica Estadual de Educação Profissional e Tecnológica de Sinop-MT.

Gabriel Sousa Almeida Assunção, Kamila Alves da Silva Ferreira, Vilian Veloso de Moura Fé, Vitória Paglione Balestero de Lima, Fernanda Antunes Dias, Túlio Couto Medeiros, Karen Nayara de Souza Braz, Rodrigo Augusto Rosa Siviero

Undergraduate student of Medicine at the Universidade Federal de Mato Grosso (UFMT), Campus Universitário de Sinop.

Pâmela Alegranci

Bachelor of Biomedicine from the Universidade de Araraquara. Master in Clinical Analysis and Ph.D. in Biosciences and Applied Biotechnology Pharmacy, Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Professor of the Immunology of the Medical Course, Universidade Federal de Mato Grosso (UFMT), Instituto de Ciências da Saúde (ICS), Sinop-MT.

Eveline Aparecida Isquierdo Fonseca de Queiroz

Bachelor of Pharmacy and Biochemistry from the Universidade Estadual de Londrina (UEL). Master and Ph.D. in Pharmacology from the Universidade de São Paulo (USP). Professor of the Pharmacology of the Medical Course, Universidade Federal de Mato Grosso (UFMT), Instituto de Ciências da Saúde (ICS), Sinop-MT.

Computer Aided Diagnostic System for Diabetic Retinopathy Detection using Image Processing and Artificial Intelligence

Anitha T Nair
Department of CSE
FISAT
Ernakulam, India
anitha.mrt@fisat.ac.in

Arun Kumar M N
Department of CSE
FISAT
Ernakulam, India
akmar_mn11@fisat.ac.in

Anitha M L
Department of CSE
PES College of Engg.
Mandya, India
anithamuralikrishna@gmail.com

Anil Kumar M N
Department of ECE
FISAT
Ernakulam, India
mn_anilkumar@fisat.ac.in

***Abstract*—The number of individuals who develop Diabetic Retinopathy (DR) has increased significantly in recent years. Early detection and diagnosis is essential to prevent the vision loss. Ophthalmologist need to analyze mass retinal images to discover the anomalies, for example, spilling veins, retinal swelling(macular edema), greasy stores on the retina (exudates), and changes in the veins. Early detection of DR from retinal images is a challenging task. Medical image examination is the most effective method for diagnosis of DR. Computer Aided Diagnosis (CAD) systems, which can be used in clinical environments assists an ophthalmologist in diagnosing and detecting DR. This paper aims to investigate, the state of art regarding CAD for DR. The review focus on major techniques in image processing and data mining that are employed for developing a CAD system for DR. This survey also comes up with a common analysis of the current CAD system according to the employed modalities for DR diagnosis or detection. Future research works are discussed to develop efficient CAD systems for DR diagnosis or detection.**

Index Terms*— *Computer Aided Detection, Classification, Diabetic Retinopathy, Feature Extraction, Image Processing, Preprocessing.

I. INTRODUCTION

Diabetic Mellitus is a chronic disease caused due to excessive level of sugar content in the blood. It mainly affects kidneys, nerves, heart and minute blood vessels in the eyes[1]. DR is an eye disease, which can cause damage to the retina. A vascular eye disease will eventually cause blindness in people and can be of two types, Non-Proliferative DR (Early DR)[2] and Proliferative DR (Advanced DR). These days DR is a significant reason for visual impairment in individuals with diabetic. Therefore, constant eye check-up and timely treatment is required. However, the dearth of experts along with related higher medical prices makes regular check up pricey. To fill this opening, development of low cost CAD systems, which can be employed in clinical environments, have gained far more attentiveness in recent years.

*Address correspondence to this author at the Department of Computer Science and Engineering, Federal Institute of Science and Technology, Ernakulam-683577, India; E-mails: amrakmar.mn11@gmail.com

In this period, individuals with diabetic is more and ophthalmologist need to look at mass retinal pictures to discover the irregularities, for example, leakage of blood vessels, deformation of retina(macular edema) and small deposits known as exudates.Early detection of DR is a challenging task in ophthalmology. Most of the CAD systems use some computerized feature extraction and classification algorithms to detect DR. These can be a better tool or an intelligent diagnostic system for an ophthalmologist in detecting or diagnosing the DR[3]. Many efforts has been made to develop CAD systems, which are based on the breakthrough or advances in digital image processing, data mining techniques and pattern recognition. Development of a DR-CAD system[4, 5] is a tough task in the field of ophthalmology. Automatic detection systems were utilized different advances beginning with image processing technologies of retinal data[6] and upgraded to AI approaches such as machine learning and deep learning[7]. Optical coherence tomography and fundus image analysis[8] are mainly used as imaging techniques to draw out the characteristics associated with the retina in the diagnosis of various retinal diseases. Several methods were employed to develop CAD system that uses various datasets, feature vectors and different methodologies for classification[9-11]. Due to the technological development, numerous applications were suggested for the development of DR-CAD system. Earlier days CAD framework were employed with the support of image processing techniques for the mass screening of retinal images[1, 12]. Retinal images were segmented using segmentation algorithms, which will identify optic disc, blood vessels and fovea localization[13, 14] etc. Geometric relationship of different features and lesions can be used along with some morphological operations[15] to obtain a better framework for analyzing the retinal images. Image processing techniques can be effectively applied on retinal images for the effective segmentation[16]. Soft computing techniques[17] employ as a proficient method for the recognition of blood vessels in digital retinal images.

With the introduction of AI based approaches CAD system acquired more accuracy than the previous methods. Automatic detection systems for DR using machine-learning approaches given a new look to the CAD system[18]. Era of deep learning approaches[19, 20] provides desirable and improved results for the detection of DR. In the field of ophthalmology, application of deep learning algorithms in retinal imaging is an upcoming research area[21, 22]. Hybrid solution including image processing and AI approaches[23] is another versatile method for developing CAD system with good accuracy. Voets et. al.[24] overcomes the issues of deep neural network by incorporating new methodologies. .In this paper, we present some of the important methods, which have been employed in developing the CAD system for DR.

A.List Of Abbreviations

AHE	Adaptive Histogram Equalization
AI	Artificial Intelligence
BDT	Binary Decision Tree
BPNN	Back Propagation Neural Network
CAD	Computer Aided Diagnosis/Detection
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CUHK	Chinese University of Hong Kong
DNN	Deep Neural Network
DR	Diabetic Retinopathy
DWT	Discrete Wavelet Transform
FFT	Fast Fourier Transform
FIRE	Fundus Image Registration
GLCM	Gray Level Co-occurrence Matrix
HE	Histogram Equalization
LESH	Local Energy-based Shape Histogram
LPBPC	Local Property-Based Pixel Correction
LFSA	Local Feature Spectrum Analysis
LTP	Local Ternary Pattern
MA	Micro Aneurysm
NPDR	Non-proliferative DR
PDR	Proliferative DR
PNN	Probabilistic Neural Network
SERI	Singapore Eye Research Institute
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
STARE	Structured Analysis of Retina
QDA	Quadratic Discriminant Analysis

II. RELATED WORKS

Many research works have been developed to improve the diagnostic accuracy of DR screening[30]. Xiao et.al.[10] presented an overview of the automatic screening systems such as Iowa DR, Tennessee Ocular Telehealth Network (OTN) etc. Our paper attempts to elaborate different life cycle stages and the various methodologies involved in each stage of the CAD system. [Fig. 1]. shows the life cycle stages of DR detection.

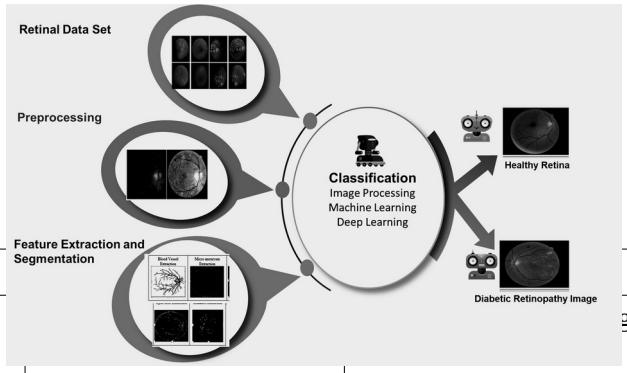


Fig. (1). Life Cycle Stages of DR Detection

A. Preprocessing of the Retinal Image

Preprocessing eliminates unwanted elements and defects from the images and resolves the problems of lighting, illumination, contrast and resolution. Preprocessing of images [31] will improve the quality of retinal images in further processing of a CAD system. Both early and modern CAD system adopted different approaches according to the requirement of the user. DR detection using image processing techniques uses different preprocessing methods to improve the quality of images. Gray scale conversion was performed in most of the images and image enhancement methods such as HE, AHE, and CLAHE were applied to it. Resizing the images to different dimensions and applying morphological operations really improved the performance of the system by reducing the processing time. Green channel extraction was performed to identify the prominent blood vessels in the retinal images. Different filtering methods including matched filter, median filter, Gaussian filter, sober filter and Gabour filter were used to reduce the noise.

Table 1. Details of available dataset.

Sl.No			Remarks
1		g various cameras.	Patients were selected from 15 National Health Service hospitals in the United Kingdom
2	DRIVE [14,21,99].	8 bits/color plane at 565 × 584 pixels.	40 fundus images with 33 typical normal images and 7 DR affected images.
3	STARE [14,21,45 ,66,79.	700 x 605 pixels.	20 retinal fundus images.
4	SERI, CUHK [22].	It was captured with a CIRRUS SD-OCT device.	128 cross-sectional scans with a resolution of 512 × 1, 024 pixels.
5	A2A SD-OCT (Duke dataset) [22].	1, 000 × 512 pixels.	384 SD-OCT volumes: 269 AMD and 115 control or normal eyes.
6	Retinopathy Online Challenge [25].	768 × 576, 1058 × 1061 and 1389 × 1383 pixels.	100 color image of the retina.
7	Messidor [26][36].	440 X960, 2240 X 1488, and 2304 X1536 pixels.	1200 images.
8	KAGGLE [27,98].	High-resolution fundus images.	It contains an aggregate of 35,126 fundus images.
9	E-Ophtha [28].	NA.	381 compressed images of which 148 have MAs presents and 233 depict healthy.
10	DIARETDB1[6][9]][28].	1500 x 1152 with 500 field of view (FOV).	28 training and 61 testing images captured at 50 ° FOV.
11	FIRE [29].	Utilizing a Nidek AFC-210 fundus camera with resolution of 2912 × 2912 pixels and 45 ° field of view (FOV).	Publicly accessible retinal image registration dataset with ground truth annotation.
12	CHASE.[14,79]	1280 X 960 pixels resolution.	28 images.

In NPDR detection, Zhentao Gao et al.[32] discussed different techniques for preprocessing such as normalization, color decomposition, space conversion techniques and contrast enhancement. To extract the structural features from the data sets Chetoui Mohamed et al.[33] used gray scale conversion methods. Khoeun Ratanak, et al. proposed a method to detect micro aneurysm using modified matched filtering with an accuracy of 90.7%. Methodologies in[34] used matched filter and[35] used Gaussian filter as a preprocessing methods for enhancing the images. Morphological operation, binarization and histogram matching were used in[36] to enhance retinal images. Table 2 provides the main themes used in the preprocessing stage.

Table 2. Main themes in preprocessing techniques.

Sl. No	Techniques	Reference(s)
1	AHE, HE, and CLAHE.	[2, 6,15, 17,23,26,31,36 [42-44], [47-49], 55, [59-63], 66,68, 70, 71, 72,74, 77, 80, 81, 86, 88, 94]
2	Resizing.	[2, 13,19,21,23,31,39, 51, 54, 55,68, 73, 74, 88,100]
3	Normalization.	[2, 13,20,23,32,38,44,98]
4	Top-hat form filter. Matched filtering.	[8,75] [8,48,86]
5	Shaded correction.	[8, 15, 47, 77]
6	Spatial Normalization. Global Contrast Normalisation	[15] [14]
7	Green channel extraction.	[15-16,28,31,[37-41], 63,67, 68,71, 74, 86,92]
8	Local-phase method	[16]
9	Median filtering.	[15, 22,25,37,57,43, 47, 62, 66,67, 89, 97]
10	Color Normalization.	[25],[32][53, 54]
11	Gray scale conversion.	[17,26,44, [48-52] , 62, 70, 71, 80, 81, 89, 93]
12	Morphological Operation.	[28,36,38, 39, 42, 43, 59, 63, 80, 94]
13	Binarization, Image cropping, Erosion.	[36, 38, 59, 64, 81, 94]
14	Fuzzy filtering(Median filter), Fuzzy HE, Fuzzy edge detection.	[40, 93]
15	Canny edge detection.	[41, 55]
16	Intensity Inversion.	[42, 43]
17	Adaptive Weiner filter. Homomorphic filtering.	[43] [45]
18	DWT.	[44,48]
19	AM-FM Decomposition.	[46]
20	Gabor filter.	[52, 84]

B. Image Segmentation

Great efforts have been put forward to perform segmentation on retinal images. Segmentation methods attempt to extract the required part of image for further processing. It actually detects the different objects present in the retinal images. In the field of medical imaging many segmentation techniques has been developed[56]. Traditional image processing methods uses some mathematical functions to perform segmentation. Lesion detection is the necessary preliminary phase for DR detection. This will provide added advantage to the later stages of the CAD system. Automatic segmentation of retinal images is very crucial to find out the exudates[57], micro aneurysms, optic disc and extraction of blood vessel, which in turn identifies DR[58]. Classification accuracy of DR can be improvised with the support of image segmentation[59]. Different models designed for the identification of the characteristics present in the retinal image is shown in [Fig.2]. Features used for the segmentation of DR include micro aneurysms, exudates[60], optic disc and hemorrhages[61].

Segmentation models of retinal images include optic disc extraction[34, 62], exudate detection[63], blood vessel extraction[34, 64-65] hemorrhages and micro aneurysm detection. Results of retinal image segmentation is grouped as shown in the [Fig.3].

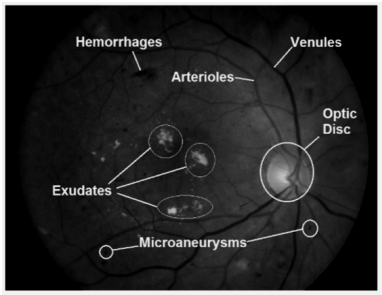


Fig. (2). Features of color retinal image

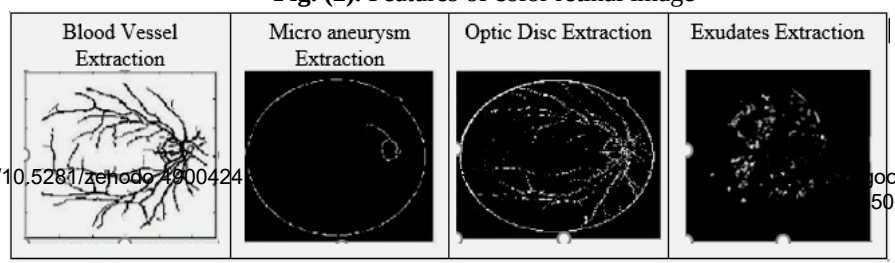


Fig. (3). Result of retinal segmentation models

Shyam et. al.[67] used Hough transform to segment optic disc in the fundus image. Prasad et al.[68] proposed a work to improve the accuracy of exudate detection. This work used top hat platform to perform segmentation with an accuracy of 90%. Tanapat et al.[72] discussed localization of optic disc using morphological operation and convex hull techniques. Method performs localization of fovea and macula using morphological erosion operation, which reported an accuracy of 97%. Siva Sundhara Raja and Vasuki proposed an automatic segment of blood vessels through the elimination of optic disc[76]. Non linear filter, anisotropic diffusion filter and morphological operations are used for detecting the retinal blood vessels. It outperforms other segmentation methods with an accuracy of 98.08%.Waleed Al-Nuaimy et al.[77] focused on the detection of exudate by exploiting characteristics of distinct borders using morphological gradient and rectified the unilluminating defects by using region based segmentation. Sensitivity of 93.1% is reported in this method. Zhou et al.[78] used LFSA to reconstruct new optic disc images and a generated spectrum was used for classification. Classification accuracy of 99% reported on Messidor dataset.

Liskowski et al.[79] proposed a supervised segmentation technique that uses a DNN trained on a huge collection of preprocessed images. It used a supervised segmentation technique using DNN with a sensitivity of 87%. It is an effective method for detecting the blood vessels in the retinal images. Prasad et.al.[80] extracted blood vessels using morphological processing and filtering methods. Kabir Md Ahasan[81] proposed LPBPC algorithm to remove the false blood vessels from the segmented vessels. Tan et al.[82] proposed a 10-layer convolutional neural network to automatically segment and discriminate exudates, hemorrhages and micro-aneurysms. Table 3 provides state of art related to retinal segmentation techniques.

Table 3. State of art related to retinal segmentation techniques.

Sl.No.	**Techniques**	**References**	**Dataset**	**Remarks**
1.	Region growing method.	[4][6][30]	MESSIDOR,DIARETDB1 .	Vessel segmentation.
2.	Deep Neural Networks.	[14][73]	DRIVE, STARE, and CHASE databases. DDR.	Vessel segmentation. Lesion detection .
3.	Morphological component analysis.	[15][57] [68]	MESSIDOR.	Vessel segmentation. Exudate sementation.
4.	Adaptive Thresholding	[16]	140 images	Blood vessel extraction
5.	Gaussian mixture model.	[27]	KAGGLE.	Blood vessel extraction. Exudate detection.
6.	Patch based analysis.	[35]	DIARETDB1 and e-Ophtha.	Identification of exudates, hemorrhages, and microaneurysms.
7.	Rule based and machine learning methods.	[65]	DRIVE, STARE.	Retinal vessel segmentation techniques using Rule based: Automatic local thresholding techniques. Kernel-based techniques(Matched filtering based). Mathematical morphology-based. Multi-scale Techniques. Model-based. Vessel-tracking.
8.	Density analysis and bounding box technique.	[66]	30 images. -blood	v Vessel extraction.
9.	Hough transform, Top hat transform. Watershed transform.	[36,37, [67-69,100]	DIARETDB1, HRF.	Blood vessel segmentation. Optic disc extraction.

				Micro aneurysm extraction.
10.	Thresholding	[6,12,17,25,30,31,38,40,45,57, 60,68,69,99]	DIARETDB1, HRF. Retinopathy Online Challenge dataset and E-Ophtha-MA dataset. DRIVE.	Blood vessel extraction. Detection of Microaneurysms. Detection of hard exudates.
11.	Intensity based properties.	[70]	DIABETDB1.	Exudate. Blood vessel extraction. Microaneurysms. Hemorrhages.
12.	Canny edge detection.	[71, 72]	STARE. and CHASE_DB1.	Detect normal and abnormal blood vessels. Detect MA and Hemorrhage
13.	U-net model.	[74]	DIARETDB1, HEI-MED, MESSIDOR.	Detection of exudates.
14.	Saliency method.	[75]	DIARETDB1 .	Detection of optic disc and exudates.
15.	Anisotropic diffusion filter.	[76]	DRIVE and STARE.	Detection of optic disc from the green channel retinal image. Lesion segmentation.
16.	Fuzzy C-means clustering.	[48]	100 images(Resolution of 1280 x 1024 or 700 x 605 Pixels).	Detection of blood vessels from the sample fundus images.
17.	Splat segmentation.	[91]	MESSIDOR.	Detection of retinal hemorrhages.

C. Feature Extraction

Feature extraction[83] is the core phase in the detection process of DR. Classification accuracy of the CAD system can be improved by the use of advanced feature extraction methods[84]. It is very essential to identify the features that are used for classification. This preliminary phase is crucial for the later stages of the detection of DR. Many features such as blood vessels, exudates, micro aneurysms and location of optic disc are extracted and analyzed for the classification purpose. Faust O et al.[11] described detailed view of the retinal features. Chaudhuri[12] used two dimensional matched filters to extract the blood vessels to differential the severity level of DR detection. Priya and Aruna[44] extracted statistical features such as radius, diameter, and area, arc length to find hemorrhage or exudate. Sarah Barman et al.[47] applied region growing method and mathematical morphology techniques to extract image components. Multi scale correlation coefficients were used to detect bright spot in the image. In another method Chand et.al.[53] used gray level co-occurrence matrix to extract the textual features from the retinal images. Pratt et al.[54] introduced recent CNN based algorithms to pull out the features accurately with an accuracy of 75%.

Bhargavi R et al.[85] used SIFT algorithm to identify the key points in an image. SIFT is one of the best among local invariant feature descriptors. By using the extracted features, the system can categorize the severity of the diseases as mild, moderate and severe. Deep neural networks extract a set of retinal features from the training data set. In addition, it performs feature selection based on deep neural network model, repeatedly removing features for the classifier model, one by one, until it reaches a stable performance measure. Nisha et al.[87] proposed a method that combines texture and vessel feature extraction based on Gabor wavelet methods. This work used entropy filter and range filters which were used to extract the neighborhood features from retinal images. Table 4 provides state of art feature extraction techniques.

Table 4. Feature extraction techniques.

Feature extraction techniques	**References**	**Remarks**
Matched Fillter	[12,66,92]	Blood vessel extraction.
Morphological transform and Gradient operator.	[16][36] [40]	Optic disc removal and blood vessel extraction.
Saddle and D-saddle Feature detector	[29]	Vessel feature extraction.
Local Energy model and LESH.	[33]	It gathers the attributes at points of an image where the local frequency components are maximized in term of phase.

	[33]	LTP uses a fixed threshold to make binary patterns extraction more robust.
Local Ternary Co-occurrence Pattern.		Uses a fixed threshold to make binary patterns extraction stronger.
SURF.	[18,41, 96]	Detector and a descriptor for points of interest in images where the image is transformed into coordinates.
Thresholding.	31,[44]	Exudates and optic disc extraction. A straightforward shape extraction technique. Method of producing regions of uniformity within an image based on some threshold criterion.
Top hat transform.	[37.99]	Micro aneurysm and blood vessel extraction.
Circular hough transform. Region of Interest (ROI)	[40]	Optic disc detection.
DWT.	[44, 84]	Extraction of wavelet based features such as aspect ratio, eccentricity, entropy etc.
GLCM.	[49, 53, 68, 84, 88]	Extraction of texture and intensity features.
FFT.	[51]	Transform domain feature extraction.
Principal Component Analysis	[27,71, 83]	Blood vessel extraction. Extraction of statistical features.
Local Binary Pattern.	[67,80,90]	Extraction of texture features sensitive to noise.
Spherical directional local ternary pattern (SDLTP)	[83]	Extraction of statistical features based on direction.
SIFT.	[27,85]	Local invariant feature descriptor extraction.
Spat based feature extraction	[91]	Detection of Hemorrhage .

D. Classification

Main goal of automated classifier is to classify the images in to a particular category based on the features extracted from them. Earlier work on DR mainly focused on image processing techniques, which were initially used as the technology for the classification of retinal images. Other than these conventional methods, neural networks provide promising result for classification tasks. New generation algorithms use statistical, machine learning, visualization, and other deep learning techniques[89] for the classification purpose.

SVM classifier[2] is employed to classify the images into normal, mild, moderate, severe and proliferative categories. This system extracted blood vessels and exudates for the detection of DR. Accuracy of the system is 96.4%. It uses small data sets such as DIARETDB0 and DIARETDB1databases. Neural network multi-layer perceptron system for classification of retinal images was a powerful mechanism[38] to extract the lesions. Accuracy of the method is 91%. For good accuracy, brighter lesion identification algorithms are needed. More number of features can be added to improve the efficiency of the system. Preprocessing techniques including image enhancement will improve the accuracy of the system. Priya et al.[44] models PNN, Bayesian Classification and SVM and their performances are compared. Authors claimed SVM outperforms other two methods with an accuracy of 95.3%. Agurt et al.[46] uses a cross validation approach, which classifies retinal images into different grades. K-means clustering was used for grouping the information and a linear regression model, Partial Least Square (PLS) was used to identify the classes with an accuracy of 92%. Velázquez-González et al.[49] classified the severity of DR into Normal DR, Light NPDR, moderate and severe DR using BPNN. Accuracy of the system is 92%. Accuracy coukd be increased by employing more retinal images and strong algorithmic implementations. Exudate detection from normal and abnormal image was scrutinized using SVM classifier[53] by Chand et.al.. Method always provided better results when it used with high dimensional data sets. Exudates were detected using SVM classifier with an accuracy of 92%. Prasad et al.[55] proposed one rule classifier and BPNN for the detection of DR. Data set used is DIARETDB1 database. Method gives an accuracy of 93.8% for BPNN and 97.75% for one rule classifier. Ratanak Khoeun et.al.[86] proposed unique approach using Matched Filters and Area Based Noise Removing to extract the micro aneurysms in the retina images. Mahiba C et .al[90] proposed a multi class DR classification methodology using hybrid color, texture features and modified CNNs. The overall classification accuracy of the proposed system is 98.41%.

In[91] classification carried out using QDA in which over fitting may occur. Class labels were identified Using K-Nearest Neighbour. Accuracy achieved by the system is 87%. Maher et.al.[92] described a comparison between different classification methods and their accuracies. Rahim et al.[93] used decision tree and the k-nearest neighbor classifiers to classify the images in to different categories. It classifies data in to two classes with the accuracy of 74% for BDT and 78% for k-nearest neighbor. Table 5provides summary of the detection of DR. Table 6 provides Advantages/Disadvantages of DR detection/classification schemes.

Table 5. Summary of the detection of DR.

Ref. No	Major steps of methodology	Features used	Dataset used	Performance/Remarks
[16]	i)Green channel extraction. ii)Local-phase method is emloyed . iii)Morphological operator used for vessel extraction. iv)Gradient operator used for removing optic disc. v)ANOVA test for classification	Blood vessels,optic disc	140 images collected at Department of Ophthalmology, Kas-turba Medcial College, Manipal, India.	Accuracy of :91%for normal, 92.7% for PDR, and 87.8% for NPDR images. This method outperforms [12]
[26]	i) Gray scale conversion is performed. ii) CLAHE is employed in next phase. iii) Error-correcting output codes (ECOC) has been used for classification.	Statistical features of Visibility graph representation.	MESSIDOR.	Accuracy of :92%. Sensitivity :95.83%. Specificity : 98.61%. This method can effectively applied for DR grading and give better performance than [35,36,40,48]
[33]	i) Gray scale conversion is performed. ii) Local Energy model and Local Energy-based Shape Histogram (LESH) used for extracting the point of interests. iii) SVM with a Radial Basis Function kernel (SVM-RBF) is employed.	Texture features.	MESSIDOR.	Accuracy:90.04%. It gives best performance due to LESH features and perform better than [48,91,93,94,97,99]
[35]	i) Contrast enhancement is employed. ii) Patch preparation and image analysis is performed. iii) CNN is applied for DR detection.	Exudates, hemorrhages and micro aneurysms.	DIARETDB1. e-Ophtha.	Accuracy :DIARETDB1 97.3%. 86.6% for e-Ophtha. Patch and image-based analysis gives better accuracy than other studies such as [36,37,40,48,54,2,74,84,96,97, 98,99]
[36]	i) Morphological processed image is binarized. ii) Noises are reduced. iii) SVM and decision tree is used for classification.	Blood vessel density. Actual number of micro aneurysms. Density of hard exudates. Quantitative features.	MESSIDOR.	SVM Accuracy:92.4%. Decision-tree Accuracy:92%. It provides an optimized better result than [48,68,97] using similar classifiers.
[37]	i) Median filtering is applied. ii) Green channel extraction is performed. iii) K Nearest Neighbor is used in the final stage.	Micro Aneurysm.	1441images.	Sensitivity : 85.4%. Specificity : 83.1%. This work detects MA only.
[40]	i) Green channel extraction and Fuzzy filtering is performed. ii) Fuzzy Histogram Equalization is employed. iii) Classifiers such as k-Nearest Neighbour, Polynomial Kernel SVM, RBF Kernel SVM and Naïve-Bayes are	Exudates. Blood vessels,optic disc.	600 images collected at the Hospital Melaka, Malaysia.	K nearest neighbor Accuracy:93%. Polynomial Kernel SVM Accuracy:70%. RBF Kernel SVM Accuracy:93%. Naïve-Bayes

	used for DR detection/classification.			Accuracy:75%. This work out performs [46] which uses AM-FM features.
[48]	(i)AHE,Matched Fitering,Gray scale conversion is performed. ii)DWT is employed. iii)Fuzzy C-means clustering is used for segmentation. iV)PNN and SVM are used for classification.	Statistical features.	100 images(Resolution of 1280 x 1024 or 700 x 605 Pixels).	Accuracy:87.68% Sensitivity : 81.42% Specificity : 100%. DR system using SVM gives more accuracy than PNN (70%). It outperforms [97]
[51]	i) Image is resized. ii) Gray scale conversion is done. iii) Feature vector is formed using FFT. iv)Multi Layer Perceptron is employed.	Statistical parameters.	DIARETDB0.	Accuracy:99%. MLPNN classifier perform well with 09 hidden PEs, learning rule momentum, transfer function tanh and step size 0.1.FFT based feature extraction gives best accuracy than [52,54,70,74,84,88,94-98]
[52]	i) Grayscale conversion is performed on image. ii) Gabor filter is applied. iii) Mathematical morpology is used for image segmentation. iv) Probabilistic,geometric,tree based and KNN classification methods are applied for DR detection/classification.	Statistical and geometric features.	MESSIDOR. DIARETDB. DRIVE. VDIS HRF HRIS e-ophtha	Average accuracy: 98.58%. This method is evaluated on 7 data sets . DIARETDB gives an accuracy of 96.6 for KNN,96.8 %for SVM,96.6 %for Bayesian and 91.7 for ensemble classifier.
[70]	i) Green channel extraction is performed. ii) HE is employed. iii) Morphological operations are applied. iv) SVM is used for classification.	Number and area of MA.	DIABETDB1.	Sensitivity:96%. Specificity:92%. Proposed method gives better sensitivity and specificity for NPDR detection than [84,97]
[74]	i) Image is resized. ii) Green channel extraction is performed. iii) Adaptive contrast enhancement technique is applied. iv) Conditional GAN is used in the final stage of the method developed.	Exudates.	e-Ophtha_EX. DIARETDB1. HEI-MED. MESSIDOR.	Accuracy:95.45%. With CGAN Specificity 92.13%, Sensitivity:88.76%, and F1score:89.58%, without cGAN thhe sores are 86.36%, 87.64%, 76.33%, 86.42% respectively. System performance is better than [95,98]
[84]	i) Candidate region is extracted. ii) Contrast enhancement is performed. iii) Morphological operations are applied to enhance MAs. iv) Multi layered feed forward neural network (FFNN) and support vector machine classifiers were applied.	Micro aneurysm Statistical and wavelet features	DIARETDB1.	SVM Accuracy, Sensitivity, Specificity: 95%, 76%, 92% respectively. Multi layered feed forward neural network (FFNN) classifier Accuracy, Sensitivity, Specificity:92%, 79%, 90% respectively.
[88]	i) Image is resized. ii) CLAHE is applied. iii) SVM is employed for classification.	Blood vessel area, Micro aneurysms area, Exudates area and Texture features	DIARETDB0. DIARETDB1.	Accuracy:96.67%. SVM classifier reports good results.
[94]	i) HE is employed. ii) Morphological operations are applied. iii) Image is binarized.	Area of veins, hemorrhages and micro aneurysms	124 retinal images.	Sensitivity:90%. High computational cost.

	iv) Back propagation algorithm is used for DR detection/classification.			
[95]	i) Data augmentation is performed. ii) CNN is employed for classification.	Hard exudates, red lesions, micro aneurysms and blood vessel	KAGGLE.	Accuracy:94.5%. CNN gives better accuracy than [38,91,93,96,9,99].This method gives more accuracy than [54] on kaggle data set.
[96]	i) Green channel extraction is performed. ii) Image is resized. iii) AHE and morphological operations are performed. iv) SVM is used in the final stage of DR detection.	Lesions	DRIDB0,DRIDB1, MESSIDOR, STARE and HRF.	Accuracy:94.4%. This method outperforms [48] .Adopted SIFT for feature extraction.
[97]	i) Daubechies Wavelet transform is employed. ii) Background normalization is performed. iii) Median filtering is applied. iv) SVM classifier is employed in the final stage.	Micro aneurysm	LaTIM (Laboratoire de Traitement de lInforma-tion Médicale). E-ophtha,ROC database.	Sensitivity: 62% (LaTIM). Sensitivity:66% (eophtha). Sensitivity:32 %. (ROC) MA are detected easily by extracting laws texture based features from small regions at a time.
[98]	i) Normalization and Data augmentation is performed. ii) Non-Local Means Denoising is performed iii) CNN is used for classification.	Automatic feature Identification	KAGGLE.	Accuracy:95.68%. Transfer learning and hyper parameter-tuning methods are applied to improve the performance of this system. It out performs [95] and [54]
[99]	i) Discrete Curvelet Transform is applied. ii) Morphological operations are applied. iii) Simple thresholding is performed. iv) Connected component analysis is performed.	Blood vessels and statistical features	DRIVE.	Accuracy: 94%. Contrast of the images can be improved by FDCT.

III. Result and Discussions

This study of DR detection techniques reveals the requirement for numerous preprocessing techniques needed for the noise reduction. The efficiency of the CAD system can be improved by applying preprocessing techniques effectively at the early stages. Most of the papers in this survey applied contrast enhancement and filtering techniques for preprocessing. Importance of various techniques in the survey is shown in the [Fig.4]. Various extracted features from the retinal images are the input to the classification algorithms. Different authors were used different features of the retinal images including blood vessels, exudates, micro aneurysms, hemorrhages, optic disc and other statistical features.

Among the methods mentioned in this paper, work[26] reported an accuracy of 92%, which adopted statistical features and error correcting output codes. Due to FFT based feature extraction on DIARETDB dataset Bhatkar et.al. Method[51]outperforms other methods with 99% success rate. Method[97] exhibited poor accuracy because of its restricted use of features extraction and preprocessing techniques . Amin et. al.[52] evaluated the performance on 7 data sets and applied various classification techniques for DR detection. Among the classification techniques employed, KNN outperforms other methods.

Optic disc localization using morphological erosion provided an accuracy of 97%. Blood vessel extraction is the main segmentation scheme used in most of the papers in this survey. This can achieve accuracy around 98%. Segmentation of exudates is another technique used with an accuracy of 93%. The frequency of segmentation methods used in the classification algorithms are depicted in the [Fig.5]. According to the survey conducted, machine learning and deep learning techniques can improve the accuracy of segmentation techniques.

This survey focuses on the different classification algorithms used in the detection and prediction of retinal images. The study includes the various image processing methods used in classification process. With the introduction of AI, classification process of CAD system improves its performance. Image processing techniques along with machine learning techniques evolved as a technological development in the diagnosis system for DR. SVM and Naive Bayes classifiers perform well in the data sets of DR detection with an accuracy of 92%. Deep learning technique such as CNN is the common method used in the field of medical imaging. It can provide promising results with larger data sets. Wan, Shaohua et.al.[98] usesd CNN algorithm on Kaggle dataset and reported an accuracy of 95.86%. It used preprocessing techniques such as augmentation and normalization on the datasets to improve the accuracy. Requirement of a large dataset is the great challenging task associated with the DNN. Synthetic images can be generated using the generative networks like GAN. Application of

GAN networks in retinal data sets can be used to generate images and can be used for noise reduction. AI techniques such as machine learning and deep learning algorithms effectively classify the retinal images.

SVM-RBF followed by LESH performed better in the classification of DR. Method reported an encouraging result. Fuzzy Histogram Equalization preceded by SVM-RBF reported an accuracy of 93%in DR detection. Multilayer perceptron which uses features from FFT reported an accuracy above 98% on DIARETDB0 data set. In the detection of DR, KNN followed by the probabilistic geometric tree method showed a better performance with an average accuracy of 98.95% over 7 data sets.

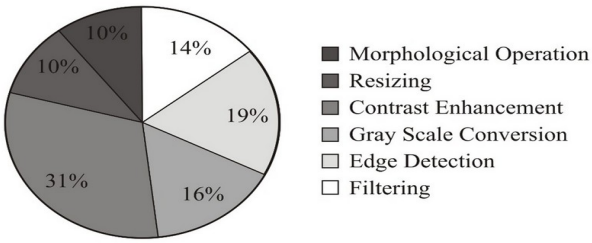
Performance of CAD system for DR detection can be further improved by the application of relevant techniques. Deep neural networks along with transfer learning will improve the accuracy of the systems. Transfer learning can be used for different strategies and can be combined (hybridized approach) to get better performance for a CAD system. In some cases, methods dealing with imbalanced data classification will improve the performance of the CAD system. Application of relevant under sampling or oversampling techniques will solve the imbalanced dataset issue and will improve the performance of the CAD system. Performance of blood vessel and lesion segmentation methods can be further improved by employing relevant dimensionality reduction techniques. It is a challenging task to compare various methodologies in terms of its accuracy, because most of the researchers have used different techniques and various datasets. Detection of different size, shape and position of the lesions can be a major challenge in the classification of DR. Transfer learning based CNN [50] can be applied to improve the performance of CAD system.

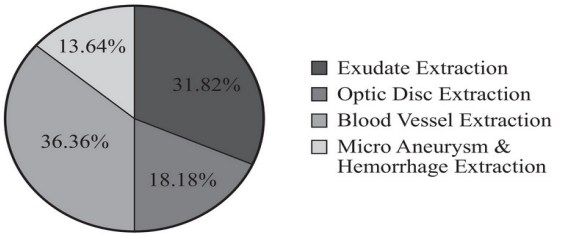
Fig. (4).Review of various preprocessing techniques used in the survey.

Fig. (5). Review of various segmentation methods

Table 6. Advantages/Disadvantages of some DR detection/classification schemes.

Work	Advantages(A) /Disadvantages (D)	Work	Advantages(A) /Disadvantages (D)	Work	Advantages(A) /Disadvantages (D)
[16]	(A) The proposed method uses local-phase method.	[32]	(A) Novel dataset with new labeling scheme is useful for ce. false negative graded as erity class. from more re not included.	[33]	(A) Use of Local Ternary Pattern (LTP) and Local Energy-based Shape Histogram (LESH) captures the connection between neighboring pixels and features, which are less sensitive to variation in illumination, color and noise. It can easily be differentiated between DR and non-DR images. (D) Suitable only for small data sets.
	available databases with the training done on one and the testing on both with similar outcomes. (D) Unfit to separate among		n to detection of performs grade classification. The outcome was additionally optimized by choosing the most relevant features and classification parameters. number of micro as not identified images. Texture l soft exudates considered for	[39]	(A) Proposed method is robust to noise. Does not depend on the specific features of lesions present in the retinal images. (D) Performance of blood vessel and lesion segmentation methods are not satisfactory.
[4]			ic and statistical sed for better . Multiple data evaluation. of DR grading is le with only	[53]	(A) SVM along with textual features provides good results to extract the exudates in DR images. (D) Utilize only optic disc segmentation.





	retinal structures must be removed. Hence, strong preprocessing methods are needed.		exudate segmentation.		
[62]	(A) Good performances on low quality images. (D) Used as a primary diagnosis tool, needs the support of ophthalmologist.	[74]	(A) Methodology applied a cGAN framework to solve the imbalance data classification problem. It also generate synthetic images to improve the generalization property of the network. (D) In the proposed methodology extra training phase is required for cGAN.	[80]	(A) Combination of neural network and fuzzy classification improves the accuracy of multi stage DR classification. The main advantage of the system is the learning capability and its simplicity. (D)Efficient segmentation methods not used for detecting retinal features.
[84]	(A) Proposed system make use of an optimal feature set for accurate detection and classification of DR. Accurate blood vessel extraction improves the accuracy of the system. (D) Other lesions present in the retinal images such as exudates were not utilized.	[87]	(A) Time constraints is reduced by the proper usage of feature reduction techniques. (D) Automatic segmentation techniques not used in this paper.	[88]	(A) Apart from the lesion features of retinal images, proposed system also includes texture features for classification. This will improve the classification accuracy. (D) Only 4 texture features and small set of images considered for evaluation.
[91]	(A) Splat-based image representation provides an efficient and optimized way to model uneven shaped abnormalities in medical images. It also took less time for classification. (D) Reference model was used only part of the dataset and took review only from a single expert.	[96]	(A) This method is a feasible, efficient, and time saving way of DR detection. Proposed method discriminates retinal images without earlier requirement for segmentation of blood vessels and optic disc. (D) Limited number of retinal images were used.	[97]	(A) Higher-order statistics method, such as Laws masks combined with conventional methods used to detect MA with a promising result. (D) Two trains and two test phases were needed. Still preprocessing was not so effective.
[100]	(A) Different specific features and datasets combined for the detection of DR. (D) Datasets with only 516 images used for the evaluation.	[101]	(A) Deep transfer learning is an appropriate method for multi-categorical classification of fundus images. (D) Due to the small data sets deep learning technique for DR detection is not so effective.	[102]	(A) Custom morphology based multi scale descriptors for lesion analysis improved the performance of the system. (D) Study only included darkly pigmented Asian Indian patients. Therefore, the result can't be generalized.

IV. CONCLUSION

Nowadays, retinal imaging and its diagnosis is an emerging field in medical imaging. This survey presented a detailed study of CAD system for DR. This paper presented methods that are used for building a CAD system for diabetic retinopathy screening. This work will be helpful for the researchers and technical persons who want to utilize the ongoing research in this area. A researcher to get an insight in to the various segmentation, feature extraction and classification methods can use our paper. The accuracy of various classification algorithms were compared by analyzing their performances. Efficiency of the CAD system can be improved by the development of deep neural networks on large data sets. New annotated data sets can be created for efficient clinical applications. Synthetic retinal images can be created by applying networks like GAN and can be used for the processing of deep neural networks. DR detection systems still need improvements in grading the severity of the retinal diseases. CAD system for DR can be further upgraded by adapting new acquisition and optimized pre-processing methods in retinal data sets.

REFERENCES

[1] Larsen M, Godt J, Larsen N, Lund-Andersen H, Sjølie AK, Agardh E, Kalm H, Grunkin M, Owens DR. Automated detection of fundus photographic red lesions in diabetic retinopathy. Investigative ophthalmology & visual science. 2003 Feb 1; 44(2):761-6.

[2] Maher RS, Kayte SN, Meldhe ST, Dhopeshwarkar M. Automated diagnosis of non-proliferative diabetic retinopathy in fundus images using support vector machine. International Journal of Computer Applications. 2015 Jan 1; 125(15).

[3] Torre J, Valls A, Puig D. A deep learning interpretable classifier for diabetic retinopathy disease grading. Neurocomputing. 2020 Jul 5; 396:465-76.

[4] Abramoff MD, Niemeijer M, Russell SR. Automated detection of diabetic retinopathy: barriers to translation into clinical practice. Expert review of medical devices. 2010 Mar 1; 7(2):287-96.

[5] Pires R, Avila S, Wainer J, Valle E, Abramoff MD, Rocha A. A data-driven approach to referable diabetic retinopathy detection. Artificial intelligence in medicine. 2019 May 1; 96:93-106.

[6] Palavalasa KK, Sambaturu B. Automatic diabetic retinopathy detection using digital image processing. In 2018 International Conference on Communication and Signal Processing (ICCSP). 2018 Apr 3 (pp. 0072-0076). IEEE.

[7] Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Progress in retinal and eye research. 2018 Nov 1; 67:1-29.

[8] Abràmoff MD, Garvin MK, Sonka M. Retinal imaging and image analysis. IEEE reviews in biomedical engineering. 2010 Dec 10; 3:169-208.

[9] Fadafen MK, Mehrshad N, Razavi SM. Detection of diabetic retinopathy using computational model of human visual system.(2018).

[10] Xiao D, Bhuiyan A, Frost S, Vignarajan J, Tay-Kearney ML, Kanagasingam Y. Major automatic diabetic retinopathy screening systems and related core algorithms: a review. Machine Vision and Applications. 2019 Apr 15; 30(3):423-46.

[11] Faust O, Acharya R, Ng EY, Ng KH, Suri JS. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. Journal of medical systems. 2012 Feb 1; 36(1):145-57.

[12] Chaudhuri S, Chatterjee S, Katz N, Nelson M, Goldbaum M. Detection of blood vessels in retinal images using two-dimensional matched filters. IEEE Transactions on medical imaging. 1989 Sep; 8(3):263-9.

[13] Tan JH, Fujita H, Sivaprasad S, Bhandary SV, Rao AK, Chua KC, Acharya UR. Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. Information sciences. 2017 Dec 1; 420:66-76.

[14].Liskowski P, Krawiec K. Segmenting retinal blood vessels with deep neural networks. IEEE transactions on medical imaging. 2016 Mar 24; 35(11):2369-80.

[15] Kasurde SD, Randive SN. An automatic detection of proliferative diabetic retinopathy. In 2015 International Conference on Energy Systems and Applications 2015 (pp. 86-90). IEEE.

[16] Yelampalli PK, Nayak J, Gaidhane VH. Blood vessel segmentation and classification of diabetic retinopathy images using gradient operator and statistical analysis. In Proceedings of the World Congress on Engineering and Computer Science 2017 (Vol. 2).

[17] Adalarasan R, Malathi R. Automatic detection of blood vessels in digital retinal images using soft computing technique. Materials Today: Proceedings. 2018 Jan 1; 5(1):1950-9.

[18] Costa P, Galdran A, Smailagic A, Campilho A. A weakly-supervised framework for interpretable diabetic retinopathy detection on retinal images. IEEE Access. 2018 Mar 15; 6:18747-58.

[19] Li YH, Yeh NN, Chen SJ, Chung YC. Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network. Mobile Information Systems. 2019 Jan 1; 2019.

[20] Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Procedia Computer Science. 2016 Jan 1;90:200-5.

[21] Zhao H, Li H, Maurer-Stroh S, Cheng L. Synthesizing retinal and neuronal images with generative adversarial nets. Medical image analysis. 2018 Oct 1; 49:14-26.

[22] Perdomo O, Rios H, Rodríguez FJ, Otálora S, Meriaudeau F, Müller H, González FA. Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography. Computer methods and programs in bio medicine. 2019 Sep 1; 178:181-9.

[23] Hemanth DJ, Deperlioglu O, Kose U. An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network. Neural Computing and Applications. 2020 Feb 1; 32(3):707-21.

[24] Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. PLOS one. 2019 Jun 6; 14(6):e0217541.

[25] Eftekhari N, Pourreza HR, Masoudi M, Ghiasi-Shirazi K, Saeedi E. Microaneurysm detection in fundus images using a two-step convolutional neural network. Biomedical engineering online. 2019 Dec 1; 18(1):67.

[26] Mohammadpoory Z, Nasrolahzadeh M, Mahmoodian N, Haddadnia J. Automatic identification of diabetic retinopathy stages by using fundus images and visibility graph method. Measurement. 2019 Jul 1; 140:133-41.

[27] Mansour RF. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. Biomedical engineering letters. 2018 Feb 1; 8(1):41-57.

[28] Chudzik P, Majumdar S, Calivá F, Al-Diri B, Hunter A. Microaneurysm detection using fully convolutional neural networks. Computer methods and programs in biomedicine. 2018 May 1; 158:185-92.

[29] Ramli R, Idris MY, Hasikin K, Karim NK, Abdul Wahab AW, Ahmedy I, Ahmedy F, Kadri NA, Arof H. Feature-based retinal image registration using D-Saddle feature. Journal of healthcare engineering. 2017 Jan 1; 2017.

[30] Valverde C, Garcia M, Hornero R, Lopez-Galvez MI. Automated detection of diabetic retinopathy in retinal images. Indian journal of ophthalmology. 2016 Jan; 64(1):26.

[31] Sisodia DS, Nair S, Khobragade P. Diabetic retinal fundus images: Preprocessing and feature extraction for early detection of diabetic retinopathy. Biomedical and Pharmacology Journal. 2017 Jun 20; 10(2):615-26.

[32] Gao Z, Li J, Guo J, Chen Y, Yi Z, Zhong J. Diagnosis of diabetic retinopathy using deep neural networks. IEEE Access. 2018 Dec 19; 7:3360-70.

[33] Chetoui M, Akhloufi MA, Kardouchi M. Diabetic retinopathy detection using machine learning and texture features. In 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE) 2018 May 13 (pp. 1-4). IEEE.

[34] Qureshi I, Ma J, Abbas Q. Recent development on detection methods for the diagnosis of diabetic retinopathy. Symmetry. 2019 Jun; 11(6):749.

[35] Khojasteh P, Aliahmad B, Kumar DK. Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. BMC ophthalmology. 2018 Dec; 18(1):1-3.

[36] Carrera EV, González A, Carrera R. Automated detection of diabetic retinopathy using SVM. In 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON) 2017 Aug 15 (pp. 1-4). IEEE.

[37] Fleming AD, Philip S, Goatman KA, Olson JA, Sharp PF. Automated micro aneurysm detection using local contrast normalization and local vessel detection. IEEE transactions on medical imaging. 2006 Aug 21; 25(9):1223-32.

[38] Handsková V, Pavlovicova J, Oravec M, Blasko R. Diabetic rethinopathy screening by bright lesions extraction from fundus Images. Journal of Electrical Engineering. 2013 Sep 1; 64(5):311.

[39] Imani E, Pourreza HR, Banaee T. Fully automated diabetic retinopathy screening using morphological component analysis. Computerized medical imaging and graphics. 2015 Jul 1; 43:78-88.

[40] Rahim SS, Palade V, Shuttleworth J, Jayne C. Automatic screening and classification of diabetic retinopathy and maculopathy using fuzzy image processing. Brain informatics. 2016 Dec 1; 3(4):249-67.

[41] Mangrulkar RS. Retinal image classification technique for diabetes identification. In 2017 International Conference on Intelligent Computing and Control (I2C2) 2017 Jun 23 (pp. 1-6). IEEE.

[42] Sreng S, Takada JI, Maneerat N, Isarakorn D, Varakulsiripunth R, Pasaya B, Panjaphongse MR. Feature extraction from retinal fundus image for early detection of diabetic retinopathy. In 2013 IEEE Region 10 Humanitarian Technology Conference 2013 Aug 26 (pp. 63-66). IEEE.

[43] Ramasubramanian B, Selvaperumal S. A comprehensive review on various preprocessing methods in detecting diabetic retinopathy. In 2016 international conference on communication and signal processing (ICCSP) 2016 Apr 6 (pp. 0642-0646). IEEE.

[44] Priya R, Aruna P. Diagnosis of diabetic retinopathy using machine learning techniques. ICTACT Journal on soft computing. 2013 Jul; 3(4):563-75.

[45] Estabridis K, de Figueiredo RJ. Automatic detection and diagnosis of diabetic retinopathy. In 2007 IEEE International Conference on Image Processing 2007 Nov (Vol. 2, pp. II-445). IEEE.

[46] Agurto C, Murray V, Barriga E, Murillo S, Pattichis M, Davis H, Russell S, Abràmoff M, Soliz P. Multiscale AM-FM methods for diabetic retinopathy lesion detection. IEEE transactions on medical imaging. 2010 Feb 2; 29(2):502-12.

[47] Sopharak A, Uyyanonvara B, Barman S. Automated microaneurysm detection algorithms applied to diabetic retinopathy retinal images. Maejo International Journal of Science and Technology. 2013 May 1; 7(2):294.

[48] Sayed S, Kapre S, Inamdar D. Detection of diabetic retinopathy using image processing and machine learning. International Journal of Innovative Research in Science, Engineering and Technology. 2017 Jan; 6(1):99-107.

[49] Velázquez-González JS, Rosales-Silva AJ, Gallegos-Funes FJ, Guzmán-Bárcenas GD. Detection and classification of non-proliferative diabetic retinopathy using a back-propagation neural network. Revista Facultad de Ingeniería Universidad de Antioquia. 2015 Mar(74):70-85.

[50] Gour N, Khanna P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. Biomedical Signal Processing and Control. 2020 Dec 9:102329.

[51] Bhatkar, Amol P, and Govind K. FFT based detection of diabetic retinopathy in fundus retinal images. Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2016.

[52] Amin J, Sharif M, Yasmin M, Ali H, Fernandes SL. A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions. Journal of Computational Science. 2017 Mar 1; 19:153-64.

[53] Chand CR, Dheeba J. Automatic detection of exudates in color fundus retinopathy images. Indian Journal of Science and Technology. 2015 Oct; 8(26).

[54] Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Procedia Computer Science. 2016 Jan 1; 90:200-5.

[55] Prasad DK, Vibha L, Venugopal KR. Early detection of diabetic retinopathy from digital retinal fundus images. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) 2015 Dec 10 (pp. 240-245). IEEE.

[56] Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, Massin P, Cochener B, Gain P, Tang L, Lamard M. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA ophthalmology. 2013 Mar 1; 131(3):351-7.

[57] Partovi M, Rasta SH, Javadzadeh A. Automatic detection of retinal exudates in fundus images of diabetic retinopathy patients. Journal of Research in Clinical Medicine. 2016 May 9; 4(2):104-9.

[58] Mookiah MR, Acharya UR, Chua CK, Lim CM, Ng EY, Laude A. Computer-aided diagnosis of diabetic retinopathy: A review. Computers in biology and medicine. 2013 Dec 1; 43(12):2136-55.

[59] Sathananthavathi V, Indumathi G, Vishalini R, Alphia J. Automatic detection of microaneurysms in retinal images for diabetic retinopathy. International Journal of Pure and Applied Mathematics. 2018; 119(15):1349-55.

[60] Sindhura A, Kumar SD, Sajja VR, Rao NG. Identifying exudates from diabetic retinopathy images. In 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) 2016 May 25 (pp. 132-136). IEEE.

[61] Qureshi I, Ma J, Shaheed K. A hybrid proposed fundus image enhancement framework for diabetic retinopathy. Algorithms. 2019 Jan; 12(1):14.

[62] Wisaeng K, Hiransakolwong N, Pothiruk E. Automatic detection of exudates in diabetic retinopathy images. Journal of Computer Science. 2012 Aug 1; 8(8):1304.

[63] Long S, Huang X, Chen Z, Pardhan S, Zheng D. Automatic detection of hard exudates in color retinal images using dynamic threshold and SVM classification: algorithm development and evaluation. BioMed research international. 2019 Jan 23;2019.

[64] Sundaram R, KS R, Jayaraman P. Extraction of blood vessels in fundus images of retina through hybrid segmentation approach. Mathematics. 2019 Feb;7(2):169.

[65] Almotiri J, Elleithy K, Elleithy A. Retinal vessels segmentation techniques and algorithms: a survey. Applied Sciences. 2018 Feb; 8(2):155.

[66] Verma K, Deep P, Ramakrishnan AG. Detection and classification of diabetic retinopathy using retinal images. In 2011 Annual IEEE India Conference 2011 Dec 16 (pp. 1-6). IEEE.

[67] Shyam L, Kumar GS. Detection of glaucoma and diabetic retinopathy from fundus images by blood vessel segmentation. 2016.

[68] Prasad DK, Vibha L, Venugopal KR. Early detection and multistage classification of diabetic retinopathy using random forest classifier. 2018.

[69] Verma SB, Yadav AK. Detection of hard exudates in retinopathy images. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal. 2019; 8(4):41-8.

[70] Kumar S, Kumar B. Diabetic retinopathy detection by extracting area and number of micro aneurysm from colour fundus image. In 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN) 2018 Feb 22 (pp. 359-364). IEEE.

[71] Kaur, Sukhpreet, and Kulwinder Singh Mann. Optimized technique for detection of diabetic retinopathy using segmented retinal blood vessels. 2019 International Conference on Automation, Computational and Technology Management (ICACTM). IEEE, 2019.

[72] Ratanapakorn T, Daengphoonphol A, Eua-Anant N, Yospaiboon Y. Digital image processing software for diagnosing diabetic retinopathy from fundus photograph. Clinical Ophthalmology (Auckland, NZ). 2019; 13:641.

[73] Li T, Gao Y, Wang K, Guo S, Liu H, Kang H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences. 2019 Oct 1; 501:511-22.

[74] Zheng R, Liu L, Zhang S, Zheng C, Bunyak F, Xu R, Li B, Sun M. Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network. Biomedical optics express. 2018 Oct 1; 9(10):4863-78.

[75] Nur N, Tjandrasa H. Exudate segmentation in retinal images of diabetic retinopathy using saliency method based on region. In Journal of Physics: Conference Series 2018 Nov (Vol. 12110).

[76] Siva Sundhara Raja D, Vasuki S. Automatic detection of blood vessels in retinal images for diabetic retinopathy diagnosis. Computational and mathematical methods in medicine. 2015 Feb 24; 2015.

[77] Jaafar HF, Nandi AK, Al-Nuaimy W. Detection of exudates from digital fundus images using a region-based segmentation technique. In 2011 19th European signal processing conference 2011 Aug 29 (pp. 1020-1024). IEEE.

[78] Zhou W, Wu H, Wu C, Yu X, Yi Y. Automatic optic disc detection in color retinal images by local feature spectrum analysis. Computational and mathematical methods in medicine. 2018 Jun 14; 2018.

[79] Liskowski P, Krawiec K. Segmenting retinal blood vessels with deep neural networks. IEEE transactions on medical imaging. 2016 Mar 24; 35(11):2369-80.

[80].Prasad DK, Vibha L, Venugopal KR. Multistage classification of diabetic retinopathy using fuzzy neural

network classifier. ICTACT Journal on Image & Video Processing. 2018 May 1;8(4).

[81] Kabir MA. A rule based segmentation approaches to extract retinal blood vessels in fundus image. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS). 2020 Mar 30; 66(1):202-24.

[82] Tan JH, Fujita H, Sivaprasad S, Bhandary SV, Rao AK, Chua KC, Acharya UR. Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. Information sciences. 2017 Dec 1; 420:66-76.

[83] Randive SN, Senapati RK, Bhosle N. Spherical directional feature extraction with artificial neural network for diabetic retinopathy classification. In 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) 2018 Dec 1 (pp. 152-157). IEEE.

[84] Shirbahadurkar SD, Mane VM, Jadhav DV. Early stage detection of diabetic retinopathy using an optimal feature set. International Symposium on Signal Processing and Intelligent Recognition Systems 2017 Sep 13 (pp. 15-23).

[85] Bhargavi R, Rajesh V. Exudate detection and feature extraction using active contour model and SIFT in color fundus images. J. Eng. Appl. Sci. 2015:2362-5.

[86] Khoeun R, Rasmequan S, Chinnasarn K, Rodtuk A. Microaneurysm candidate extraction using modified matched filter. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) 2016 Jul 13 (pp. 1-5). IEEE.

[87] Chandran A, Nisha KK, Vineetha S. Computer aided approach for proliferative diabetic retinopathy detection in color retinal images. In 2016 International Conference on Next Generation Intelligent Systems (ICNGIS) 2016 Sep 1 (pp. 1-6). IEEE.

[88] Harini R, Sheela N.
In 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP) 2016 Aug 12 (pp. 1-4). IEEE.

[89] Dutta S, Manideep BC, Basha SM, Caytiles RD, Iyengar NC. Classification of diabetic retinopathy images by using deep learning models. International Journal of Grid and Distributed Computing. 2018 Jan 1; 11(1):89-106.

[90] Mahiba C, Jayachandran A. Severity analysis of diabetic retinopathy in retinal images using hybrid structure descriptor and modified CNNs. Measurement. 2019 Mar 1; 135:762-7.

[91] Tang L, Niemeijer M, Reinhardt JM, Garvin MK, Abramoff MD. Splat feature classification with application to retinal hemorrhage detection in fundus images. IEEE Transactions on Medical Imaging. 2012 Nov 15; 32(2):364-75.

[92] Raju Maher SK, Dhopeshwarkar DM. Review of automated detection for diabetes retinopathy using fundus images. International Journal of Advanced Research in Computer Science and Software Engineering. 2015 Mar; 5(3).

[93] Rahim SS, Palade V, Jayne C, Holzinger A, Shuttleworth J. Detection of diabetic retinopathy and maculopathy in eye fundus images using fuzzy image processing. In International Conference on Brain Informatics and Health 2015 Aug 30 (pp. 379-388).

[94] Yun WL, Acharya UR, Venkatesh YV, Chee C, Min LC, Ng EY. Identification of different stages of diabetic retinopathy using retinal optical images. Information sciences. 2008 Jan 2; 178(1):106-21.

[95] Xu K, Feng D, Mi H. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. Molecules. 2017 Dec; 22(12):2054.

[96] Islam M, Dinh AV, Wahid KA. Automated diabetic retinopathy detection using bag of words approach. Journal of Biomedical Science and Engineering. 2017 May 10; 10(5):86-96.

[97] Veiga D, Martins N, Ferreira M, Monteiro J. Automatic microaneurysm detection using laws texture masks and support vector machines. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2018 Jul 4; 6(4):405-16.

[98] Wan S, Liang Y, Zhang Y. Deep convolutional neural networks for diabetic retinopathy detection by image classification. Computers & Electrical Engineering. 2018 Nov 1; 72:274-82.

[99] Miri MS, Mahloojifar A. Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction. IEEE Transactions on Biomedical Engineering. 2010 Dec 10; 58(5):1183-92.

[100] Ravishankar S, Jain A, Mittal A. Automated feature extraction for early detection of diabetic retinopathy in fundus images. In 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009 Jun 20 (pp. 210-217). IEEE.

[101] Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, Rim TH. Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. PLOS one. 2017 Nov 2; 12(11):e0187336.

[102] Wang K, Jayadev C, Nittala MG, Velaga SB, Ramachandra CA, Bhaskaranand M, Bhat S, Solanki K, Sadda SR. Automated detection of diabetic retinopathy lesions on ultra wide field pseudocolour images. Acta Ophthalmologica. 2018 Mar; 96(2):e168-73.

Strategies for Correlating DB2 & CICS SMF Records to aid Problem Determination

Dr. Latha Sadanandam,
Senior Cloud Modernization Architect,
Hybrid Cloud Centre of Competency
IBM India Pvt Ltd.,
Bangalore, India

Atul Misra,
Executive IT Enterprise Architect
Hybrid Cloud Center of Competency
IBM India Pvt Ltd.,
Bangalore, India

James Roca,
WW Technology Partner Architect
IBM Cloud & Cognitive Software
IBM Services, Austin, TX ,
United States

Abstract— **Proactive problem prevention in association with quick remediation—even automated—can solve issues before they impact critical business applications. Systemsare becoming smarter, using analytics to find out preferences and make real-time suggestions for remediation. Today's Performance Management solutions require variety of components to support this plethora of systems. SMF records of IBM Z systems plays a vital role in analysis of performance records from end to end transaction perspective.**

IBM Z, SMF 101, SMF 110, Db2, CICS, Performance Analysis, CPU time, authorization ID. Etc.,

I. INTRODUCTION

Companies need an answer that keeps an eye fixed on critical z/OS metrics using intelligent and dynamic thresholds to attenuate false positives. Any basic solution must provide operators and administrators with real-time visibility into mainframe application performance and transaction flows, But an efficient solution should also help simplify mainframe performance management and increase the effective use of mainframe system resources. Proactive problem prevention in association with quick remediation—even automated—can solve issues before they impact critical business applications. Systems are becoming smarter, using analytics to find out preferences and make real-time suggestions for remediation. Today's Performance Management solutions require variety of components to support this plethora of systems. Transactions are often examined as they flow across distributed and mainframe environments to ascertain what nodes more easily are the sources of any performance impact. Hence, it becomes necessary to correlate Performance records from end to end perspective in different address spaces to create visibility in application performance and transaction flow of mainframe applications.

- System Management Facility (SMF)

IBM System Management Facility (SMF) is a component of IBM's z/OS for mainframe computers, providing a standardised method for writing out records of activity to a file (or data set to use a z/OS term). SMF provides full "instrumentation" of all baseline activities running on that IBM mainframe operating system, including I/O, network activity, software usage, error conditions, processor utilization, etc., [1]

The DB2 instrumentation Facility Component provides a powerful trace facility that can be used to record DB2 data and events. Accounting trace types collects information related to application program. Performance trace collects a lot of detailed information about any type of DB2 event. It is used to identify the causes of performance problems [4]. The DB2 accounting trace produces hundreds of valuable metrics that are contained within SMF 101 records. It includes Plan name, Authorization id, Connection id, Correlation id etc., Plan Deadlocks, timeouts, lock escalations suspensions etc., Package name, collection id, CPU time, Elapsed time and details of executed SQL.

System Management Facilities (SMF) record type 110 data is generated by CICS Transaction server for z/OS for monitoring region-wide transaction statistics, transaction-level statistics, and storage exceptions [5]. This data highlights potential problems in CICS system operation and can help you identify system constraints that affect the performance of online CICS transactions.

II. NEED FOR RELATION

Transaction from End-to-end perspective flows through various address spaces. Starting from Online transaction server (CICS) to Database (DB2). Because CICS and Db2 have different accounting needs, it is not always easy to match up Db2 accounting records and CICS performance class records [3].

Three issues while trying to correlate the CICS and Db2 accounting and performance records.

There is no a one-to-one relationship between the CICS performance class records and the Db2 accounting records. A Db2 accounting record can contain information about one CICS transaction, multiple CICS transactions, or part of a CICS transaction.

The Db2 accounting records do not have a field that matches exactly with the corresponding CICS performance records.

This article helps to understand the following points

- Ways to configure DB2 accounting records and CICS performance Class records to achieve co-relation for better debugging and performance perspective.
- Strategies for identifying the Key attribute or parameter for mapping records between Db2 and CICS in different scenarios.

III. GOVERNING THE RELATION

Naturally, Db2 puts down its bookkeeping accounts at thread end, or at the sign on of another approval ID that is reusing the thread. This means that each Db2 accounting record for the thread can contain information about multiple CICS transactions. In addition, if different types of CICS transactions use the same transaction ID to access Db2, the Db2 accounting record can contain information about different types of CICS transactions.

Three ways to influence the relationship between Db2 accounting records and CICS performance class records, to deal with these issues

Design CICS applications so that each CICS transaction ID and Db2 authorization ID always represents the same piece of work, that consumes the same resources. This ensures that each Db2 accounting record contains either a single piece of work, or more than one occurrence of the same piece of work, and it will not contain different items. If such a Db2 accounting record contains multiple items because the items are identical divide the resources used equally between them. The drawback of this configuration is need for re-design the application for accounting purposes.

Avoid reusing threads. This ensures that the thread terminates, and Db2 writes an accounting record, after each task, so each Db2 accounting record represents a single task. However, by doing this, the drawback is loss in the significant performance advantages of thread reuse.

Configure Db2 to produce an accounting record each time a CICS task finishes using Db2 resources. This is achieved by configuration in DB2CONN or DB2ENTRY definition. Use ACCOUNTREC(TASK) option instead of ACCOUNT(UOW). It ensures that there is at least one identifiable Db2 accounting record for each task. In case of different type of transaction using the same transaction ID to access Db2, CICS passes it network protocol LU6.2 token id to Db2 to be included in the accounting record. Correlation can be done using this token with relevant CICS transaction. Using ACCOUNTREC(TASK) is generally the most practical and complete solution to control the relationship between Db2 accounting records and CICS performance class records. It carries an overhead for each transaction, but its usefulness for accounting purposes normally outweighs this overhead.

IV. STRATEGIES FOR MAPPING

There isn't one ideal method of coordinating Db2 bookkeeping records and CICS execution class records. In a couple of cases, it very well may be difficult to make the coordinating right since transactions are being run simultaneously. Much of the time in any case, there are strategies that you can use to coordinate the two kinds of records with sensible precision.

The two main factors that determine what strategies to be used

- CICS Transaction ID: Identify if each CICS Transaction ID represents only one possible transaction path or different transaction paths share the same CICS transaction id.
- DB2 Accounting ID: Each DB2 Accounting record relates to one transaction or part of one transaction or more than one transaction.

These factors combine to create four typical scenarios that helps in matching Db2 accounting records and CICS performance class records. The Figure.1 represents strategies for matching the records in each case.

Strategy A

In this scenario, each Db2 bookkeeping record contains data identifying with a solitary, recognizable CICS transaction. On the off chance that the transaction got to Db2 assets more than once, Db2 may have made more than one bookkeeping record for it. Map the Db2 accounting records relating to the transaction, to the CICS performance record for the transaction.

CICS Performance record contains the CICS activities related to a transaction. Db2 accounting records that apply to this transaction can be identified by using any one of the data items.

The CICS LU6.2 token – Configuring ACCOUNTREC(TASK) or ACCOUNTREC(UOW) in the DB2ENTRY or DB2CONN, CICS passes its LU6.2 token to Db2 to be included in the Db2 trace records as in Table 1. (Highlighted in orange colour).

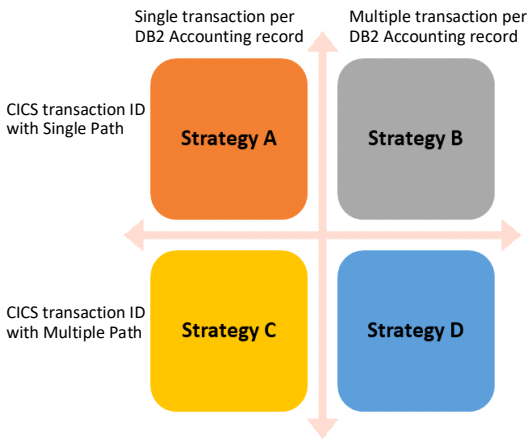


Figure 1. Strategies for matching the SMF records

The authorization ID field. The authorization ID that a CICS transaction uses is determined by the AUTHID or AUTHTYPE parameter in the DB2CONN or DB2ENTRY as in Table 1 (Highlighted in Yellow color).

The timestamp fields. The start and end time for the thread can help identify the CICS transactions that the Db2 accounting record covers as in Table 1. (Highlighted in blue color).

TABLE 1 STRATEGY A MAPPING

SMF 110 CICS Record Type

LPAR	SSID	TRAN ID	PROGRAM NAME	AUTH ID	CPU TIME	ELAPSED TIME	START TIME	END TIME	LU6.2 TOKEN ID
SYS1	CICS3	CRSQ	DFHCRQ	PFPACSR	0.4389	0.012345	09-02-2020 12:34:50:00	09-02-2020 12:34:51:00	6590876
SYS1	CICS4	SBQE	SMM01008	W24AA01	0.5498	0.00234	09-02-2020 17:14:40:00	09-02-2020 17:114:40:30	9761237
SYS1	CICS4	SBQV	SMV01009	W24AB01	0.38546	0.00184	09-02-2020 18:30:30:00	09-02-2020 18:30:30:53	9807123

SMF 102 Db2 Record Type

LPAR	SSID	AUTH ID	ESTIMATED TIME	CPU TIME	SUSPEND TIME	PLAN	TIMESTAMP	LU6.2 TOKEN ID
SYS1	DBS1	PFPACSR	0.589501	0.024994	0.0013	DISTSERVER	09-02-2020 12:34:51:00	6590876
SYS1	DBS1	W24AA01	0.38924	0.0017834	0	DISTSERVER	09-02-2020 17:114:40:30	9761237
SYS1	DBS1	W24AB01	0.18527	0.0013612	0	DISTSERVER	09-02-2020 18:30:30:53	9807123

Strategy B

In this scenario, each Db2 accounting record can contain information relating to more than one transaction, so matching each Db2 accounting record directly to the relevant CICS performance record is not possible. However, the types of transaction that are present in the Db2 accounting record can be identified, because each transaction ID only refers to one type of transaction.

If only one type of CICS transaction is present in a particular Db2 accounting record, then for accounting purposes, the resources consumed in Db2 can be split equally between each transaction. This is reasonable because the transactions are almost identical. The number of commits and backouts in the Db2 accounting record indicates the number of units of work covered in this record.

However, units of work in the same transaction might use a different thread, and so not be present in the same Db2 accounting record.

If two or more different types of CICS transaction are present in a particular Db2 accounting record as represented in Table 2 (Highlighted in Yellow color), the method of distributing the resources equally cannot be used. In this case, create model transactions by periodically measuring the amount of Db2 resources used by each type of CICS transaction. Take these measurements by temporarily disallowing thread reuse, and looking at the resulting Db2 accounting records, which contains information relating to only one transaction. Use these model transactions to map the resources to the transaction But periodically validate the correctness of the model transactions.

TABLE 2 STRATEGY B MAPPING

SMF 110 CICS Record Type									
LPAR	SSID	TRAN ID	PROGRAM NAME	AUTH ID	CPU TIME	ELAPSED TIME	START TIME	END TIME	SERVER CLASS
SYS1	CICS3	CRSQ	DFHCRQ	W24AA01	0.4389	0.012345	09-02-2020 12:34:50:00	09-02-2020 12:34:51:00	A
SYS1	CICS4	SBQE	SMM01008	W24AA01	0.5498	0.00234	09-02-2020 17:14:40:00	09-02-2020 17:114:40:3(	A
SYS1	CICS4	SBQV	SMV01009	W24AA01	0.38546	0.00184	09-02-2020 18:30:30:00	09-02-2020 18:30:30:53	B

SMF 102 Db2 Record Type							
LPAR	SSID	AUTH ID	ESTIMATED TIME	CPU TIME	SUSPEND TIME	PLAN	TIMESTAMP
SYS1	DBS1	W24AA01	1.164011	0.458053	0.0013	DISTSERVER	09-02-2020 18:30:30:53

Strategy C

In this scenario, each Db2 accounting record contains information relating to a single CICS transaction, but because several types of transaction use the same transaction ID, identifying which type of transaction is shown by a particular Db2 accounting record is impossible.

Mapping a set of records for one instance of a transaction and then reuse those figures, as in Scenario A is not correct method. Mapping all the individual CICS performance records with their corresponding Db2 accounting records is mandate. Until then, what type of transaction is represented by each Db2 record is not predictable.

Map each of the Db2 accounting records to the relevant CICS performance record by using the data items

- The CICS LU6.2 token – Configuring ACCOUNTREC(TASK) or ACCOUNTREC(UOW) in the DB2ENTRY or DB2CONN, CICS passes its LU6.2 token to Db2 to be included in the Db2 trace records as in Table 3. (Highlighted in orange colour).
- The authorization ID field. The authorization ID that a CICS transaction uses is determined by the AUTHID or AUTHTYPE parameter in the DB2CONN or DB2ENTRY as in Table 3. (Highlighted in yellow colour).
- The timestamp fields. The start and end time for the thread can help identify the CICS transactions that the Db2 accounting record covers as in Table 3. (Highlighted in blue colour).

TABLE 3 STRATEGY C MAPPING

SMF 110 CICS Record Type									
LPAR	SSID	TRAN ID	PROGRAM NAME	AUTH ID	CPU TIME	ELAPSED TIME	START TIME	END TIME	LU6.2 TOKEN ID
SYS1	CICS3	SBQE	DFHCRQ	W24AA01	0.4389	0.012345	09-02-2020 12:34:50:00	09-02-2020 12:34:51:00	6590876
SYS1	CICS4	SBQE	SMM01008	W24AA01	0.5498	0.00234	09-02-2020 17:14:40:00	09-02-2020 17:114:40:30	9761237
SYS1	CICS4	SBQE	SMV01009	W24AA01	0.38546	0.00184	09-02-2020 18:30:30:00	09-02-2020 18:30:30:53	9807123

SMF 102 Db2 Record Type								
LPAR	SSID	AUTH ID	ESTIMATED TIME	CPU TIME	SUSPEND TIME	PLAN	TIMESTAMP	LU6.2 TOKEN ID
SYS1	DBS1	W24AA01	0.589501	0.024994	0.0013	DISTSERVER	09-02-2020 12:34:51:00	6590876
SYS1	DBS1	W24AA01	0.38924	0.0017834	0	DISTSERVER	09-02-2020 17:114:40:30	9761237
SYS1	DBS1	W24AA01	0.18527	0.0013612	0	DISTSERVER	09-02-2020 18:30:30:53	9807123

Strategy D

In this scenario, each Db2 accounting record can contain information relating to more than one transaction, and also cannot use the transaction IDs to identify which types of transaction are present in the accounting record.

The present circumstance is best kept away from because it is probably not going to have the option to coordinate records precisely. In this situation, the best solution is to create model transactions, as described for Scenario B. Next, find a custom way to mark the CICS performance records with an identifier that is unique to each transaction. For example, the user could supply information in a user field in the performance records that identifies the transaction being executed. Now this field can be used to identify which of the model transactions should be used for accounting in this case.

TABLE 4 STRATEGY D MAPPING

SMF 110 CICS Record Type									
LPAR	SSID	TRAN ID	PROGRAM NAME	AUTH ID	CPU TIME	ELAPSED TIME	START TIME	END TIME	SERVER CLASS
SYS1	CICS3	SBQE	DFHCRQ	W24AA01	0.4389	0.012345	09-02-2020 12:34:50:00	09-02-2020 12:34:51:00	A
SYS1	CICS4	SBQE	SMM01008	W24AA01	0.5498	0.00234	09-02-2020 17:14:40:00	09-02-2020 17:114:40:30	B
SYS1	CICS4	SBQV	SMV01009	W24AA01	0.38546	0.00184	09-02-2020 18:30:30:00	09-02-2020 18:30:30:53	A
SYS1	CICS4	SBQV	SMX01009	W24AA01	0.38546	0.00184	09-02-2020 18:30:30:00	09-02-2020 18:30:30:53	A

SMF 102 Db2 Record Type							
LPAR	SSID	AUTH ID	ESTIMATED TIME	CPU TIME	SUSPEND TIME	PLAN	TIMESTAMP
SYS1	DBS1	W24AA01	1.164011	0.458053	0.0013	DISTSERVER	09-02-2020 18:30:30:53

V. CONCLUSION

During Performance analysis, mapping the Db2 accounting records and CICS performance class records becomes mandate. In case of accounting exact mapping or relating the record types are optional. There is no one ideal way of matching Db2 accounting records and CICS performance class records. In a few cases, it might be impossible to make the mapping correct, because CICS transaction IDs are being run concurrently. In most circumstances the strategies can be applied to map the two types of records with reasonable accuracy. But in complex cases the approach could be hybrid approach where more than one strategy has to be applied.

https://doi.org/10.5281/zenodo.4900564 ISSN 1947-5500

ACKNOWLEDGMENT

The authors would like to express our gratitude to everyone who directly or indirectly has supported our paper in these challenging times. We are grateful to Howard Hess, Distinguished Engineer, IBM, and Ian J Mitchel, Distinguished Engineers, IBM Services, for their valuable support and inputs on the paper.

The authors would like to thank the Management team, IBM Services for providing support and providing permission to publish the paper. The author would also like to acknowledge and thank the authors and publishers of referenced papers for making available their invaluable work which served as excellent input for this paper.

REFERENCES

[1] https://watsonwalker.com/wp-content/uploads/2020/11/SMF-Reference-20201107.pdf

[2] https://www.ibm.com/docs/en/zos/2.3.0?topic=smf-records

[3] https://www.precisely.com/resource-center/ebooks/understanding-smf-records-and-their-value-to-it-analytics-security

[4] https://www.ibm.com/docs/en/zoa/4.1.0?topic=insights-smf-100-data

[5] https://www.ibm.com/docs/en/zos/2.1.0?topic=sr-record-type-110-6e-cics-ts-zos-statistics

DISCLAIMER
The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of IBM. IBM takes no responsibility for any errors or omissions in, or for the correctness of, the information contained in papers and articles.

CLOUD-BASED ENTERPRISE RESOURCE PLANNING FOR SUSTAINABLE GROWTH OF SMES IN THIRD WORLD COUNTRIES

Anthony I. Otuonye

Department of Information Technology,

Federal University of Technology Owerri, Nigeria.

Email: anthony.otuonye@futo.edu.ng , ifeanyiotuonye@yahoo.com

Phone: +234(0)8060412674

ABSTRACT

In this research paper, we have developed a new cloud-based Enterprise Resource Planning System using the Linked Servers technology. Our focus is on Small and Medium Scale Enterprises (SMEs), seeking for ways of assisting them enjoy the huge benefits of Enterprise Resource Planning and the internet cloud. Our ERP system can integrate key business processes into a single software solution, and enables seamless flow of information from heterogeneous data sources across all functional areas of the organization and beyond. The new system was developed using the prototyping methodology, considered as best suited for a project of this kind due its advantage of active user involvement throughout the development process. Using the questionnaire as our major data gathering instrument, user requirements were gathered from selected SMEs across the south-east geopolitical region of Nigeria to aid our design. A thorough Requirement Analysis was carried out to ensure proper design of the prototype system, and, using the system development life cycle approach, we carried out design of the architectural frameworks for the cloud based ERP system, to illustrate basic layout of application deployment and the synchronization mechanism for data exchange between vendor servers and the central server. Result obtained from our study shows that the new ERP software solution provides improved operational efficiency and customer satisfaction. We recommend the creating of enabling environment by appropriate government authorities, and a systematic implementation of the findings of this research paper by corporate business owners.

Keywords: Internet-cloud, ERP, Vendor servers, SMEs, Data integration.

1. INTRODUCTION

1.1 The Concept of Enterprise Resource Planning (ERP)

Enterprise Resource Planning (ERP) is an attempt to integrate key business processes of a firm into a single software solution, whose aim is to enable easy flow of information throughout the organization. Ordinarily, such systems focus more on internal processes, though in some cases, may include transactions with customers, business partners, and vendor companies.

Normally, a large business organization will have dissimilar information systems or software applications built around different functional areas, organizational levels, and business processes, which ordinarily cannot easily exchange information, especially at the point where such information is needed for quick decision making. Most times also, each of the different software application will require a fragmentation of data in hundreds of separate databases, and this can equally degrade efficiency and business performance. A good example is the problem of a Sales Personnel who may not be able to tell (at the time a customer places an order) whether or not the ordered items exist in the inventory store. Also, the Manufacturing unit may find it difficult to carry out their duty of planning for new production due to their inability to easily access sales information at the time they need it.

A good Enterprise Resource Planning system can solve this problem of integration and seamless data exchange across various levels of the organization, and across various functional areas of the business corporation. Such a system will receive data from various key business processes in Production and Manufacturing, Finance and Accounting, Sales and Marketing, Human Resources, and so on, into a central data repository. From the central data repository, information that was previously fragmented in different systems can be shared across all parts of the firm. Undoubtedly, an efficient coordination of these parts will lower a business's operational cost, while increasing customer satisfaction.

A seemingly simple process such fulfilling a customer order as mentioned above, will require an involvement of several functional areas of the business and a free flow of data across the firm, between business partners, and vendors, as well as customers.

There are existing ERP software solutions which are mainly products of such Software Vendors like SAP, Oracle, and Microsoft. Some of these software include: SAP's Business Suite, Oracle's e-Business Suite, Microsoft's Dynamic Suite, and so on. Most of these software solutions however, target very large multinational business corporations. Over the years, not much attention has been given to Small and Medium scale Enterprises (SMEs) to find ways of helping them enjoy the huge benefits of ERP.

In a bid to solve the data exchange and web application integration problem using Enterprise Resource Planning software, software vendors have always made use of Web Services and a collection of web services popularly known as Service-Oriented Architecture (SOA) to integrate various software systems into one. Service-Oriented Architecture is a set of self-contained services that communicate with each other to create a working software application. On the other hand, a web service is a collection of open protocols and standards used for data exchange between applications or systems. Furthermore, these web services are self-contained, modular applications that can be described, published, located and invoked over a network, generally the World Wide Web [1].

Although the use of web services can provide a near flexible solution to the problem of application integration [9], the service architecture is confronted with a number of stubborn problems including the issue of data security [7], and the challenge of synchrony of web services [11]. With the use of WSDL and UDDI in web services, an attacker can access any publicly available WSDL file and tamper with it. Such attacks can be in the form of WSDL Scanning or WSDL tampering. The former scans the WSDL file and exposes some operational and even confidential information. According to [1], protection against these threats is not easy with typical methods like authentication and authorization.

There is need to design efficacious ERP solution that can enable business owners to streamline and automate tedious back office tasks. Apart from streamlining and automating back office tasks, managers can equally get real-time visibility into the inner workings of their operations and motivate employees to be more productive, focused, and successful towards performing their roles. A focus on cloud-based ERP will encourage small and medium scale enterprises to take advantage of cloud computing which delivers computing services such as servers, storage, database, networking, software, as well as analytical and business intelligence to corporate business owners at a very low cost. With the internet cloud, a small scale company will not have to purchase the software, hardware, servers and facilities necessary to run her ERP and there will be no need to train and maintain an IT team that is responsible for the software. All that is needed are computers that can access the internet. Since the cloud host/vendor offers the maintenance of infrastructure, there will be reduced operational cost for such SMEs.

In this research paper therefore, we shall design a cloud-based Enterprise Resource Planning System for efficient data exchange and web application integration using the Linked Servers technology. Our focus will be on Small and Medium Scale Enterprises (SMEs) to seek ways of assisting them enjoy the huge benefits of such technology-based systems, with very little operational costs. Being on the cloud, the system will be directly managed off-site by a third party provider and with the ability to address diverse data sources similarly.

1.2 Objectives of Study

In this research paper, the Linked Servers technology is adopted in the design of Cloud-based Enterprise Resource Planning which aims to guarantee efficient data exchange and web application integration to ensure rapid and sustainable growth of Small and Medium Scale Enterprises (SMEs) in developing countries. The study will achieve the following:

i. Discover major weaknesses of existing ERP systems in meeting business needs of SMEs.
ii. Propose a Cloud-based ERP model using Linked Severs technique to provide efficient and secure data synchronization between heterogeneous vendor database servers.
iii. Ensure distributed queries, updates, and transactions on dissimilar data sources across the enterprise.
iv. Make recommendations to assist local SMEs enjoy the huge benefits of ERP and the competitive business advantages of the internet cloud.

https://doi.org/10.5281/zenodo.4900658 ISSN 1947-5500

2. LITERATURE REVIEW

This section seeks to review existing literature on concepts, theories and empirical studies that border on Enterprise Resource Planning, Executive Support Systems, Cloud Computing, and the Linked Servers Technology Architecture. We will begin with a conceptual framework to discuss various perceptions relating to our topic of study.

2.1. Definition of Enterprise Resource Planning

According to [15], an Enterprise Resource Planning (ERP) a business process management software that enables an organization to use a system of integrated applications in managing her business. Normally, such a system will make use of tools and applications that cover all areas of the business and enables potential communication from various sources.

The software will also seeks to automate many back-office functions related to technology, services and human resources. With ERP Software, all essential business functions such as estimation, finance, human resources, production, marketing, sales, purchasing, and others can be collected at a central source. From here, data can be easily accessed by the concerned persons and departments. Enterprise resource planning system can also streamline the assemblage, storage and usage of an organization's data in a most unified way. ERP system efficiently intermingles all components of business procedures and methods, which consist of development, product planning, manufacturing, sales and marketing, and others, in a single database, application and user interface. Such Enterprise Resource Planning system is reckoned to be a type of enterprise application that is created to be used by large-scale businesses and it oftentimes requires devoted teams to analyze and customize the data to handle deployment and upgrades of the software [15].

Today, majority of corporate organizations implement ERP systems for a number of reasons. In fact, a study conducted in 2016 by Panorama Consulting Solutions, LLC, reveals some reasons why organizations implement ERP systems. Their report shows the following reasons, by percentage:

- To replace out-of-date ERP software (49%)
- For replacing accounting software (15%)
- To replace other non-ERP systems / had no system (20%)

- To replace homegrown systems (16%) (Panorama Consulting Solutions, 2016).

According to the report also, organizations that have never implemented ERP System will surely need one for the following reasons:

- To improve internal business processes
- In order to improve company performance
- For reducing IT expenses and labor costs
- To improve interactions between internal employees and external organizations. (Panorama Consulting Solutions, 2016).

However, ERP applications for small-scale businesses can be designed as a lightweight business management software which can be customized for particular business industry. With a functioning Enterprise Resource Planning (ERP) in place, crucial issues can easily be resolved which could be damaging to the organization if left unchecked.

2.2. Why ERP is Indispensable for Growth of Small and Medium Scale Enterprises (SMEs)

Before now, an all-in-one solution software was not in existence which could help organizations in managing their business process, but as technology grows, Software Developers has seen the need to introduce the ERP technology. ERP solution can be a wonderful management tool when it comes to maintaining and managing businesses in a most efficient way, and this ideology can only get better with time. A well-designed Enterprise Resource Planning, such as the one being proposed in this research project, can perfectly bring key business processes together and allow small and medium scale enterprises as well as large business organizations make data-driven decisions, improve collaboration, and strengthen business productivity. As a matter of fact, ERP can be made to covers all functional areas of a company, which include the following: finance, human resources, customer relationship management, manufacturing, and supply chain.

2.2.1. Finance

Modern Enterprise Resource Planning systems offer dashboards that give users an overview of their current financial status, and with this they can tap into real-time business data anytime, anywhere. For small businesses, ERP can assist managers track income and expenses and record transactions and account structures. It can also create financial documents like profit and loss statements and balance sheets.

2.2.2. Human Resources (HR)

Some modern ERP software can enable SMEs to manage company data and help in streamlining employee management tasks such as hiring, payroll, and other duties. Small businesses can take advantage of ERP to automate payroll processing, track employee attendance to work, and manage employee records like performance reviews, payroll benefits, and scheduling. Self-service functionalities can equally allow employees to request time off or view their attendance record. One good thing about an ERP is that it can help managers to save time, energy, and the risk-factor because they could track each employee's performance and pinpoint HR problems before they start to happen.

2.2.3. Customer Relationship Management

A good ERP can help SMEs manage customer contact information, order histories, invoices, and quotes.

2.2.4. Manufacturing

A well designed Enterprise Resource Planning solution can optimize project and cost management as well as production planning. This feature can improve factors that play prominent part in automating

daily processes and business communication. It can offer manufacturers the ability to manage resources by accessing real-time data and fulfil customer needs by providing fast and reliable services.

2.2.5. Supply chain

ERP is a good manager of the Supply Chain. It can help manage the flow of goods and services from raw material acquisition to delivery of the finished product to the customer. Most SMEs today are still in the habit of entering information by hand while striving to track down the inventory present in her warehouse. In the case of such SMEs, implementing an ERP is quite indispensable. With such a smart solution, business owners can save money and time by automating all its major business processes. Modern solutions also offer dashboards and business intelligence to help these SMEs handle most of their inventory management problems.

2.3. Cloud-based ERP Systems for Small Businesses

Some ERP solutions are marketed as being only for small businesses. These solutions can be on-premise or web-based; though, web-based is more common for small businesses due to the generally lower upfront cost. ERP software for only small businesses is less complex and has limited functionality to cut costs and tailored to meet the needs of smaller companies.

Research has shown that most ERP solutions represented as free on the WWW are meant to be more of a demo than a permanent solution. There are usually some costs associated with these systems to add functionality that are required for interested businesses or installation and maintenance fees. According to Panorama Consulting Solutions (2016), some examples of free and open source small business ERP software include:

- Odoo: The free plan, Odoo Community, is an open-source ERP software that incorporates the following applications: CRM, invoicing, expense tracking, e-Commerce, appointment scheduling, and POS. With a special link, this ERP software can equally integrate with heavy applications like eBay or USPS.
- Dolibarr: A free open source ERP system that features CRM, HR management, CMS, inventory control, marketing automation, and project management.

2.4. The Linked Servers Technology

According to [5], Linked Servers technology makes it possible for an SQL Server to “talk” to another ODBC compliant database, such as another SQL Server instance or an Oracle database, with a direct T-SQL query. It can enable one to execute distributed queries against tables stored in a Microsoft SQL Server instance and another data store. It is easy to use the Microsoft SQL Server Management Studio to link a data store to an SQL Server instance and then execute distributed queries against both data stores.

Linked servers can be created using the SQL Server Management Studio. From the Object Explorer pane, expand the "Server Objects" section, right click on "Linked Servers" and choose "New Linked Server" from the menu.

SQL Server can be linked to any server provided it has OLE-DB provider from Microsoft to allow a link. Example, Oracle has an OLE-DB provider for oracle that Microsoft provides to add it as linked server to SQL Server group.

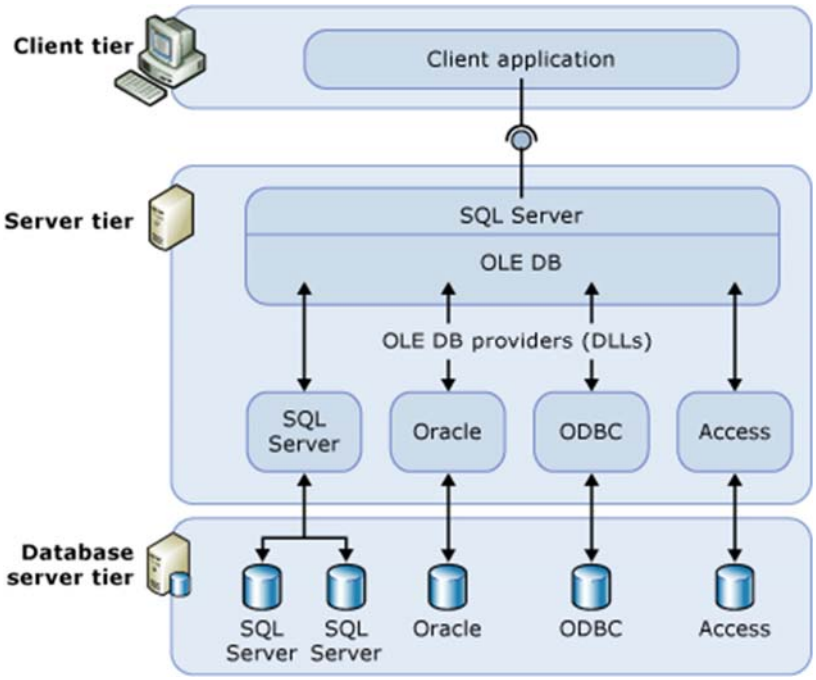


Figure 1: Basic Linked Servers Configuration Architecture (source: Eric Blinn in Edgewood Solutions Guide, 2021).

2.5. Improving Privacy and Data Security Level in Cloud-Based ERP Systems

Nowhere is security more of a concern than with cloud-based ERP applications. Many businesses assume that after they move their ERP system to cloud, the security is beyond control. But, there are proactive steps that businesses can take to help them gain more control over their cyber security. Despite the security risks, businesses have good reasons for moving their ERP systems to the cloud. In addition to 24 hours cloud-based access to data across multiple departments and geographical locations. According to [13], competitiveness is another clear reason.

3. METHODOLOGY AND USER REQUIREMENT ANALYSIS FOR THE NEW SYSTEM

For this study, the prototyping model was used for early development of our proposed cloud based enterprise resource planning. Prototyping is a software development model in which a software prototype is developed, evaluated, and re-designed until an acceptable prototype is achieved. The accepted prototype is then used as the basis of development of the final system.

https://doi.org/10.5281/zenodo.4900658 ISSN 1947-5500

Software prototyping works best in situations where user requirements are not fully known at the onset. User involvement, especially the continuous interaction between system users and system developers makes it the most preferred method for the task of this study.

The figure 2 shows the prototyping model, which allows for quick design of an early system after initial requirements gathering from system users. The prototype is then subjected to an evaluation process using appropriate test data. The evaluation aims at allowing users to put the system into use in order to identify additional user requirements or any need to improve on existing ones. Evaluation and testing continues until all identified user requirements are fully met.

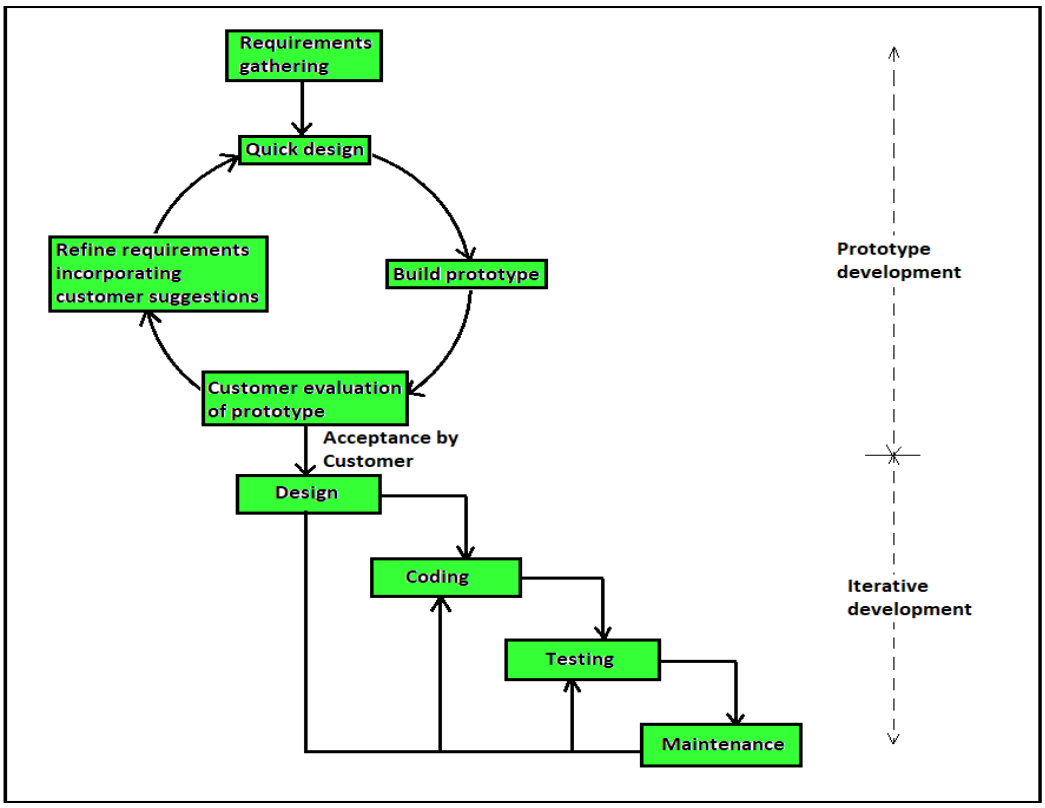


Figure 2: A popular prototyping model (source: www.geeksforgeeks.org)

Finally, the prototyping model will be integrated into the wider system development approach which include the different phases of software development life cycle, and will be adopted for this study. The major phases of development include the following: System definition, data gathering / analysis, system design, implementation, testing, and system maintenance phase.

3.1 Data Gathering and Requirement Analysis

The questionnaire method was used to gather data for development of our proposed system. The purpose of the data gathering, however, is to generate user requirements to aid our design.

About 300 copies of questionnaire were administered to Business Entrepreneurs, as well as management staff of selected Small and Medium size Enterprises in the south-east geopolitical region of Nigeria.

Some of these SMEs include Hotels, grocery stores, garages, etc., that serve a hyper local target audience, and operate with less than a certain level of workforce and assets. Our goal was to ensure accurate and honest information gathering and user requirements to aid our system design process.

In order to gather relevant user requirements for our proposed ERP system, structured questions were posed to elicit information in the following areas:

i. **Need for accessibility of common company data**: Most SMEs in developing countries are accustomed to spreadsheet accounting and some manual business processes using outdated data sources. Therefore, the following questions were posed:
   - Do you need a system that compiles and stores company common data as well as make it readily accessible to you?
   - Do you wish to gain real business insight for informed decision making?
   - Would you like to view, manage, and track core business processes and resources in real-time using a single data source?

ii. **Integration of core business functions**: The following questions were posed:
   - Do you need a system of shared database to support and connect multiple business activities?
   - Do you wish to integrate business functionality for accounting, inventory management, order processing, human resources etc.?
   - Do you require integration and collaboration across departments?
   - Do you require improved inventory management?
   - Do you think your company requires standardized business processes?

iii. **Cloud-based ERP system**: Business owners and top managers were asked the following question: Whom will you select as your preferred software partner or vendor, or do you prefer a cloud-based system?
   - Do you wish to leverage on global markets?

iii. **ERP components and basic features:** The following questions were posed:
   - Which processes of your business do you wish to incorporate into an ERP system?
   - What features do you think you will need your ERP to have as a start?

v. **Legacy systems**: In order to gather user requirements in the area of legacy systems and applications, the following questions were asked:
   - Do you wish to replace your legacy systems?
   - Do you wish to reposition your company for sustainable growth and development?
   - Do you require improved customer service and operational efficiency?

vi. **Operational cost and business scalability**: The following question was asked in the area of operational cost and business scalability:
   - Do you want to experience reduced cost of business operations?
   - Do you require an ERP system that scales up as your business grows?
   - What are your projected performance enhancements based on your expectations and user preferences?

Ninety five percent (95%) of all questions were answered in the affirmative, showing the need for cloud based enterprise resource planning systems for sustainable growth and development of small and medium scale enterprises in third world countries.

All information and user requirements were gathered from business entrepreneurs, top management cadre of country SMEs, directors, section mangers, and supervisors, and from the information gathered during this stage, we were able to establish the following:

a. Cloud-based ERP is required for sustainable growth of SMEs in developing countries.
b. Existing ERP systems and businesses have major weaknesses in meeting the needs of SMEs.
c. There is need to ensure data synchronization between heterogeneous vendor database servers in order to ensure efficient information exchange.
d. There is need to implement distributed database that updates data in remote vendor databases in an efficient and secure manner.
e. There is need for automated data back-up in the cloud as safeguard against data loss due to vendor services malfunction or crash of central server machine.

4. System Design, Result and Discussions

In this section, we present the technical architecture which basically defines layout of layers of application deployment between servers and client computers, interfaces, platforms and emerging technologies that will provide technical functionality for our proposed cloud based enterprise resource planning system.

4.1. Architecture of the Proposed Cloud based ERP System for Sustainable Growth of SMEs

The figure 4 depicts the architecture of the proposed cloud based ERP system for sustainable growth of SMEs, especially in developing countries. It is a conceptual framework the shows basic layout of application deployment and various activities involved in the new system, and illustrates the synchronization mechanism for data exchange between vendor servers and the central server.

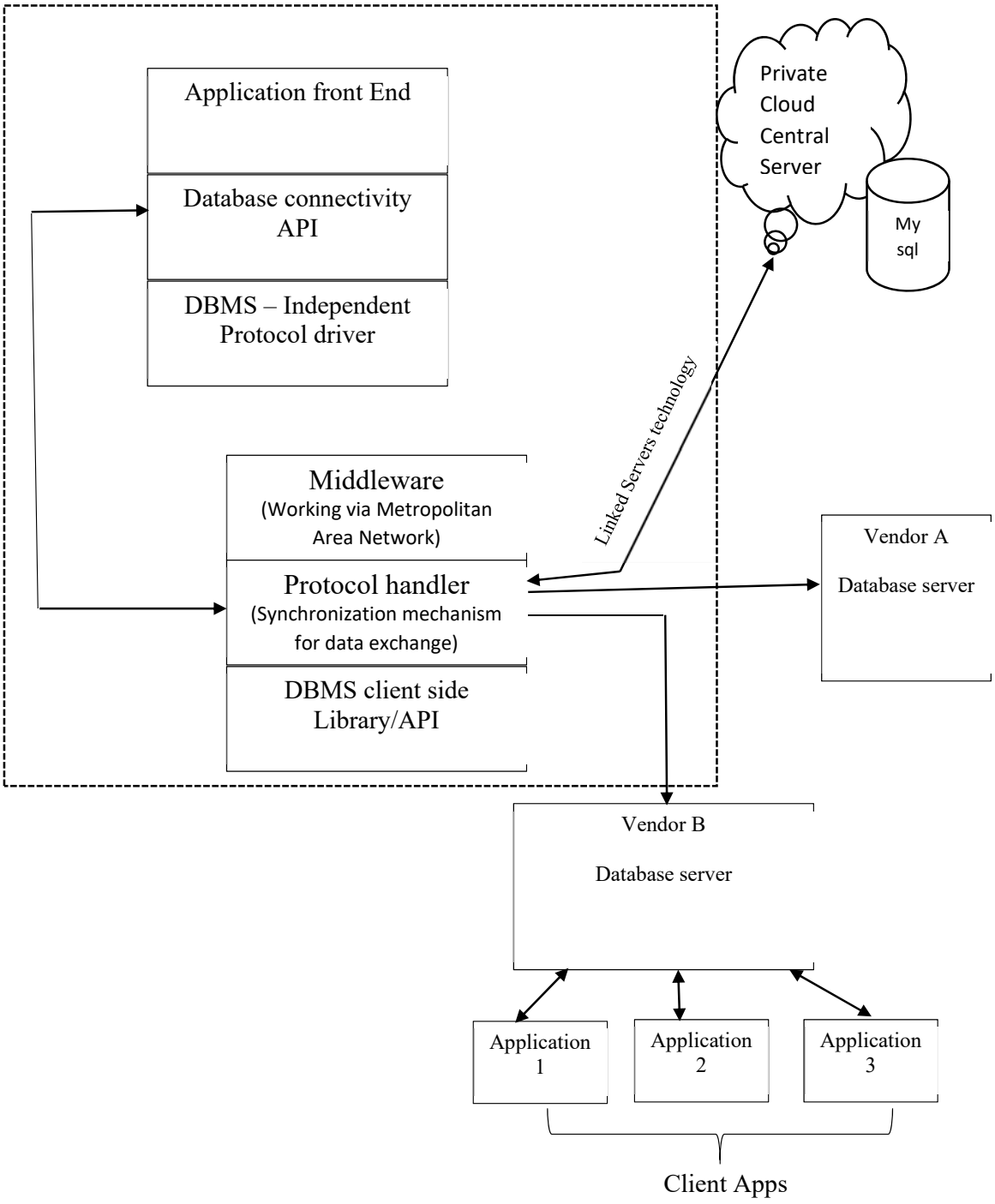


Figure 4. Architecture of the Proposed Cloud based ERP System for SMEs.

https://doi.org/10.5281/zenodo.4900658

4.2. Algorithms for Inventory Management Module of the proposed ERP system

The basic algorithms for inventory management module of our proposed ERP system is shown in figure 5.

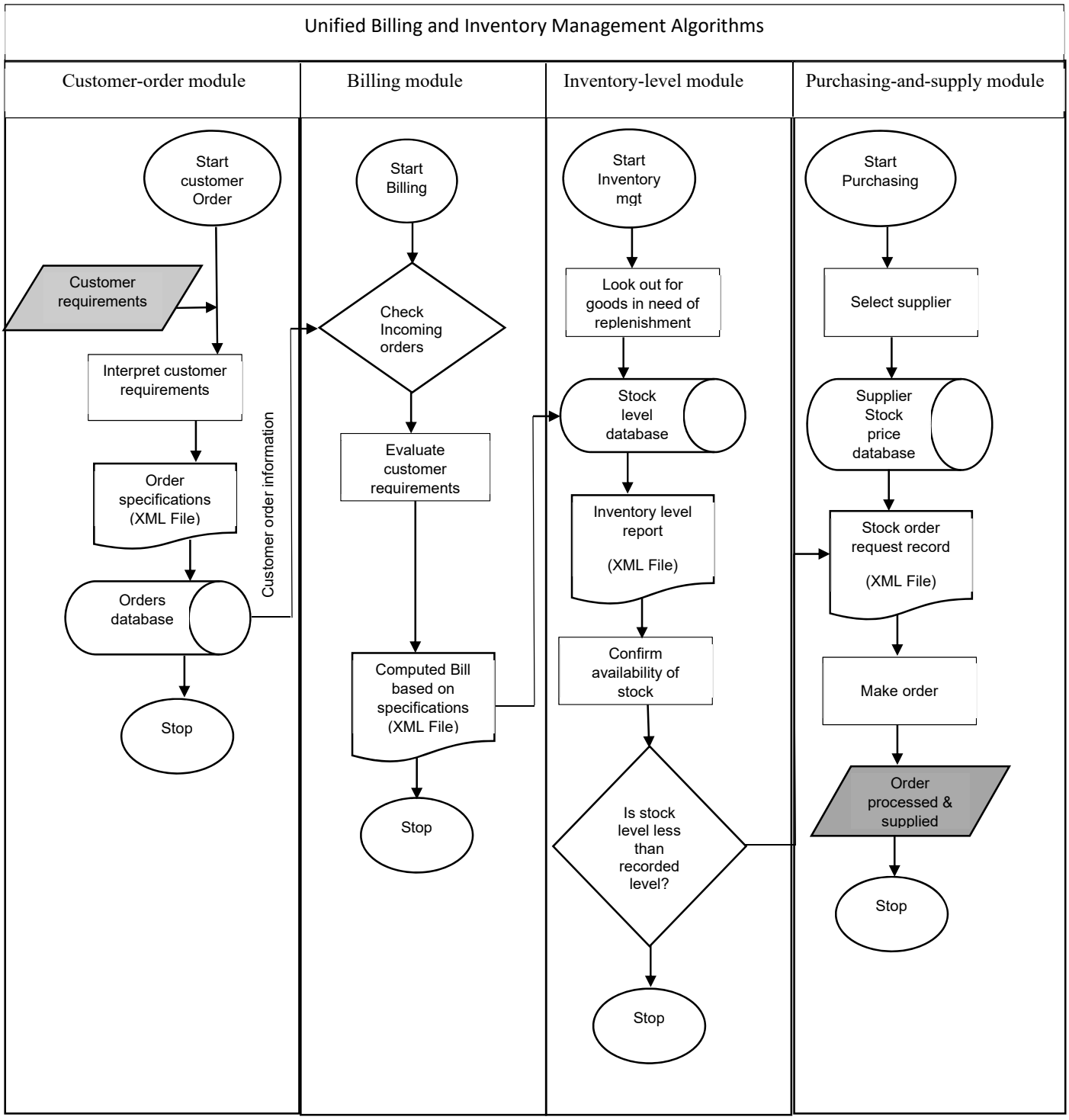


Figure 5. Unified Inventory management algorithms for the proposed based ERP system

4.3. High Level Model for Accounting Management Subsystem

The High Level Model for Accounting Management subsystem of our proposed ERP system is shown in figure 6.

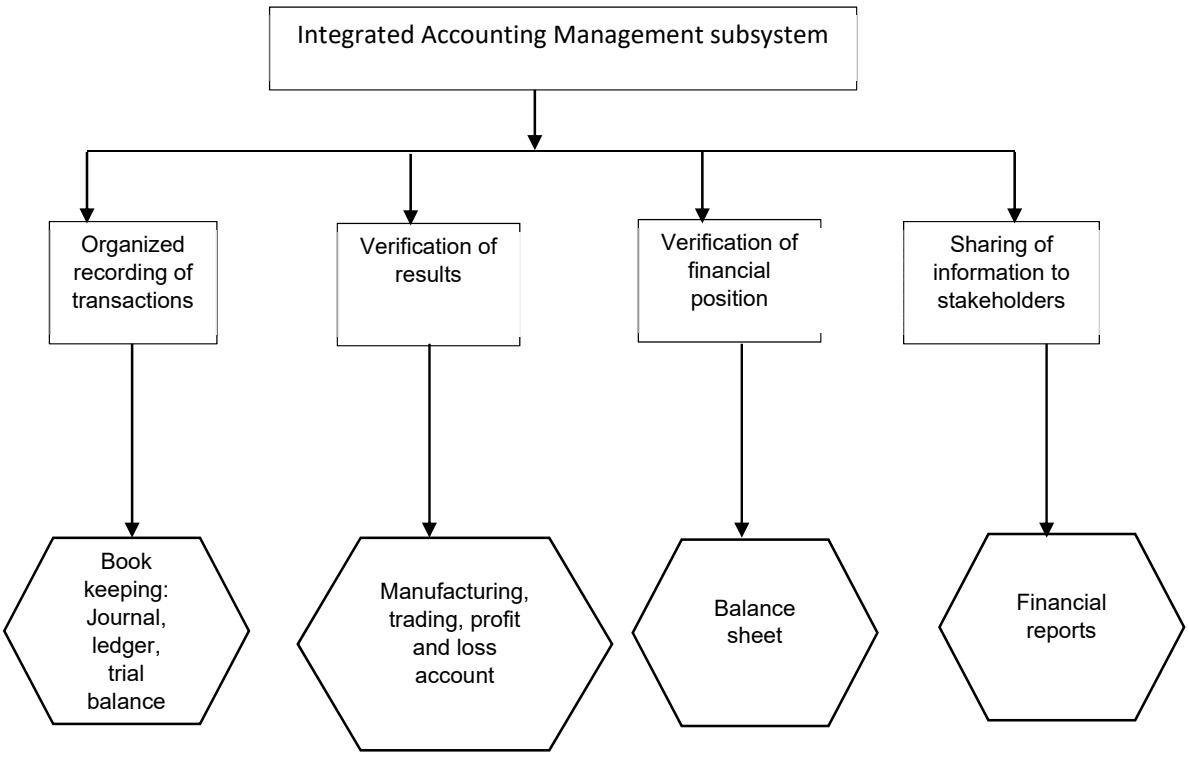


Figure 6. Integrated accounting model for the proposed based ERP system

4.4. System Implementation and Discussions

Here, we discuss some of the development tools used in the implementation of our new design to guarantee efficient and secure data synchronization among vendor databases from remote locations. The following sections of the new system were implemented using the under listed programming tools:

Database Synchronization and Exchange: Linked Servers Technology was used to create synchronization mechanism and database connectivity for exchange between vendor servers and the central server.

User Management Module: The ASP.NET was used to create the user management session for administering the central server.

Central server database: My-SQL server technology was adopted for design of the central server database.

Vendor server 1: My-SQL, an open source database management system, was used to develop a relational database to store data for the Inventory management subsystem, which included such information as Inventory Management System Report Types such as list of stock items, list of sold items, list of returned items, report date filter, category, price, and quantity.

Vendor server 2: MS-SQL was used to develop a relational database to store data for the Accounting management subsystem, which included such information as Account report category and information including Account Expenses, Purchases, Sales, Point-of-Sale records, and summary report information. Others include dates of transaction, transaction description, transaction receipt number, and sales personnel's remarks.

Data Migration from various platforms to the remote server: My-SQL workbench, an open source development tool was used to export databases from other platforms to the My-SQL-based remote server. Another is the My-SQL connector for ODBE, which is a connector tool used to create an interface between the central MY-SQL server and all other database platforms that are compatible with open database connectivity technology.

Inventory Management subsystem: Visual Studio 2015 was used to implement the Inventory Management System. Some of the tools include: C# for the front end and application logic development, and Ado.NET for relational database connectivity.

Accounting Management Subsystem: VB.NET was used to implement the Accounting Management subsystem.

Web Hosting: The IIS and the APACHE web servers were adopted for hosting our web applications built with the .NET technology.

4.4.1. Results

The output screens in figure 7 and figure 8 show some of the results of the implementation.

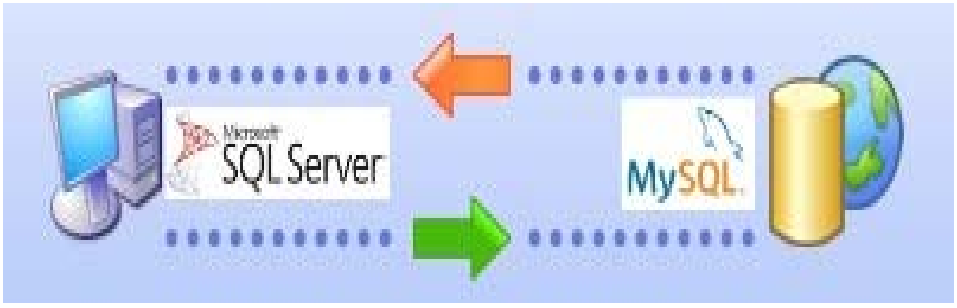


Figure 7. Output screen illustrating the synchronization mechanism for data exchange between vendor servers and the central server.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500
https://doi.org/10.5281/zenodo.4900658

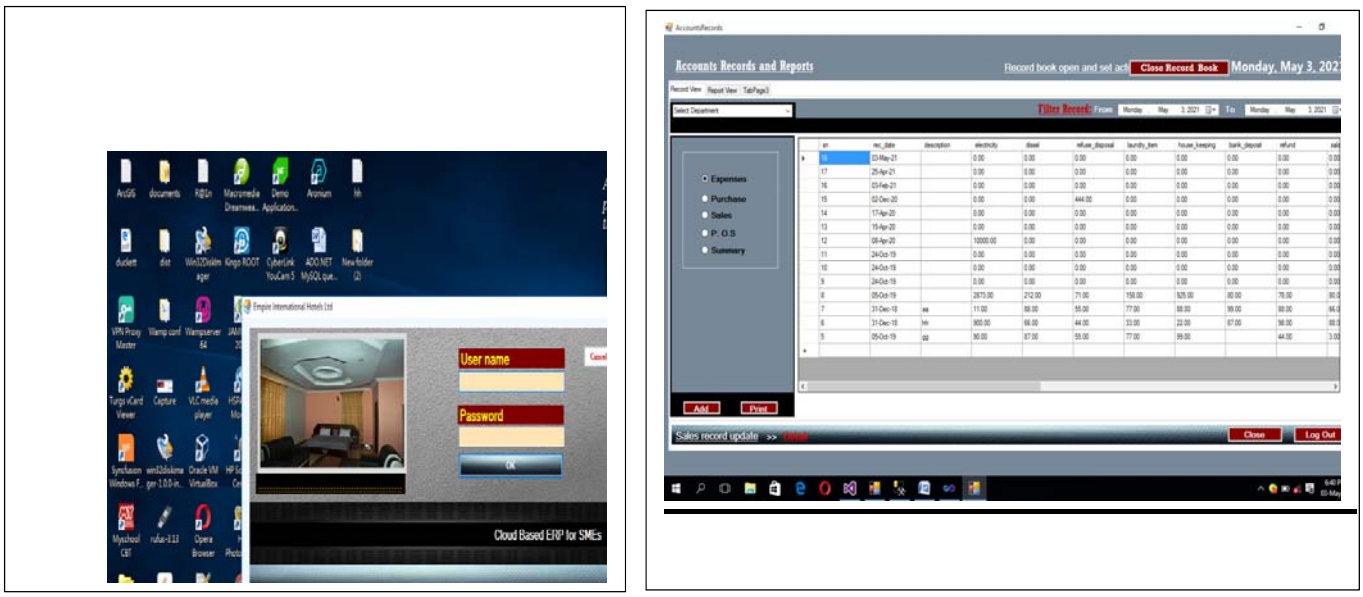


Figure 8. Some output screen from the accounting management subsystem and the inventory management subsystem.

5. CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

In this research paper, we have developed a cloud-based Enterprise Resource Planning System for efficient data exchange and web application integration using the Linked Servers technology. Our focus was on Small and Medium Scale Enterprises (SMEs), seeking for ways of assisting them enjoy the benefits of Enterprise Resource Planning. With such an integrated system on the cloud, there will be very little operational cost on the part of the SMEs because most of the institutional and IT infrastructure will be directly managed off-site by a third party provider.

https://doi.org/10.5281/zenodo.4900658 https://sites.google.com/site/ijcsis/ ISSN 1947-5500

Again, the system will encourage small and medium scale enterprises to take advantage of cloud computing which delivers such computing services as servers, storage, database, networking, software applications, and analytical and business intelligence to corporate business owners at a very low cost.
Our system can integrate key business processes of a firm into a single software solution, which enables seamless flow of information throughout the organization and beyond. It can solve the problem of integration and seamless data exchange across heterogeneous data sources at various levels of organization, and across all functional areas of a business corporation. It has the ability to receive data from key business processes in Production and Manufacturing, Finance and Accounting, Sales and Marketing, Inventory and Human Resources, and so on, into a single central data repository, from where information can be shared among various stakeholders as the need arises.
The new system was developed using the prototyping methodology, considered as best suited for a project of this kind due to its iterative nature and user involvement. The questionnaire method was used to gather all user requirements from selected SMEs across the south-east geopolitical region of Nigeria to aid our design. A thorough Requirement Analysis was carried out to ensure proper design of the prototype system.
Result obtained from our study shows that the new ERP software solution provides improved operational efficiency and customer satisfaction.

5.2. Recommendations

We recommend the creation of enabling business environment by appropriate government authorities, and a systematic implementation of the findings of this research paper by corporate business owners. The system is a new attempt to ensure rapid growth and sustainability of our Small and Medium sized Enterprises by harnessing the benefits of ERP and the internet cloud. There is need for a concerted effort by all stake holders, government agencies, corporate organizations, and business owners to take advantage of the opportunities offered by the emergence of the internet and Information Technology to improve the business sector.

References

[1] Aruna S. (2016): Security in Web Services-Issues and Challenges, International Journal of Engineering Research & Technology, Vol 5, issue 09, September 2016.

[2] Baburajan, Rajani (2011): "The Rising Cloud Storage Market Opportunity Strengthens Vendors". It.tmcnet.com. Retrieved 2018-12-02.

[3] Daniel M., Adnan M., & Rakibul H. (2015): “Enterprise Resource Planning (ERP) Systems: Design, Trends, and Deployment, The International Technology Management Review, Vol. 5 (2015), No. 2, 72-81.

[4] Davenport T. (1998). Putting Enterprise into the Enterprise System, Harvard Business Review, 76(4),121-131

[5] Eric Blinn (2021): Understanding SQL Server Linked Servers, Edgewood Solutions, LLC , 2021

[6] Fripp C. (2010): Cloud vs. Hosted services, what's The Difference? IT News Africa Available via http://www.itnewsafrica.com/2011/04/cloud-vs-hosted-services/ [accessed October 30, 2020].

[7] Hongbing Wang, and Joshua Zhexue Huang (2004): Web services: problems and future directions**,** Journal of Web Semantics**,** Volume 1, Issue 3, April 2004, Pages 309-320

[8] Jaideep M., Ram Subramanian & Pradeep G. (2021): "Critical factors for successful ERP implementation: Exploratory findings from four case studies", Elsevier B.V., vol. 56, Issue

[9] Kennet C. Laudon and Jane P. Laudon (2010): Management Information Systems, eleventh edition.

[10] Lin, A. And Chen, N. C. (2012): Cloud computing as an innovation: Perception, attitude and adoption International Journal of Information Management. Vol. 4 No. 1

[11] Miko Matsumaru (2006): Ten Top Web Services "Issues",

[12] Moller, C. (2005). ERP II: a conceptual framework for next-generation enterprise systems? Journal of Enterprise Information Management, 18(4), 483-497

[13] Owens S.J., (2018): Planning for Attack: Security and cloud based ERP. Enterprise Software by ERP Desk.

[14] Raheela Nasim et al (2020: "A Cloud-Based Enterprise Resource Planning Architecture for Women's Education in Remote Areas", Electronics, September 2020

[15] Rafi Ahmed K. (2017): "Detailed Introduction to Enterprise Resource Planning", Computer Information Services (CIS), Software House, 2017.

[16] Wang (2012): "Enterprise Cloud Service Architectures". Information Technology and Management. **13** (4): 445–454. doi:10.1007/s10799-012-0139-4. S2CID 8251298.

[17] www.geeksforgeeks.org

PCA, SPCA & Krylov-based PCA for Image and Video Processing

[1] Amanda Zeqiri, [2] Markela Muca, [3] Arben Malko

[1, 2] Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Albania Tirana, Albania
amanda.zeqiri@fshn.edu.al
markela.muca@fshn.edu.al
[3] Lev Tech, Software Development Company
arben1malko@gmail.com

***Abstract*— Processing types of data like noise, images and videos, which are raw data collected from technological or medical devices, is a challenge since numerical representation of them are very large datasets. Subtracting valuable information for surveillance, detection or biomedical purposes, consists in a pre-processing phase of the original dataset that includes reducing the number of variables without losing any important information or properties. Once the dimensionally reduction techniques are applied, more complex strategies and methods can be used to further process the data. Background subtraction techniques are necessary to separate moving objects from the steady ones, using a reference background frame.**

This paper describes a collection of popular and effective methods used in image/video processing, particularly for background estimation. It highlights advantages, limitations, modifications and efficiency, starting from the standard approach (PCA) up to innovative methods using Krylov subspaces, associated with background estimation experiments.

Keywords— Principal components, Sparse PCA, Block Krylov subspace, SVD, Background subtraction.

I. INTRODUCTION

Real modern world problems generate huge amount of data every day. These raw data is very hard to manipulate or to get valuable information from. Data processing translates raw data into meaningful information by performing different techniques. Traditional approaches based on using the entire dataset become very impractical as the dimensions and number of variables increase therefore before applying different methods datasets are transformed into a suitable form by reducing dimension or the number of variables. Dimension reduction can be applied by keeping the most relevant variables only or by transforming the real dataset into a smaller dataset of new variables, that contain meaningful information and properties of the real dataset.

Dimensionally reduction can be very useful in many fields such as signal processing, image processing or video processing. Security and surveillance systems need image/video processing in order to detect, define or count people and other moving objects. To identify these elements background subtraction is applied i.e., separate the static elements called *background* from the other non-static elements referred to as *foreground*. Foreground do not have a significant contribution to the background estimation as long as they are small and change position at different times.

One of the most widely used linear dimensionality reduction techniques, meaning each new variable being a linear combination of real variables, is Principal Component Analysis (PCA). This paper gives a detailed description of standard PCA (advantages and limitations), some modified versions (SPCA and Elastic net) and up to date Krylov subspace methods also used for dimensionality reduction and principal components approximation.

We conclude with background estimation experiments on different image datasets corresponding to three videos, to highlight differences between methods.

II. STANDARD PCA

Considering $X \in R^{n\times p}$ the real data matrix of dimensions n and p, where n rows represent the number of observations and the p columns represent the number of features/variables. The rows of matrix X are referred as $x_i = [x_{i1}\ x_{i2}\ \dots x_{ip}]$ than X can be written as $X = [x_1\ x_2\ \dots x_n]^T$.

Before processing the data, the columns of X are centered meaning from each column it is subtracted its sample mean $\bar{x}_j = \sum_{i=1}^{n} x_{ij}\ /n$ for $j = 1,2,..,p$. Sometimes it is necessary to scale the columns of X, that means to divide each variable by its sample variance, in order that each new variable has a sample variance of 1. This process is not always necessary, especially when the variables are all measured in the same units.

The main purpose of the dimensionally reduction techniques is to generate a $(n \times k)$ dimensional representation of the data $X_{n\times p}$, where k is much smaller than p, that would preserve the information present in the variables space. Linear PCA finds the best fit for this k-dimensional subspace by generating new variables as a linear combination of the original variables, such that the new variables capture maximal variance.

So, the new dimension-reduced data points are stored in the z_i rows of matrix $Z = YV^T \in R^{n\times p}$ and are called projected points. Matrix is $Y = XV \in R^{n\times k}$, where rows $x_i \in R^p$, for $i = 1,\dots,n$ of matrix $X_{n\times p}$, are transformed into rows $y_i \in R^k$, for $i = 1,\dots,n$ with less variables. Columns of $Y_{n\times k}$ are the k

principal component scores obtained by projecting X onto orthonormal vectors $v_j \in R^p$, for $j = 1, \ldots, k$, which means they are orthogonal $v_i^T v_j = 0$ for $i \neq j$ and they are unit vectors $v^T v = 1$.

Vectors $v_1, v_2, \ldots v_k \in R^p$ are set as columns of matrix $V_{p\times k}$, they contain the loadings/directions/ coefficients of the k principal components and altogether explain the most variance in the data. Generally, the variance explained by the first k principal component directions should be at least 70%-80% of the total variance. The variance explained by principal component direction k is $d_k^2/n = \frac{1}{n} v_k^T (X^T X) v_k$, where $d_k = \sqrt{v_k^T (X^T X) v_k}$ and $S_n = \frac{1}{n} X^T X$ the sample covariance. So, in order to choose a suitable dimension, values close to one of function $\sum_{j=1}^{k} d_j^2 / \sum_{j=1}^{p} d_j^2$ mean the data can be explained by a small number of principal components. Sometimes the normalized scores of the principal components u_j are used. $u_j = (X v_j)/d_j \in R^n$ for $j = 1,2,..,k$ are also orthonormal vectors $u^T u = 1$.

One way of computing the first k principal component directions is by using Singular Value Decomposition (SVD) on the centered (standardized) data matrix $X_{n\times p}$. In contrast to Eigen-decomposition method which can only be applied to square metrices (sometimes does not exist even if the matrix is square), SVD exists for any type of rectangular matrices and does not need the calculation of $p \times p$ covariance or correlation matrix in order to evaluate the principal component directions v_j.

Matrix $X_{n\times p}$ can be decomposed as product of three matrices $X = U \cdot D \cdot V^T$ where $U_{n\times k}$ the matrix that contains as columns the normalized scores u_j, for for $j = 1,2,..,k$; $D_{k\times k}$ a diagonal matrix $D = diag(d_1, d_2, \ldots, d_k)$ with $d_1 \geq d_2 \geq \cdots \geq d_k \geq 0$ and $V_{p\times k}$ the loadings matrix. The variance explained by principal component direction k is D_{kk}^2/n.

SVD provides simultaneously directions, scores and variances for the principal components [4]. $V_{p\times k}$ gives the eigenvectors of matrix $X^T X$ (right singular vectors of X) and $D_{k\times k}$ gives the square roots of the eigenvalues of matrix $X^T X$ (singular values of X), and hence the directions and standard deviations of the principal components for the data covariance matrix S_n. The principal component scores are $Y = UD$ and $U_{n\times k}$ has as columns the eigenvectors of matrix XX^T (scaled left singular vectors of X).

Standard PCA guarantees minimal loss of information during transformation since the principal components (PCs) sequentially capture the maximum variability among the columns of data matrix $X_{n\times p}$. Also, one single PC ca be addressed without referring to the other PCs since PCs are not correlated.

On the other hand, in standard PCA each principal component is a linear combination of all variables, the loadings are typically non-zero which make it difficult to interpret the new acquired information from the PCs and it is very sensitive to highly corrupted observations. Different techniques were proposed for sparse loadings/directions of the PCs and rotations techniques where the first to be applied as described by I.T Jolliffe [5]. Different thresholding techniques like Simple Thresholding, Diagonal Thresholding, Iterative Thresholding, Covariance Thresholding and Hard Thresholding can help to enhance the sparsity of the loadings/directions [6]. Another very interesting approach is to write PCA as a regression-type optimization problem.

Furthermore, in high-dimensional problems such as image processing and microarray information, where p is much greater than n, standard PCA has inconsistent results [7]. Computing all pairs of eigenvalues and corresponding eigenvectors (λ_i, v_i) for $i = 1,2, \ldots, p$ can be challenging. Estimating only the k most 'important' eigenpair can be achieved by projecting the original eigenvalue space onto a k-dimensional subspace which includes an invariant subspace associated with the 'important' eigenvectors. Krylov subspace methods as projection schemes can be used to generate these low-dimensional subspaces [8].

III. SPARSE PCA

There This paper addresses different modified versions of PCA as a regularized regression problem with a weighted L_1 penalty term added to the traditional least-squares criterion, in order to get sparse PC directions. As addressed in SVD, standard PCA can be formulated as a least-squares problem, meaning, minimizing the sum of squared residual errors between the input data and the projected data:

$$\min_A \|X - XAA^T\|_F^2$$
$$\text{subject to } A^T A = I$$

The matrix of the right singular vectors V meets this least-squares criterion, hence $A = V$ has exactly the first k loadings of the standard PCA.

Now, considering the linear regression model $Y = \sum_{j=1}^{p} X_j \beta_j + \beta_0$ with n observations and p predictors. Let $Y = (y_1, y_2, \ldots, y_n)^T$ be the measurement/response vector and $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$ for $j = 1,2, \ldots, p$ the predictors, which are centered vectors. Let $X_{n\times p}$ be the matrix containing X_j as columns and β_j the regression coefficients vectors. The Least Absolute Shrinkage and Selection Operator (LASSO) method imposes a constraint on the L_1-norm of β_j as the sum of the absolute values of them. The Lasso sparse coefficients estimates are obtained by minimizing the Lasso criterion [9]:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \|Y - \sum_{j=1}^{p} X_j \beta_j\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where λ is a non-negative tuning/penalty parameter.

The lasso coefficients estimates shrink toward zero and depending on the nature of the L_1 penalty, some coefficients will be exactly equal to zero if λ is large enough. Lasso estimates make it easier to select variables but the number of variables selected are limited by n. In case where $p >> n$, the LASSO

method can only select at most n features/variables when the number of observations is less than a thousand. In order to overcome this limitation of LASSO, H. Zou and T. Hastie [10] proposed the naïve elastic net model which is a regularized regression method that combines the Lasso penalty L_1 and ridge penalty L_2. The sparse coefficients estimates are obtained by minimizing the naïve elastic net criterion:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|_2^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\}$$

where λ_1 and λ_2 are any non-negative tuning/penalty parameters. For $\lambda_1 = \lambda$ and $\lambda_2 = 0$ the lasso estimates $\hat{\beta}_{lasso}$ are obtained. Note that minimizing the naïve elastic net criterion is equivalent to a lasso-type optimization problem. For $\lambda_2 = \lambda$ and $\lambda_1 = 0$ the ridge estimates $\hat{\beta}_R$ are obtained by minimizing the ridge criterion:

$$\min_{\beta} \left\{ \|Y - \sum_{j=1}^{p} X_j \beta_j\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 \right\}$$

Penalty L_2 is not used to penalize the regression coefficients but it ensures the reconstruction of PCs. The approximations of the loading vectors $\hat{V}_j$ of PCA equal normalized regression coefficients $\hat{\beta}_j / \|\hat{\beta}_j\|$ and $X\hat{V}_j$ are j-th approximated PC.

Viewing naïve elastic net as a regression-type optimization problem the function $(1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha \sum_{j=1}^{p}|\beta_j|^2$ is considered the elastic net penalty, where $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$ is a convex combination of the lasso penalty and the ridge penalty. For $0 < \alpha < 1$ the elastic net has the characteristics of both the ridge and lasso regression.

Since the naïve elastic net finds an estimator with a two-stage procedure which includes first finding $\hat{\beta}_R$ for each fixed λ_2, and then apply a lasso type shrinkage, it faces over shrinkage. Over shrinkage does not help reduce the variances much and leads to poor predictions. To avoid this the coefficient estimates of the naïve elastic net are scaled by a scaling factor $(1+\lambda_2)$, hence the elastic net coefficient estimates are $\hat{\beta}_{en} = (1+\lambda_2)\cdot\hat{\beta}$ [10].

Elastic net is an automatic variable selection method and by choosing the tuning parameter $\lambda_2 > 0$ it overcomes the difficulty appearing in case where $p \gg n$. Elastic net can also select groups of correlated variables.

Sparse PCA based on [11] transforms the standard PCA problem to a regression-type problem and then uses Lasso or Elastic net to produce sparse coefficients/loadings by adding the Lasso penalty or the elastic net penalty into the least-squares criterion:

$$\min_{A,B} \left\{ \sum_{i=1}^{n} \|\mathrm{x}_i - \mathrm{AB}^T \mathrm{x}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|_2^2 + \sum_{j=1}^{k} \lambda_{1,j} \|\beta_j\|_1 \right\}$$
$$\text{subject to } \mathrm{A}^T\mathrm{A} = I_{k\times k}$$

A general alternating algorithm is used to minimize this optimization problem [11] and in the case of image processing or microarray data, where $p >> n$, the computational cost is very expensive since it requires a large number of nonzero loadings by solving each elastic net problem, for a positive finite λ. Another special case of the elastic net is when $\lambda \to \infty$ [10] and it gives a convenient solution when $p >> n$. This adapted algorithm is given below:

Step 1 Let A start at $V[\,;1,2,\ldots,k]$, the loadings of the first k ordinary principal components.

Step 2 Given a fixed $A = [\alpha_1, \alpha_2, \ldots, \alpha_k]$, solve the following elastic net problem by applying soft thresholding:

$$\hat{\beta}_j = \arg\min \left\{ -2\alpha_j^T (X^T X)\beta_j + \|\beta_j\|^2{}_F^2 + \lambda_{1,j}\|\beta_j\|_1 \right\}, \quad \text{for } j = 1,2,\ldots,k$$

which has the explicit form solution

$$\beta_j = \left(\left|\alpha_j^T \mathrm{X}^T \mathrm{X}\right| - \frac{\lambda_{1,j}}{2} \right)_+ \mathrm{sgn}\left(\left|\alpha_j^T \mathrm{X}^T \mathrm{X}\right|\right) \text{ for } j = 1,2,\ldots,k.$$

Step 3 For a fixed $B = [\beta_1, \beta_2, \ldots, \beta_k]$, compute the SVD of $\mathrm{X}^T\mathrm{XB} = \mathrm{UDV}^T$, then update $A = \mathrm{UV}^T$.

Step 4 Repeat Steps 2–3, until convergence.

Step 5 Normalization: $\hat{V}_j = \frac{\beta_j}{\|\beta_j\|}$ for $j = 1,2,\ldots,k$.

Some disadvantages of Sparse PCA are that it only approximates a single PC at a time and sometimes explains less variability in the data than standard PCA. The deflation process used after estimating each PC has consequences regarding loss of orthogonality and multiple tuning parameters. To pick good tuning parameters that give a good compromise between variance and sparsity several combinations of $\lambda_{1,j}$ are tested.

All the above methods for the sparse analysis can be found in R package *elasticnet* and Matlab package *SpaSM*. They can be downloaded respectively at the links:

https://cran.r-project.org/web/packages/elasticnet/
https://www.jstatsoft.org/article/view/v084i10.

IV. Principal Component via Limited Memory Krylov Subspace Optimization

SVD decomposition was presented as a convenient tool for Standard PCA and to accelerate the performance of this method, a subspace optimization technique using limited memory block Krylov subspace was proposed by X. Liu [3]. This technique is based on the Simple Subspace Iteration (SSI), a generalization of the power method, and acceleration of SSI is achieved by reducing the number of iterations without having additional matrix-block multiplications of form $X^T X\, W$. Limited memory block Krylov accelerates the computation of the k-largest singular value decomposition of matrix $X_{n\times p}$, also the k-largest eigenvalues and leading eigenvectors of matrix $X^T X$ which maximize the Rayleigh-Ritz criterion:

$$\max_{W} \|XW\|_F^2$$
$$\text{subject to } W^T W = 1, W \in R^{p\times k}$$

First, we apply the usual SSI, starting from an initial point $W^{(0)}$ and compute the next iterate $X^{(i+1)} \in orth\left(X^T X W^{(i)}\right)$. To reduce the number of iterations in the process, the last iterate

$W^{(i)}$ is updated by an intermediate iterate $\widehat{W}^{(i)}$ and the next iterate is computed as $X^{(i+1)} \in orth\left(X^T X \widehat{W}^{(i)}\right)$. The subspace optimization problem is the following:

$$\widehat{W}^{(i)} = \arg\min_{W} \|XW\|_F^2$$
$$\text{subject to } W^T W = 1, W \in \mathcal{S}^{(i)}$$

for a chosen subspace $\mathcal{S}^{(i)}$ with a block Krylov subspace structure. $\mathcal{S}^{(i)}$ is the subspace spanned by the current i-th iterate and the previous s iterates: $\mathcal{S}^{(i)} \coloneqq span\left\{W^{(i)}, W^{(i-1)}, \cdots, W^{(i-s)}\right\}$ with block size $s \geq 0$. The block size is very important for this technique and usually at each iteration it is set a value s_{max}. The greater the value of s_{max}, the smaller the number of iterations. On the other hand, by increasing s_{max} the computational costs per iteration also increases. Different experiments indicate that 2, 3 and 4 are good suitable values for the block size/memory length but there are also other strategies adapted for selection (see, [3]).

The current and s iterate blocks obtained are placed into matrix $\boldsymbol{W} = W_s^{(i)} \in R^{p \times q}$ where $q = k \cdot (1 + s)$ is the number of columns in $\boldsymbol{W}_s^{(i)}$. SSI simultaneously computes blocks $XW_s^{(i)}$ and they are as placed in matrix $\boldsymbol{Y} = Y_s^{(i)} \in R^{n \times q}$. Note the difference between collection matrix (bold) and its blocks. Matrix $W \in \mathcal{S}^{(i)}$ if and only if $W = \boldsymbol{W}V$ for some $V \in R^{q \times k}$, so the subspace optimization problem is transformed into the generalized eigenvalue decomposition problem:

$$\max_{V} \|(X\boldsymbol{W})V\|_F^2$$
$$\text{subject to } V^T(\boldsymbol{W}^T\boldsymbol{W})V = I$$

The last problem is transformed into an equivalent and more specific decomposition problem since matrix $\boldsymbol{W}^T\boldsymbol{W}$ can become numerically rank deficient:

$$\max_{V} \|\boldsymbol{R}V\|_F^2$$
$$\text{subject to } V^T V = I$$

where $\boldsymbol{R} = R_s^{(i)} \coloneqq XQ_s^{(i)}$ and $\boldsymbol{Q} = Q_s^{(i)} \in orth\left(W_s^{(i)}\right)$ an orthonormal basis for $\mathcal{S}^{(i)}$ such that matrix $W \in \mathcal{S}^{(i)}$ is expressed as $W = \boldsymbol{Q}V$ for some $V \in R^{q \times k}$.

To compute $\boldsymbol{Q}$ and $\boldsymbol{R}$ even when the matrix $\boldsymbol{W}$ is numerically rank deficient it is used the following procedure. The last block $W^{(i)}$ is kept intact and the rest of the blocks are projected onto the null space of ${W^{(i)}}^T$. This way matrix $\boldsymbol{P}_W = P_W^{(i)} \coloneqq (I - W^{(i)}(W^{(i)})^T)\left[W^{(i-1)}, W^{(i-2)}, \cdots, W^{(i-s)}\right]$ is obtained with dimensions $p \times (k \cdot s)$. Next, stabilization is performed by deleting the columns of $\boldsymbol{P}_W$ whose Euclidean norms are below a threshold $\varepsilon_1 > 0$.

Also, the same columns deleted for $\boldsymbol{P}_W$ are deleted for matrix $\boldsymbol{P}_Y = P_Y^{(i)} \coloneqq \left[Y^{(i-1)}, \cdots, Y^{(i-s)}\right] - Y^{(i)}(W^{(i)})^T)\left[W^{(i-1)}, W^{(i-2)}, \cdots, W^{(i-s)}\right]$. Computing the eigenvalue decomposition of the Gram matrix for $\boldsymbol{P}_W$:

$$\boldsymbol{P}_W^T\boldsymbol{P}_W = U_W \Lambda_W U_W^T \in R^{(k \cdot s) \times (k \cdot s)}$$

orthogonal matrix U_W and diagonal matrix Λ_W are obtained. Afterwards, a stabilization step is performed by deleting the eigenvalues in Λ_W, and corresponding columns in U_W, that are less than $\varepsilon_2 > 0$ in order to shrink them. To generate matrix $\boldsymbol{R}$ first we need to stably construct matrix $\boldsymbol{Q}$ as $\boldsymbol{Q} = Q_s^{(i)} \coloneqq [W^{(i)}, \boldsymbol{P}_W U_W \Lambda_w^{-1/2}]$. Now that $\boldsymbol{R}$ is available we can compute the k leading eigenvectors of matrix $\boldsymbol{R}^T\boldsymbol{R} \in R^{q \times q}$, a small symmetric positive definite matrix. Let $\hat{V}$ be the solution to the generalized eigenvalue decomposition problem $\max_V \|\boldsymbol{R}V\|_F^2$ s.t $V^T V = I$, then $\widehat{W}^{(i)} = \boldsymbol{Q}\hat{V}$ and $\hat{Y}^{(i)} = \boldsymbol{R}\hat{V}$. To conclude, the next iterate of SSI is $W^{(i+1)} \in orth\left(X^T\hat{Y}^{(i)}\right)$ and $Y^{(i+1)} = XW^{((i+1)}$.

This algorithm is called the LMSVD algorithm and it is available as a Matlab directory at the the following link: https://www.caam.rice.edu/~yzhang/LMSVD/lmsvd.html#download.

V. KRYLOV PCA

This method is used to estimate the dimension of the principal/dominant subspace of the covariance matrix and approximate this dominant subspace at the same time by using Krylov subspace-based methods. As pointed out in the previews section Block Krylov methods are very useful and give optimal PCs for different types of high dimensional data matrices $X_{n \times p}$. These methods have the advantage to avoid forming the sample covariance matrix $S_n = \frac{1}{n}X^T X$ and computation of its full eigenvalue decomposition. In order to compute the top k eigenvalues of S_n and corresponding eigenvectors, the Block Lanczos algorithm can be used [2].

The criterion proposed by Sh. Ubaru [1] for selecting an appropriate dimension q of dominant subspace is derived using random matrix perturbation theory results and also contains a penalty parameter:

$$\mathrm{q} = \arg\min_{k} \left\{ \frac{n}{2\sigma^2} \sum_{i=k+1}^{p} (\lambda_i - \sigma)^2 - C_n \frac{(p-k)(p-k-1)}{2} \right\}$$

where λ_i for $i = 1,2,..,p$ are the eigenvalues of S_n, σ the noise variance related to the remaining $(p - k)$ non dominant eigenvalues and penalty C_n. Different methods can be used to estimate σ such as the thresholding method in image processing and techniques for approximating spectral densities of the matrix.
The penalty parameter has the following bound $C_n > \frac{(p+2\sqrt{np})^2}{n(p-q-1)}$.

Approximations θ_i and y_i (for $i = 1{:}k$) of the k dominant eigenvalues and corresponding eigenvectors (loadings) of S_n are obtained by using m steps of the Block Krylov Subspace $\mathbb{K}^{(m)}$. We recall that the m-th Block Krylov Subspace is $\mathbb{K}^{(m)}(A, V) = span\{V, AV, \ldots, A^{m-1}V\}$ for a symmetric matrix A and $V \epsilon R^{p \times k}$ a random matrix such that $V \nsubseteq null(A)$. For a

given error ϵ the number of optimal Block Krylov steps is $m = \log(p)/\sqrt{\epsilon}$. The error ϵ is related to the spectral gap of S_n (or singular gap of $X_{n\times p}$) so it can be equal to $(\lambda_k/\lambda_{k+1}) - 1$ and selected as the threshold. The detailed algorithm explaining this process is called Krylov PCA and it is the following (Sh. Ubaru et al., 2018):

Inputs: Transpose data matrix $X_{p\times n}$, noise σ, penalty C_n and tolerance ϵ
Step 1 Set $IC = zeros(p,1)$, $Q = [\,], k = 1$ and $m = log(p)/\sqrt{\epsilon}$
Step 2 Compute norm $\phi = \frac{1}{n^2}\|X\|_F^4 - \frac{2\sigma}{n}\|X\|_F^2 + p\sigma^2$
For $k = 1,2,\dots,p$ do
Step 3 Generate a random vector v_k with $\|v_k\|_2 = 1$.
Step 4 $K = \frac{1}{n}[Xv_k, (XX^T)Xv_k, \dots (XX^T)^{m-1}Xv_k]$
Step 5 $Q = orth([Q,K]), Q = Q(:,1{:}k)$.
Step 6 $T = \frac{1}{n}Q^T XX^T Q$.
Step 7 Compute eigen decomposition: $[V,\Theta] = eig(T)$.
Step 8 $IC(k) = n\left(\phi - \sum_{i=1}^{k}(\theta_i - \sigma)^2\right) - C_n\frac{(p-k)(p-k-1)}{2}$
Step 9 If $(k > 1)$ && $IC(k) > IC(k-1))$ then
break;
end if
End For
Output: Estimated dimension $q = k - 1$ and $Y = QV$ containing the approximated loadings as columns.

Steps 4 to 7 can be replaced by a modified version of the regular Lanczos algorithm [8], which builds an orthonormal basis for the Krylov subspace using only matrix-vector multiplications and updates the previous subspace Q and the tridiagonal matrix T. For the experiments used on image and video processing in this paper, Krylov PCA was implemented in Matlab Software *function [q,Y,lambda]=Krylov_PCA(X, sigma, epsilon, penalty, maxit)*.

VI. Background Estimation

In PCA analysis the principal components of a video are the elements of the matrix representation of n frames of size $a \times b$ that remain relatively constant over n frames i.e., background components. This means removing non static elements i.e., foreground, which can be seen as a noise added to the ground truth (GT) background. GT can be selected from one or a group of frames from the video without moving objects. The video database for experiments is SBI dataset available at https://sbmi2015.na.icar.cnr.it/SBIdataset.html. Three image sequences of different scenarios (less or more activity) have been selected to use the above methods for background estimation and compare them to the corresponding GTs. General information about selected datasets is presented in Table I.

TABLE I. Dataset general information.

Dataset	Resolution	n	p
CaVignal	136x200	258	27200
Foliage	144x200	394	28800
HighwayII	240x320	500	76800

At first, each selected frame (image) of a dataset is transformed from three dimensional arrays $a \times b \times c$ for $c = 1,2,3$ (corresponding to RGB color channels) into three 2-dimensional arrays $a \times b$, for a constant c. Then the elements of each two-dimensional array are rearranged as row vectors of length $p = a \times b$. For each dataset, using n frames from the sequence, are consequently formed three $n \times p$ data matrices. All three matrices corresponding to the sequence are centered (standardized) before using Krylov PCA and other methods include these options as arguments of the algorithms used in R or Matlab packages.

After getting PCs approximation Y_i for $i = 1{:}3$ and store projected points in matrices Z_i for $i = 1{:}3$ for each sequence, the projected data go through de-standardization since the input matrices were standardized. As a final step, we reconstruct the estimated image back to 3D. Background images estimated from SPCA, LMSVD and Krylov PCA method are listed below beside their GTs.

Figure 1. CaVignal Dataset

https://doi.org/10.5281/zenodo.4900667 https://sites.google.com/site/ijcsis/ ISSN 1947-5500

Figure 2. Foliage Dataset

Figure 3. HighwayII Dataset

In the first video CaVignal the foreground object is just one (a man), that appears in the same location in more than half of the frames used and then moves slowly through the frames. Even though there is only one object, the persistent presence affects the background subtraction, so the object significantly appears in all estimations (Fig. 1). The second video (Foliage) has more foreground objects (leaves) that obstruct the background but are not as firm during the sequence as the object in CaVignal video. All three methods in this case give a satisfactory result in background estimation (Fig. 2). Whereas the last video sequence (HighwayII) as expected, gives a very good estimation (Fig. 3). Although there is a lot of activity (moving cars), foreground objects change locations very quickly and the background is exposed most of the time.

In order to compare GTs and background images obtained from each method, peak signal-to-noise ratio (PSNR) is shown in Table II.

TABLE II. PSNR FOR EACH METHOD OF ALL DATASETS.

Method / Data	SPCA	LMSVD	Krylov PCA
CaVignal	20.01	23.64	20.24
Foliage	12.26	28.05	12.26
HighwayII	32.45	33.18	32.39

As seen in the pictures and Table II, LMSVD prevails over other methods for background estimating regardless of the activity in datasets. Meanwhile, SPCA and Krylov PCA have close accuracy to one another. When foreground objects continually move and the background is very exposed SPCA gives better estimations than Krylov PCA. Note that in this case (HighwayII) all methods give higher PSNR compared to other datasets. When foreground objects are more persistent but do not obstruct most of the background Krylov PCA gives better results than SPCA. In case where most of the background is not exposed during the sequence SPCA and Krylov PCA make no difference.

VII. CONCLUSIONS

Although Standard PCA is a great tool for PCs estimation, when it comes to high-dimensional applications such as image processing where $p \gg n$, it has inconsistent results. Also, when increasing the number of variables p, Standard PCA makes it difficult to name and interpret the new groups of variables. Different sparse PCA techniques have been used to resolve deficiencies of the standard approach. SPCA introduced by H. Zou [11] shows optimal results (a modified version of LASSO and Elastic Net) and only a few slightly-modified algorithms have been proposed using the same approach. On the other hand, these sparse techniques depend on multiple sparsity penalties/parameters which require other methods to be evaluated or different experimental tests for selection. Krylov subspace-based methods (including Block Krylov subspace) like LMSVD and Krylov PCA are contemporary methods and of high interest in data processing. Background estimation using Krylov subspace methods is very convenient due to low memory cost and their efficiency in case of high number of foreground elements (noise).

REFERENCES

[1] Sh. Ubaru, A. Seghouane and Y. Saad, “Find the dimension that counts: Fast dimension estimation and Krylov PCA”, *arXiv, Cornell University*, 2018.

[2] C. Musco and C. Musco, “Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition”, Advances in Neural Information Processing Systems, 2015.

[3] X. Liu, Z. Wen, and Y. Zhang, “Limited Memory Block Krylov Subspace Optimization for Computing Dominant Singular Value Decomposition,” *SIAM Journal on Scientific Computing*, 2013.

[4] I.T. Joliffe, *Principal Component Analysis, Second Edition*, Springer, USA, 2002.

[5] I.T. Jolliffe, “Rotation of principal components: choice of normalization constraints”, *Journal of Applied Statistics*,1995.

[6] M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation”, *Journal of Electronic Imaging*, 2004.

[7] M. Johnstone and A.Y. Lu, “On Consistency and Sparsity for Principal Components Analysis in High Dimensions”, *Journal of the American Statistical Association*, 2009.

[8] Y. Saad, Iterative Methods for Sparse Linear Systems, Second Edition, Society for Industrial and Applied Mathematics, 2003.

[9] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society*, 1996.

[10] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net”, *Journal of the Royal Statistical Society Series B*, 2005.

[11] H. Zou, T. Hastie and R. Tibshirani “Sparse Principal Component Analysis”, *Journal of Computational and Graphical Statistics*, 2006.

Backpropagation and fuzzy algorithm Modelling to Resolve Blood Supply Chain Issues in the Covid-19 Pandemic

Aan erlansari[1], Rusdi Efendi[2], Funny Farady C[3], Andang Wijanarko[4], Reza Herliansyah[5], Boko Susilo[6]
Faculty of Engineering, University of Bengkulu

aan_erlanshari@unib.ac.id, r_efendi@unib.ac.id, ffarady@unib.ac.id, andang@unib.ac.id, reza@gmail.com, bsusilo@unib.ac.id

Abstract

Bloodstock shortages and its uncertain demand has become a major problem for all countries worldwide. Therefore, this study aims to provide solution to the issues of blood distribution during the Covid-19 Pandemic at Bengkulu, Indonesia. The Backpropagation algorithm was used to improve the possibility of discovering available and potential donors. Furthermore, the distances, age, and length of donation were measured to obtain the right person to donate blood when it needed. The Backpropagation uses three input layers to classify eligible donors, namely age, body, weight, and bias. In addition, the system through its query automatically counts the variables via the Fuzzy Tahani and simultaneously access the vast database.

Keywords: Blood Supply Chain, Backpropagation, fuzzy tahini.

I. PRELIMINARY

Blood is one of the important tissues in the human body that has lots of specific functions. [1][2]. An example include transporting oxygen (O2) and nutrients and releasing toxins within the body. Furthermore, this explains the theory which states that the human body contains around 7 to 8% of blood.

This tissue is made up of three components, namely erythrocytes, leukocytes, thrombocytes, and plasma[3]. These components has a specific function to maintain human health. Furthermore, it can be transfused from one person circulatory system to another due to certain medical conditions such as trauma, surgery and shock. There are eight ABO blood types, and it is preferable to transfuse blood between patients of the same blood match in order to prevent the immune system from attacking the tranfused red blood cells.

The blood supply chain (BSC) manages the flow of blood products from donors to patients through five echelons, namely donors, mobile collection sites (CSS), blood centers (BCs), demand nodes, and patients. The demand nodes include hospitals, clinics, or other transfusion points. Furthermore, mobile CSS, BCs, and demand nodes need to be coordinated in other to perform the six main processes associated with blood donation. They include collection, testing, component processing, storage, distribution, and transfusion[4].

The shortage of bloodstock and its uncertain demand has become a significant problem for all countries worldwide. A sufficient population of donors need to be available in order meet the needs for transfusion within a reasonable period of time.. Inadequate bloodstock during the Covid-19 pandemic was mentioned many times at Bengkulu Province. Furthermore, between 2019 to 2020[5], the blood unit's realization was around 16.000 blood packs, and it was still far enough from the target of 22.000 blood units[6]. Moreover, the Red Cross faced a problematic situation in order to arrange a massive donation from a donor mobile. It is believed that when this current situation extends for a long period of time, Bengkulu would face a serious problem with no blood supply in their blood bank.

The Indonesian Red Cross chapter at Bengkulu began to collect donors data few years ago in order to manage their activities, schedule, and blood type. However, the data collected was not able to handle inadequate bloodstock during the Covid-19 Pandemic. Figure 1 [7] displays the current distribution scheme, whereby donors (suppliers) periodically come to the Indonesian Red Cross center (PMI) to donate blood or via the Mobile Collection sites (CSS). PMI is the legal organization processing the blood from the donor into final products such as whole blood, plasma, thrombocyte, and others. Furthermore, PMI distributes the final products to hospitals or other health facilities.

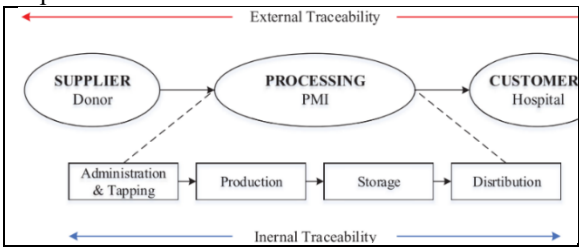


Fig 1 Blood distribution model

This problem was solved by developing a blood supply management system using Backpropagation algorithm and BSC evaluation using Fuzzy Tahani.

II. BACKPROPAGATION AND FUZZY ALGORITHM

A. *Backpropagation*

This has become the most popular method of training neural networks due to the underlying simplicity and relative power of the algorithm.

Backpropagation[8] is a learning algorithm used in reducing the error rate by adjusting the weight based on the difference in the output and the desired target. It is also defined as a Multilayer training algorithm because of its three layers, namely input, hidden, and output. Backpropagation involves developing a single-layer network with two layers, namely the input and output. Fig one shows the schematic diagram of a two-layered feed-forward network employing full connectivity between adjacent layers.

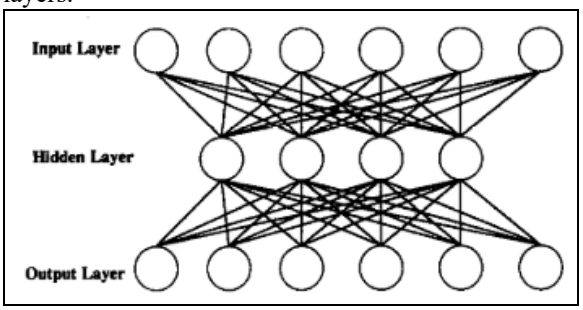


Fig. 2 Backpropagation network layer Full connection

Based on the to figure above, the 'input layer' performs no processing on its inputs and mainly distribute them to the first processing layer. While the hidden layer, receives no input and produces no output. Finally, the 'output layer' produces the output results of the network for the user. The number of input nodes is fixed by the number of input variables provided for the task. While the number of output nodes is fixed by the number of values that are desired.

B. *Fuzzy Tahani*

Fuzzy logic is the best way of mapping an input into an output area for any complicated issue. Its basic concept is to perform a calculation on input variables based on its disguised value.

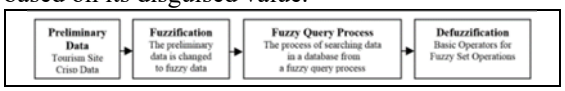


Fig 3. Fuzzification process

The figure above shows all data related to parameters and alternatives from blood donor attraction and preparedness. Furthermore, the membership function is a curve that shows the use of data and has an interval between 0 and 1 in its membership value. It is also present in a fuzzy system which has a combination membership degree between the left shoulder, triangle, and right shoulder curves. Each function's domain starts from 0 to ∞ (infinite), in order for the procedure's domain to become more flexible.

Fuzzification is the conversion of crisp values to that of fuzzy. Furthermore, a fuzzy Inference System (FIS) is in charge of making conclusions from a set of rules. Therefore,, the FIS results in this study would be used to determine the value of recommendations from age, distance, and time attractions. Query fuzzification is assumed to be a conventional query database management system that creates and implements a basic system of fuzzy query logic.

III. RESULT AND ANALYSIS

In order to overcome the issues explained in section 1, The backpropagation algorithm that evaluates donation possibility was used in solving the issues in section 1. The donor were initialized O_1 and O_2 for possible and impossible donors. Furthermore, two types of input combined with cross-validation from the table below were used.

Table 1. cross-validation

Attribute	Value
Fold number	10
Training cycle	100
Learning rate	0.001
Hidden layer	1
Neuron	3
Momentum	0.9
Error epsilon	0.001
Activation	Sigmoid (0-1)

The results from the above table is described in the figure below..

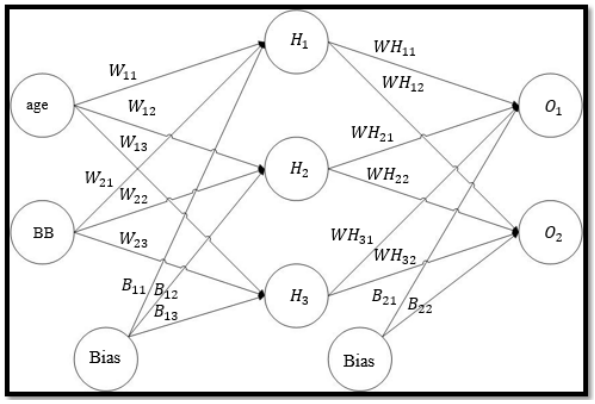


Fig 4. Backpropagation evaluation

Using this equation:

$$h_1 = i_1 \times wi_{11} + i_2 \times wi_{21} + b_1 \quad (1)$$

Initiation from Figure above we have:

$Input$: $Usia = i_1 = 0{,}827; BB = i_2 = 1$

Bias :

$B_{11} = -2.557;\ B_{12} = -2.436;\ B_{13} = -1.608$

$B_{21} = -3.690;\ B_{22} = 3.7$

Weight :

$W_{11} = 2.646;\ W_{12} = 2.530;\ W_{13} = 1.785;\ W_{21} = 2.581;\ W_{22} = 2.462;\ W_{23} = 1.676;\ WH_{11} = 3.360;\ WH_{12} = -3.327;\ WH_{21} = 3.093;\ WH_{22} = -3.066;\ WH_{31} = 1.580;\ WH_{32} = -1.658;$

And,

$$H_1 = (0{,}827 \times 2.646) + (1 \times 2.581) - 2.557$$

$$out\ h_1 = \frac{1}{1 + e^{-h_1}} = \frac{1}{1 + e^{-0.3775}} = 0.593269992$$

$$out\ h_2 = 0.596884378$$

$$o_1 = out\ h_1 \times wo_{11} + out\ h_2 \times wo_{21} + b_2$$

$$o_1 = 0.3775 \times 0.40 + 0.596884378 \times 0.45 + 0.16 = 1.105905967$$

$$out\ o_1 = \frac{1}{1+e^{-h_1}} = \frac{1}{1+e^{-1.105905967}} = 0.75136507$$

$$out\ o_2 = 0.772928465$$

$$O_1 = 3{,}4606$$

$$Out(O_1) = \frac{1}{(1 + e^{-O_1})} = \frac{1}{(1 + e^{-3.4604})} = 0.9663$$

$$O_2 = H_1 \times WH_{12} + H_2 \times WH_{22} + H_3 \times WH_{32} + B_{22}$$

$$O_2 = 0.9012 \times -3.327 + 0{,}8926 \times -3.066 + 0.8239 \times -1.658 + 3.7$$

$$O_2 = -3.517$$

$$Out(O_2) = \frac{1}{(1 + e^{-O_2})} = \frac{1}{(1 + e^{+3.517})} = 0.0338$$

Based on the calculation above, the result of confident O_1 was 0.9663 and confident O_2= 0.0338. In Tahani databases, initially, a fuzzy set formed with its membership function. In order to access the available donor, several categories were constructed, namely ages, distance, and time.

The database and its sturucture used in this study was Tahani, and relational. Furthermore, the selected data was processed using the Fuzzy Tahani method with the parameters desired by the donor. Each variable fuzzy membership function used a left shoulder, triangle, and right shoulder curves for three fuzzy sets.

A. *Age*

Figure 4 presents the membership parameter in the domain of functions 1 (Age) which is divided into three parts, namely Muda (young) which is between 0 to 33, Paruh baya (middle age) which is between 17 to 60, and Tua (elderly) which is between 33 to 60.

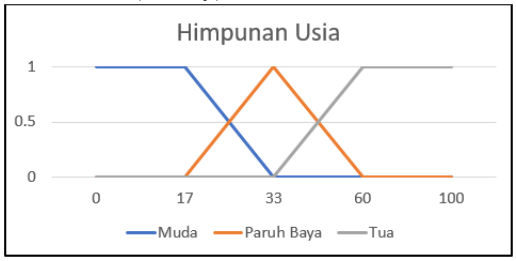


Fig 5. Set of age

The age parameter stated that the higher value was 1 when the age was at 60.

Therefore, the calculation from the existing formula for age parameter is presented below.

$$\mu_{Tua}[X] = \begin{cases} 0; & x \le a \\ \frac{x-a}{b-a}; & a \le x \le b \\ 1; & x \ge b \end{cases} \quad (2)$$

$$\mu_{Tua}[Usia] = \begin{cases} 0; & x \le 33 \\ \frac{x-33}{60-33}; & 33 \le x \le 60 \\ 1; & x \ge 60 \end{cases}$$

Table 2. age data set

Donor	age (y)	Derajat Keanggotaan		
		Muda	Baya	Tua
Person 1	38	0	0.815	0.182
Person 2	42	0	0.667	0.333
Person 3	37	0	0.852	0.148

B. Distance

Figure 5 present the membership parameter in the domain of functions 2 (distance) which is divided into three parts, namely jauh (far) 0 to 10000m, dekat (near) 5000m, agak jauh (a bit far)

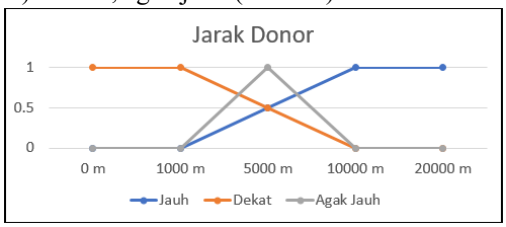


Fig 6. Set of distance

The distance parameter stated the higher value was 1 when the distance was at 10000

The fuzzy Tahani equation used in calculating the prior distance is shown below.

$$\mu_{Jauh}[X] = \begin{cases} 0; & x \leq 1000 \\ \dfrac{x-1000}{10000-1000}; & 1000 \leq x \leq 10000 \\ 1; & x \geq 10000 \end{cases}$$

Furthermore, the result of the distance parameter calculation is shown in table 3.

Table 3. distance data set

Name	Distance	Degree membership		
		Near	A bit Far	Far
Person 1	1302	0.966	0.076	0.034
Person 2	4835	0.574	0.959	0.433
Person 3	8109	0.210	0.378	0.7899

C. Donor Time

The membership parameter in the domain of function three (donor time) is presented in Figure 6, and is divided into two parts, namely *lama* (old) 0-90 days, *baru* (recent time), agak lama (short-term)

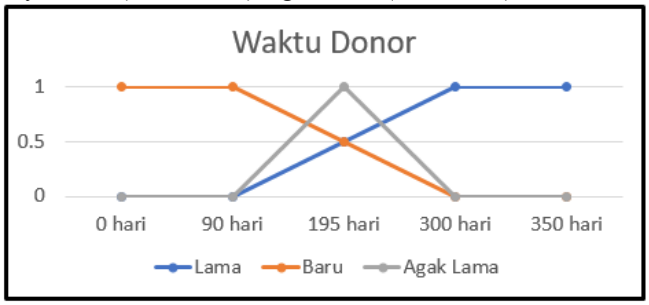


Figure 7. set of donor time

The distance parameter stated the higher value was 1 when the donor time was at 300 days or less than 90

Table 4. time data set

Name	Time	Degree membership		
		Recent time	Short-term	Old
Person 1	270	0.143	0.286	0.857
Person 2	158	0.676	0.648	0.324
Person 3	320	0	0	1

All the data set formed and calculated are shown in table 2, 3, and 4. Furthermore, Tahani was able to perform the evaluation process using the query from the database:

*Select * From Data Pendonor Where (Jarak = "Dekat") And (Usia = "Baya") And (Waktu Donor = "Lama")*

Furthermore, the results in the table below were obtained using these parameters and query.

Table 5. Evaluation using Tahani and Backpropagation

Nama Pendonor	Dekat	Baya	Lama	Prioritas
Erikson	0.966	0.815	0.857	0.857
Deddy dinpansyah	0.574	0.667	0.324	0.324
Yetti Sukmawati	0.210	0.852	1	0.210

https://doi.org/10.5281/zenodo.4901213 https://sites.google.com/site/ijcsis/ ISSN 1947-5500

IV. Conclusion

This study described the whole process of implementing Backpropagation in order to specify eligible donors. Based on Backpropagation's three input layers, the donor was identified by its blood type, age, and weight. The selection criteria was for ages between 17 to 60 years, and body weight above 40 kg. Outside these category donors were automatically eliminated by the system.

The Fuzzy Tahani algorithm classified the potential donors based on age, distance, and time to last donation. Furthermore, in order to support mobility of blood supply chain management this algorithm was embedded directly into the system using a specific query to access the database. The effectiveness of using the Backpropagation and Fuzzy Tahani produced 99.5% accuracy when selecting eligible and potential donors.

REFERENCES

[1] E. Roki, "PMI : Target Pengumpulan Darah Sepanjang Tahun 2019, Belum Terpenuhi," 2020. https://rri.co.id/bengkulu/daerah/767986/pmi-target-pengumpulan-darah-sepanjang-tahun-2019-belum-terpenuhi (accessed Oct. 10, 2020).

[2] R. Ramezanian and Z. Behboodi, "Blood supply chain network design under uncertainties in supply and demand considering social aspects," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 104, pp. 69–82, 2017, doi: 10.1016/j.tre.2017.06.004.

[3] F. Lestari, F. Ulfah, N. Nugraha, and B. Azwar, *Transactions on Engineering Technologies*. Springer Singapore, 2019.

[4] A. Pirabán, W. J. Guerrero, and N. Labadie, "Survey on blood supply chain management: Models and methods," *Comput. Oper. Res.*, vol. 112, 2019, doi: 10.1016/j.cor.2019.07.014.

[5] Antaranews, "Pandemi COVID-19, PMI Bengkulu kesulitan cari pendonor darah," 2020. https://bengkulu.antaranews.com/berita/99650/pandemi-covid-19-pmi-bengkulu-kesulitan-cari-pendonor-darah.

[6] G. T. P. P. Covid, "Peta Sebaran Covid-19." .

[7] I. Vanany, A. Maryani, B. Amaliah, F. Rinaldy, and F. Muhammad, "Blood Traceability System for Indonesian Blood Supply Chain," *Procedia Manuf.*, vol. 4, pp. 535–542, Jan. 2015, doi: 10.1016/j.promfg.2015.11.073.

[8] B. J. Wythoff, "Backpropagation neural networks. A tutorial," *Chemom. Intell. Lab. Syst.*, vol. 18, no. 2, pp. 115–155, 1993, doi: 10.1016/0169-7439(93)80052-J.

INTEGRATING NEUROLOGICAL EXAMINATION WITH RADIOLOGY DIAGNOSIS THROUGH ONTOLOGY

Suela Maxhelaku[1], Alda Kika[2], Ridvan Alihmehmeti[3]

[1, 2]*Computer Science Department, Faculty of Natural Sciences, University of Tirana, Albania*

[3]*Departament of NeuroSciences, University of Medicine, Service of Neurosurgery, University Hospital Center Mother Teresa, Albania*

Abstract: **Nowadays integration of health information is facing many challenges. There is a need in different sectors or departments at the hospital to exchange information in order to offer better healthcare for the patient. This paper describes the integration process of radiology diagnoses and findings according to the neurological examination of the patient. Also, this ontology can be used in facilitating, exchanging knowledge, offering interoperability between these sectors and it is designed in a manner that can help in the decision-making process.**

Key Words: ontology, neurology, radiology, mapping, integration, interoperability.

I. INTRODUCTION

There is a need for exchanging information between medical sectors in the hospital in order to provide better healthcare for the patients. University Hospital Center "Mother Teresa" is the biggest hospital in Albania. For instance, in the neurology and radiology department, there is a need to facilitate, exchange patient information and knowledge between them. Neurologic patients may need to be imagery diagnostification in order for the doctor/physician to understand their health situation. The patient may take different neurological examinations, such as EEG, Long Term Video Monitoring (LTVM) or medication. After certain situations the patients may be sent to the radiology department in order to take resonance or scanner according to the health situation of each of them.

All the information about the patient is very important and it is needed in a very short time because of the situation of the patient. The neurology doctor first makes the neurologic examination of the patient and then may be needed for example the resonance from the radiology department in order to identify the exact diagnosis, treatment and provide better healthcare services for the patient. So, neurology and radiology are very connected to each other aiming at identifying the diagnoses of the patient in the hospital.

This research paper gives the solution to these problems by offering an ontology that will integrate neurological examination with the radiology diagnosis. This ontology can be used in:

- the decision-making process in order to provide better healthcare for the patients;
- writing health reports;
- as a dictionary providing synonyms for different diseases, symptoms, etc.;
- providing interoperability using SNOMED CT;

In this ontology we will reuse existing ontologies like NEO Ontology, GAMUTS Ontology and SNOMED CT.

This research paper is organized in 5 sections. Section 2 presents the state of the art. Section 3 analyses the methodology, section 4 presents the Ontology and in section 5 are the conclusions and our future work.

II. STATE OF THE ART

Using ontology in the medical domain offers the opportunity to efficiently manage various sources in this domain [1]. In this research, we were focusing on ontologies used in neurology, radiology, their application and their use in these fields.

Radiology Gamuts Ontology (RGO) is a knowledge model of differential diagnoses in radiology, diseases, interventions and imaging findings [2]. In [3] it is constructed as a decision support system for potential radiological diagnoses in response to one or more user-specified imaging observations. The GAMUTS ontology has been integrated with other ontologies like Orphaned Rare Disease Ontology [4], Disease Ontology and Human Phenotype Ontology [5], RadLex, SNOMED CT and ICD-10-CM [6].

TRIES ontology is the most important component of the TRIES system that is a system for information extraction that is designed to process free texts radiology reports in order to extract and convert the available information into a structured information model [7].

RadiO is a prototype application ontology for radiology reporting tasks, which aligns a controlled imaging vocabulary (RadLex) to a reference ontology for anatomy (FMA) in order to support the structured reporting of image observations and to build a knowledge base concerning how image entities are used in the process of diagnosing diseases [8].

Neurology Disease Ontology provides a formal foundation for the representation of clinical and research data pertaining to neurological diseases [9].

III. METHODOLOGY

A. Main Ontologies and standards used in the Ontology

In order to maximize the integration of data using the ontology first we consider reusing other ontologies. In the ontology we have imported NEO and GAMUTS ontologies, as described at the state of art.

NEO (Neurological Examination Ontology) Ontology is an ontology describing neurological findings (signs and symptoms). It consists of 1273 classes, 1276 declaration axioms, 1262 logical axioms [10]. Radiology Gamuts Ontology (RGO) ontology as described at the state of art is an ontology of diseases, interventions, and imaging findings that was developed to aid in decision support, education, and translational research in diagnostic radiology [11].

In the Annotation Properties we can distinguish the relationships of the ontology to other ontologies or standards. The relationships of NEO and GAMUTS Ontology with other ontologies using sameAs axiom are shown in the figure above.

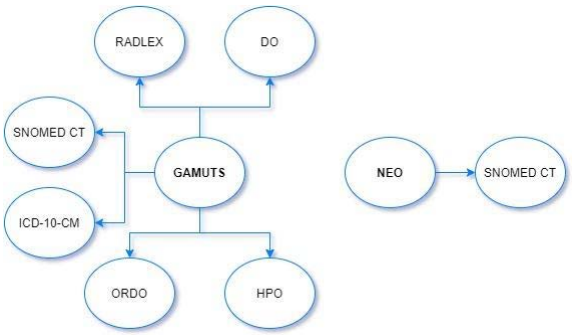


Figure 1. Relationships of NEO and GAMUTS Ontologies to other Ontologies

The NEO and the GAMUTS Ontologies are accessed using the BioPortal of the Center for Biomedical Ontology [12] [13]. The integration process between NEO and GAMUTS is done according to the ontology mapping. So the concepts of NEO will be mapped to the GAMUTS Ontology in order to gain knowledge from the may_cause and may_caused_by relationships.

SNOMED CT is the most comprehensive clinical terminology in use around the world in order to offer the integration and facilitate the exchange of clinical medical information. The advantages of using SNOMED CT range from increased opportunities for real-time decision support to more accurate retrospective reporting for research [14]. It has a hierarchical structure and includes clinical findings, anatomy, test findings, and morphological connections [15].

B. Linked data for the application ontology

According to the [16] mapping is the oriented, or directed, version of an alignment and it maps the entities of one ontology to at most one entity of another ontology. There are different tools and techniques used in matching and mapping the entities from on ontology to the other ontology.

In [17] the authors propose measures for describing the similarity of different ontology parts at the lexical and conceptual level. In the process of integrating the ontologies we will use a hybrid mapping as described below.

Identifier SNOMED CT is being used in both ontologies. SNOMED CT is the most comprehensive clinical healthcare terminology [18]. As we can see from figure 1 both

ontologies have used SNOMED CT (when it is possible). In NEO there is only the SNOMED CT code for the concept (for example 52008007) while in GAMUTS the whole URI. So the first step towards finding the mappings between these ontologies is by using SNOMED CT Code. The SNOMED CT codes are accessed from the URI in the GAMUTS Ontology and are compared to the SNOMED CT codes in NEO Ontology using SQL Server Database. In the table below are identified the same concepts in both ontologies using SNOMED CT code.

TABLE I Mappings using SNOMED CT

GAMUTS Ontology	***NEO Ontology***
Myoclonus	Myoclonus
Nystagmus	Nystagmus
Papilledema	Papilledema
Paraplegia	Paraplegia
Parinaud syndrome	Parinaud syndrome
Proptosis	Proptosis

The same concept can be expressed in different ways. In the table below are some of the concepts that have the same meaning.

TABLE II Mappings using SNOMED CT

GAMUTS Ontology	***NEO Ontology***
Alopecia	Baldness
blepharospasm_oromandibular_dystonia	Meige syndrome
Pyrexia	Fever
Facial nerve paralysis	Facial paralysis
Miosis	Small Pupil
Microphthalmos	Nanophthalmos

But, it was impossible to find all matches using SNOMED CT code because the code was missing in either NEO or GAMUTS ontology.

According to [19] it has been done a lexical comparison between the list of the terms in NEO Ontology and the list of the terms in GAMUTS Ontology. The terms are enriched with synonyms when the synonyms were missing or with other synonyms. All the terms were

imported into SQL Server Database. After this process, each synonym class/concept from NEO Ontology is compared with the GAMUTS Ontology. Then all the matched terms are mapped to each other in the Protégé. Some of the mapped terms are illustrated in the above table.

TABLE III Mappings using synonyms

GAMUTS Ontology	***NEO Ontology***
Intellectual deficit	Mental retardation
Deafness	Hearing loss
Diabetes mellitus	Diabetes
Dysmorphic facies	Dysmorphic face

In table number IV is being illustrated some examples of the mapped terms that are found between two ontologies. They have the same terms used in the concept of the class but in different positions.

TABLE IV Mappings through changing the positions of the terms

GAMUTS Ontology	***NEO Ontology***
Optic ataxia	Ataxia optic
Oromandibular dystonia	Dystonia oromandibular
Essential tremor	Tremor essential
Ptosis bilateral	Bilateral ptosis
Weakness generalized	Generalized weakness
Pain neck	Neck pain

In the table below are illustrated some of the mappings that are done through comparing classes in terms of the similarity in NEO and GAMUTS Ontology.

Table V Mappings through similarity comparisons

Gamuts Ontology	***NEO Ontology***
Saccadic intrusion	saccadic intrusions
Horner syndrome	Horner's syndrome
cervical spine injury	injury of cervical spine

In the table below are illustrated some of the mappings that are done through removing words like left or right in the NEO Ontology.

TABLE VI Mappings through removing words

Gamuts Ontology	*NEO Ontology*
Eyelid ptosis	Ptosis left eyelid
Eyelid ptosis	Ptosis right eyelid
Proptosis	Right proptosis

IV. INTEGRATION THE ONTOLOGIES

As stated at the beginning of this paper, we will import NEO and GAMUTS Ontology in the new ontology. In order to save information on the origin URI of the following ontologies, we have created the relevant prefixes in Protégé. For example, for the http://www.gamuts.net/entity# URI we will create the gamuts prefix and the neo prefix for the http://www.semanticweb.org/danielhier/ontologies/2019/3/untitled-ontology-57# URI. In the Owl file of the ontology we can distinguish the terms/classes from the following ontologies and their relationships. In the example below the two terms, diabetes and diabetes_mellitus are mapped with each other using the sameAs annotation.

```
<AnnotationAssertion>
  <AnnotationProperty IRI="sameAs"/>
  <AbbreviatedIRI>neo:diabetes</AbbreviatedIRI>
<AbbreviatedIRI>gamuts:diabetes_mellitus</AbbreviatedIRI>
</AnnotationAssertion>
```

From the figure below we can distinguish the imported NEO classes used in the ontology. There are connected with the Patient class using the relevant object properties like has_symptom, has_headfinding, has_motor_finding etc.

Figure 2. The relationship of the patient class with neurological examination classes

After mapping the terms from NEO Ontology to GAMUTS Ontology, the doctor can found the may_cause and may_cause_by findings of a certain term using the knowledge from the Gamuts Ontology.

V. CONCLUSIONS AND FUTURE WORK

In this paper, it is proposed the integration between neurological examination of the patient (NEO Ontology) with the radiological diagnoses (GAMUTS Ontology) in order to offer interoperability between neurology and radiology departments by using SNOMED CT. Also the ontology describes the may_cause and may_cause_by findings (GAMUTS Ontology) in radiology according to the neurologic examination of the patients. This ontology can be used as a semantic dictionary for the neurological examination, common diseases in neurology and the related causes in radiology.

The ontology is designed in a way that it can help in easily sharing and gaining knowledge according to the patient's data. The records of the patients can be used in gaining more knowledge and relationships according to the signs, symptoms, causes or diseases.

As a future work of this research we will add neurological diagnoses and their relationships in the overall ontology.

VI. REFERENCES

https://sites.google.com/site/ijcsis/
ISSN 1947-5500
https://doi.org/10.5281/zenodo.4919416

[1] M. Zouri, N. Zouri and A. Ferworn, "ECG Knowledge Discovery Based on Ontologies and Rules Learning for the Support of Personalized Medical Decision Making," 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2020, pp. 0701-0706, doi: 10.1109/IEMCON51383.2020.9284951.
[2] Budovec JJ, Lam CA, Kahn CE Jr, Radiology Gamuts Ontology: differential diagnosis for the Semantic Web. RadioGraphics 2014; 31:254-264. Doi: https://doi.org/10.1148/rg.341135036.
[3]Kahn CE Jr. Ontology-based diagnostic decision support in radiology. Stud Health Technol Inform. 2014;205:78-82. PMID: 25160149.
[4]C. E. Kahn Jr., "Integrating ontologies of rare diseases and radiological diagnosis," Journal of the American Medical Informatics Association, vol. 2, pp. 1164-1168, 2015.
[5]M. T. Finke, R. W. Filice and C. E. Kahn Jr., "Integrating ontologies of human diseases, phenotypes, and radiological diagnosis," Journal of the American Medical Informatics Association, vol. 26, pp. 149-154, 2019.
[6] R. W. Filice and C. E. Kahn, "Integrating an Ontology of Radiology Differential Diagnosis with ICD-10-CM, RadLex, and SNOMED CT," J Digit Imaging, vol 32, pp. 206–210, 2019. https://doi.org/10.1007/s10278-019-00186-3
[7] E. Soysal, I. Cicekli, N. Baykal. Design and evaluation of an ontology based information extraction system for radiological reports in Computers in Biology and Medicine, Volume 40, Issues 11–12, 2010, Pages 900-911, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2010.10.002.
[8] Marwede, D., Fielding, M., & Kahn, T. (2007). RadiO: a prototype application ontology for radiology reporting tasks. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2007, 513–517.
[9]Jensen, M., Cox, A.P., Chaudhry, N. et al. The neurological disease ontology. J Biomed Semant 4, 42 (2013). https://doi.org/10.1186/2041-1480-4-42.
[10] Hier, D.B., Brint, S.U. A Neuro-ontology for the neurological examination. BMC Med Inform Decis Mak 20, 47 (2020). https://doi.org/10.1186/s12911-020-1066-7
[11] Kahn CE Jr. Transitive closure of subsumption and causal relations in a large ontology of radiological diagnosis. Journal of Biomedical Informatics 2016; 61:27-33, doi: https://doi.org/10.1016/j.jbi.2016.03.015
[12] Bioportal, https://bioportal.bioontology.org/ accessed on March 2021.
[13] N.F. Noy, NH Shah, P.L. Whetzel, B. Dai B, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.A. Storey, C. G. Chute, M. A. Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;37:W170–3.
[14] SNOMED International International Health Terminology Standards Development Organization: Snomedct ontology. http://www.snomed.org/, 2021.
[15] P. Deshpande, A. Rasin, J. Son, et al. (2020). Ontology-Based Radiology Teaching File Summarization, Coverage, and Integration. J Digit Imaging 33, pp.797–813. https://doi.org/10.1007/s10278-020-00331-3.
[16] J. Euzenat and P. Shvaiko. 2013. Ontology Matching (Second Edition). Springer, Verlag, Heidelberg.
[17] A. Maedche and S. Staab, Measuring Similarity between Ontologies. In: Gómez-Pérez A., Benjamins V.R. (eds) Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. EKAW 2002. Lecture Notes in Computer Science, vol. 2473, Springer, Berlin, Heidelberg., 2002. https://doi.org/10.1007/3-540-45810-7_24.
[18] T.Benson and G.Grieve, Principles of Health Interoperability Fourth Edition, Springer, 2021.
[19]X. Hu and J. Liu, "Ontology Construction and Evaluation of UAV FCMS Software Requirement Elicitation Considering Geographic Environment Factors," in IEEE Access, vol. 8, pp. 106165-106182, 2020, doi: 10.1109/ACCESS.2020.2998843.

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2021-2022
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/