Computational Science and Engineering

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

- 9th International Conference on Computational Science and Engineering (CSE 2021)
- 9th International Conference of Artificial Intelligence and Fuzzy Logic (AI & FL 2021)
- 2nd International Conference on NLP Trends & Technologies (NLPTT 2021)
- 2nd International Conference on Software Engineering, Security and Blockchain (SESBC 2021)

Published By

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-58-9
DOI: 10.5121/csit.2021.112401- 10.5121/csit.2021.112405

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

9th International Conference on Computational Science and Engineering (CSE 2021), December 24 ~ 25, 2021, Sydney, Australia, 9th International Conference of Artificial Intelligence and Fuzzy Logic (AI & FL 2021), 2nd International Conference on NLP Trends & Technologies (NLPTT 2021) and 2nd International Conference on Software Engineering, Security and Blockchain (SESBC 2021) was collocated with 9th International Conference on Computational Science and Engineering (CSE 2021). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CSE 2021, AI & FL 2021, NLPTT 2021 and SESBC 2021 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, CSE 2021, AI & FL 2021, NLPTT 2021 and SESBC 2021 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CSE 2021, AI & FL 2021, NLPTT 2021 and SESBC 2021.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

General Chair

Organization

David C. Wyld, Southeastern Louisiana University, USA
Dhinaharan Nagamalai (Eds) Wireilla Net Solutions, Australia

Program Committee Members

Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Abdellatif I. Moustafa, Umm AL-Qura University, Saudi Arabia
Abderrahim Siam, University of Khenchela, Algeria
Abderrahmane Ez-zahout, Mohammed V University, Morocco
Abdullah, Adigrat University, Ethiopia
Addisson Salazar, Universitat Politècnica de València, Spain
Adrian Olaru, University Politehnica of Bucharest, Romania
Ahmad A. Saifan, Yarmouk university, Jordan
Ahmed Naseh Ahmed Hamdan, Assistant professor, Iraq
Akhil Gupta, Lovely Professional University, India
Ali Abdrhman Mohammed Ukasha, Sebha University, Libya
Alireza Valipour Baboli, University Technical and Vocational, Iran
Amari Houda, Networking & Telecom Engineering, Tunisia
Amizah Malip, University of Malaya, Malaysia
Ana Luísa Varani Leal, University of Macau, China
Anand Nayyar, Duy Tan University, Viet Nam
António Abreu, ISEL, Portugal
Aridj Mohamed, Hassiba Benbouali University, Algeria
Basank Kumar Verma, G H Raisoni College of Engineering, India
Bouchra Marzak, Hassan II University, Morocco
Cagdas Hakan Aladag, Hacettepe University, Turkey
Cheng Siong Chin, Newcastle University, Singapore
Chittineni Suneetha, R.V.R & J.C. College of Engineering, India
Christian Mancas, Ovidius University, Romania
Dário Ferreira, University of Beira Interior, Portugal
Dharmendra Sharma, University of Canberra, Australia
Domenico Rotondi, FINCONS SpA, Italy
Dongping Tian, Baoji University of Arts and Sciences, China
Francesco Zirilli, Sapienza Universita di Roma, Italy
Grigorios N. Beligiannis, University of Patras, Greece
Hala Abukhalaf, Palestine Polytechnic University, Palestine
Hamid Ali Abed AL-Asadi, Basra University, Iraq
Hamid Khemissa, USTHB University Algiers, Algeria
Hiba Zuhair, Al-Nahrain University, Iraq
Ilango velchamy, CMR Institute of Technology, India
Ilham Huseyinov, Istanbul Aydin University, Turkey
Isa Maleki, Islamic Azad University, Iran
Iyad Alazzam, Yarmouk University, Jordan
Jawad K. Ali, University of Technology, Iraq
Jia Ying Ou, York University, Canada
Juntao Fei, Hohai University, P. R. China
Kamel Benachenhou, Blida University, Algeria
Kamel Hussein Rahouma, Minia University, Egypt
Kazuyuki Matsumoto, Tokushima University, Japan

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

9th International Conference on Computational Science and Engineering (CSE 2021)

9th International Conference of Artificial Intelligence and Fuzzy Logic (AI & FL 2021)

2nd International Conference on NLP Trends & Technologies (NLPTT 2021)

2nd International Conference on Software Engineering, Security and Blockchain (SESBC 2021)

AN INNOVATIVE METHOD TO EXTRACT DATA IN A REAL-TIME DATA WAREHOUSING ENVIRONMENT

Flavio de Assis Vilela[1] and Ricardo Rodrigues Ciferri[2]

[1]Departament of Computing, Federal Institute of Goiás, Jataí, GO
[2]Departament of Computing, Federal University of São Carlos, São Carlos, SP

ABSTRACT

ETL (Extract, Transform, and Load) is an essential process required to perform data extraction in knowledge discovery in databases and in data warehousing environments. The ETL process aims to gather data that is available from operational sources, process and store them into an integrated data repository. Also, the ETL process can be performed in a real-time data warehousing environment and store data into a data warehouse. This paper presents a new and innovative method named Data Extraction Magnet (DEM) to perform the extraction phase of ETL process in a real-time data warehousing environment based on non-intrusive, tag and parallelism concepts. DEM has been validated on a dairy farming domain using synthetic data. The results showed a great performance gain in comparison to the traditional trigger technique and the attendance of real-time requirements.

KEYWORDS

ETL, real-time, data warehousing, data extraction.

1. INTRODUCTION

In recent years, the collect of data that support the decision-making process has become fundamental for different applications domains, such as health care systems, road control systems, retail systems and smart farm systems [1], [2], [3], [4]. These data are located in operational sources (OLTP systems), that is, in the placement where occurs transactions issued by business applications. Furthermore, the operational data are available in different formats from information providers (data sources) that are autonomous, heterogeneous, and distributed. Thus, there are a lot of ways to store the data of interest for the decision-making process, such as in files stored in computer desktops, in relational databases using centralized or distributed servers, in mobile applications, in semi-structured and non-structured sources, in text files, XML or JSON files, in spreadsheets and other formats, and storage medias [5], [6].

The decision-making process is performed based on the data of interest gathered from the operation sources (data sources). Therefore, to provide value for the operational data and effectively aid the decision-making process, it should be used the ETL (Extract, Transform, and Load) process. This process is responsible for dealing with the operational data and performing the extraction, cleaning, processing, and integration of data. Thus, the data of interest available in the operational environment are sent to a data warehousing environment and stored into a multidimensional, homogeneous, and integrated database called data warehouse [7], [8], [5].

David C. Wyld et al. (Eds): CSE, AI & FL, NLPTT - 2021
pp. 01-17, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112401

In the traditional ETL process, the process for gathering data from the data sources frequently takes place in a specific and pre-defined period of time, in a loading time window or according to the organizations business rules. Generally, the frequency of updates is on a daily, weekly, monthly basis and occurs in off peak hours. Furthermore, it takes place in a predefined frequency that is suitable to the data requirements for the decision-making process in organizations [9], [5], [10]. When using the loading time window for gathering data, both operational and informational environments should be offline to perform the process. It means that while updating, the OLAP (on-line analytical processing) applications cannot access any up-to-date data. This fact narrows the ETL process to the highest time of the loading time window. Furthermore, it imposes challenges to the organizations that have branch offices around the world, which time zone prevent to defining the same loading time window period. According to Muddasir e Mohammed [11], by using the loading time window period is just suitable to organizations which allow this feature or that focus on long run goals, which data can be updated in a less frequency. So, the usage of loading time window period to update data is enough to the majority applications, as it had been widely studied and applied in OLAP applications with low cost when is used the incremental data updating approach [12].

On the other hand, from the evolution in the way that data are generated and handled for decision-making process, which are generated quickly, a lot of applications has been emerged in which needs gathering operational data from operational sources in a real-time, such as health care systems [13] and systems that handle data from sensors [14]. Moreover, some applications, such as digital agriculture systems [3], [15], [16] and road traffic systems [17], not always generate large volume of data, but they need to guarantee that the data will be available in real-time into data warehouse to making-decision process[12], [18].

The real-time requirement can be classified as hard real-time or soft real-time. We adopt the soft real-time, where the failure to meet the timing constraint requirements does not cause serious damages. So, performing all response time rigorously is not the only aspect to be considered. One example of the system that applies this concept is on-line transaction systems, which performs a lot of tasks in a real-time.

According to Sabtu et al. [19], the updating feature of traditional ETL process can negatively and critically affect the data analysis process for the applications that have this real-time requirement. This is because the results of OLAP queries that were obtained from data warehouse throughout the day can return inconsistent data in relation to the current organization situation. In other words, data warehouse is updated few times a day and the operational sources generate data continuously [20], [11]. So, this fact makes narrows to adopt the traditional ETL process in the context of data warehousing real-time requirements.

This paper proposes a new innovative method named Data Extraction Magnet to solve the extraction phase of ETL process in real-time data warehousing environments. DEM method was designed to provide low response time, scalability, decoupling between ETL process, operational environment, and soft real-time requirements. Moreover, from all available techniques to traditional ETL process, DEM is a change of paradigm as a solution to perform extraction phase of ETL process in real-time data warehousing environments, since DEM is not intrusive to gather the data of interest in data sources. The experimental tests showed a great performance gain of DEM in comparison to the traditional trigger technique.

This paper is organized as follows: Section II presents the concepts of traditional, on-demand, near real-time and real-time approaches for performing ETL process. Section III summarizes the most important related work, Section IV presents the proposed Data Extraction Magnet (DEM),

Section V discusses the performed experiments to validate the DEM and Section VI concludes the paper.

2. ETL

A data warehousing environment is the most important component of an informational environment, which supports data management and decision-making processes. Its main goal is to promote new decision markets, identify possible failures in the enterprise organizational process, plan new tasks, define analysis of tendency and other types of analysis [21]. The data warehousing environment is composed of architectures, algorithms, tools, and systems that make possible to obtain data from information providers to make available them in a dimensional, homogeneous, and integrated repository called data warehouse [21]. Furthermore, the data warehousing environment is characterized by dealing with historical data, that is, it makes use of the facts that already happened between ten years before or more to support the decision-making process. As a consequence of data accumulated during a large previous period of time, the data volume handled by the data warehousing environment is huge and increasing quickly, and it can easily reach Petabytes of storage [21], [8], [11].

ETL process is an inseparable part of the data warehousing environment, and it is composed and represented by a workflow of tasks. These tasks are logically split in three steps: 1) the first one is called Extraction (E). The main goal is to extract or capture the data of interest from data sources. This process is performed by using a wrapper or some techniques inherent to data storage systems, such as trigger or log files. In this step, it takes place the communication to operational sources, where the data are stored in the operation environment, each data source with its specific format and access path; 2) the next step is called Transformation (T), where is performed the data cleaning, data processing and data integration. In this step, the data is cleaned and, therefore, it is eliminated the conflicts of value, structural conflicts, and semantic conflicts for the data, such as synonyms, namesake, and inconsistent values. The cleaned and transformed values are stored in a temporally data source called data staging area (DSA), which standardizes the data format, which is independent of the source formats; 3) the third step is called Loading (L), which is responsible for load data to data warehouse from DSA [9], [22]. The result of the ETL process is to make the data of interest of the operational environment available to the decision-making process, that is, storing operational data into the data warehouse in an integrated and consistent way [9], [23].

The ETL process is based on data collected from data sources in the operational environment. These data can be generated in different ways, such as: 1) manually, through users interactions with OLAP systems, such as management control systems, social media and Web systems; 2) automatically, by systems, using batch processing; 3) automatically, through sensors, such as presence sensor, road sensors and sensors connected to people or animals body; 4) automatically, by means of satellite, which sending data constantly to your stations (i.e., climate data and vehicle traffic).

In addition, there are three main ETL approaches to update data from data sources to the data warehouse: 1) on-demand: the process of extracting and loading are performed, and operational data are stored directly into the data warehouse without applying transformation, cleaning, processing, or integration data processes. For each requested OLAP query, the transformation, cleaning, processing, and integration processes are performed on the fly and the query results are shown to the user. Thus, the main part of the ETL process takes place in posteriori, that is, at query processing moment. In this way, the drawback is the high response time to answer the OLAP query; 2) near real-time: the data are stored into the data warehouse in a lower response time than traditional way. In other words, there are organizations that, not necessarily, needs data in a real-time, but just in a lower response time. Thus, through this approach, the ETL process is

performed at a higher frequency throughout the day. In this case, it is acceptable to have a certain latency, such as each hour or each minute. According to Muddasir and Mohammed [11], Kakish and Kraft [23], Langseth, J. [24], Sabry and Ali [8], there are three different techniques to be applied to the ETL process to reach the near real-time approach goals. These techniques are: *Direct trickle feed*, *Trickle and flip* and *External real-time data cache*. Basically, the most difference in each technique is just the way to handle the historical and new data. However, all techniques are performed using an intrusive strategy. An intrusive strategy needs to directly access the operational sources and also deal with the way to connect to the data sources and the heterogeneity of the operational sources.; 3) real-time: this approach stores the data into the data warehouse in a real-time manner. The data is selected from data sources and stored in the data warehouse immediately after some insert event occurs in the data source or in as low response time as possible. However, there is an implicit overhead in data exchange between the operational environment and data warehousing environment. It is due to hardware limitations, network protocol, limitations of transfer rate of data or whatever type of restrictions that act to increase the time to send data to the target data warehouse. [7], [18], [11], [9], [25].

The immediate solution to meet real-time is to increase the data warehouse data updating frequency, that is, the ETL process can be performed more than once a day and in each application idle time [20], [7], [9]. However, this solution can have three main drawbacks: 1) data reading from operational sources: the data sources from the operational environment are used to serve a group of users that performs transactions continuously. By using traditional ETL process, all available techniques are intrusive, that is, they need do access directly the operational sources to gather the data. So, it causes a query overloading every time that ETL process access the operational sources to extract data stored into these sources. This fact can cause a performance degradation of operational sources; 2) data writing into an informational environment: the data warehouse is used to serve a group of users that execute OLAP queries over the data. These queries are complex and involve a large data volume which in turn demand an excessive time to be performed. Thus, the data warehouse will be unavailable to be updated until the end of query processing; 3) updating time and data volume: when both environments are idle, the updating process can be performed into the data warehouse. However, if there if a large data volume to be updated from operational sources to data warehouse, this process takes time to be executed and all users should wait until the process ends to have access to both environments [11]. Therefore, the use of the near real-time approach and its variants are not suitable to comply with real-time requirements in a data warehousing environment. In other words, although the ETL process has suffered a lot of adaptations to be possible to apply it in a real-time environment, it is not yet sufficient to deal with real-time data warehousing environments.

3. RELATED WORK

In this section we show the proposed method, which deals with the extraction phase of ETL process in a real-time data warehousing environment. We consider the following three groups: 1) meeting real-time data warehousing requirements; 2) applied techniques to data extraction; 3) applied techniques to deal with data extracted from sensors. Regarding to meeting real-time data warehousing requirement, Bruckner et al. [26], Naeem et al. [27], Javed, M. and Nawaz, A. [28], YiChuan and Yao [29], MadeSukarsa et al. [30], Jain et al. [13], Jia et al. [31], Obali et al. [32], Mao et al. [33], Li and Mao [20] and Muddasir and Raghuveer [9] deals with low latency. Further, just Bruckner et al. [26] and Viana et al. [34] deal with availability requirements.

Although there are a lot of works that approach different aspects of real-time environment, few works deal with more than one of these requirements. Just Bruckner et al. [26] approaches the availability and low response time requirements, and Viana et al. [34] approaches the availability and scalability requirements. Just Guerreiro et al. [1] pointed out the fact that ETL process should

deal with Big Data in a real-time data warehousing. Chieu, T. and Zeng, L. [35] deal with ETL process in a near real-time environment. However, authors do not show which real-time requirements are considered.

Regarding techniques to data extraction, in all works they did: 1) create a trigger inside the table of operational environment. By creating the trigger, it is possible to define the data of interest into trigger and store them into a target data warehouse whenever table receive an insert command; 2) fetching data of interest in log files of databases in an operational environment. It is possible to access these log files, fetching all data of interest and storing them into the target data warehouse. Regarding techniques to deal with data extraction from sensors, just Guerreiro et al. [1] show a means to extract data from it. However, the authors do not describe how the extraction is made. Moreover, to simulate the environment to generate and extract data, the authors made use of data stored into CSV files.

Although the aforementioned studies describe its methods to perform extraction phase of ETL process in real-time data warehousing environment, they differ from DEM. The main difference is that these methods do not decouple the operational environment and ETL process. In other words, for all methods, to apply ETL process, we should be intrusive, that is, we should access the operational source directly by using a wrapper, trigger, access a log file or other means to fetch data of interest. This fact causes a coupling between operational sources and ETL process, which in turn causes a performance decrease and a high cost to develop and maintain the data warehousing environment. Also, all works do not consider all requirements that DEM was built over.

4. DATA EXTRACTION MAGNET

In this section, we present DEM, a new and innovative method to comply with the extraction phase of ETL process in a real-time data warehousing environment. The Figure 1 shows the workflow of DEM and its components. This figure is composed of a representation of Operational Environment, Management Environment and ETL process. These components will be detailed in this section. To better explain the DEM workflow, we consider the following scenario: in the operational environment there are several heterogeneous data sources, which maintain operational data. It can be represented in Figure 1 by the Operational Environment component, which in turn is composed of operational sources. Moreover, data from operational sources are generated in real-time, for instance using sensors. These generated data should be extracted from the operational data sources in the ETL process.

Although there are a lot of methods to support the extraction phase of ETL process in real-time, all previous methods are based on traditional approaches. In other words, the data of interest are available to ETL process by using some CDC techniques (log files or triggers) and in intrusive way. By intrusive, these methods need to directly access the operational sources. This access is made through the wrapper connected to a log file or using triggers into relational tables. Also, these methods should deal with the details of connection to the data sources and the data heterogeneity of the operational sources. In an innovative way, DEM was developed to perform the extraction phase of ETL process in real-time in a data warehousing environment without the need to directly access the operational data sources, that is, in a non-intrusive way. To do this, DEM was built over the following properties:

- **Non-intrusive**: DEM does not access the operational sources to get data of interest using triggers, log file or some CDC techniques (that is why we call it non-intrusive). Instead, DEM is prepared to receive data of interest from operational sources by using the tag concept. So, this property causes a decoupling between the ETL process and the operational

environment. Thus, this property aids to increase the performance of the extraction phase of ETL process, as this phase turns less expensive to be performed as DEM does not need to deal with the way to access the operational sources nor deal with the heterogeneity of data. Besides, DEM was designed to get data from operational sources using a specific format, which aid to solve the data heterogeneity problem. In other words, the operational source is responsible to share information about the structure and meaning of data using the specific format of DEM to receive data.

- **Tag**: DEM is a method to receive data (instead of seeking data into operational sources) from operational sources and make them available for the remaining of the ETL process. Specifically, DEM makes use of the tag concept. Tag is an indicator that an item of data (some data from operational sources like name, number of phones, city and so on) will be used in the future into a data warehouse to the decision-making process. Thus, the tag is placed on an item of data from operational sources. By using the tag, DEM is able to receive the item of data from operational sources and automatically make it available for some data warehouse that has interest in storing data that is associated with a tag.

- **Parallelism**: as an innovative way to think about the extraction phase of ETL process, DEM was built over a parallelism concept, that is, it makes use of a lot of parallel processes to support this concept. DEM is able to receive data of interest from the operational environment in parallel, which means that data of interest does not need to wait for a long time into a queue to be processed as the number of data that is received for ETL process increases.

By using these three aforementioned properties, DEM is able to extract data of interest from operational sources in a real-time non-intrusive parallel way. However, there should be a way to allow sharing data in a heterogeneous environment. Specifically in a data warehousing environment, the operational sources can be composed of a lot of data sources like relational databases, XML files, spreadsheets, NoSQL databases and so on. In the same way, a data warehousing environment is commonly composed of a relational database. To solve this problem, DEM uses a publish/subscribe system to deliver data from operational sources to the extraction phase of the ETL process [36], [37], [38]. This component can be seen in the Figure 1 by the Pub/Sub component.

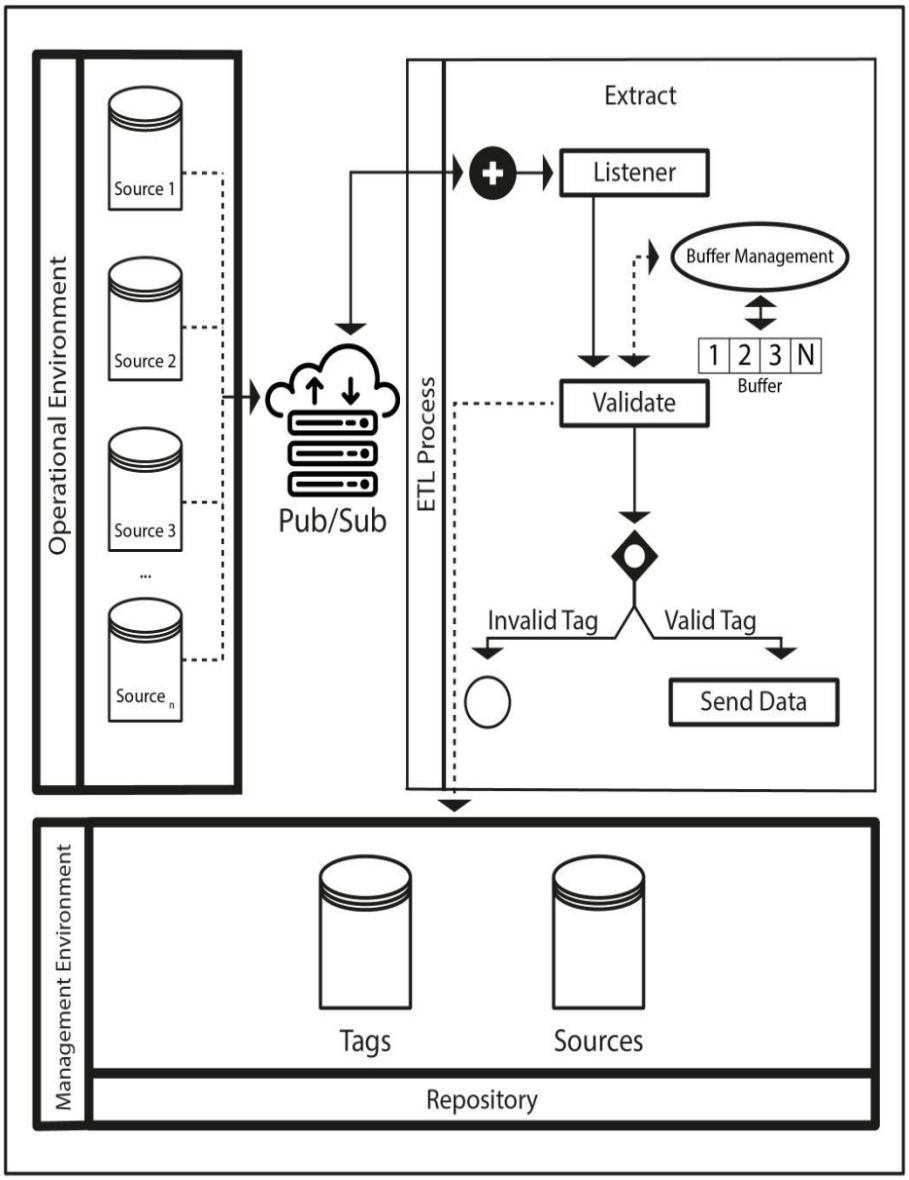


Fig. 1. Workflow of Data Extraction Magnet

The publish/subscribe system is composed by three main components:

- **Broker**: this component represents a cloud server, in which is responsible for receiving the data of interest from the publisher, storing them temporarily and making these data available to subscribers.
- **Publisher**: this component represents a data generator and in turn is responsible for sending data of interest to a broker. The data generator can be a data source, computer, smartphone, sensors, and any type of device that is capable to sending data to a broker.
- **Subscriber**: this component represents the entities that consume data sent by publishers. These entities can be relational databases, NoSQL databases, XML files and so on. The subscriber subscribes itself to the broker and expresses interest in receiving data from a

specific publisher. Whenever data is sent to broker by a publisher, subscriber is notified, and a new data is ready to be consumed. After that, the subscriber gets the data, sends a received notification, and waits for new data from a publisher.

By adopting the publish/subscribe system into DEM as a means of sharing data between the operational environment and ETL process, it is possible to integrate two different environments and overcome the problem of communication between heterogeneous environments [37]. Also, by applying the publish/subscribe system and those three aforementioned properties, we solved the problem to extracting data from one heterogeneous environment to another in real-time. Moreover, we consider that data generated by the operational environment has a strong link with the data warehouse that will store them in the future, and that is why we make the strong relation by using tag concept.

A. Repository of configurations

As previously pointed out, we need to configure tags so that we can receive data from the operational environment. So, these settings should be stored into a repository permanently and subsequently DEM can access them whenever needed to get all tags and in turn make data available for the remaining of ETL process. The repository of these settings can be seen in Figure 1 by the Repository component. The required configurations are: 1) tags; 2) data of interest from operational source and its respective tag. There are two repositories:

- **Tags**: this database stores the name of a tag that is defined by the administrator user.
- **Sources**: this database keeps relation between a tag (stored into Tags database) and a data of interest from operational sources.

It is worth highlighting that these configurations should be performed just one time by an administrator user. Once configured, DEM is able to access all repository whenever needed. This feature causes DEM act in a non-intrusive way and in turn contribute to reach the main feature of DEM.

B. Workflow of DEM

From the configurations defined by the administrator user, the following topics represents all components that DEM is built and in the order that these components are performed:

- **Operational Environment**: represents the operational data sources. In other words, this component represents a group of processes responsible for generating data in an operational environment. Also, it represents the environment that send (publishes) data to the broker to be consumed by the DEM. The operational source acts in an autonomous way and requests to send data to the data warehouse in a desirable moment. This fact allows the operational environment to adjust itself to the real-time requirements. So, the operational sources can opt to send data immediately to DEM and be sure that data will be prepared to be stored into the data warehouse immediately after data is generated in the operational environment. In this way, DEM considers that operational sources are important components to guarantee the real-time requirements. In other words, the operational source itself defines how will be defined the real-time requirements, due to the non-intrusive property.

- **Pub/Sub**: this component represents the publish/subscribe system, which is responsible to receive data from operational sources and make it available to be consumed by ETL process. It is important to note that this component represents just a way to share data between two environments. So, other techniques can be used to replace the

publish/subscribe system. However, in our perspective, the publish/subscribe system is actually the better choice.

- **ETL Process**: represents the ETL process that performs extraction, transformation and loading phases. An important aspect to highlight is that the ETL process was built over a parallelism concept. In other words, whenever new data is ready to be received from the publish/subscribe system to ETL process, the Listener component takes the data and makes it available uniquely for the rest of ETL process. It causes the ETL process to work with a lot of data received parallelly.

  - **Extraction**: it is responsible to receive data from the operational environment. To do so, it is composed by a component named Listener.
  - **Listener**: this component connects to a broker and waits for new data notifications sent by the operational environment. When data notification is received, Listener gets this data and makes it available for the Validate process of extraction phase.
    It is worth to say that DEM should receive data in a specific standard format, that is, it requires that operational data sources send data to be stored into a data warehouse in a format that is recognized and accepted by the DEM. The standard data format is: ID(Data, Date, Time). From this component, all ETL process are started parallelly by using a lot of parallel processes. By doing it, DEM can deal with a lot of data of interest in parallel and this data does not need to wait into a queue to be managed.

  - **Validate**: it is responsible for validating the data that was sent by the operational environment. These validation processes include checking whether the data has a valid tag, that is, if there is a tag associated to the item of data sent from the operational environment. Other validation is the format of the sent data. If the validation is not true, the process is stopped. Otherwise, the data is sent to the Send Data component. This validation is performed automatically by DEM and by means of a seek into repositories.

  - **Buffer Management**: it represents a buffer that store temporarily the validations that had been made by the Validate process. Once validated, the validated data is stored into Buffer so that can be accessed in the future. If the validated data is already into Buffer, we do not need to access the repository and in turn provides a performance gain.

  - **Send Data**: this component represents the end of the extraction phase. Once data is prepared to be available for the remainder of the ETL process, this component is ready to send data to the target data warehouses that have interest in storing the validated data of interest.

5. EXPERIMENTAL EVALUATION

In this section, we describe the experimental tests used to validate the feasibility and the functionalities of DEM. This experimental test was performed in a dairy farming domain. We choose this domain because its characteristics are suitable and expected to be applied to DEM. In this domain, milk data from cows is collected three times a day, which are generated by sensors connected to the cows. These milk data values are extracted and used later for dairy farmers to making-decision process.

A. Performance Analysis

The elapsed time for sending data from operational sources to the ETL process is a good way to measure the performance and scalability of DEM. In this sense, we report the following measures for the elapsed time: minimum, maximum, average, median, percentile (p) and standard deviation. It is important to highlight that to be able analyze the median and percentile, we should keep the list of all performed operations sorted by response time from the fastest to the slowest.

The median is the value of time of the operation that is in the middle of the list of operations (p50). It means that median indicates the value of time spent to perform an operation that is in the middle of the list. In turn, the percentile is important to show the values of time spent to perform other operations that are above average time and p50. As the list of operations is sorted by average response time and in ascending order of average, by analyzing percentile, it is possible to expand the performance analysis to show how far the application can perform any operations in a suitable response time. So, the percentile is obtained from the value of spent time that corresponds to a 95% e 99% operation of the list of all performed operations.

The standard deviation (SD) is a measure that indicate the degree of dispersion of a data set. In other words, the standard deviation indicate how homogeneous are the collected data. We can consider how near zero is the standard deviation, the more homogeneous the data are. So, we can use the same unit of measurement to indicate the standard deviation, and, to this experiment, we will use seconds.

Beyond average, median, percentile and standard deviation, it is important to show the minimum and maximum elapsed time. The minimum elapsed time is important to show the best case in performing some operation. Otherwise, the maximum elapsed time is important to show the worst case in performing some operation.

We also gathered storage costs for maintaining Tag and Source databases. In fact, the Data Magnet only spent 0.2 MB.

Above all techniques to measure the behavior of DEM, we should test and compare its performance results against another commonly used technique to solve the extraction phase of the ETL process in a real-time data warehousing environment. In this way, we could show how good DEM is in comparison to the other techniques. In this sense, we will compare the results of performance and scalability of DEM against the trigger technique, which is one of the most used techniques to support the extraction phase of the ETL process.

We developed a prototype to allow us to generate milk data with the same characteristics and properties of real data. However, we use trigger to perform ETL process and in turn to get the measures.

B. Experimental Test

The experimental test was made using synthetic data, that is, the milk data produced by the experiment has the same semantic, properties and characteristics of data generated by real data. However, we can control the frequency and volume in which data is generated. The synthetic data was generated by using a software named Synthetic Data Generator (SDG), which was developed just to generate synthetic data. The experimental test was built to validate the availability, low response time and scalability of DEM when increasing the volume of data generated by sensors. By doing it, we can measure the elapsed time and in turn compare DEM against the trigger technique.

The data generated synthetically by SDG has the same schema of data generated by real domain, that is, data are generated into a JSON file, encompassing fields that represents the number of cow earring, date, time, and milk volume produced by the herd a real-time data warehousing environment. So, we can define values to these fields that represents the same range of values collected in the real case. By doing this, we can keep the same characteristics of generated data and their values and control the volume and frequency that milk data are produced. Thus, we can analyze the behavior of DEM and in turn we show availability, low response time, scalability requirements and its feasibility to apply it in environments with these characteristics. So, the section V-B1 depicts the settings to be able to perform the tests, section V-B2 shows the performed tests and results.

1) Workbench

To perform the experiments with synthetic data, we used the following set of software: 1) DEM, which was running in MacBook Pro with MacOS Catalina 10.15.7, 1.4 Intel Core i5, 8GB RAM; 2) SDG, which was running in Dell Inspiron 15R, Intel Core i5, 8GB RAM and Windows 10 64 bits SO; 3) Local MySQL database as a metadata repository, which was running in the same computer that was running DEM; 4) Eclipse Mosquitto as a publish/subscribe system, which is a free service provided by Eclipse. Moreover, the Internet used in this experimental test was a shared home Internet of 10MB and one router TP-Link TL-WR840N.

Regarding configuration of software, two main configurations were performed to allow us to run the tests. The first one is regarding DEM itself. To perform the experimental test with synthetic data, the settings are: 1) define all data of interest from operational sources, in this case, data generated synthetically by SDG; 2) define tags to be assign to data of interest; By doing it, DEM is ready to perform its task in a real-time parallel and non-intrusive way.

To generate synthetic data, it is needed to set how many synthetic cows, how many synthetic sensors and the frequency in which data will be generated by SDG. Thus, the generating of data simultaneously was guided by the proportion of 8 cows managed by 1 sensor, 2 synthetic sensors managing milk data of 16 synthetic cows, 3 synthetic sensors managing milk data of 24 synthetic cows and so on. So, the second configuration is the setting of SDG, which in turn allows define how many synthetic sensors, how many synthetic cows and frequency in which milk data are generated by each sensor.

2) Scalability of the Number of Sensors

The following three tests were performed into SDG, in which each test had its synthetic sensors sending data simultaneously: 1) 32 synthetic sensors managing milk data of 250 cows; 2) 63 synthetic sensors managing milk data of 500 cows; 3) 125 synthetic sensors managing 1,000 cows. The frequency of collecting of data was performed every thirty seconds. Further, all tests were performed using DEM itself and the trigger technique.

All performed tests reflect the current scenario of all dairy farming domains. In other words, currently, there are dairy farming with around five hundred cows in lactation. However, besides considering the current scenario, we went further, and we performed tests with 125 sensors to 1,000 cows in the worst case. By doing it, we can show the scalability of DEM as the number of sensors increases.

The Figure 2 shows the results of aforementioned measures obtained after performing the test 1. Also, this figure is composed by DEM and trigger measures. As we can see, the response time of DEM for all measures was lower than trigger. It means that throughout the test DEM behavior

kept stumbling and its value of response time kept constant. We can prove it by analyzing Avg, p50, p95, p99 and SD measures.

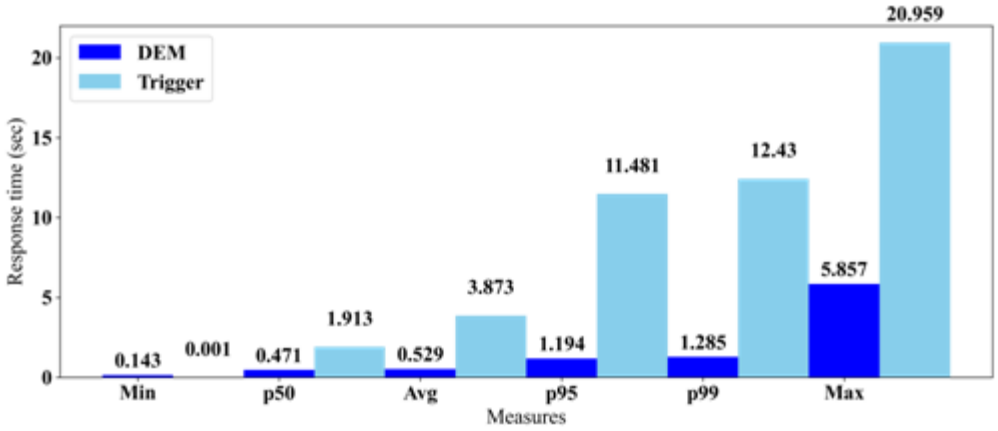


Fig. 2. Graphic of measures to show the response time of DEM and trigger for test 1. SD was 0.33 for DEM and 4.41 for Trigger.

Regarding to Avg value, DEM obtained 0.53 seconds whereas this value for trigger was 3.87 seconds. Regarding to p50, p95 and p99, the behavior of DEM is far better than trigger. The value of p50 for DEM was 0.47 seconds and for trigger was 1.91 seconds. The p95 and p99 value for DEM was 1.19 seconds and 1.29 seconds respectively, whereas the value of these measures for trigger was 11.48 seconds and 12.43 seconds. Regarding to SD, its value for DEM was 0.33 seconds whereas this value for trigger was 4.41 seconds. It means that SD value for DEM remained near zero. These values indicate that the elapsed time for DEM remained homogeneous from all collected data and for all performed tests.

Based on these analyzes, we can show a comparison of both techniques regarding performance gain and performance loss for measures. For almost all measures, DEM obtained a performance gain over trigger. For p50 measure, DEM obtained its lower performance gain, which was 306%, whereas for p99 measure, DEM obtained its higher performance gain, which was 867%. However, just for Min value DEM obtained a performance loss over trigger, which was 43%. The performance loss was caused by the overhead of managing parallel processes and tags.

The Figure 3 shows the results for the test 2. Also, this figure is composed by DEM and trigger measures. As we can note, the elapsed time of DEM for all measures was lower than trigger. It means that throughout the test DEM behavior kept stumbling and its value of elapsed time kept constant. We can analyze it by Avg, p50, p95, p99 and SD measures. Regarding to Avg value, DEM obtained 0.86 seconds whereas this value for trigger was 5.73 seconds. Regarding to p50, p95 and p99, the behavior of DEM was far better than trigger. The value of p50 for DEM was 0.49, whereas for trigger is 3.81 seconds. The p95 value for DEM was 1.21 seconds and for trigger is 17.23 seconds. The p99 value for DEM was 13.03 seconds for trigger was 20.07 seconds. Regarding to SD value for DEM, its value was 2.27 seconds, whereas this value for trigger was 5.57 seconds. It means that SD value for DEM remained near zero. These values indicate that the elapsed time for DEM remained homogeneous from all collected data and for all performed tests.

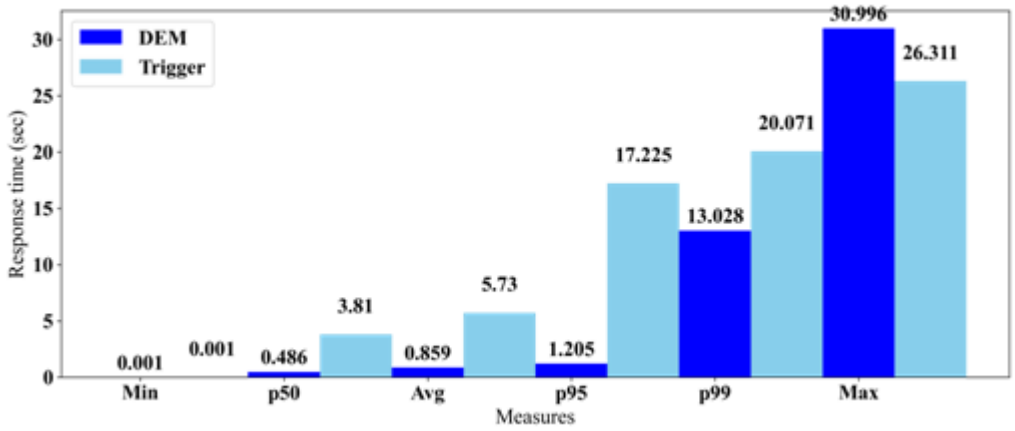


Fig. 3. Graphic of measures to show the response time of DEM and trigger for test 2. SD was 2.27 for DEM and 5.58 for Trigger.

Despite the Max value of DEM being higher than for trigger, it is not a strong measure to show how good one technique is in comparison to another technique. This is because it represents just a small fraction in relation to all collected data. Overall, by analyzing SD and Avg measures we can see that elapsed time for DEM kept constant and stable, whereas the elapsed time for trigger had a linear increase. Based on these analyzes, we can show a comparison of both techniques regarding the performance gain and the performance loss for each measure. For almost all measures, DEM obtained a performance gain over trigger. For p99 measure, DEM obtained its lower performance gain, which was 54%, whereas for p95 measure, DEM obtained its higher performance gain, which was 1,329%. However, just for Max value DEM obtained a performance loss over trigger, which was 17%.

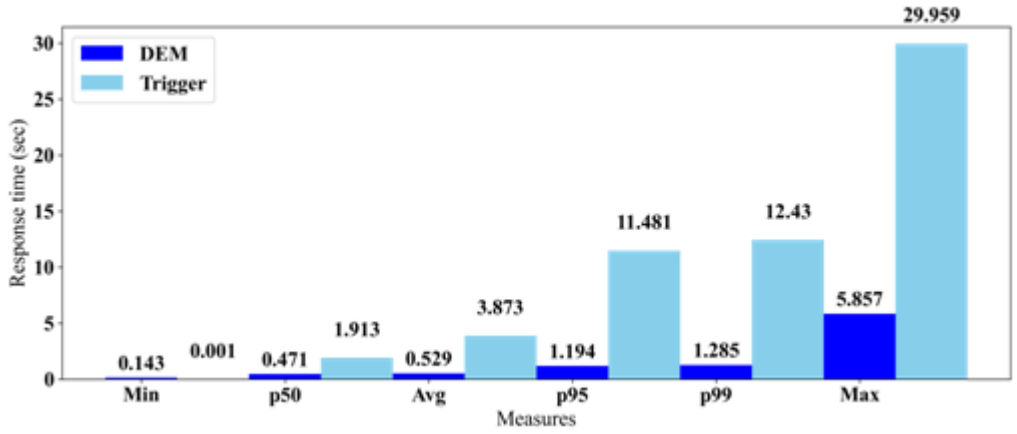


Fig. 4. Graphic of measures to show the response time of DEM and trigger for test 3. SD was 0.33 for DEM and 4.41 for Trigger.

The Figure 4 shows the results for the test 3. Also, this figure is composed by DEM and trigger measures. As we can note, the elapsed time of DEM for almost all measures was lower than trigger. It means that throughout the test DEM behavior kept stumbling and its value of response time kept constant. We can analyze it by Avg, p50, p95, p99 and SD measures. Regarding to Avg value, DEM obtained 0.47 seconds, whereas trigger obtained 3.87 seconds. Regarding to p50, p95 and p99, the behavior of DEM is far better than trigger. The value of p50 for DEM was 0.47,

whereas for trigger was 1.91 seconds. The p95 value for DEM was 1.19 seconds, whereas for trigger, this value was 11.48 seconds. The p99 value for DEM was 1.29 seconds and for trigger was 12.43 seconds. Regarding to SD value for DEM, its value was 0.33 seconds, whereas this value for trigger was 4.41 seconds. As the SD value for DEM remained near zero, we can conclude that its response time remained homogeneous from all collected data and for all performed tests.

Despite the Min value of DEM being higher than trigger, in the same way of test 2, it is not a strong measure to show how good one technique is to the other. This is because it represents just a small fraction in relation to all collected data. Further, the Min value of DEM is 0.14 seconds, whereas this value for trigger is near 0.001 seconds, but even so, Min value of DEM is great. As we can note in test 2, by analyzing SD and Avg measure to test 3, we can see that the behavior and elapsed time for DEM kept constant and stable, whereas the elapsed time for trigger had a linear increase. So, regarding the performance gain and performance loss for almost all measures, DEM obtained a performance gain over trigger. For Max measure, DEM obtained its lower performance gain, which was 75%, whereas for p95 measure, DEM obtained its higher performance gain, which was 190,404%. However, just for Min value DEM obtained a performance loss over trigger, which was 117%.

6. CONCLUSION AND FUTURE WORK

This paper presented the proposal of a new and innovative method to perform the extraction phase of the ETL process in a real-time data warehousing environment. Unlike the related work, DEM introduces the concepts of tags, parallelism, and non-intrusive properties into a real-time data warehousing context,and it shows a new way to accomplish the real-time ETL process. DEM defines the concept of real-time and discusses its applicability and feasibility for most applications that make use of the real-time data warehousing concept.

By applying non-intrusive property, DEM does not need to access the operational sources to get data of interest, conversely as is performed using trigger, log files and CDC techniques. Instead, DEM is designed to receive data of interest from operational sources. To do this, DEM makes use of another property called Tag. Basically, a tag is an indicator that an item of data from operational sources will be used in the future into a data warehouse to the decision-making process. So, this property causes a decoupling between the ETL process and the operational environment. Thus, this property aids to increase the performance of the ETL process, as the extraction phase turns less expensive to be performed, because DEM does not need to deal with the way to access the operational sources and also does not need to deal with the heterogeneity of the data.

From the experimental tests performed by using synthetic data, we could validate the low response time, scalability, and the decoupling between ETL process and operational environment. DEM provided low elapsed times even when the volume of data increased. Furthermore, DEM was better than the traditional trigger technique, where for almost all measures used to compare the two techniques, DEM showed lower response time and higher performance as data volume increased.

The main limitation observed during testing and evaluating the proposed DEM was the degradation of the performance for the max value. DEM showed in some tests a higher elapsed time for the max value then the trigger technique. This performance degradation for the Max value is because of the implicit overhead imposed by the whole environment, that is, two heterogeneous environments being integrated by a publish/subscribe system over the Internet. However, this performance loss represents a tiny quantity of all data gathered throughout the

experiment. Also, for the most configurations and measures, the Data Magnet obtained a great performance gain over the trigger technique.

For future work, we intend to extend the performed tests by applying DEM into a real case of the dairy farming domain. Besides, for the load phase of the ETL process, we intend to build an innovative way to store data into a data warehouse, that is, once data is validated and it is prepared to be sent to target data warehouse, we will consider the tag concept and create a way to data warehouse recognize the data that is associated to a tag. So, this data can be sent directly to a target data warehouse.

REFERENCES

[1] G. Guerreiro, P. Figueiras, R. Silva, R. Costa, and R. Jardim-Goncalves, "An architecture for big data processing on intelligent transportation systems. An application scenario on highway traffic flows," *2016 IEEE 8th International Conference on Intelligent Systems, IS 2016 - Proceedings*, pp. 65–71, 2016.

[2] P. Figueiras, R. Costa, G. Guerreiro, H. Antunes, A. Rosa, and R. Jardim-Goncalves, "User interface support for a big ETL data processing pipeline an application scenario on highway toll charging models," in *2017 International Conference on Engineering, Technology, and Innovation: Engineering, Technology and Innovation Management Beyond 2020: New Challenges, New Approaches, ICE/ITMC 2017 - Proceedings*, 2017.

[3] M. A. Zamora-Izquierdo, J. Santa, J. A. Martínez, V. Martínez, and A. F. Skarmeta, "Smart farming IoT platform based on edge and cloud computing," *Biosystems Engineering*, vol. 177, pp. 4–17, 2019.

[4] S. Fuentes, C. Gonzalez Viejo, B. Cullen, E. Tongson, S. Chauhan, and F. Dunshea, "Artificial intelligence applied to a robotic dairy farm to model milk productivity and quality based on cow data and daily environmental parameters," *Sensors*, vol. 20, 05 2020.

[5] R. Mukherjee and P. Kar, "A comparative review of data warehousing ETL tools with new trends and industry insight," in *Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017*, 2017.

[6] B. Yadranjiaghdam, N. Pool, and N. Tabrizi, "A survey on real-time big data analytics: Applications and tools," in *Proceedings - 2016 International Conference on Computational Science and Computational Intelligence, CSCI 2016*, 2017, pp. 404–409.

[7] H. Chandra, "Analysis of Change Data Capture Method in Heterogeneous Data Sources to Support RTDW," in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8510574/

[8] F. Sabry and E. Ali, "A Survey of Real-Time Data Warehouse and ETL," *International Journal of Scientific & Engineering Research*, vol. 5, no. 7, 2014. [Online]. Available: http://www.ijser.org

[9] M. N and R. K, "Study of Methods to Achieve Near Real Time ETL," in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, 2017, pp. 436–441.

[10] A. Wibowo, "Problems and available solutions on the stage of Extract, Transform, and Loading in near real-time data warehousing (a literature study)," *2015 International Seminar on Intelligent Technology and Its Applications, ISITIA 2015 - Proceeding*, pp. 345–349, 2015.

[11] Muddasir, Mohammed and Raghuveer, K, "CDC and Union based near real time ETL," in *2nd International Conference on Emerging Computation and Information Technologies (ICECIT)*, 2017, pp. 1–5.

[12] A. Sabtu, N. F. M. Azmi, N. N. A. Sjarif, S. A. Ismail, O. M. Yusop, H. Sarkan, and S. Chuprat, "The challenges of Extract, Transform and Loading (ETL) system implementation for near real-time environment," *International Conference on Research and Innovation in Information Systems, ICRIIS*, pp. 3–7, 2017.

[13] T. Jain, R. S, and S. Saluja, "Refreshing Datawarehouse in Near Real-Time," *International Journal of Computer Applications*, vol. 46, no. 18, pp. 975–8887, 2012.

[14] M. Mesiti, L. Ferrari, S. Valtolina, G. Licari, G. Galliani, M. Dao, and K. Zettsu, "StreamLoader: An event-driven ETL system for the on-line processing of heterogeneous sensor data," *Advances in Database Technology - EDBT*, vol. 2016-March, pp. 628–631, 2016.

[15] C. Kulatunga, L. Shalloo, W. Donnelly, E. Robson, and S. Ivanov, "Opportunistic Wireless Networking for Smart Dairy Farming," *IT Professional*, vol. 19, no. 2, pp. 16–23, 2017.

[16] M. Ryu, J. Yun, T. Miao, I. Y. Ahn, S. C. Choi, and J. Kim, "Design and implementation of a connected farm for smart farming system," *2015 IEEE SENSORS - Proceedings*, pp. 1–4, 2015.

[17] P. Figueiras, R. Costa, G. Guerreiro, H. Antunes, A. Rosa, R. Jardim-gonc¸alves, and D. D. Eng, "User Interface Support for a Big ETL Data Processing Pipeline," pp. 1437–1444, 2017.

[18] K. V. Phanikanth and S. D. Sudarsan, "A big data perspective of current ETL techniques," *Proceedings - 2016 3rd International Conference on Advances in Computing, Communication and Engineering, ICACCE 2016*, pp. 330–334, 2017.

[19] A. Sabtu, N. F. M. Azmi, N. N. A. Sjarif, S. A. Ismail, O. M. Yusop, H. Sarkan, and S. Chuprat, "The challenges of extract, transform and load (ETL) for data integration in near real-time environment," *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 22, pp. 6314–6322, 2017.

[20] X. Li and Y. Mao, "Real-Time data ETL framework for big real-time data analysis," in *2015 IEEE International Conference on Information and Automation, ICIA 2015 - In conjunction with 2015 IEEE International Conference on Automation and Logistics*, Lijiang, China, 2015, pp. 1289–1294.

[21] C. D. d. A. Ciferri, "Distribuição dos dados em ambientes de data warehousing: o Sistema WebD 2W e algoritmos voltados˜ a` fragmentação horizontal dos dados," Ph.D. dissertation, Universidade Federal de Pernambuco, 2002.˜

[22] R. Kimball, J. Caserta, R. Kimball, and J. Caserta, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data*, 2004.

[23] K. Kakish and T. A. Kraft, "ETL Evolution for Real-Time Data Warehousing," *Conference on Information Systems Applied Research*, vol. 5, no. 2214, 2012. [Online]. Available: www.aitp-edsig.org

[24] J. Langseth, "Real-Time Data Warehousing: Challenges and Solutions," 2004.

[25] C. Thomsen, T. B. Pedersen, and W. Lehner, "RiTE: Providing On-Demand Data for Right-Time Data Warehousing," *ICDE 2008*, vol. 00, pp. 456–465, 2008.

[26] R. M. Bruckner, B. List, and J. Schiefer, "Striving towards Near Real-Time Data Integration for Data Warehouses," *LNCS*, vol. 2454, pp. 317–326, 2002. [Online]. Available: http://www.ifs.tuwien.ac.at/ bruckner/pubs/dawak2002 data integration.pdf

[27] M. A. Naeem, G. Dobbie, and G. Weber, "An event-based near real-time data integration architecture," in *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC*, 2008, pp. 401–404.

[28] M. Y. Javed and A. Nawaz, "Data load distribution by semi real time data warehouse," *2nd International Conference on Computer and Network Technology, ICCNT 2010*, pp. 556–560, 2010.

[29] S. YiChuan and X. Yao, "Research of Real-time Data Warehouse Storage Strategy Based on Multi-level Caches," *Physics Procedia*, vol. 25, pp. 2315–2321, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.phpro.2012.03.390

[30] I. MadeSukarsa, N. Wayan Wisswani, I. K. Gd. Darma Putra, and L. Linawati, "Change Data Capture on OLTP Staging Area for Nearly Real Time Data Warehouse Base on Database Trigger," *International Journal of Computer Applications*, vol. 52, no. 11, pp. 32–37, 2012.

[31] R. Jia, S. Xu, and C. Peng, "Research on real time data warehouse architecture," *Communications in Computer and Information Science*, vol. 392 PART I, pp. 333–342, 2013.

[32] M. Obali, B. Dursun, Z. Erdem, and A. K. Gorur, "A real time data warehouse approach for data processing," in *2013 21st Signal Processing and Communications Applications Conference (SIU)*. IEEE, April 2013, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/document/6531245/

[33] Y. Mao, W. Min, J. Wang, B. Jia, and Q. Jie, "Dynamic mirror based real-time query contention solution for support big real-time data analysis," *Proceedings of 2nd International Conference on Information Technology and Electronic Commerce, ICITEC 2014*, pp. 229–233, 2014.

[34] N. Viana, R. Raminhos, and J. Moura-Pires, "A real time data extraction, transformation and loading solution for semi-structured text files," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3808 LNCS, pp. 383–394, 2005.

[35] T. C. Chieu and L. Zeng, "Real-time performance monitoring for an enterprise information management system," in *IEEE International Conference on e-Business Engineering, ICEBE'08 - Workshops: AiR'08, EM2I'08, SOAIC'08, SOKM'08, BIMA'08, DKEEE'08*, 2008.

[36] J. Walkenbach, *Kafka the Definitive Guide*, 2010.

[37] M. Toshev, *Learning RabbitMQ*. Birmingham, UK: Packt Publishing, 2015.
[38] E. Onica, P. Felber, H. Mercier, and E. Riviere, "Confidentiality-preserving publish/subscribe: A survey,"` *ACM Computing Surveys*, vol. 49, no. 2, pp. 1–41, 2016.

MAJRADOC AN IMAGE BASED DISEASE DETECTION APP FOR AGRICULTURAL PLANTS USING DEEP LEARNING TECHNIQUES

Sara Saleh Alfozan and Mohamad Mahdi Hassan

Department of Computer Science, College of Computer,
Qassim University, Buraydah, Saudi Arabia

ABSTRACT

Infection of agricultural plants is a serious threat to food safety. It can severely damage plants' yielding capacity. Farmers are the primary victims of this threat. Due to the advancement of AI, image-based intelligent apps can play a vital role in mitigating this threat by quick and early detection of plants infections. In this paper, we present a mobile app in this regard. We have developed MajraDoc to detect some common diseases in local agricultural plants. We have created a dataset of 10886 images for ten classes of plants diseases to train the deep neural network. The VGG-19 network model was modified and trained using transfer learning techniques. The model achieved high accuracy, and the application performed well in predicting all ten classes of infections.

KEYWORDS

Plant diseases, plant diseases diagnosis, deep learning, VGG19 CNN, mobile application.

1. INTRODUCTION

Agriculture is one of the vital sources of economic development [1]. Healthy plants are essential for human survival. It ensures higher productivity and quality of crops, fruits, and vegetables. However, infections due to various diseases negatively impact both productivity and quality of yields that may cause food insecurity [2].

Disease identification in a plant is very important in a successful farming system. In general, a farmer discovers disease symptoms in plants through naked-eye observations, which demands constant monitoring [3]. Moreover, with low education levels of farmers coupled with limited awareness and lack of access to plant pathologists, human-assisted disease diagnosis is not effective and cannot keep up with the exorbitant requirements [4]. It is not an easy task to diagnose plant disease through optical observation of the symptoms on plant leaves. This method incorporates a significantly high degree of complexity. Even skilled agronomists and plant pathologists frequently fail to detect specific illnesses due to this intricacy and the enormous number of grown plants and their existing phytopathological problems, leading to incorrect conclusions and treatments [5][6].

A smart mobile application for the detection and diagnosis of plant diseases is proposed in this work to assist farmers in detecting plant diseases. This work would be of great use to agronomists who are asked to do such diagnostics by optical observation of infected plant leaves. Instead of optical observations, the farmer can diagnose plant disease by feeding the application an image of leaves of a specific infected plant. The application will give the correct disease diagnosis with

David C. Wyld et al. (Eds): CSE, AI & FL, NLPTT - 2021
pp. 19-31, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112402

high accuracy. The mobile application was developed using an image based convolutional neural network (CNN) which can easily detect and give a faster diagnosis of plant diseases which may help develop an early treatment technique. A pre-trained deep learning model (VGG19) was trained to recognize 6 different diseases of 4 crop species, namely pepper powdery mildew, cucumber downy mildew, zucchini powdery mildew, tomato mosaic virus, tomato bacterial spot, and tomato spotted Spider mite. The model was trained and tested using a data set collected by a professional person. The VGG19 model architecture and the dataset are discussed in section 3. Section 2 presents related work. Section 4 comprehensively discusses the simulation analysis and results to ensure good performance. Finally, in section 5, the conclusions are presented.

2. RELATED WORK

Researchers who are located at varied places on the earth, alongside experts in artificial intelligence and botanists, have explored variant techniques in order to classify plant ailments. A study by Rumpf et al. has provided early findings regarding the diagnosis and the identification of sugar beet diseases using Support Vector Machine (SVM), based on the spectrogram of plant indexes [7]. Using the K-Means clustering method on color and texture extracted features, with Artificial Neural Network (ANN) Al-Hiary et al. has conducted segmentation of diseased areas for a set of five plant diseases [8]. Ravathi was able to identify 6 observed diseases from cotton leaves by Cross Information Gain Deep Forward Neural Network. The inputs of the NN were a set of vectorized information obtained from the images, such as texture, color, edge-based features, in addition to Particle Swarm Optimization for feature selection [9]. A further related study, by Mokhtar et al., has been implemented using SVM, on Tomato leaf ailments for the purpose of recognition of two viruses; the widespread tomato yellow leaf curl virus (TYLCV), and tomato spotted wilt virus (TSWV) [10]. And many other researches were conducted on different types of plants leaves and agronomic diseases such as vine, and wheat [11][12].

It is a fact that all the above-mentioned studies have successfully achieved good classification accuracies for well-known agronomic defects, within a range between 75% - 92%, by the use of regular machine learning algorithms. Nonetheless, they require manual pre-processing and multi-feature extraction techniques in a way that is expensive computationally and has excessive processing time [13].

Most of the recent researches show competitive exertions and achievements regarding plant diseases recognition with well-developed automated systems. Hitherto, the continuous underpinning of artificial intelligence technologies provided researchers community significant results with low-time consumption. Moreover, the main highlight of recently utilized deep neural network models is the exclusion of features extraction manually [14].

Thaiyalnayaki and Joseph have conducted classification of soybean diseases (absent and colored brown) using Probabilistic Neural Network (PNN), a custom net multilayer perceptron, with a large database, consisting of 683 instances (36 attributes). The PNN recruited structure has achieved 94.1435% classification accuracy with 19 output neurons and 84 input neurons. SoftMax and ReLu activation functions were chosen for the layers aside, with negative loglikelihood loss optimization function and SGD. However, the classic SVM has just achieved 88.7262% accuracy [15]. Moreover, another considerable performance for the recognition of plant diseases by image-based naive networks (Relu rectifier linear unit), and transfer learning (VGGNet, ResNet50, and Inception-v3). VGG16 had the most relevant result with 93.5 classification accuracy [16]

Rangarajan et al. (2018) have examined the separability of pre-trained AlexNet and VGG16 net among six health classes of tomato crops (13,262 segmented images from PlantVillage dataset).

When the models were at their best state by tuning the hyperparameters, changing minibatch size, and the number of images used with the models, 97.29% and 97.49% classification accuracy was achievable for AlexNet and VGG16 respectively [17]. A further automated diseases identification AI model was implemented to detect 5 different diseased mediums of leaves, by the use of CNN (convolutional neural network); 1) Early Blight, 2) Late Blight, mostly found in Potatoes, and 3) Esca, 4) Isariopsis, 5) Black Rot, in Grapes. The recruited model of CNN concluded that 87.47% and 91.96% of classification accuracies were possibly obtained for Potatoes and Grapes respectively [18].

Hallau et al. (2018) built a system to identify sugar beet leaf diseases based on captured images using smartphone cameras. The system can identify five categories of leaf diseases of sugar beet-Cercospora or leaf spot, beet rust, bacterial blight, ramularia leaf spot, andphoma leaf spot. The system goes through the following steps: 1) Infected region detection, 2) Feature extraction, and 3) Class prediction by using Support Vector Machine (SVM). They find out multiple diseases can occur simultaneously on the same plant which complicates the identification process [19].

Mohanty et. al (2016) used the PlantVillage data set which contains 38 class labels (i.e., infections) of 14 crops. They built a model using a deep convolutional neural network (CNN) to diagnose 26 diseases of those 14 crops or their absence. According to them, CNN is applicable to image classification problems even without any feature engineering. Their trained model correctly classifies crop disease with an accuracy of 99.35%. Training the model took a significant amount of time but during testing, it was very fast. Based on their tested model they developed an app for smartphones. They found some limitations though, like when tested on a set of images taken under conditions different from the images used for training, the model's accuracy reduced substantially[20].

Alvaro et al. (2017) presented a deep learning approach to detect diseases and pests in tomato plants using images captured by portable camera devices with various resolutions. Their dataset was collected from different farms on Korean Peninsula. It consists of about 5000 images, categorized and annotated into nine tomato diseases as Gray mold, Leaf mold, Low temperature, Plague, Leaf miner, Whitefly, Canker, Nutritional excess or deficiency, and Powdery mildew. Since they had a relatively smaller dataset, they applied extensive data augmentation to avoid overfitting. They considered three main detectors: Faster Region-based Convolutional Neural Network, Single Shot Multibox Detector, and Region-based Fully Convolutional Network. Finally, combined each of these meta-architectures with deep feature extractors. Their experimental results suggest various deep-meta-architectures with feature extractors can successfully and accurately recognize nine different categories of tomato diseases [21].

3. MATERIAL AND METHODS

3.1. VGG19 Network

VGG-19 [22] is a convolutional neural network that is 19 layers deep based on the stacked architecture of AlexNet with more numbers of convolution layers added to the model. More than a million images from the ImageNet database were used for training VGG-19 CNN. This network can classify photos into 1000 different object categories. It has 13 convolution layers, each of which is followed by a ReLU layer. Similar to AlexNet, some of the convolution layers are followed by max-pooling to minimize the dimension. The convolutional kernel is 3*3 in size, while the input is 224*224*3.

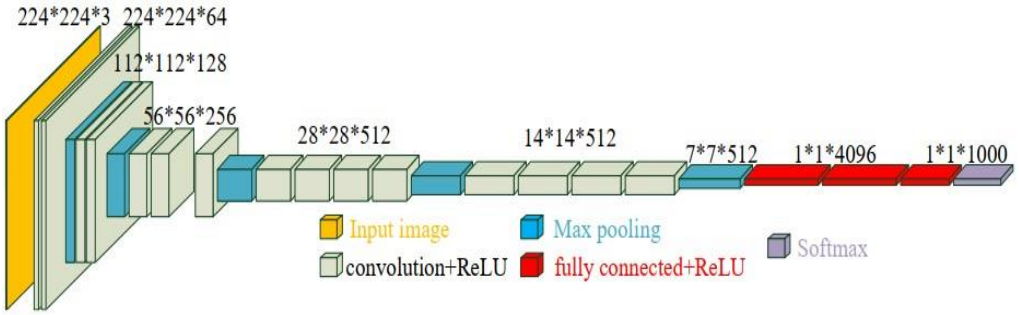


Figure 1. VGG19 network architecture 19 [22]

Figure 1 illustrates VGG-19 network model structure with the size of each layer .VGG-19 CNN is used as a pre-processing model. The network depth has been improved in comparison to traditional convolutional neural networks. It is better than a single convolution because it employs an alternating structure of several convolutional layers and non-linear activation layers. The multiple layer structure can better extract image features, use Maxpooling for downsampling, and modify the linear unit (ReLU) as the activation function, that is, choose the largest value in the image area as the area's pooled value. The downsampling layer is primarily used to increase the network's anti-distortion capabilities to the image while preserving the sample's key features and minimizing the number of parameters[23].

3.2. VGG19 CNN Implementation

In this paper, VGG19 CNN was implemented using Keras library and python language. The application of pre-trained deep learning models for classifying new classes of objects is employed in this work, which is referred to as transfer learning. In VGG-19, the parameters are concentrated in three FC layers. The parameters of the network were originally designed for 1000 classification, but this article only focuses on the classification of 10 categories (Healthy Pepper, Pepper powdery mildew disease, Healthy Cucumber, Cucumber downy mildew disease, Zucchini powdery mildew disease, Healthy Zucchini, Healthy Tomato, Tomato mosaic virus, Tomato bacterial spot, Tomato spotted Spider mite). Therefore, the last layer has been replaced with the output layer, which is equal to the number of classes. In addition, VGG-19's three fully connected layers were replaced with a single Flatten layer and three fully connected layers. Because the convolution layer and the Dense fully connected layer cannot be connected directly, a Flatten layer is added. The modified architecture of VGG19 is shown in figure 2.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv4 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv4 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv4 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
spec (Flatten)	(None, 25088)	0
dense_1 (Dense)	(None, 100)	2508900
dense_2 (Dense)	(None, 100)	10100
dense_3 (Dense)	(None, 100)	10100
dense_4 (Dense)	(None, 12)	1212

Total params: 22,554,696
Trainable params: 2,530,312
Non-trainable params: 20,024,384

Figure 2. The modified VGG19 architecture

3.3. Dataset

Training and validation datasets were collected with the help of a specialist inplant diseases at Qassim University. We were able to obtain 10886 images for 10 classes, namely Healthy Pepper, Pepper powdery mildew disease, Healthy Cucumber, Cucumber downy mildew disease, Zucchini powdery mildew disease, Healthy Zucchini, Healthy Tomato, Tomato mosaic virus disease, Tomato bacterial spot disease, Tomato spotted wilt disease. The images were divided into two datasets, the training dataset, and the validation dataset, by randomly splitting the 10886 images, so the data are allocated into training and testing set in the ratio of 80:20. In other words, 80% of the data are selected to conduct the training process while 20% are for testing purposes. The Pareto principle is a common rule of thumb to divide the dataset into two sub-sets; training and testing data. This is also called the 80/20 rule[24]. 80% of the images were used for training, and 20% of the images were used for validation. The number of images in each category are shown in figure 3.

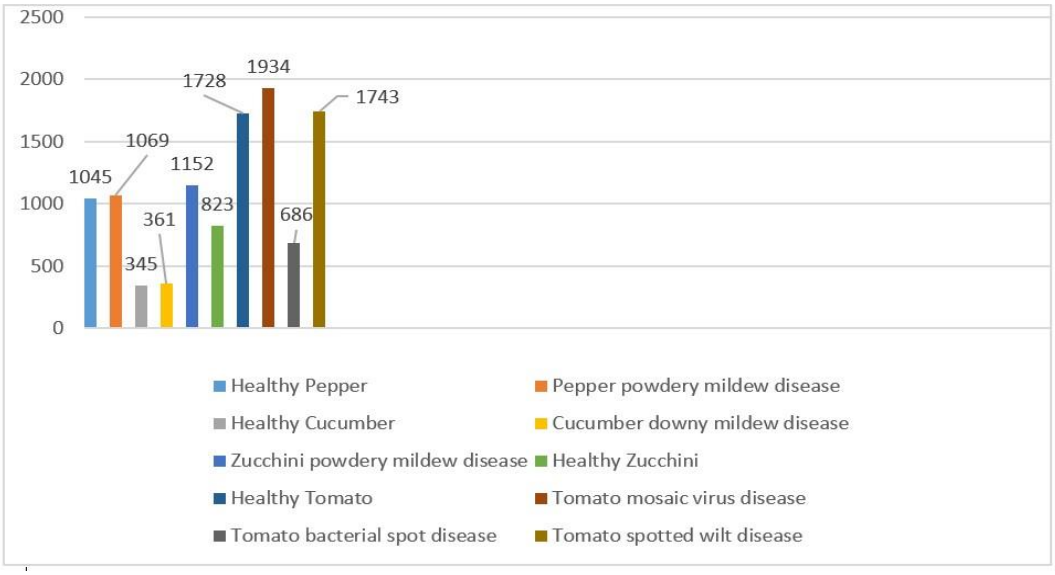


Figure 3. Number of images in each class in the dataset

Figure 4. Samples of the dataset (a) Health Pepper(b) Healthy Zucchini (c) Healthy Tomato (d) Zucchini powdery mildew disease

3.4. Mobile Application

For building the iOS mobile application, we had to convert the trained model from Keras format into a suitable format that can be used in Xcode for developing the plant disease detection and diagnosis mobile app. The tool that was used for converting the model is coremltools. Xcode was used for developing the mobile application using the converted model.

3.4.1. Coremltools

Coremltools is a Python package that allows you to: (i) convert trained models from common machine learning tools to Core ML format (. mlmodel), (ii) write models to Core ML format

using a simple API, and (iii) make predictions using the Core ML framework (on certain platforms) to check conversion. Apple's Core ML framework makes it simple for developers to integrate machine learning (ML) models into their apps. iOS, iPadOS, watchOS, macOS, and tvOS all support Core ML. Deep neural networks (convolutional and recurrent), tree ensembles (boosted trees, random forest, decision trees), and generalized linear models are among the ML

Figure 5. Keras model to Core ML model with coremltools

approaches that Core ML presents as a public file format (.mlmodel). Within XCode, you can embed core ML models straight into apps. Since our trained model was in Keras format, it was necessary to convert it into Core ML in order to be used for building an iOS mobile app using XCode. In Figure 5 we show the overall conversion process.

3.4.2. XCode

Xcode is Apple's integrated development environment for macOS, used to develop software for macOS, iOS, iPadOS, watchOS, and tvOS. It is the only officially supported way to develop iOS and other Apple OS apps. We used Xcode for developing the iOS mobile app using the model that was converted from Keras format into Core ML by coremltools Python package. The developed mobile application is easy to use and can be used by farmers for detecting the plant diseases that were mentioned above.

4. RESULTS AND DISCUSSION

The modified VGG19 model was trained using the training dataset which is 80% of the whole dataset. The learning rate was set to 0.001 and the number of epochs was set to 150. During the training process, the training accuracy, validation accuracy, training loss, and validation loss were plotted for the whole 150 epochs to show the performance of the model. Figure 6 shows that the training loss and validation loss decreases over the 150 training epochs, while the training accuracy and the validation accuracy increase consistently. The test dataset was used to find the test accuracy of the model. It was found to be 99.29% at epoch 150. The trained model was saved at this point. This high accuracy indicates that our trained model will perform well on classifying the 10 classes that it was trained to recognize, namely Healthy Pepper, Pepper powdery mildew disease, Healthy Cucumber, Cucumber downy mildew disease, Zucchini powdery mildew disease, Healthy Zucchini, Healthy Tomato, Tomato mosaic virus disease, Tomato bacterial spot disease, and Tomato spotted wilt disease.

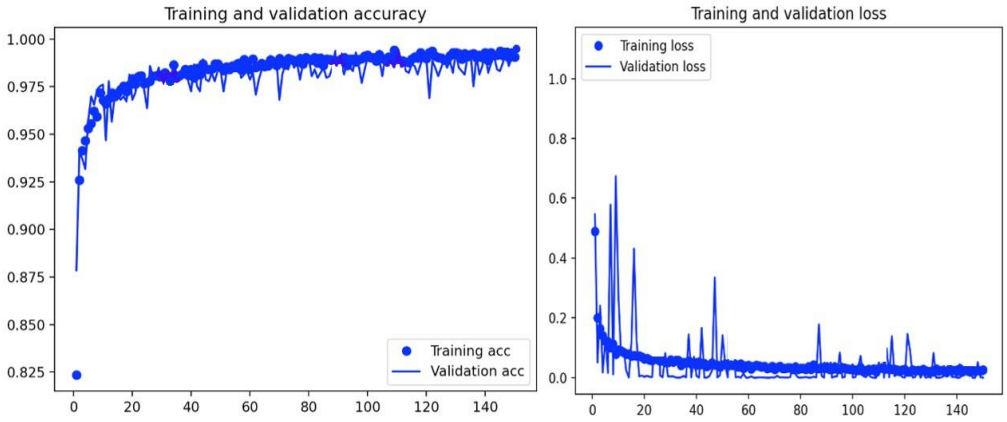


Figure 6. Accuracy and loss for training and validation of the modified VGG19 CNN model

4.1. Testing using Majradoc Mobile App

To evaluate the final performance of our proposed system for recognizing plant diseases, the mobile app that was built using the trained model was installed on iOS mobile. Random images were chosen for which the trained model has never seen. These images were captured with different orientations and illuminations and at different distances from the camera to check the robustness of the trained model. The images were fed to the mobile application one after the other. The application was able to predict the classes of the images correctly. All the following images are screenshotting for MajraDoc app using iPhone 11.

i. Healthy-Plants Images

Figure 7 illustrates the performance of the trained model for three healthy plants images were chosen randomly. It shows some results of diagnosis using MajraDoc application. Where (a) Healthy Tomato leaf (b) Healthy Cucumber leaf (c) Healthy Zucchini. The three images were fed to the mobile application. It was able to predict the classes of the images correctly.

Figure 7. Testing the performance of the model using MajraDoc application for healthy plants

ii. Sick-Plants Images

Figure 8 shows prediction to three trained plants diseases, in each prediction the model was able to give the correct diagnosis with all of these sick plants images. It shows some results of diagnosis using MajraDoc application Where (a) Zucchini powdery mildew disease (b) Pepper powdery mildew disease (c) Cucumber downy mildew disease. It was able to predict the diagnoses of the images correctly.

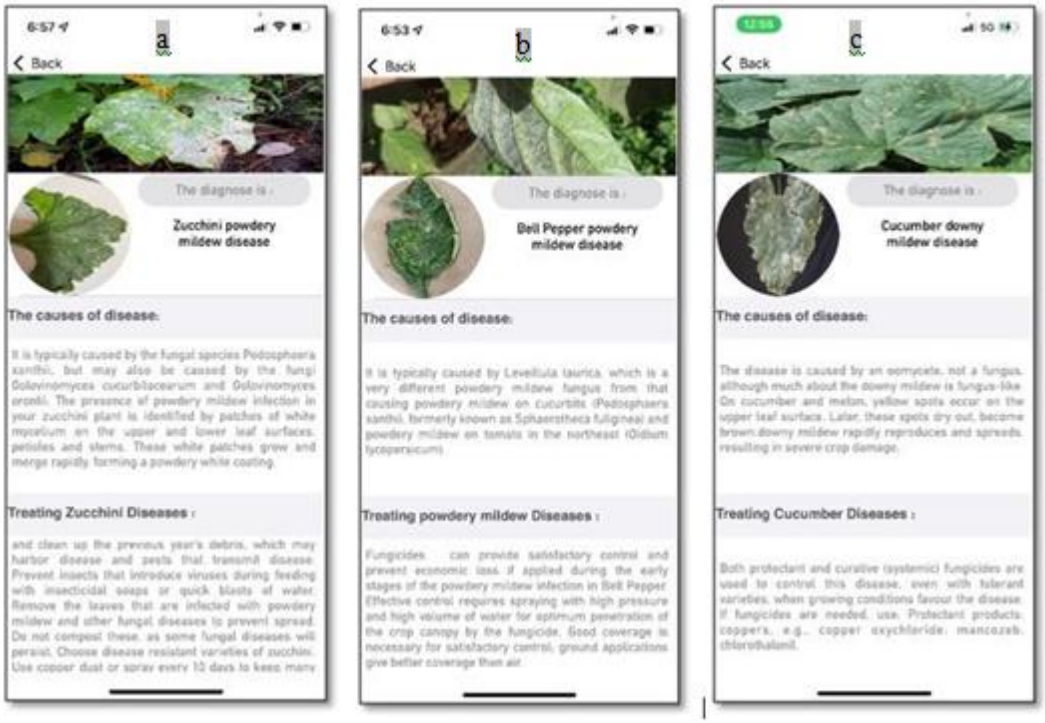


Figure 8. Testing the performance of the model using MajraDoc application for sick plants

iii. Untrained-Plants Images

Figure 9 illustrates the performance of the trained model where not trained plants were captured. It shows results of diagnosis using iPhone 11 in MajraDoc application. It was able to know the image is not from the trained plants' diseases classes.

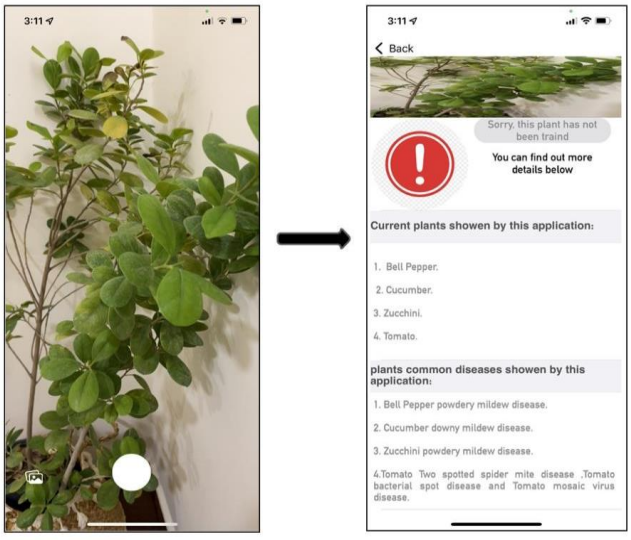


Figure 9. Testing the performance of the model using MajraDoc application for random plant

iv. Not-Plants Images

Figure 10 illustrates the performance of the trained model where no plants were captured. It shows results using iPhone 11 in MajraDoc application. It showed the MajraDoc app able to handle images, not for plants or a wrong image.

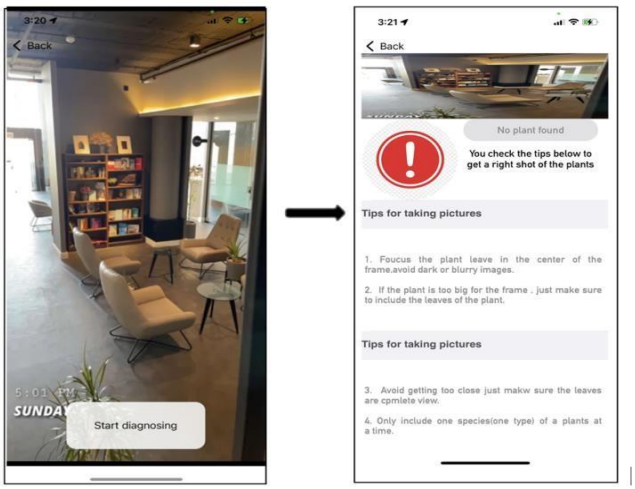


Figure 10. Testing the performance of the model using MajraDoc application for not plant

The results and discussion in this section demonstrate that our proposed system that was constructed using our modified VGG19 CCN illustrated in a previous section and the MajraDoc iOS mobile application is performing well in recognizing plant diseases for which the model was trained to recognize. In addition, performed well in distinguishing between trained and untrained plants or a random image not containing plant. This application will be beneficial for farmers in diagnosing crop disease easily.

Compared with the previous researches, our paper achieved high accuracy, as well as works on a real mobile environment .the reasons, are due to the trained dataset that was used, in addition to using a powerful and suitable Al model for the plant diseases detection. The dataset had images were captured in several directions and modify them under different lights, which gave better results.

5. CONCLUSIONS

In this work, the performance of the model was evaluated by using the test dataset that we collected. The obtained test accuracy for the model was 99.29%. An iOS mobile application was developed for the classification of plant diseases using the trained model that was converted to Core ML format (. mlmodel). The final mobile application was evaluated by feeding new images to the application. All the images were classified correctly. This study proves that deep learning architectures such as VGG-19 are suitable for the detection of plant diseases. As well as, the quality of the dataset used to train the model is very important to give correct results and achieve high accuracy. However, in this study, the model was trained to recognize only 6 plant diseases and 4 healthy plants, it can be improved to recognize more diseases by training the model on a large dataset that contains images of more plant diseases. The developed application will help farmers in detecting plant disease easily without the need of consulting plant pathologists.

REFERENCES

[1] U. Shruthi, V. Nagaveni, and B. K. Raghavendra, "A Review on Machine Learning Classification Techniques for Plant Disease Detection," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Mar. 2019, pp. 281–284, doi: 10.1109/ICACCS.2019.8728415.

[2] K. K. Singh, "An Artificial Intelligence and Cloud Based Collaborative Platform for Plant Disease Identification, Tracking and Forecasting for Farmers," in *Proceedings - 7th IEEE International Conference on Cloud Computing in Emerging Markets, CCEM 2018*, 2019, pp. 49– 56, doi: 10.1109/CCEM.2018.00016.

[3] V. Singh and A. K. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," *Inf. Process. Agric.*, vol. 4, no. 1, pp. 41–49, 2017, doi: 10.1016/j.inpa.2016.10.005.

[4] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput. Electron. Agric.*, vol. 145, pp. 311–318, 2018, doi: 10.1016/j.compag.2018.01.009.

[5] M. Jeger *et al.*, "Global challenges facing plant pathology: multidisciplinary approaches to meet the food security and environmental challenges in the mid-twenty-first century," *CABI Agric. Biosci.*, vol. 2, no. 1, 2021, doi: 10.1186/s43170-021-00042-x.

[6] S. R. Maniyath *et al.*, "Plant disease detection using machine learning," in *Proceedings - 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control, ICDI3C 2018*, 2018, pp. 41–45, doi: 10.1109/ICDI3C.2018.00017.

[7] T. Rumpf, A. K. Mahlein, U. Steiner, E. C. Oerke, H. W. Dehne, and L. Plümer, "Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance," *Comput. Electron. Agric.*, vol. 74, no. 1, pp. 91–99, 2010, doi: 10.1016/j.compag.2010.06.009.

[8] H. Al Hiary, S. Bani Ahmad, M. Reyalat, M. Braik, and Z. ALRahamneh, "Fast and Accurate Detection and Classification of Plant Diseases," *Int. J. Comput. Appl.*, vol. 17, no. 1, pp. 31–38, Mar. 2011, doi: 10.5120/2183-2754.

[9] P. Revathi and M. Hemalatha, "Cotton disease identification using proposed CIG-DFNN classifier," *Asian J. Sci. Res.*, vol. 7, no. 2, pp. 225–231, Mar. 2014, doi: 10.3923/ajsr.2014.225.231.

[10] U. Mokhtar, M. A. S. Ali, A. E. Hassanien, and H. Hefny, "Identifying two of tomatoes leaf viruses using support vector machine," in *Advances in Intelligent Systems and Computing*, 2015, vol. 339, pp. 771–782, doi: 10.1007/978-81-322-2250-7_77.

[11] X. E. Pantazi, D. Moshou, A. A. Tamouridou, and S. Kasderidis, "Leaf disease recognition in vine plants based on local binary patterns and one class support vector machines," in *IFIP Advances in Information and Communication Technology*, 2016, vol. 475, pp. 319–327, doi: 10.1007/978-3-319-44944-9_27.

[12] A. Johannes *et al.*, "Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case," *Comput. Electron. Agric.*, vol. 138, pp. 200–209, Jun. 2017, doi: 10.1016/j.compag.2017.04.013.

[13] T. M. M. S. Mohamed *et al.*, "The Identification of Significant Mechanomyography TimeDomain Features for the Classification of Knee Motion," in *Lecture Notes in Electrical Engineering*, 2022, vol. 730, pp. 313–319, doi: 10.1007/978-981-33-4597-3_29.

[14] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4. pp. 230–243, Dec. 2017, doi: 10.1136/svn-2017-000101.

[15] K. Thaiyalnayaki and C. Joseph, "Classification of plant disease using SVM and deep learning," in *Materials Today: Proceedings*, May 2021, vol. 47, pp. 468–470, doi: 10.1016/j.matpr.2021.05.029.

[16] A. V Panchal, S. C. Patel, K. Bagyalakshmi, P. Kumar, I. R. Khan, and M. Soni, "Image-based Plant Diseases Detection using Deep Learning," *Mater. Today Proc.*, Aug. 2021, doi: 10.1016/j.matpr.2021.07.281.

[17] A. K. Rangarajan, R. Purushothaman, and A. Ramesh, "Tomato crop disease classification using pre-trained deep learning algorithm," in *Procedia Computer Science*, 2018, vol. 133, pp. 1040–1047, doi: 10.1016/j.procs.2018.07.070.

[18] A. Ghosh and P. Roy, "AI Based Automated Model for Plant Disease Detection, a Deep Learning Approach," in *Communications in Computer and Information Science*, 2021, vol. 1406 CCIS, pp. 199–213, doi: 10.1007/978-3-030-75529-4_16.

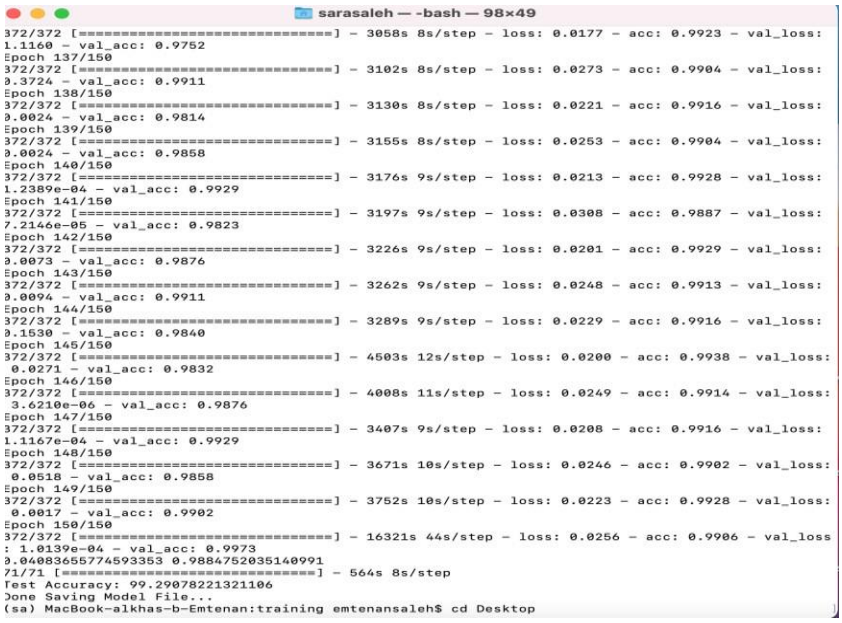
[19] L. Hallau *et al.*, "Automated identification of sugar beet diseases using smartphones," *Plant Pathol.*, vol. 67, no. 2, pp. 399–410, Feb. 2018, doi: 10.1111/ppa.12741.
[20] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front. Plant Sci.*, vol. 7, no. September, Sep. 2016, doi: 10.3389/fpls.2016.01419.
[21] A. Fuentes, S. Yoon, S. Kim, and D. Park, "A Robust Deep-Learning-Based Detector for RealTime Tomato Plant Diseases and Pests Recognition," *Sensors*, vol. 17, no. 9, p. 2022, Sep. 2017, doi: 10.3390/s17092022.
[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
[23] A. Ajit, K. Acharya, and A. Samanta, "A Review of Convolutional Neural Networks," 2020, doi: 10.1109/ic-ETITE47903.2020.049.
[24] H. B. Harvey and S. T. Sotardi, "The Pareto Principle," *J. Am. Coll. Radiol.*, vol. 15, no. 6, p. 931, 2018, doi: 10.1016/j.jacr.2018.02.026.

Authors

Sara Saleh Alfaozan: Master Student in Qassim Unvirsity, Computer Science Department, Qassim University, Saudi Arabia.

Mohammad Mahdi Hassan: an Assistant Professor in the Computer Science, Department at Qassim University, Saudi Arabia. He received his PhD in Computer Science (Specialize in Software Engineering) from University of Western Ontario, Canada. His research interests include Software Engineering, Testing, Mining Software Data, Intelligent System Development and Application of Blockchains.

The accuracy result of the trained Model



SOFTWARE ENGINEERING AND ARTIFICIAL INTELLIGENCE: RE-ENHANCING THE LIFECYCLE

Sabeer Saeed[1] and Asaf Varol[2]

[1]Department of Software Engineering, Firat University, Elazig City, Turkey
[2]Department of Computer Engineering,
Maltepe University, Istanbul City, Turkey

ABSTRACT

As automation is changing everything in today's world, there is an urgent need for artificial intelligence, the basic component of today's automation and innovation to have standards for software engineering for analysis and design before it is synthesized to avoid disaster. Artificial intelligence software can make development costs and time easier for programmers. There is a probability that society may reject artificial intelligence unless a trustworthy standard in software engineering is created to make them safe. For society to have more confidence in artificial intelligence applications or systems, researchers and practitioners in computing industry need to work not only on the cross-section of artificial intelligence and software engineering, but also on software theory that can serve as a universal framework for software development, most especially in artificial intelligence systems. This paper seeks to(a) encourage the development of standards in artificial intelligence that will immensely contribute to the development of software engineering industry considering the fact that artificial intelligence is one of the leading technologies driving innovation worldwide (b) Propose the need for professional bodies from philosophy, law, medicine, engineering, government, international community (such as NATO, UN), and science and technology bodies to develop a standardized framework on how AI can work in the future that can guarantee safety to the public among others. These standards will boost public confidence and guarantee acceptance of artificial intelligence applications or systems by both the end-users and the general public.

KEYWORDS

Software Engineering (SE); Artificial Intelligence (AI); Machine Learning (ML); Deep Learning (DL); Cross Section between Artificial Intelligence and Software Engineering.

1. INTRODUCTION

Today's world activities are driven by technology. Technology is an indispensable tool as it affects the way we work, the way we play, and the way we live generally in today's society. Technology is the ability that human beings have with tools to sustain our environment and the world at large. Decades after decades, several new technologies emerge. The invention of the integrated circuit chip by Jack Kilby, the work of Bardeen and Walter House Brattain has brought about what is called transistor which was among the reasons for personal computers [1].

As a result, individuals can acquire a computer system, which led to high demand for software which offers software developers the myth like 'We already have standby framework that is full

David C. Wyld et al. (Eds): CSE, AI & FL, NLPTT - 2021
pp. 33-44, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112403

of principles, guidelines, and approaches for developing software. Why would my people 'profile' again after having the entirety of all the required know-how to do?'[2]. Instead of without concern whether: those frameworks or standards can be applied or alive? Software professionals are mindful of the framework? Does it mirror the present or latest software engineering profession and implementation? Is it conclusive? Is it favourable to organizations? Is it more efficient to increase delivery-time while also continuing to work on effectiveness? In most situations, the response to all these problems is 'no,' those were among the events that could not be answered decades back that was called 'software crises' [2].

As computers, physical systems (hardware) results are, therefore, subject to the laws of physics, which are what eventually influence what they can or cannot do, at least in theory [1]. The introduction of engineering concepts to software development (which give birth to Software Engineering (SE)) by NATO to solve the software crises was a great achievement that today automating business is done not just done successfully, but profitably. Product (i.e. software) is software-engineered that is nearly different with respect to malleability [3].

Despite the progress made in software engineering, some researchers argued 'is yet to clearly give formal methods to software design and development', perhaps, which AI might assist in conventionalization of the software design and development [3]. Artificial intelligence (AI) keeps rising from giving the ability to computers to act based on the knowledge rules set to it, to Machine Learning where a computer is trained to learn in order to perform given tasks.

'During the 1950s when the researchers were working on the new area called AI in computer science with the purpose of 'can computer think like human beings?' or can computers perform a function like human beings?' They accepted the fact that computers can perform functions that can be discharged by humans [4]. Machine learning (ML) is the process of training a computer to learn from experience in n set of tasks. Deep learning (DL) is an advanced ML as AI encompasses both DL and ML. The value of AI is worthwhile to research to see how influential it is to SE while avoiding the pitfall. For AI to be acceptable, both engineers, lawyers, philosophers, environmental experts, and the government need to set a clear regulation. The most researched areas in AI are concerned with its abilities for mathematics and algorithm that made it capable to understand, perform, and improve from experience [5]. While the research in SE deals with enabling human beings with easiness 'tools' or 'process' to create a software that meets the need of business; to perform the task efficiently and effectively; to be easy to operate, use and maintain; and to accept modification change and do those changing for better and future enhancement [2].

Traditional software follows rigorous stages of software development life cycle from requirement elicitation to testing and implementation but AI mostly is an environmental trial. Both fields are impacting each other in the world of technology by giving tremendous progress to humanity. The study will review works to see how they can contribute to each other.
The study able to:

- ⇒ Outline how Artificial Intelligence can contribute more to Software Engineering.
- ⇒ Highlight how Software Engineering can contribute more to Artificial Intelligence.
- ⇒ Emphasize to stakeholders in the SE and AI sector not just to work on remaking the safety of autonomous vehicle (AI) but also to work on software theory for all technology such as AI that requires software to operate
- ⇒ Proposed for the need for professionals' bodies from philosophy, law, doctors, engineers, government, international communities (such as NATO, UN) and, science and technology communities to develop a professional approach, set up a high standard (on

how AI can work in the future that can guarantee safety)and be committed to the public interest.

2. ARTIFICIAL INTELLIGENCE

AI emergence begins in the 1950s, as a concept that gives the ability for computers to act like humans. It was expressed by John McCarthy (Lisp programming language pioneer [7]). McCarthy and his colleagues' work on AI can be stated in four words: Intelligent performance requires knowledge [7]. In 1843, when lady Ada Lovelace made a statement about her colleague, Charles Babbage's work on 'Analytical Engine,' she said, the manifestation of the 'engine was to assist us, humans, in doing what we know 'how' but quicker. That is, the machine can only do anything we know how to command it to do.' So, she rejected the idea of artificial intelligence. But she was contradicted by Alan Turing (AI pioneer) on his paper titled 'Computing Machinery and Intelligence' during the 1950s.

The AI pioneer revealed the Turing test as well as key technicalities that would come to entangle as AI [4]. The Pioneer, after going over the Ada Lovelace quotes while considering the alternative general-purpose-machine could be able to learn and think independently, and he came to the conclusion that they could [4]. AI investigators attain to believe intelligence by pleasing suitably to deliver the cause of intelligence as an intelligent act. Among the principle of AI procedure and collection is that advancement is 'sought' by developing a system that does 'synthesis' before 'analysis' [5].

But building the AI system by synthesizing before analysis rounding up to start operation may have the chance to end up like that of Amazon when they decided to develop an AI system that would help them 'filter CVs of potential candidates. The AI was selecting some set of CVs while also rejecting some. The 'Algorithm was tested to be biased' against females. They found out that, the reason for the algorithm rejecting women was due to 'certain traits' known to 'netball' in the hobbies segment. So, the training set was unfair which led to the scrapping of the project [8].

Nevertheless, AI is bringing more types of knowledge to witness. This means it reflects changes concurrently because, AI, of last decades, is not the AI of today. To mention a few AI is continuing to make impacts in healthcare services, email spam management, natural language processing (NLP), speech recognition, etc. In SE, AI is transforming natural language requirement (NL) into design and specification that can be used for description and data types in programming [9].

Also, in software requirements [9], the author develops a class-model builder (CM Builder) that helps in building a class diagram specified in Unified Modeling Language (UML) from the NL requirement document. AI Knowledge Base System (KBS) includes three oriented layers: knowledge acquisitive layer (which can be human-familiarized, sensible, informal), the depiction layer (convention, logical), and the implementation layer (machine-familiarized, data structures, and systems) are used to capture design relatives upon the development of the needed input and output of the system activity [7, 9].

Researchers also initiated the Ontology-Based Software Development Environment (ODE) from the software process existence [9]. The concept of Computational Intelligence (CI) techniques was built to help and encourage requirements engineering through intelligence computing. AI concepts for software architecture development, the Robyn Lutz as cited in [9], applied Genetic Algorithms (GAs) to seek the interval of potential ranking 'decompositions of a system.' She also further her investigation on Product Line Architectures (PLA) to place different points are clearly

and obviously defined to improve 'reusability and editability' of mention architecture that likely to be applied to exemplify the class set of architectures.

Experts system can help programmers in coding and testing programming process. But automate programming process could be applied by using AI techniques such as analogical reasoning to software reuse [9]. Equivalence of analogical reasoning is Case-Based Reasoning (CBR) that solves similar problems with similar solutions. Another is Experience Factory (EF), also, for reuse and management of experience, knowledge, process, and product. EF is likewise popularly known as the Learning Software Organization (LSO). There are reusable techniques such as Knowledge Acquisition (KA) and Domain Modeling (DM).

Another AI knack is Constraint programming that is applied in software engineering. For example, it is expanded to program the PTIDEJ system (Pattern Trace Identification, Detection, and Enhancement in Java), an automated system created to recognize micro-architectures mirror-like design form models in object-oriented source code. A research field appear to emphasis on expressing viewpoints of Software Engineering as problems that may be answered using meta-heuristic search algorithms built-in AI is Search-Based Software Engineering (SBSE). This SBSE is the reestablishment of software engineering function as an optimization to solving problems. Such examples of the optimization and search abilities that can be applied are genetic algorithms [9].

As for testing**,** still remains an expensive function in the Software development life cycle. AI researchers have been committed to the application of constraint solving concepts in the automation of program testing (Constraint-based testing). CBT refers to as the method of generating test scenarios or cases from software or models by applying the constraint programming technology. Handling up-to or more than hundreds of thousands of lines of code, with vigorous built structure particularly like enormous dynamic data structures, in the company of non-sequential numerical constraints obtained from ambiguous statements or facts are some of the issues we have to face with. Expandability is the major question that CBT tools have to answer for us [9].

Such AI techniques can distinguish parallel features with many dimensions, including the scope of the purpose to address? What power level of automation? What portion level of the course from not formal requirements to ML is mechanized? Purpose of what portions of the system lifecycle is addressed? Knowledge of what types of know-how information is used by the machine?

AI capabilities that have been tested to be serving a purpose in SE research and implementation can be referred to as 'Probabilistic SE', 'Classification, Learning, and Prediction for SE', and 'Search-Base SE.' Figure 1 Showing AI and its sub-field.

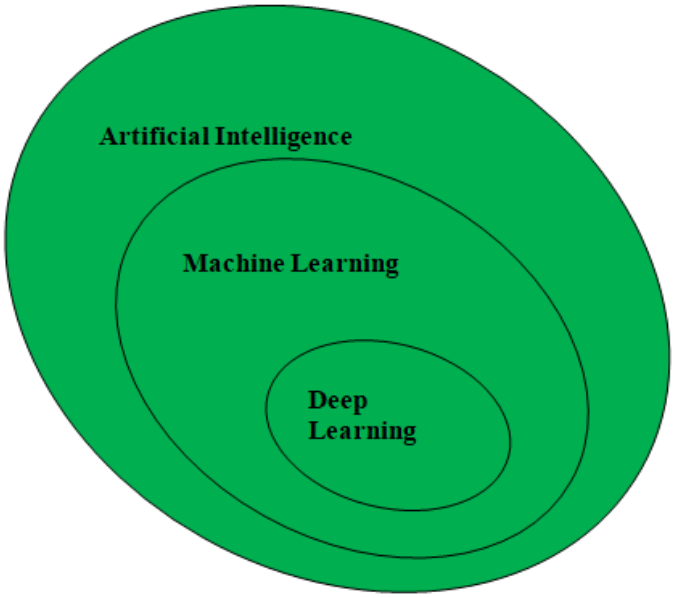


Figure. 1. adopted from [4]. Showing AI and its intra areas

2.1. Why AI Needs Research From SE?

In order to standardize AI, we need research from SE to systematically support the production of AI systems and applications, the operations in the real ecosystem, together with examining, maintaining, and modifying to increase more features and for future enhancement. And, to the discovery of the best among the better alternatives in terms of scientific & systematic approaches and methods so that we can have conventional models and architectures for AI application systems.

SE domain can enable the description of AI application methods and can serve as an information dictionary to all elements used in the development process [7]. These can make an easy way in the future to have a universal software theory that could support the development of AI or other technology so that it will give room for the legal framework in case if there is for breach of contract (or warranty) between customers and manufacturers/developers.

3. SOFTWARE ENGINEERING

Software Engineering it the systematic approach, application of mathematical technique, and the scientific method for developing, operating, maintaining, enhancing, and retiring software at an affordable cost to the customer. The discipline was introduced in the 1960s by NATO following the 'software crises.' The study found out that as a result of innovations of microchips transistors and integrated circuits that led to the affordability of personal computers to individuals, which in turn led to high demand for software. Also, there was inadequate standard to software development, and programmers at that time thought that '*After we develop the software and set it to operate, our task is finish' or 'the solely deployable function service for an achievement software project is the functioning software.'* While in real existence, the functioning software is merely one segment of a software arrangement or setup that brings many components. A specific type of service products (e.g., models, documents, plans) support for building successful engineering, more substantial, and control for software support.

[2] Cited that 'A software philosopher once stated that earlier beginning to developing software code, then, the extension of the period to complete it done. Also, research from industry information discovered that from 60 to 80 percentages of all tasks expended on software expect to be spent after it is deployed to the client for the first time' [2]. The NATO discussion at Garmisch-Partenkirchen, Germany, the Major aim is to have an efficient and effective production way of high-standard and mostly huge software systems. The objective is to help and encourage software engineers and managers to develop credible software in a quicker way with tools, methods, discipline, and processes by adopting engineering techniques such as:

- **Tools**. Which can be automated or semi-automated enablers for software processes and methods.
- **Method**. The procedural technical 'know-how' for developing software.
- **Process**. The 'glue' between that technological layers together and support rational and development to the software.

Therefore, SE is the standard development of software not theory. There are many other arguments for the use of theories. Moreover, 'theories provide common frameworks that allow the structuring of knowledge in a precise and concise path, that facilitates the communication of ideas and knowledge independent space and time. 'Nonetheless, the usefulness of theories for software development is at this moment a discussion issue and the current use of theories in this discipline is not well known' [1]. Even thou Simon presents that 'more and more, computers will program themselves' during a time of the great expectation for AI, this expectation did not come to fruition at that time [11].

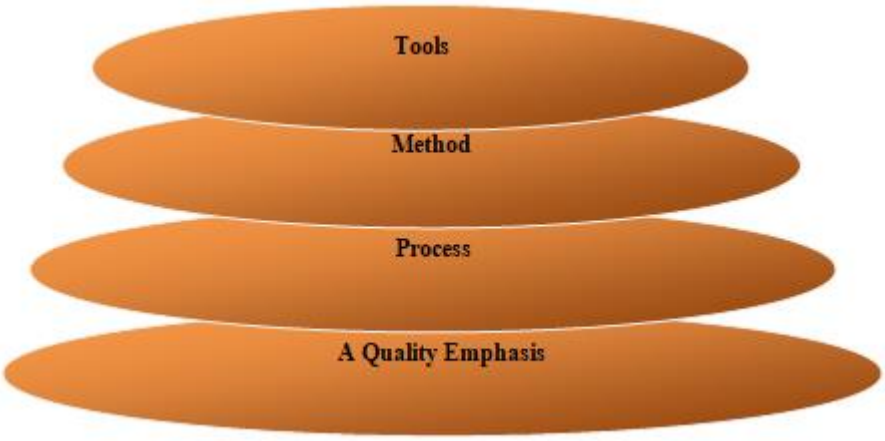


Figure 2. Showing software engineering standards structured [2]

3.1. Why SE need Research from AI?

Why SE aimed at research from AI? As we are growing into the internet-of-things (IoT), AI can make the inclusion of those technologies more productive as well as to serve as an ecosystem for testing software and models. This research can enhance the civic-legal framework to raise up user acceptance. SE methods may be re-engineered using AI concepts, which can lead to perceivably robust technology. And, can help in the preparation of a data dictionary for describing the meaning of the respective SE methods and tools. In annex, other new AI techniques that are extending or trending in a virtual environment can be applied for a better variation of the SE approaches [7].

4. CROSS-SECTION BETWEEN SE AND AI

As the cross-section of AI and SE is presently uncommon, but increasingly are getting higher. Nowadays, pathways and methods from both major encourage research and implementation relative to each other research area. Figure 3 below from [5] represent some research fields in AI and SE sector as well as their cross-section.

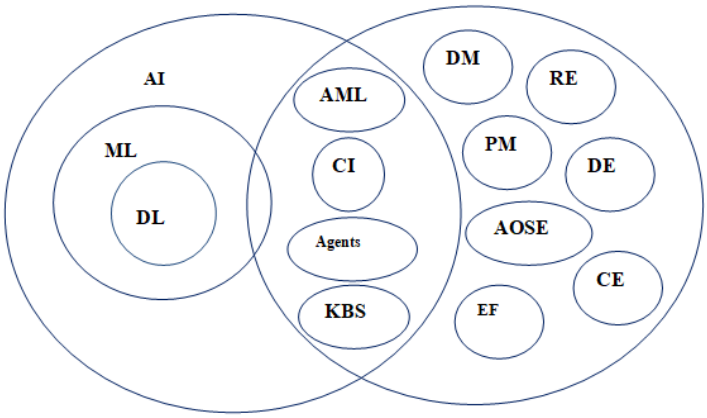


Figure 3. Showing research fields in AI and SE and their inter-sections [5].

AI system or application is similar to many programs and applications; merely more, therefore, the building of AI software can be figured as a progressive form of ideal maintenance often needed to transform specification. Traditional SE procedure produce both cohesion and adhesion, function unwell in such frequent specification changing condition, so we need to create more SE possibilities techniques that can enable rapid application development and evolution [7].

We apply the concept of SE to the development of production project management (P-PM), creation of requirement engineering (RE), designing engineering (DE), software source code engineering (CE), and method models to enable the development of an intelligent system as well as using advanced data analysis. KA advancement can simplify to makes EF and reasoning atmosphere system like DM advancement which enables the development of models for generating the requirement for system programs and services product lines [5].

Between CBR and EF, the former is applied for accessing an information to the latter. SE applied the concept of information agent to model (or replicate) production processes or to share and show substitute requests [5].

The sophistication of AI is getting more value, that is worth vital to examine how it is of assistance to SE while preventing possible drawbacks. AI development which can lead to a high sum of preferences, reasoning, and autonomy path to crucial law and ethical issues whose solution affects both experts and customers of AI technology [6]. These issues cline between civil code of conduct, rules of law, public matters, the practitioner ethics, philosophical policy, and the likes that will require intelligentsias from AI scientists (or Engineers), software engineers, legal practitioners, political analysts, and advocators to answer [6].

Even nowadays a customer is limited to sue for any damage to breach-of-contract in commercial-off-the-shelf software, in fact, most cases end up discouraging the customer due to inadequate professional standards in the industry. So how significant are we trying to ensure those sets of principles are put in place for customers to have confidence in the system for AI?

Vladeck brought an example in the year 2014, that, if the autonomous driving vehicle is deployed in a country like the US to reduce approximately 40,000 yearly traffic inevitabilities and crashes in one of two equal figures, then, the manufacturers could get more than 20,000 damages lawsuits instead of thousand feedback messages of appreciation notes from the customers. But the question is 'in what legal regulation or legislation capacity can guarantee the certainty of the autonomous self-driving car or drone flying vehicle to be best possibly achieved?' [6].

Because if not so, that can be repudiated due to silence by the eyes of the law of the customer right. Therefore, shall the legal issues concerning AI to be managed or regulated by the current software & internet cyberlaw, or should they be managed differently? (Calo 2014b) cited in [6]. Differently in the sense that, should we enact new laws and create a new regulator for it? But with the profession of SE, enacting the law and creating the code of conduct bureau can be easier especially in the future if we are able to have global software system theory.

Some researchers proposed that society will ignore autonomously technologies except there are some trustworthy ways of propelling them to wherewithal safe. SE for AI can help to ascertain the indefinite areas in which an AI may inadequately act as desired equivalent to various fields of robustness research. An example of such SE concept to successful design and implementation of software is verification and validation.

4.1. Verification

Verification is the process of confirming in order to substantiate that system is able to act what it formally requested to perform. Did we develop the system correctly? Verification measure mean approaches that assure a highly reliable set of formal conditions that AI software must certify to give room for a legal framework and the likes. If so, then, the persuasion for AI and other automation technologies to scale critical safety measures and to reach satisfactory state-of-affairs so that they can be acceptable to the society will be victorious.

4.1.1. Existing Standard on Verification: Its Relevance to AI System

Previously, the standard verification of software accelerated significantly. A notable example of that is the work of Klein and his colleagues in the year 2009, on a fully common purpose operating kernel named as SeL4kernel. Klein and his colleague mathematically examined the device base on the standardized specification to assert an assurance against dangerous operations, mortality, and crashes. The verification technique and approach methods are to set up high guarantee software tools (Fisher 2012) [6]. With the help of SE, we should not just only develop an AI system that can be verified on the underlying layer, but rather we should also verify the design of the AI system especially, if the designs follow a componentized architecture, in such a way it guarantees that every single part-contained component can be integrated according to the functions and connections to reflect the attributes of the main system [6].

4.1.2. Verification Difference Between SE and AI

Research shows that approximately when developing a technological device, the noticeable difference between SE and AI in terms of verification is the outcome of SE which is software is implemented with the thoughtfulness of a known mathematical and machine model, while in contrast to the AI systems particularly robot that function in the ecosystem that is relatively known by the system experts. In today's turbulent environment, SE can come up with a model to give AI full support for universal acceptance due to convention verification.

4.2. Validation

The is the state of being authentic. Validation is to measure the system to see how it corresponds to its formal requirement by not having unacceptable features with repercussions. Did we develop the correct system? Validity measures: is to affirm and ground unwanted attributes that may exist/happen regardless of the system is formally proper. In the future terms, research shows that AI systems influentially can have more capabilities and self-control, in such instance unjustifiable validity could progress to huge damages. Healthy and wealthy assurance for machine-learning methods, a field 'represented as for not-long-term validity research, will also be significant for long-term validity for safety' [6].

4.3. Other Requirement and Support for Creation of Legal Framework for AI

SE verification and validation can only allow for the development of AI whether autonomously (or not) to an absolute correctness level. But in case of if there could be 'hidden' unrecovered 'defect' that can only be detected when the customer has almost spent more than a month using it. That defect could cause economic damages, crashes, and/or fatalities. How can a customer seek restitution of such damages? Vendors of COTS software is only with 'hard to convince' to be liable in the view of the law for fixing any fault, defects, and failure no matter how substantial the damage is as far as COTS license and law are concerned. Even if the vendor can be proven of the negligence of duty in the development process customer can only be difficult to claim his/her right because of a lack of a strong professional body that certifies the software engineering profession.

Also, lack of professional government agencies in some countries that regulate and standardize software related issues. But had it been it is in the building sector where there are building and construction regulators that regulate and develop the necessary standards governing all the construction affairs. They make sure the building meets the required features and standards while building practitioners must be certified by the building professional organ. The building professional body evaluates the civil engineers, architectures, quantity surveyors, regional planners to list a few, in order to make sure that individual does not claim to be knowledgeable about the field but has acquired the technical knowledge of know-how about the field. So, for a vendor or developer of an AI to be able to be responsible for breach of warranty or contract then there is a need for such a system of the construction sector to be initiated in AI or ICT in general.
We need to have a universal SE and AI professional body that certifies who wants to pursue a career in the sector so that we can have a chartered software engineer or chartered artificial intelligence engineer and so on. To have an SE and AI regulatory commission that will regulate the activities of the technology, and also, to have legislation or law that can protect the interest of both customers and engineers. For the law to be acceptable in society there is a need for collaboration from the legislators, advocators, lawyers, philosophers, political analysts, psychologists, media practitioners, academicians, innovators, entrepreneurs, engineers, scientists, and the likes.

5. RECOMMENDATIONS

The study encourages that the future of AI can be ascertained with the help of SE as SE already relevantly benefitted from AI. Let make an example of how the world came to accomplish automobile car into existence in the society in the US. The revolution of the automobile car was a success to the individuals, economics, and change to countries landscape at large. But it also has it on ups-and-down at its inception stage.

Around the first decades of the twentieth century, the automobile was recognized with the rise of fatal and accidental death and injuries that suggested an expression of great concern. A community to dialogize the issues was inaugurated which involves safety advocates, engineers, physicians, media practitioners, and others related to contribute their various opinion about the causes of the fatalities, injuries, and accidents. They revealed up that what causes these tragedies was from the driving ethics, design of automobiles, to environmental hazards and highway road engineering.

So, to minimize those tragic consequences and to retain the precedence of automobile effort began from regulating driver character, redesigning of the automobile, to the reformation of highway road and environment. The automobile took almost or more than a decade to recognize, prioritize, and regulate these risk agents.

Thus, this study recommends a similar approach to avoid such tragedies and gain public confidence in adopting AI in society. As humankind, we should try to get the possible best of AI before deploying it to our ecosystem. AI should or must be at least 99.98% accurate that can give assurance to the people. All stakeholders from the industry, academia, government, research agencies, and other related should able to answer questions like 'can this AI (like an autonomous vehicle) be adopted in any environment whether rural or urban?' 'Does it require more sophistication knowledge from the user be he/she can use it?' Or 'anybody with or without literacy can use it? 'What navigation system does it use? Is it a radioactive signal? If it is a radioactive signal then what will happen to the product supposing if it is limited or unavailable?
As the comprehensive testing of AI sometimes is complex and difficult to put into practice or time-consuming application of automation using simulation can be helpful. One vital aspect of SE that can help to accomplish quality assurance to provide guaranteed AI products to the public is verification and validation (V&V). Once if the stakeholders can ascertain AI quality assurance according to the ecosystem and public character or behavior then the deployment of AI can begin getting people's attention and confidence.

Although these days there has been an increase in research grants on software theory which can give of a phenomenal guarantee of software life spam maybe even on Autonomous AI i.e. if the research reach its peak. The study hopes so for the impact of software theory especially on AI can reduced some resources spending and give a platform for universal AI where every citizen can have a right from confidence level to a legal aspect in case if he or she is not satisfied.

6. CONCLUSIONS

As the technique of AI to SE makes an impact achievement to SE, then we need to work more on SE for AI. Especially the concept of verification and validation from the techniques of SE can make the autonomous vehicle easy to implement to the market as it can give AI quality assurance to the public. As people expect the ideal factor AI to be a perfect autonomous vehicle that varied from what it is in reality the study cross-check between AI and SE and find out the possible feature of each segment that can be solution to boost each other.

For us to be able to create an autonomous AI that can work robustly accurate, then all AI communities must of course need to come up with good attributes that can be applied to every domain. And the standard issues need to also find the solution to questions like what SE formal abilities are currently available? How reliable are they? What kind of research should undertake to find more abilities and concept? And what concept can we interchange such that all fields of computer science and broader AI expertise can be made useful worthy? To exemplify this is the work of Wallach and Allen (2008) that concern for substantial relevant factor is the logical value of a variety set of formal standards (or ethical code of conduct): we may need to apply little

worth of approximations if the formal standard cannot be organized well enough to influence behaviour in the safety-critical state of affairs.

'There is nothing so practical as a good theory' [1]. theories explain one specific aspect of reality, according to what is known from it up to the present time, and they enable predicting future events [1]. Today's software architecture is largely described as 'emulate to adopt', so there is no accepted software architecture widely, not to talk of software architecture for AI. So, researchers need to do more in between the SE and AI to reduce AI development time, effort, and cost. Also, to have an acceptance SE model to AI in such a way that is not just only SE standard to AI, but conventional software theory and software architecture that can standardize the development of AI and other technologies.

ACKNOWLEDGEMENTS

We will like to thank the Almighty God for this priceless opportunity called life, because with life then there is hope contribute to human progress through academics.

REFERENCES

[1] Andrade, J. Ares, García, R. Martínez, M. Rodríguez, S. & Suárez, S. (2011). Recent Advances in Signal Processing, Computational Geometry and Systems Theory. Retrieved from https://www.researchgate.net/publication/262212304_About_theory_in_software_development? On May, 5th 2020

[2] Pressman, R.S. Software Engineering A Practitioner's Approach. USA: The Mcgraw-Hill, 2010

[3] Partridge D. Artificial Intelligence and Software Engineering Understanding the promise of the future Routledge Taylor and Francis group, 1998 https://books.google.com.tr/books?id=ncBmAgAAQBAJ&printsec=frontcover#v=onepage&q&f=false

[4] F. Chollet, Deep Learning with Python. USA: Manning Publications, 2018

[5] Jörg, R. Klaus-Dieter, A. (2004). Retrieved from https://www.researchgate.net/publication/220633840_Artificial_Intelligence_and_Software_Engineering_Status_and_Future_Trends on April, 27th 2020

[6] Russell, S. Dewey, D. Tegmark, M. (2005). Research Priorities for Robust and Beneficial Artificial Intelligence, Copyright © 2015, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602 Retrieve from https://futureoflife.org/data/documents/research_priorities.pdf on April, 24th 2020

[7] Mostow J. (1985). What is AI? And What Does It Have to Do with Software Engineering? IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. SE-li, NO. 11, NOVEMBER 1985. Retrieve from https://s3.amazonaws.com/ieeecs.cdn.csdl.content/trans/ts/1985/11/01701944.pdf? on May, 2th 2020

[8] Retrieved from https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e on May, 11th 2020

[9] Hany, H. A. Walid, Abdelmoez. & Mohamed S. H. (2012) Software Engineering Using Artificial Intelligence Techniques: Current State and Open Problems. ICCIT 2012 Retrieve from https://d1wqtxts1xzle7.cloudfront.net/30814527/p24-ammar.pdf?1362360177=&response-content-disposition=inline%3B+filename%3DSoftware_Engineering_Using_Artificial_In.pdf on May, 2nd 2020

[10] Mark H. (2012). The Role of Artificial Intelligence in Software Engineering. 2012 IEEE, RAISE Zurich, Switzerland. 978-1-4673-1753-5/12/$31.00 c? Retrieve from https://www.computer.org/csdl/pds/api/csdl/proceedings/download-article/ on May, 4th 2021

[11] Sunil, M. Thomas, K. & Jonathan, W. (2018) Artificial Intelligence and IT Professionals Published by the IEEE Computer Society 1520-9202/18/$33.00©2018 Retrieve from https://s3.amazonaws.com/ieeecs.cdn.csdl.content/mags/it/2018/05/mit2018050006.pdf? On May, 4th 2020

[12] K. Goertzel, “Legal Liability for Bad Software” Research Gate, USA. Accessed on April, 2nd 2021 https://www.researchgate.net/publication/309429972_KM_Goertzel_Legal_Liability_for_Bad_Software_CrossTalk_Vol_29_No_5_2016-0910

[13] Nees, M. (2016). Acceptance of Self-driving Cars: An Examination of Idealized versus Realistic Portrayals with a Self- driving Car Acceptance Scale. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 60. 1449-1453. 10.1177/1541931213601332.Retrieved from https://www.researchgate.net/publication/308179113_Acceptance_of_Self-driving_Cars_An_Examination_of_Idealized_versus_Realistic_Portrayals_with_a_Self-_driving_Car_Acceptance_Scale on June, 12th 2021

[14] National Museum of American History. Retrieved from https://americanhistory.si.edu/america-on-the-move/essays/automobile-safety on June 20th 2021

[15] T. Mustafa and A. Varol, "Review of the Internet of Things for Healthcare Monitoring," 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 2020, pp. 1-6, doi: 10.1109/ISDFS49300.2020.9116305.

AUTHORS

Sabeer Saeed is, currently, a Master's student in the department of Software Engineering at Firat University Elazig, Turkey. Before his coming to Firat University, He obtained International Advance (and Diploma) Diploma in computing from Informatics Academy. He also obtained B.Tech. in Management Information Technology (MIT) at Abubakar Tafawa Balewa University (ATBU), Bauchi, Bauchi Nigeria. Sabeer Saeed is a winner of National Information Technology Development Agency (NITDA)’s study scholarship to Firat University, Turkey where is studying Master’s in Software Engineering. He is a humanitarian par excellence, especially, on technology and gender diversity. The same humanitarian passion drove made him to establish an ICT Centre for School for Continuing Education (Women) at the Zamfara State, Nigeria during his one-year mandatory National Youth Service.

FUTURE SALES ESTIMATION USING PATENTS

Koichi Kamijo

Department of Information Technology, International Professional University of Technology in Tokyo, Shinjuku-ku, Tokyo, Japan

ABSTRACT

We propose a model to improve estimation accuracy of the future sales volume, focusing on pharmaceutical products, from their patents. Our approach is based on an analysis of patents obtained in the early development stages of the products. The development of pharmaceuticals often takes a long time (up to several decades in some cases), and the costs are huge, even exceeding one billion USD for just one product. Therefore, it is strongly desirable to estimate future sales volume at an early stage. One piece of information potentially useful for the estimation is the brand, i.e., the name of the developing company. Our model learns the sales volume and words used in multiple patent specifications and also focuses on the extent to which "seasonal" words are used. Experiments showed that our model much improved the accurately of the sales volume estimation compared with the case of just estimating from its brand name.

KEYWORDS

Sales Estimation, Pharmaceuticals, Patents, Natural Language Processing, Deep Learning.

1. INTRODUCTION

As COVID-19 vaccines are being announced by several pharmaceutical companies, the world's attention is currently focused on vaccines. These vaccines are being sold worldwide, and their sales volume is bound to be huge. For example, Pfizer expects robust COVID-19 vaccine demand in the current year and estimates a sales volume of 26 billion USD [1].

In a case such as the COVID-19 pandemic, the demand for vaccines is huge, spanning almost the entire global population, and this demand is bound to result in huge profits for pharmaceutical companies. However, excluding such special cases, estimating the sales volume of a new pharmaceutical in its development stage is not easy because we cannot accurately predict how demand will evolve.

Estimating the sales volume of new products or services in the early stages of development is very important when formulating a marketing strategy, especially for pharmaceuticals, as pharmaceuticals have a longer development period than other products and many of them have the potential to make huge profits.

Currently, estimating pharmaceutical sales volume requires knowledge of the disease and pharmaceutical market, a comprehensive understanding of the domestic and global regulatory environment, and market access. Therefore, this task is often performed by a company with appropriate expertise, such as Clarivate [2]. However, outsourcing such estimations is typically very expensive and requires pharmaceutical companies to share sensitive information with the contracting company. Accordingly, it would be desirable to make sales volume forecasts in-house without such exposure.

David C. Wyld et al. (Eds): CSE, AI & FL, NLPTT - 2021
pp. 45-58, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112404

When a company develops a new product or service, it usually applies for a patent before introducing it in the market, acquires the rights, and then starts full-fledged development and trading.

In this paper, we present a model for estimating the future sales volume of new pharmaceuticals by using specifications available in the initially acquired patents for each pharmaceutical.

Our model considers only pharmaceuticals whose prior sales volume and first patent are available. We performed morphological analysis of each patent specification provided to the patent office in the early development stage and counted usage ratio of each word in each patent. The usage ratio of each word and sales volume were used as training/test data for the model. We evaluated the model by leave-one-out (LOO) cross-validation; that is, out of *n* data, we used the *n*-1 data as training data and estimated the rest as test data. We repeated this *n* times and evaluated the sales estimation performance by averaging the *n* estimation results.

For developing the model, we also performed morphological analysis of articles related to pharmaceuticals, and the word usage ratio discussed above were weighted based on the usage of words contained in the articles.

For patent specifications, to ensure a unified patent format, we used only those patents that complied with the Patent Cooperation Treaty (PCT) [3]. We also targeted English texts only.

Section 2 of this paper presents related work, and Section 3 introduces the proposed pharmaceutical sales estimation model. Section 4 details the experiment, which are then discussed in Section 5. We conclude in Section 6 with a brief summary and mention of future work.

2. RELATED WORK

Since we could not find any research that directly discusses the relationship between sales estimation and patents, we examined studies on sales estimation and patent analysis.

2.1. Sales Estimation

Merino et al. proposed the combination of a spatial interaction model and simulation approaches for the reliable estimation of retail interactions and store sales volume on the basis of data on consumer shopping behavior in Mexico [4]. Their proposed methodology was based on the combination of a Huff model [5] and a Monte Carlo simulation [6] to reproduce shopping patterns in retail stores. Jordan et al. investigated how to improve the estimation accuracy of a firm's sales volume [7] and emphasized that rather than customer satisfaction, return on investment, or economic value, an evaluation of the quality of the firms' planning practices is the most important. Pavlyshenko et al. used machine learning for predicting sales volume and found that the use of stacking techniques could improve the performance of predictive models used for sales volume time series forecasting [8]. They noted that the use of regression approaches for sales volume forecasting could often provide better results than time series methods. Loureiro et al. investigated the use of a deep learning approach to forecasting sales volume in the fashion industry, namely, for predicting the sales volume of new individual products in future seasons, without the use of historical data [9]. They developed forecasting models by considering a wide and diverse set of variables (e.g., products' physical characteristics and the opinion of domain experts) and were able to perform highly accurate forecasting. They also found that deep neural networking outperformed other techniques such as random forest.

Note that none of the methods mentioned above are based on a patent or articles referring to the products.

2.2. Patent Analysis

Kim et al. analyzed patents to identify emerging and vacant technology areas of wireless power transfer. They extracted topic areas from patents by text mining, where topics with similar semantics were grouped together, and then applied a time series analysis and innovation cycle of technology to the grouping result [10]. The results of the clustering, time series analysis, and innovation cycle were then compared to minimize the possibility of misidentifying emerging and vacant technology areas. Guderian et al. investigated how innovation management decisions in times of crisis (e.g., the COVID-19 pandemic) could be improved through publicly available data, such as patents [11], and examined which data were valuable from the viewpoint of patent citation. Lee et al. proposed a forecasting model for new innovative product diffusion based on both technology diffusion and interest diffusion. Technology diffusion was defined on the basis of the number of patent citations, while interest diffusion was defined on the basis of web search traffic [12]. They used the model to predict the sales volume of hybrid cars and industrial robots in the US market and found that its prediction performance was better than that of the Bass model [13] and the Bass model with patent citation for both cases.

While all of the above works discuss how patent analysis can contribute to the forecasting of future business and technology trends, none of them focus on actual sales volume values and none of them use deep learning for analysis.

In a paper related to patent analysis, Suzuki et al. proposed an approach to automatically extract keywords related to novelties or inventive steps from patent claims by using the structure of the claims [14]. Hido et al. addressed the problem of assessing the quality of patent specifications on the basis of machine learning and text mining techniques. They computed a score called patentability, which indicates the likelihood of an application being approved by the patent office [15], and employed a new statistical prediction model to estimate examination results (approval or rejection) on the basis of a large dataset including 0.3 million patent specifications. While these two papers do not directly relate to sales estimation with patents, they do provide tips for analysing patents.

3. METHODOLOGY

Our objective was to construct a model that could estimate the sales volume of new pharmaceuticals on the basis of not only the names of the development companies but also patents and articles related to the pharmaceuticals. Figure1 shows our research framework. (A) through (D) below correspond to those in the list after the sentence “In total” in the next page. We obtained the pharmaceutical list and sales volume from a database of Cortellis [16] ($1^{\dagger}$and $2^{\dagger}$ in the figure). We first estimated the sales volume by using the information about the pharmaceutical, specifically, the name of the company that developed the pharmaceutical and the year in which the first patent application for the pharmaceutical was made (A). For each pharmaceutical, we collected the first patent application for the pharmaceutical that complied with the PCT and that was written in English ($3^{\dagger}$), and estimated the sales volume using the information about the patents (B). For the first patent, we used the information provided by Cortellis and Derwent [16,17], whose employees include experts on pharmaceuticals and patents. We then performed morphological analysis for each patent ($4^{\dagger}$) and estimated the sales volume

from the usage ratio of each word, that is, the number of appearances of each word divided by the number of appearances of all the words in each patent specification (C). Next, we collected articles related to the pharmaceuticals (5†), performed morphological analysis on them (6†), and calculated the

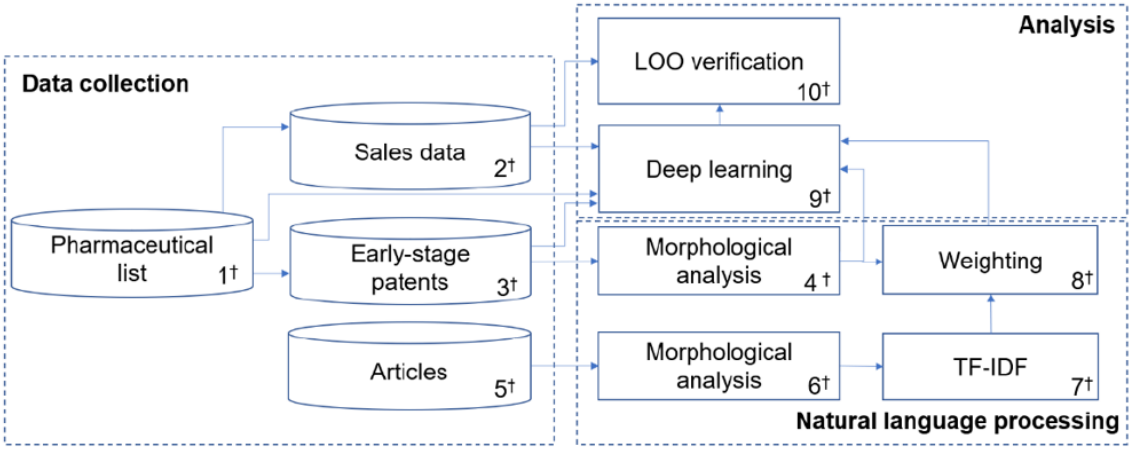


Figure 1. Research framework.

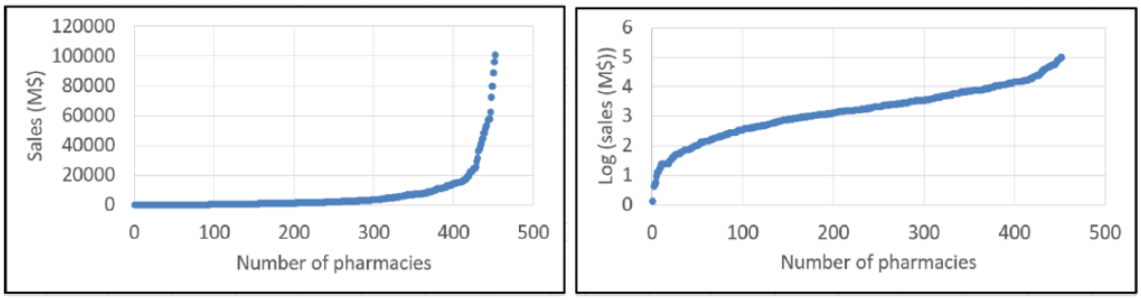


Figure 2. Plot ofthe sales volume of pharmaceuticals used in the experiment (in millions of USD, 2019 to 2027, actual sales volume + estimation by experts), smallest first (left: real; right: logarithm).

Term Frequency - Inverse Document Frequency (TF-IDF) of each word in each year (7†). We then calculated the sum of the usage ratio of each word weighted by the TF-IDF each year (8†), and estimated the sales volume (D). In our model, for sales estimation, we used deep learning (9†), and for the evaluation of the sales estimation, we used LOO cross-validation (10†).

In total, we performed the following estimations.

A. sales estimation from information about the pharmaceutical
B. sales estimation from information about the first patent of the pharmaceutical
C. sales estimation from the words used in the first patent of the pharmaceutical
D. sales estimation from C plus pharmaceutical related articles
E. sales estimation by combining A–D

For the model construction, for the sales volume s_i of pharmaceutical d_i, we used log(s_i)(base=e) instead of s_i, since the range of pharmaceutical sales volume can vary widely. Figure2 shows a plot of the sales volume of pharmaceuticals used in the experiment (in millions of USD, 2019 to 2027, actual sales volume + estimation by Clarivate [2]), smallest first. The left-side figure shows

the real data and the right-side figure shows the logarithm of the real data. Clearly, logged sales volume are well distributed and linear, and it is expected that we can make a more accurate estimation compared with the case of the real data.

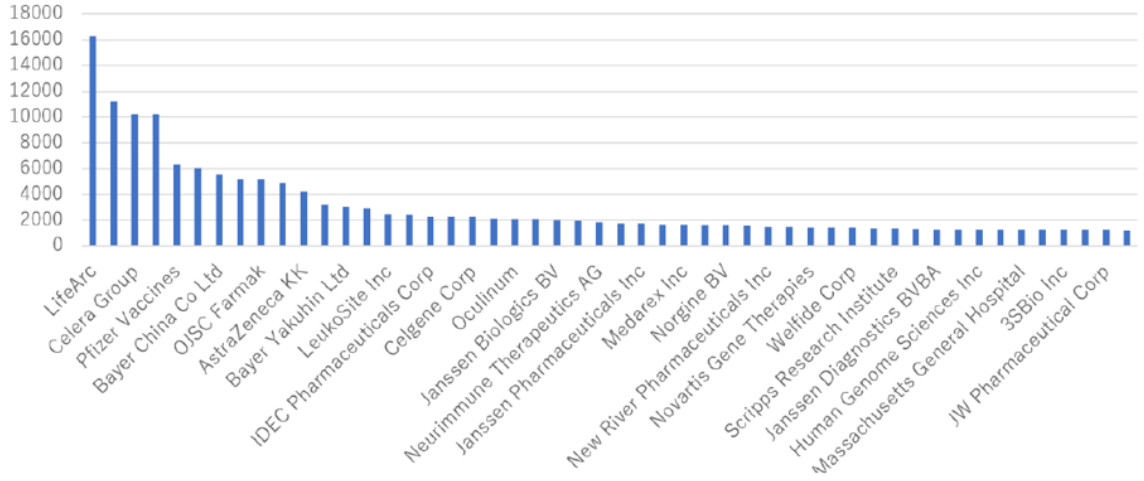


Figure 3. Sales volume values for different developing companies for the year 2019 (in millions of USD).

In this study, the sales volume we used was the sum of actual sales volume in 2019 plus expert-. estimated sales volume from 2020 to 2027.

In other words, the values to be estimated include the future sales volume value estimated by experts. This is because pharmaceuticals that were developed just before or after 2019 do not have enough sales achievement data. From the viewpoint of sales estimation research, estimating these values still has worth for research.

The following subsections describe the details of each estimation.

3.1. Sales Estimation from Information about the Pharmaceutical

In this step, we estimate sales volume by utilizing information on each pharmaceutical. This information includes the name of the company that developed the pharmaceutical and the year in which a patent application was made in the early development stage. Figure3 shows the sales volume values for each pharmaceuticals developing company for the year 2019[2]. Since the sales volume values are different for different companies, by knowing the developing company, we can roughly estimate the future sales volume of each pharmaceutical.

We could also use additional information, such as the name of the selling companies, but these companies may not have been decided at the time the pharmaceutical was developed. Therefore, we used only the developing company and the year of the first patent application.

For sales estimation from information about the pharmaceutical d_i, we used a one-hot vector $\boldsymbol{v}_i[j]$ for the developing company, defined as

$$\boldsymbol{v}_i[j] = \begin{cases} 1, & j = f_c(d_i), \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $f_c(d_i)$ is the index of a company that developed d_i. We used the year in which the first patent for the pharmaceutical d_i was applied (yp_i) as part of the training/test data. We defined a ($|\boldsymbol{v}_i|$+1) dimensional vector $\boldsymbol{x}_i^{[1]}$ as

$$\boldsymbol{x}_i^{[1]} = [v_i, yp_i]^T, \tag{2}$$

where ^{T} is a transpose.

3.2. Sales Estimation from Information about the First Patent of the Pharmaceutical

Patent specifications usually consist of a title, an abstract, claims, a detailed explanation, and so on. We analyzed the correlation between these components and sales volume and examined the sales estimation obtained with these parameters. We used the following information from the first patent of pharmaceutical d_i for sales estimation.

- pr_{1i}: Word count of title
- pr_{2i}: Word count of abstract
- pr_{3i}: Word count of the document
- pr_{4i}: Number of claims

We defined a vector $\boldsymbol{x}_i^{[2]}$ for a pharmaceutical d_i as

$$\boldsymbol{x}_i^{[2]} = [pr_{1i}, pr_{2i}, pr_{3i}, pr_{4i}]^T. \tag{3}$$

3.3. Sales Estimation from the Words used in the First Patent of the Pharmaceutical

We analyzed the correlation between words used in each patent and the sales volume of each pharmaceutical. For each patent specification a_i for a pharmaceutical d_i, we first performed morphological analysis for all the words in the patent and counted the usage ratio of each word, u_{wi}, where $\sum_w u_{wi} = 1$. We excluded stop words, numbers, and symbols in the calculation of the usage ratio, but included all other words, regardless of the part of speech to which they belonged. We defined a vector $\boldsymbol{u}_w$ for each word w as

$$\boldsymbol{u}_w = [u_{w1}, .., u_{wn}]^T. \tag{4}$$

We calculated Pearson's r-value, r_w, between $\boldsymbol{u}_w$ and $\boldsymbol{ls} = [\log(s_1), .., \log(s_n)]^T$, and then selected a set of words $\Omega(Tr, Tp)$ that satisfied the following:

$$\Omega(Tr, Tp) = \{w | |r_w| \geq Tr, p_w \leq Tp\}, \tag{5}$$

where T_r and T_p are thresholds and p_w is the p-value for r_w. Table 1 shows an example of $\Omega(0.1, 0.01)$ sorted by $|r_w|$, which are used in the experiment. Bold words in the table are related to a pharmaceutical or a disease, and "ratio" is the ratio of patents that used at least once out of 423 patents for each word.The table shows that 36% of the words were related to a pharmaceutical or a disease.

Finally, for each pharmaceutical d_i, we used the usage ratio of the words in Ω, as

$$\boldsymbol{x}_{\boldsymbol{i}}^{[3]} = [u_{w_1 i}, .., u_{w_{|\Omega|} i}]^T, \quad (6)$$

where $\Omega = \bigcup_j w_j$.

3.4. Sales Estimation from C plus Pharmaceutical Related Articles

If a patent includes "hot" keywords that reflect the high popularity of the pharmaceutical at the time an application is made for the patent, the sales volume of the pharmaceutical is likely to grow in the future. With this in mind, we weighted the word usage ratio in the patent specification based on the usage of each word in pharmaceutical related articles.

For the analysis, we first collected pharmaceutical-related articles published in year *y* and then calculated the TF-IDF of the word *w* in year *y*, tfidf(*w*,*y*), defined as

Table 1. Sample of words with |*r*|-values in 432 patents used in the experiment, and ratio of patents that used at least once out of 423 patents for each word.

word	r-value	ratio	word	r-value	ratio
coval	0.567	0.124	**epiderm**	−0.277	0.106
plasma	0.49	0.108	**cag**	0.274	0.106
tag	0.483	0.177	**cynomolgu**	−0.273	0.104
unmodi	0.482	0.119	herebi	0.27	0.195
fraction	0.427	0.261	**best**	−0.268	0.128
test	0.334	0.179	qiagen	−0.266	0.106
mainten	0.328	0.146	draw	−0.265	0.131
energi	−0.321	0.108	**precipit**	−0.262	0.122
stock	0.321	0.104	inhibitor	0.259	0.383
obtain	0.308	0.139	wherea	−0.256	0.155
lupu	−0.307	0.102	posit	0.254	0.575
copi	−0.304	0.104	examin	−0.253	0.102
germlin	0.303	0.128	complement	−0.247	0.104
second	−0.301	0.108	transform	−0.242	0.102
ctg	0.297	0.102	underlin	−0.241	0.113
dmso	0.292	0.157	prepar	−0.241	0.144
transplant	−0.287	0.111	**coloni**	−0.241	0.148
delet	0.285	0.175	**microtit**	−0.24	0.128
amino	0.283	0.126	separ	0.236	0.361
vegf	0.28	0.1	**substrat**	0.235	0.106
non-limit	0.279	0.215	principl	−0.234	0.113
isoleucin	0.279	0.117	alter	−0.234	0.162
immobil	0.278	0.177	altern	0.232	0.648
lymphoma	0.277	0.197	**polynucleotid**	−0.232	0.223
prolong	−0.277	0.10	combin	−0.231	0.142

$$\text{tfidf}(w, y) = \text{tf}(w, y)\log(\text{idf}(w)), \quad (7)$$

where tf(*w*,*y*) represents the frequency of the word *w* in articles published in year *y* and idf(*w*) denotes the inverted frequency of *w* among all of the articles for all of the years. The frequency is normalized so that we have $\sum_w \text{tf}(w, y) = 1$ for all of *y* s.

For each pharmaceutical d_i, we calculated the sum of the word usage ratio in the patent specification a_i, that is, u_{wi}, weighted by tfidf(*w*,*y*) for each year *y*, as

$$ut(y, i) = \sum_w \text{tfidf}(w, y) u_{wi.} \quad (8)$$

It is possible to increase the *ut* weighting ratio in the model, whose year of first patent specification application (yp_i) is close to the nearest article publication year. In deep learning, such a weighting ratio is automatically adjusted. From this viewpoint, we evaluated the case of

putting elements of *ut*s so that the position of each element for the year the article is published minus the year the first patent is applied (yp_i) is the same for all *i*s. Towards this, we created a new vector, $\boldsymbol{x}_i^{[4]}$, by padding 0s to the left and/or right, as

$$\boldsymbol{x}_i^{[4]} = [\mathbf{0}^{z1_i}, ut(ya_{min}, i), .., ut(ya_{max}, i), \mathbf{0}^{z2_i}]^T, \tag{9}$$

where $\mathbf{0}^j$ is a vector that consists of *j* 0s,

$z1_i = yp_{max} - yp_i, z2_i = yp_i - yp_{min}, yp_{min}$ and yp_{max} are the minimum and maximum value of yp_i, respectively, and ya_{min} and ya_{max} are the oldest and latest years of the articles for our analysis, respectively. No year is skipped between ya_{min} and ya_{max} in (9). For example, if yp_0=1996,yp_1=1999, ya_{min}=1998, ya_{max}=2020, yp_{min}=1980, and yp_{max}=2021, then,

$$\begin{aligned} \boldsymbol{x}_0^{[4]} &= [\mathbf{0}^{25}, ut(1998,0), .., ut(2020,0), \mathbf{0}^{16}]^T, \\ \boldsymbol{x}_1^{[4]} &= [\mathbf{0}^{22}, ut(1998,1), .., ut(2020,1), \mathbf{0}^{19}]^T. \end{aligned} \tag{10}$$

$\boldsymbol{x}_i^{[4]}$ included *ut*s with articles published after the first patent application. We can calculate such *ut*s some years after the first patent application year and publication year of the articles. However, immediately after the first patent application, we do not have articles published after that year, so we define another vector $\boldsymbol{x}_i^{[5]}$ that only includes *ut*s that use articles published in the year equal to or before the first patent application, as

$$\boldsymbol{x}_i^{[5]} = \begin{cases} [\mathbf{0}^{z1_i}, ut(ya_{min}, i), .., ut(yp_i, i), \mathbf{0}^{z3}]^T, & yp_i \geq yp_{min}, \\ [\mathbf{0}^{yp_{max} - yp_{min} + ya_{max} - ya_{min}}]^T, & \text{otherwise}, \end{cases} \tag{11}$$

where $z3 = ya_{max} - ya_{min}$.

3.5. Sales Estimation by Combining A–D

To obtain a more accurate estimation, we combined the training/test data defined in the previous subsections. We define

$$\boldsymbol{x}_i^{[a_1,..,a_m]} = [\boldsymbol{x}_i^{[a_1]T}, .., \boldsymbol{x}_i^{[a_m]T}]^T. \tag{12}$$

a data that combines $\boldsymbol{x}_i^{[a_1]}$ through $\boldsymbol{x}_i^{[a_m]}$, as $m \geq 1$.

4. EXPERIMENT

We evaluated the sales estimation performance of each of the methodologies discussed in the previous sections. In the experiment, we used *n*=432 pharmaceuticals whose sales volume (s_i) and first patent (a_i) (in PCT, written in English) were both available. Table 2 shows the notation used.

The left-side panel of Figure4 shows the number of companies that developed one or more pharmaceuticals; a total of 432 pharmaceuticals were considered in the experiment. Specifically, 186 companies (43%) developed only one pharmaceutical, and one company developed 13 pharmaceuticals, which was the highest number of pharmaceuticals developed by a company. The right-side panel of Figure4 shows the number of first patent applications in different years.

Since n=432 is not sufficiently large, we evaluated the accuracy of performance by using LOO cross-validation.

For the estimation model, we used Keras Regressor for multiple regression with two hidden layers, 128 nodes with relu activation each, and number of epochs = 100, unless specified otherwise. We used mean squared error for the loss and Adam for the optimizer. We normalized the input vector by z-score normalization, except for the one-hot vector. To speed up the experiments, we used a Google Colaboratory [18] TPU. For morphological analysis, stemming, and lemmatization, we used "word_tokenize" in the nltk Package [19]. In this package, sentences "We were performing maintenance. It rains cats and dogs." are converted to "We were performmainten. It rain cat and dog."Some are converted to words not in dictionaries. We dealt with case insensitive.

For each d_i, we used the following training and test sets, (X_train, Y_train), (X_test, Y_test), as

Table 2. Notation. Description of symbols and variables used in this paper.

Notation	Description
n	number of pharmaceuticals
d_i	pharmaceutical (drug) i
s_i	actual + experts' estimated sales of d_i from 2019 to 2027
a_i	first patent specification for d_i, PCT written in English
u_{iw}	ratio word w is used in a_i
yp_i	the year a_i was applied to a patent office
yp_{min}	minimum value of yp_i (=1980)
yp_{max}	maximum value of yp_i (=2021)
ya_{min}	the oldest year of the articles in the experiment (=1998)
ya_{max}	the latest year of the articles in the experiment (=2020)
v_i	one-hot vector for pharmaceutical d_i
pr_{ki}	property of a_i, $k = 1, .., 4$
r_w,p_w	r-value, p-value for word w
$r_c(T_c)$	the ratio whose difference between estimated volume and actual sales volume $\leq T_c$
$\Omega(T_r, T_p)$	set of words that satisfy $\|r_w\| \geq T_r$ and $p_w \leq T_p$
tfidf(w, y)	TF-IDF of the word w in year y
$ut(i, y)$	u_{iw} weighted by tfidf(w, y) and summed over w

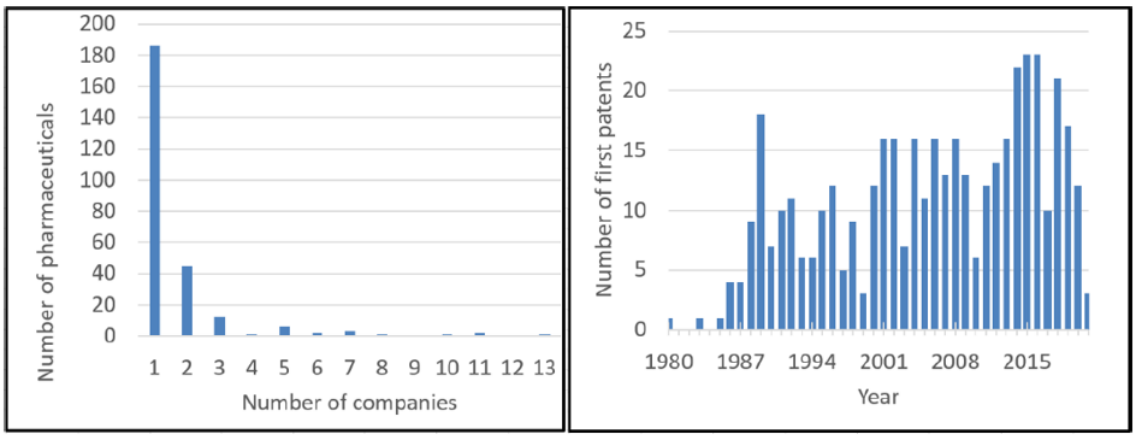


Figure 4. (Left) Number of companies that developed one or more pharmaceuticals and (right) number of first patents applied for per year, both of which were considered in our experiment.

$$(X\_train, Y\_train), (X\_test, Y\_test) = (\boldsymbol{X}_{\neg i}^{[k]}, \boldsymbol{ls}_{\neg i}), (x_i^{[k]}, \log(s_i)), \quad (13)$$

where

$$\boldsymbol{X}_{\neg i}^{[k]} = [\boldsymbol{x}_i^{[k]}, .., \boldsymbol{x}_{i-1}^{[k]}, \boldsymbol{x}_{i+1}^{[k]}, .., \boldsymbol{x}_n^{[k]}]^T,$$
$$\boldsymbol{ls}_{\neg i} = [\log(s_1), .., \log(s_{i-1}), \log(s_{i+1}), .., \log(s_n)]^T, \quad (14)$$

And k=1,..,5 or a combination of these values, as discussed in Section 3. We constructed models n times for each d_i and then calculated the root mean square error (RMSE) and mean absolute error (MAE), as

$$RMSE = (\sum_{i=1}^{n} (l\hat{s}_i - \log(s_i)^2/n)^{0.5},$$
$$MAE = \sum_{i=1}^{n} |l\hat{s}_i - \log(s_i)|/n, \quad (15)$$

Table 3. Experimental results: bold = best data, italic = worst data in RMSE,MAE, and r_c, respectively.

No.	Input(k)	No. of epochs	Node size - 1	Node size - 2	RMSE	MAE	r_c log(2)
1	1	100	128	128	2.314	1.893	*0.208*
2	2	100	128	128	*2.65*	*2.121*	0.211
3	3	100	128	128	1.921	1.489	0.289
4	4	100	128	128	2.036	1.634	0.266
5	1,3	100	128	128	1.811	1.408	0.32
6	1,3,4	100	128	128	**1.724**	**1.359**	0.326
7	1,3,5	100	128	128	1.809	1.459	0.282
8	1,3,4	10	128	128	1.994	1.604	0.241
9	1,3,4	20	128	128	1.88	1.499	0.282
10	1,3,4	50	128	128	1.773	1.378	0.317
11	1,3,4	150	128	128	1.865	1.443	0.299
12	1,3,4	200	128	128	1.774	1.361	**0.35**
13	1,3,4	100	64	128	1.802	1.38	0.34
14	1,3,4	100	256	128	1.844	1.415	0.317
15	1,3,4	100	128	64	1.749	1.365	0.319
16	1,3,4	100	128	256	1.973	1.474	0.333

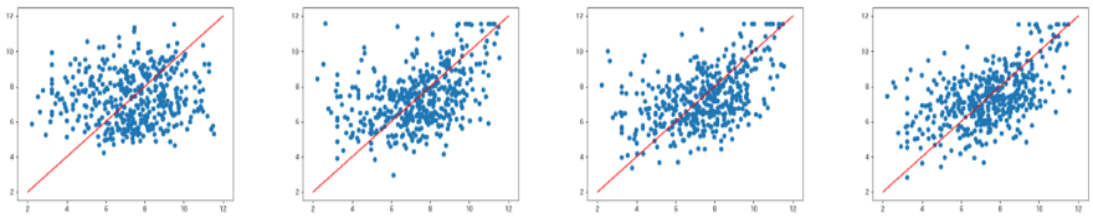


Figure 5. Scatter plots of actual (x-axis) versus estimated (y-axis) sales (logged in millions of USD, from 2019 to 2027). From left to right: Nos. 1, 3, 5, and 6 in Table 3. The diagonal line shows estimation = actual sales.

where $l\hat{s}_i$ is the estimated volume by our model for the input of (13). For outliers, we replaced $l\hat{s}_i$ with max(min($l\hat{s}_i$),max$_{j\neq i}$(log(s_j))).

We also calculated r_c, the ratio of d_i whose estimated volume was close to actual sales volume log(s_i), defined as

$$r_c(\mathrm{T}_c) = |\{i||l\hat{s}_i - \log(s_i)| \leq T_c|/n, \quad (16)$$

where T_c is a threshold.

Table 3 shows the results of all the experiments. Values in the "input(k)" column correspond to k in (14). We used T_c=log(2), which implies that the difference between the actual and estimated data is log(2); in other words, their ratio is between 0.5 and 2.0. We discuss each of the experiments below, where the various numbers (e.g., No. 1) refer to the serial number ("No.") in Table 3.

The scatter plots in Figure5 show the actual versus estimated sales for input Nos. 1, 3, 5, and 6.

Nos. 1 - 4 are the results without using any combinations. No. 5 is the results of combinations with k=1, 3, that is, information of developing company name, the year the first patent was applied for, and the words used in the first patent. Nos. 6, 7 are the results of combinations with k=1, 3, 4, and k=1, 3, 5, respectively. k=4 is the case of using words in pharmaceutical articles, and k=5 is the case where articles published after the first patent were excluded in the k=4 case.

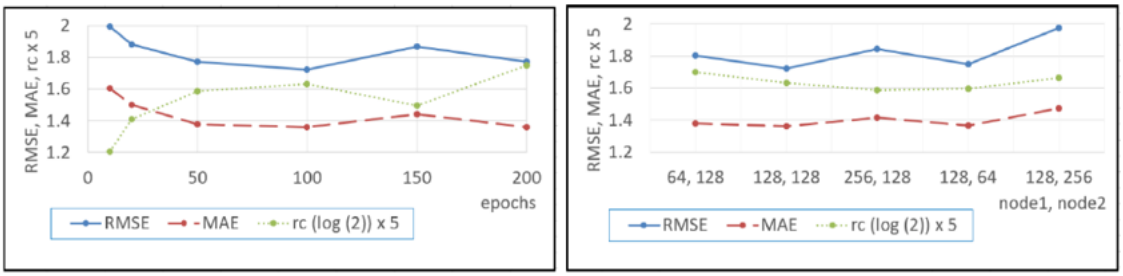


Figure 6. Estimation accuracy for different number of epochs (left) and different node sizes (right).

We performed other tests by changing the number of epochs and node sizes with combinations of k=1, 3, 4. Nos. 8 - 12 are the results when changing the number of epochs from 100 to 10, 20, 50, 150, and 200, respectively. Nos. 13 - 16 are the results when changing node sizes to (node 1, node 2) = (64,128), (256, 128), (128, 64), and (128, 256), respectively. Figure6 shows the results of Nos. 8-12 (left) and 13-16 (right).

We used pharmaceutical data from the years between 1980 and 2021, so we had yp_{min}=1980 and yp_{max} = 2021. For the 432 patent specifications, 203,632 different words were used. For k=3, we used $\Omega(0.01,0.01)$, $|\Omega|$=866, and 50 of the 866 words in Table 1. To calculate the r-value for each d_i, we used all of the patent specifications except for a_i.

As the articles for k=4, 5, we used Pharmaceutical Benefits Pricing Authority Annual Reports published from 1998 to 2020 [20, 21] and reports of the Pharmaceuticals and Medical Devices Agency from 2004 to 2018 [22]. For the years between 1998 and 2003 and between 2019 and 2020, we used reports from Pharmaceutical Benefits Pricing Authority Annual Reports. Therefore, we had ya_{min}=1998 and ya_{max} = 2020.

5. DISCUSSION

Several inferences can be made from the experimental results.

Combination of k=1, 3, 4 (No. 6 in Table 3) yielded the best performance for RMSE and MAE.

(RMSE, MAE) = (1.724, 1.359) is ×0.75 and ×0.72 of those of No. 1, respectively, which use only the developing company name and the year of the first patent application. This is a significant improvement.

One interesting fact is that, while the performance of k=4 alone (No. 4) was not as good as No. 3, it helped improve the combination of k=1, 3 (No. 5). This implies that patents containing "hot" words indicate high potential for the future sales volume of the pharmaceutical.

Using only the information of the first patent yielded the worst result for RMSE and MAE (No. 2). This implies that the length of the title, the abstract, the patent, and the number of claims contain very little or no useful information regarding future sales volume.

Using only the developing company name and the year of the first patent application (No. 1) yielded a better performance than the case of No. 2 for RMSE and MAE. This is reasonable since Figure 3 indicates that the sales volume of some pharmaceuticals depends on the developing company names. However, 43% of the pharmaceuticals were "single" pharmaceuticals with only one developing company in the training/test data, and for these, sales volume estimation was close to the average value of the rest of the companies.

In contrast, using words whose absolute*r*-values were equal to or more than 0.1 was very effective (No. 3), compared with No.1 or No. 2, even without combining with other vectors.

This indicates that patents may include words implying that the target products or services will sell in the future. It is possible that this stems from the confidence of the patent authors.

Comparing the panels in Figure5 from left to right, we can observe that dots are shifting to the diagonal line, which shows estimation = actual sales, which are the cases of k=1only (No. 1), k=3 only (No. 3), combination of k=1, 3 (No. 5), and combination of k=1, 3, 4 (No. 6), respectively.

We evaluated the case of using articles published in the same year or before the first patent application, namely, the combination of k=1, 3, 5 (No. 7). In this case, the estimation performance was worse than that for No. 6, which is reasonable since the input data were partly omitted, but the performance was still better than without using k=5 (No. 3) for RMSE.

For the case of the combination of k=1, 3, 4, we evaluated the performance by using a different number of epochs (Nos. 8-12, Figure6, left panel) and node size (Nos. 13 - 16, Figure6, right panel). For Nos. 8 - 10, the estimation performance improved for RMSE and MAE as the number of epochs increased, but the performance remained unchanged or deteriorated as the number of epochs increased beyond 100, implying that deep learning parameters overfit after 100 epochs. On the other hand, varying the node size did not influence the performance significantly, but found that node size = 256 in either node (Nos. 14, 16) yielded worse results than those with smaller nodes (Nos. 13, 15).

In this experiment, we considered only the usage of words in each patent, regardless of whether they were used in the abstract, claim, or other parts. However, as several studies have been performed on patent structure analysis [14,15,23-35] and keyword extraction analysis[14,15,26,27], there is still scope for further estimation accuracy improvement by, for example, applying weights to word usage in accordance with the location of the words (e.g., in the abstract, claims, or other parts).

6. CONCLUSION AND FUTURE WORK

We proposed a model that improves the estimation performance of future pharmaceutical sales by analyzing the first patent specification submitted for the pharmaceutical and articles related to pharmaceuticals. We performed experiments using a data consisting of 432 pharmaceuticals whose first patents and sales volume are both available and found that the best sale estimation performance was obtained by using a combination of pharmaceutical developing company name, the year the first patent was applied for, words used in the first patent specification, and TF-IDF calculated from words used in the pharmaceutical related articles to weight the word usage ratio of the first patent of the pharmaceutical.

One interesting finding was that just using words in the patent specification yielded much better estimation performance than the case of using the company name (i.e., the brand name) and the year the first patent was applied for. Also, the estimation performance was much improved by combination all of these plus pharmaceutical related articles. These are groundbreaking results because these prove that patents and related articles contain information about future pharmaceutical sales. Since patent specifications and articles can easily be obtained, this will help us significantly in building a marketing strategy.

In this paper, we focused on pharmaceuticals, but our model can be applied to other industries such as food, electrical appliances, cars, clothes, and so on.

As future work, we would like to apply NLP while taking the structure of patents and articles into consideration. We would also like to examine the use of word embedding concepts (e.g., BERT [28] or word2vec [29]) to determine similar word usage between patents and articles, and see how these concepts improve the estimation performance.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Tetsuya Nasukawa and Dr. Shoko Suzuki in IBM Research-Tokyo for their accurate advice.

This work was supported by JSPS KAKENHI Grant Number 10881998.

REFERENCES

[1] M. Erman and M. Mishra, (2021) “Pfizer sees robust COVID-19 vaccine demand for years, $26 bln in 2021 sales.” https://www.reuters.com/business/healthcare-pharmaceuticals/pfizer-lifts-annual-sales-forecast-covid-19-vaccine-2021-05-04.(Last accessed: 13.Dec.2021)
[2] Clarivate. https://clarivate.com/.(Last accessed: 13. Dec.2021)
[3] PCT – The International Patent System. https://www.wipo.int/pct/en/.(Last accessed: 13. Dec.2021)
[4] M. Merino and R. Adrian, (2016) “Estimation of retail sales under competitive location in Mexico,” *Journal of Business Research* 69.2, pp. 445-451.
[5] D. L. Huff, (1963) “A probabilistic analysis of shopping center trade areas,” *Land economics* 39.1, pp. 81-90.
[6] D. E. Raeside, (1974) “An introduction to Monte Carlo methods,” *American Journal of Physics* 42.1 pp. 20-26.
[7] S. Jordan and M. Martin, (2020) “The use of forecast accuracy indicators to improve planning quality: Insights from a case study,” *European Accounting Review* 29.2, pp. 337-359.
[8] B. Pavlyshenko, (2019) “Machine-learning models for sales time series forecasting,” *Data* 4.1, 15.
[9] A. L. D. Loureiro, V. L. Miguéis, and L. F. M. da Silva, (2018)“Exploring the use of deep neural networks for sales forecasting in fashion retail,” *Decision Support Systems* 114, pp. 81-93.
[10] K. H. Kim, Y. J. Han, S. Lee, S. W. Cho, and C. Lee, (2019) “Text mining for patent analysis to forecast emerging technologies in wireless power transfer,” *Sustainability* 11.22, 6240.

[11] C. C. Guderian, P. M. Bican, F. J. Riar, and S. Chattopadhyay, (2021) "Innovation management in crisis: Patent analytics as a response to the COVID-19 pandemic," *R&D Management* 51.2, pp. 223-239.
[12] W. S. Lee, S. C. Hyo, and Y. S. So, (2018)"Forecasting new product diffusion using both patent citation and web search traffic," *PloS one* 13.4, e0194723,
[13] F. Bass, (1969) "A newer product growth for model consumer durables," *Management Science*, Vol. 15, No. 5, January, pp. 215-227.
[14] S. Suzuki and H. Takatsuka, (2016) "Extraction of keywords of novelties from patent claims," *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*: Technical Papers, pp. 1192-1200.
[15] S. Hido, et al., (2012) "Modeling patent quality: A system for large-scale patentability analysis using text mining," *Information and Media Technologies* 7.3, pp. 1180-1191.
[16] Cortellis. https://clarivate.com/cortellis.(Last accessed: 13. Dec.2021)
[17] Derwent. https://clarivate.com/derwent/solutions/derwent-innovation/.(Last accessed: 13. Dec.2021)
[18] Google Colaboratory. https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index.(Last accessed: 13. Dec.2021)
[19] nltk Package. https://www.nltk.org/api/nltk.html.(Last accessed: 13. Dec.2021)
[20] Pharmaceutical Benefits Pricing Authority Annual (1998-2010).https://www.pbs.gov.au/pbs/industry/pricing/pbs-items/historical/pbpa-annual-reports.(Last accessed: 13. Dec.2021)
[21] Pharmaceutical Benefits Pricing Authority Annual (2011-2020). https://www.health.gov.au/about-us/corporate-reporting/annual-reports?utm_source=health.gov.au&utm_medium=callout-auto-custom&utm_campaign=digital_transformation.(Last accessed: 13. Dec.2021)
[22] Pharmaceuticals and Medical Devices Agency. https://www.pmda.go.jp/english/index.html.(Last accessed: 13. Dec.2021)
[23] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama, (2003) "Patent Claim Processing for Readability: Structure Analysis and Term Explanation," *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, 20: pp. 56-65.
[24] P. Parapatics and M. Dittenbach, (1990) "Patent Claim Decomposition for Improved Information Extraction," *Proceedings of the 2nd International Workshop on Patent Information Retrieval*: pp. 33-36.
[25] S. Sheremetyeva, S. Nirenburg, and I. Nirenburg, (1996) "Generating patent claims from interactive input,"*Proceedings of the 8th International Workshop* on Natural Language Generation: pp. 61-70.
[26] M. Verma and V. Varma, (2011) "Applying Key Phrase Extraction to Aid Invalidity Search," *Proceedings of the 13th International Conference on Artificial Intelligence and Law*: pp. 249-255.
[27] M. A. Hasan, W. S. Spangler, T. Griffin, and A. Alba, (2009) COA: Finding Novel.
[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv*:1810.04805.
[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, (2013) "Efficient estimation of word representations in vector space,"*arXiv preprint arXiv*:1301.3781.

AUTHOR

Koichi Kamijo received B.E. degree in Electrical Engineering from the University of Tokyo in 1985, M.E. degree in Computer Science from Cornell University in 1996, and Ph.D. degree in Engineering from the University of Tokyo in 2010, respectively. After having worked at IBM Research-Tokyo for more than 20 years, he is currently a professor of International Professional University of Technology in Tokyo, department of Information Technology, since 2020. He is interested in Image Processing, Human Interaction, Digital Rights Management, Machine Learning, and Natural Language Processing. He holds 98 issued patents all over the world.

SOCIAL MEDIA NETWORK ATTACKS AND THEIR PREVENTIVE MECHANISMS: A REVIEW

Emmanuel Etuh, Francis S. Bakpo and Eneh Agozie H

Department of Computer Science, Faculty of Physical Sciences,
University of Nigeria, Nsukka, Nigeria

ABSTRACT

We live in a virtual world where actual lifestyles are replicated. The growing reliance on the use of social media networks worldwide has resulted in great concern for information security. One of the factors popularizing the social media platforms is how they connect people worldwide to interact, share content, and engage in mutual interactions of common interest that cut across geographical boundaries. Behind all these incredible gains are digital crime equivalence that threatens the physical socialization. Criminal minded elements and hackers are exploiting Social Media Platforms (SMP) for many nefarious activities to harm others. As detection tools are developed to control these crimes, hackers' tactics and techniques are constantly evolving. Hackers are constantly developing new attacking tools and hacking strategies to gain malicious access to systems and attack social media network thereby making it difficult for security administrators and organizations to develop and implement the proper policies and procedures necessary to prevent the hackers' attacks. The increase in cyber-attacks on the social media platforms calls for urgent and more intelligent security measures to enhance the effectiveness of social media platforms. This paper explores the mode and tactics of hackers' mode of attacks on social media and ways of preventing their activities against users to ensure secure social cyberspace and enhance virtual socialization. Social media platforms are briefly categorized, the various types of attacks are also highlighted with current state-of-the-art preventive mechanisms to overcome the attacks as proposed in research works, finally, social media intrusion detection mechanism is suggested as a second line of defence to combat cybercrime on social media networks

KEYWORDS

Intrusion Detection System, Data Warehouse, Machine Learning, Hackers, Social Media Platform, Online Social Network Intrusion.

1. INTRODUCTION

Social media network platforms provide mechanisms that enhance the effectiveness of virtual socialization in the global village. It is a medium that enable families, friends, and associates to interact and communicate seamlessly irrespective of their locations, distances, and platforms. Online Social Networks (OSNs) are the connection and communication platform that promotes the social interaction in the virtual space [1]. [2] identified 7 categories of social networks on the Internet to include: electronic mail services like Gmail, Yahoo mail, Microsoft outlook, Hotmail, etc; Instant messengers like WhatsApp, Twitter, Yahoo messenger, Instagram, Telegram, Snapchat, etc; Blogs platforms like Blogger, Tomblr, Wix, Linda etc; Social networking sites like Facebook, TikTok, Quora, LinkedIn, etc; Multimedia sharing systems like YouTube, Skype, Flikr, etc; Auction platforms like Jumia, Alibaba, Konga, OLX, etc; and Social search engines like Google, Yahoo, Safari, etc. All these platforms enable users to socialize and stay in touch

David C. Wyld et al. (Eds): CSE, AI & FL, NLPTT, SESBC - 2021
pp. 59-72, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112405

with social reality in the virtual environment with varying functionalities. The pervasive nature of Information and Communication Technology (ICT) has greatly influenced every aspect of human activities; this has also influenced social media users to see the platform as a virtual home where they save their sensitive information on the database of these platforms.

The growing reliance on the use of social media networks worldwide has resulted in great concern for information security. One of the factors popularizing the social media platforms is how they connect people worldwide to interact, share content and engage in discussions of mutual interest that know no geographical boundaries. Behind all these incredible gains, most traditional crimes now have digital equivalence enabling criminal minded elements and hackers to exploit social media platforms for many nefarious activities to harm others. As security administrators and policy makers develop detection tools to control these crimes, hackers' tactics and techniques are also constantly evolving. Hackers are cybercriminals that specializes in virtual terrorism that endangers the legitimate users of Social Media Network Platform (SMNP) in particular and the entire virtual community in general [3] through various kinds of cyber-attacks.

These cybercrimes have significant negative impact on the social media platforms and the users in particular. Because of ease of accessibility, some of the social media users prefer to store their sensitive data on the network and when the account is hacked, this information could be used to swindle and defraud the user; also, the user's social contacts on the platform are at high risk of being defrauded by the hacker who could use their techniques by masquerading as the authorized user. High profile users like public and political leaders with private information that could tarnish their image if extracted can be used to threaten the user for ransom.

Different approaches have been used to prevent hackers' intrusion on the social media network platforms. The prevalent one is authentication using credentials like username and password, or PIN; biometric authentication like using face recognition technology, fingerprint, pattern matching, or voice recognition are all various forms of authentication. Other methods are Role Based Access Controls (RBAC), Extended RBAC, Temporal (RBAC), Risk-based access control [4]. These security methods have a lot of drawbacks like weak password which can easily be guessed by hackers using dictionary attack. In an attempt to enforce stronger password for authentication, social media users are forced to write their authentication credentials on papers which can also be stolen by hackers, these weaknesses have influenced many researchers to propose several security mechanisms to curtail the activities of these hackers. Some of these proposals include: biometric authentication, hybrid system for anomaly detection in social networks [5],Network Intrusion Detection System[6], [7], [8]etc.

All these methods are not suitable for data warehouse security. The commonly used network securities software like firewall and anti-viruses independently provide different services to network security but they can be bypassed by hackers. Hackers improvises new techniques of breaking into the social media platforms without being detected, the two close proposals on Data Warehouse Database Intrusion Detection System by [9] and [4] does not discourage resilient hackers. The limitations of all these proposed security approaches have been pointed out in Table 2.4 of the literature review. Hence, there is therefore a need for Intelligent Intrusion Detection Model (IIDM) that is efficient to disarm the hackers from carrying out their cybercrime activities against SMNP.

2. THEORETICAL BACKGROUND

Social media platforms have become an integral part of average network users in the virtual community. Billions of connected devices to the Internet operate on one social media platform or

the other. According to report in [10], over 500million IoT devices were implemented globally in 2003, 12.5 billion in 2010, and 50 billion in 2020.There are about 3.5 billion people on social media with an estimated attacks that generate over $3 billion annually for cyber criminals [11]. Online social network platform like Facebook incorporate several functionalities like product and services advertisement and sales that makes it relevant to almost all internet users either cooperate or private. This has also increased cybercriminals' activity on the platform. According to a recent survey by Computer Emergency Response Team (CERT), the rate of cyber-attacks has been doubling every year [6]. Online social network is faced with threatening security challenges [2]. Facebook is the most popular social networking site. It was launched in February 2004 [12]With roughly 2.89 billion monthly active users as at the second quarter of 2021, Facebook is the biggest social network worldwide.

The Covid19 pandemic has been instrumental to the geometric shift to virtual socialization. The technological shift to cloud computing paradigm also has positively influenced the ubiquity of social media. This shift seems to have given hacker an edge to securely carryout their nefarious acts since humans are less involved. Cloud intrusion attacks are set of actions that attempt to violate the integrity, confidentiality or availability of cloud resources on cloud SMNP. The rising drop in processing and Internet accessibility cost is also increasing users' vulnerability to a wide variety of cyber threats and attacks.

Intrusion detection is meant to detect misuse or an unauthorized use of the computer systems by internal and external elements [7]. IDS are an effective security technology, which can detect, prevent and possibly react to the attack [13], [14] opined that artificial Intelligence plays a driving role in security services like intrusion detection.

2.1. Social media network

Social media network is a platform that creates virtual environment for social interactions among circle of friends and fans of like-minds. "Social media platforms are internet-based applications focused on broadcasting user-generated Content"[15] It deals with the sharing of information and multimedia content between users on similar platforms over electronic network especially the internet and cyberspace [16]. This platform has geometrically grown to become not only an effective communication tool for personal and social use, but also an essential channel for businesses and official communication channels. There are thousands of social media platforms being used today for different purposes, few of them that are most popular are highlighted below.

- **Facebook** is an online social media platform that provides several services like social networking of friends and fans, online advertising, voice calls, instant messaging, video calls, video sharing and viewing, online market place, virtual gifts among both young and the old, private and corporate bodies. It was launched on February 4, 2004, by Mark Zuckerberg. It had over 1.18 billion monthly active users as of August 2015 [16] and 2.85 billion active users in 20201 according to statistics by [17] with an engagement of over 4 billion views of videos everyday on the network. About 2.14 billion people can be reached via advertising on Facebook [17].
- **WhatsApp** is a cross-platform internet-based instant messaging application that allows smart phone users to exchange text, image, video and audio messages for free provided the device has Internet access. It was developed in 2009 by Brian Acton and Jan Koum. WhatsApp became the most popular messaging app with about 900 million active users as at September, 2015 [16].
- **MySpace** is a social networking website offering an interactive, user-submitted network of friends, personal profiles, blogs, groups, photos, music and videos. It was the biggest social

media platform up till 2008 when it was overtaken by Facebook. It was cofounded by Chris DeWolfe and Tom Anderson

- **Twitter** is a social network platform that enables users to write and read short character messages called tweets. It revolves around the principle of followers who are equally users, who choose to follow another Twitter user and can thus view tweets sent by that user. Whereas unregistered users can read tweets, one must be registered to send tweets. It was founded in March, 2006 by Jack Dorsey [16].
- **Instagram** allows users to upload media that can be edited with filters and organized by hashtags and geographical tagging. Posts can be shared publicly or with pre-approved followers. Users can browse other users' content by tags and locations and view trending content. Users can like photos and follow other users to add their content to a personal feed. Instagram has 1.38 billion active users with 500 million daily active users of Instagram stories, 1.16 billion people can be reached through adverts on Instagram [17]
- **YouTube** is a video sharing service that allows users to watch videos posted by other users and upload videos of their own. With the ubiquitous use of smart phones this platform has become the first choice in personal broadcasting and video sharing. It was cofounded by Chad Hurley, Steve Chen, and Jawed Karim in February 2005. In November 2006, it was bought by Google and now operated by Google
- **LinkedIn** is a social media platform for professional networking. It is a social networking tool available to job seekers and professionals where users can invite other users and even non-users to connect. Inviters who get several rejections from invitees risk having their accounts restricted or closed. On this platform, users can get introduced to networks of contacts, new job and business opportunities, display products and services in their company profile pages, list job vacancies and search for potential candidates
- **Skype** is an IP telephony service provider that can be used to make free voice and video calls over the Internet to any Skype subscriber or to any other non-user at low calling rates. It is relatively simple to download and install the software, which works on most computers and phones. A dedicated Skype phone can be used or desktop computers, notebooks, tablets, mobile phones and other mobile devices fitted with a headset, speakers, microphones or USB phone. Skype also enables file transfers, texting, video chat and videoconferencing.
- **Viber** is a mobile application that allows phone calls and text messages to all other users, whether mobile or landline, for free. It is available over WiFi or 3G with sound quality much better than a regular call with mobile carrier charges applicable when used over a 3G network. Once the application is installed, calls can also be made to numbers that do not have Viber at low rates using ViberOut. Viber works on most android, iphone, blackberry, windows, mac, nokia and bada devices.
- **Tumblr** is a microblogging and social networking platform it's service allows users to post multimedia and other content to a short-form blog. Users can follow other users' blogs. Bloggers can also make their blogs private. For bloggers many of the website's features are accessed from a "dashboard" interface. It was founded by David Karp in 2007.
- **WeChat** is a Chinese multi-purpose instant messaging, social media and mobile payment app developed by Tencent. First released in 2011, it became the world's largest standalone mobile app in 2018, with over 1 billion monthly active users. It has been described as China's "app for everything" and a "super app" because of its wide range of functions. It provides text messaging, hold-to-talk voice messaging, broadcast (one-to-many) messaging, video conferencing, video games, sharing of photographs and videos and location sharing.
- **Reddit** This social media platform enables you to submit content and later vote for the content. The voting determines whether the content moves up or down, which is ultimately organized based on the areas of interest (known as subreddits). Number of active users per month: 100 million approximately.

- **Taringa** is one of the largest social networking platforms in Latin America and allows users to share their experiences, content and more. Number of active users: 75 million approximately.
- **Renren** is the largest social networking site in China and is literally a platform for everyone. It has been highly popular with the youth due to its similarity to Facebook, as it allows users to easily connect with others, quickly share thoughts and posts, and even update their moods. Number of active users per month: More than 30 million approximately

All the social media network platforms including the ones highlighted above can be categorized based on their support for the types of data they exchange or based on their aspect of support for social interactions.

Based on their support for the types of data they exchange, social media network platforms can be categorized into four main types:

i) Textual-based platform: used for text-related social communication for sending/receiving messages. A good examples are the messenger platforms.
ii) Visual-based platforms: used to for picture-related social interactions like sending and receiving images. A good example is flikr platform
iii) Audio-visual based platforms: used to for video-related social interactions like sending/receiving video data).A typical example is the YouTube,
iv) Hybrid platforms: this platform combines the functionalities of more than one of the textual, visual, and audio-visual platforms. A typical example is the Facebook

Social media platforms can also be categorized based on their aspect of support for social interactions. These categories are summarized in Table 2.1 below.

Table 2.1. Summary of Categories of Social Media Platform

Category	Usage	Examples
Electronic Mail Service Platforms	The very first social media platform that gave birth to the Internet. Used to send and receive electronic mails from friends and associates	Gmail, Yahoo mail, Microsoft outlook, hotmail, etc
Social networking websites	Mostly used to connect friends, family, brands, and to reach out to target audience.	Facebook,Twitter, Whatsapp, Instagram,
Discussion Platforms	Mostly designed for research, and focused discussion with people with common interest	Reddit, Quora, Nairaland, etc
Blogging platforms	Mostly used for news, writing of articles, and personal messages to targeted audience	Tumblr, Medium, Blogger, Wix, Linda
Instant Messenger Platforms	Used for real-time textual conversation with instant sending and receiving functionalities	Whatsapp, Twitter, Yahoo messenger, Instagram, Snapchat,
Multimedia Platforms	For sharing of videos with both subscribed and visitor social media user of this platform	YouTube, Skype Flikr, TikTok, etc
Auction Systems	For sales of goods and services	Jumia, Alibaba, Konga, OLX,
Cooperate Social Platforms	Most companies are now incorporating social flavor like blog to their virtual platform to enable them to reach their target customers with their products and services	Microsoft news, Yahoo news, BBC news, Safari etc
Educational /Professional Platforms	For sharing knowledge through, chart, uploading of resources, live meetings, and connecting with professionals in the chosen profession of the social	Elsevier, Academia, Slack. DOAJ, Google Meet, Zoom, LinkedIn

	media user	etc
Gaming platforms	Hosts gamming applications where social media users play games either for entertainment, fun, or betting	Nairabet, Naijabet, Kingsbet, etc
Search Engines	It is the major tool that enhances virtual socialization. It enable social media user to easily locate the object of search	Google, Yahoo, Microsoft edge, Firefox, etc

2.2. Social media network platforms attacks

There are several attacks launched against social media network platforms. It is important to know them because a more thorough understanding of these types of attacks equips social media user an armament of defensive measures and knowledge to lessen the likelihood of being exploited [3]. In August 6, 2009, Twitter, Facebook, LiveJournal, Google's Blogger, and YouTube were attacked by a distributed denial-of-service (DDoS) attack, in October, 2021, similar service disruption was encountered. [3] identified seven deadliest attacks on social media network platforms. These attacks are highlighted as follows

1. Social networking infrastructure attacks: here the attacker launches the attack on the platform that provides the social service with the view to disconnecting users from accessing the services provided by the platform. The major attack used against social networking infrastructure that directly affects the users is DDoS.

2. Malware attacks: in this type of attack, the hacker develop a malicious software with the intention of gaining control and utilizing the user's device to perform some malicious activities like launching DoS attack, keystrokes logging, theft of credential, credit-card number or bank details, etc. The mode of infecting user's device on social media is usually through links or images sent to the user's inbox knowing that the user will likely open since it comes from a connected social contact [18]. Once a user is infected, the hacker uses the compromised social media account to spread the worm by delivering a message to other users who are friends with the infected user containing a luring link to a third-party Web site, where they are then prompted to perform an action like "register to view full image", "update you Adobe Flash player to have a better view", etc. Once the action is performed, the worm will automatically infect the devices of all the connected friends that followed the link to the third party site. Common Malware Categories are Crimeware, Spyware, Adware, Browser Hijackers, Downloader, Toolbars, and Dialers [18]. Hackers leverage on the openness of social networking sites where users generate their contents; the large number of users; and the trust that is implied where users assume all friends are to be trusted to launch attacks to connected billions of users. The most effective method is by using Cross-Site Scripting XSS to implement their malicious codes on social networking site.

3. Phishing attacks: as the name imply is a hacking technique where the hacker lure a user using "bate" that is most appealing to the user with the intention of trapping the user. In most cases, the user is coarse into divulging sensitive information that will in turn be used to attack the user.

4. Evil twin attacks: in this type of attack, the hacker uses the target's profile to create account to mimic the authentic user. This attack can also be called cyber-impersonation. The new account is then used to send friend's request to the contacts on the social media platform just to enable the attacker to enjoy the privileges of friends and gain access to the users on the platform.

5. Identity theft: in this type of attack, the user's credential is stolen and used to securely gain access to the user's social media platform. Once the attacker successfully gains access, they launch their pre-conceived attacks while impersonating the authentic user.

6. Cyber-bullying: a way of threatening or intimidating a social media user either through messages or by posting objectionable content on the social media network to harass or intimidate the targeted user.

7. Physical threats: in this attack, the hacker launches physical attack against selected user. It could be in form of bypassing physical security of the platform through threatening the user to remove the device security.

All these threats can use any of the following attacking mode to carryout the threats.

1. Denial of Service (DoS) where an attacker tries to prevent legitimate users from using a service
2. Probe attack where an attacker tries to find information about the target host through ways such as scanning victims to get information about available services and the operating system
3. User to Root (U2R) where unauthorized access to local super-user privileges are being granted
4. Remote to Local (R2L) where unauthorized access from a remote machine through approaches such as guessing password to obtain a local account on the victim host
5. Advanced Persistent Threat (APT) is a targeted attack against a high-value asset or physical system where attackers often leverage stolen user credentials or zero-day exploits to avoid triggering alerts

2.3. Security measures for social media network

The "juicy prospect" of social media network platforms has made hackers to constantly device techniques to intrude and usurp users. They have two fold targets which are the social media users and the SMNP which they break into and control for their selfish gain [19]. On the users end, the hackers' activities make them susceptible to threats which include identity theft, evil twin, password resetting, sim cloning, brute force, fake links, phishing, information leakage, celebrity spoofing, fake account, impersonation, etc [20] [21] [22]. They also use code injection through malicious SQL script to disrupt the network. The existing security mechanism for DW include Role Based Access Controls (RBAC), Extended RBAC, Temporal RBAC (TRBAC), Risk-based access control[4] which all has to do with authentication using username and password. [9] were the first to propose Database Intrusion Detection Systems (DIDS) for DW. [4] improved it by incorporating second level authentication, instant messenger like Whatsapp also uses two-steps verification where the user is asked to supply a PIN at intervals to prevent hackers from the network any time there is a detected anomaly, but hackers still use social engineering to trick the users thereby compromising the account which go undetected by role-based access controlled systems.

To hack into a social media network, there are basic steps taken by attackers to perform their operations. These steps are:

i. Target selection: the attacker determine the victim of the social media users to attack
ii. Attack selection: The attacker determine the type of attack to launch against the target, e.g infrastructure, malware, phishing, evil-twin, identity theft, cyber-bullying, physical threat, celebrity spoofing etc.
iii. Strategy formation: the strategy is dependent on the type of attack to use. If the selected attack is to launch DDoS attack, then the attacker will have to recruit accomplices (botnet) to use in launching the attack either through Calls, E-mail, Post to user groups, or Creation of a Web page where users are redirected to for infection.
iv. Army training: Furnish the accomplices with package containing the attack, time, date, and instructions on how to perform the attack.

v. Launching attack: here the attacker will launch the attack, wait and watch the execution of the attack.

To overcome the various attacks highlighted on the social media network platform, the user should:

i. Ensure an up-to-date antivirus software on the device as a primary line of defence
ii. Not open e-mails from people you don't know.
iii. Not click on unknown links
iv. Not visit unfamiliar sites
v. Disable JavaScript.
vi. Maintain and ensure regular update on software patches.
vii. Implement browser security policies, such as blocking pop-ups and limiting the number of connections.
viii. Implement platform privacy security policies, such as "who can see my personal info", "who can post on my wall", status, etc
ix. Implement IDS/IPS as second line of defence against attackers.

2.4. Intrusion detection mechanisms

Intrusion detection system (IDS) is a device or software application that monitors a system, network or application for malicious activity or policy violations with the aim of detect them. Two major IDS emphasized in literatures are Host-based and Network-based. These IDSs are not suitable for intrusion detection in application related intrusion attacks. This gave rise to the development of application specific IDS which is application based. Fig 2.1 shows the three types of intrusion detection systems presently available for information security.

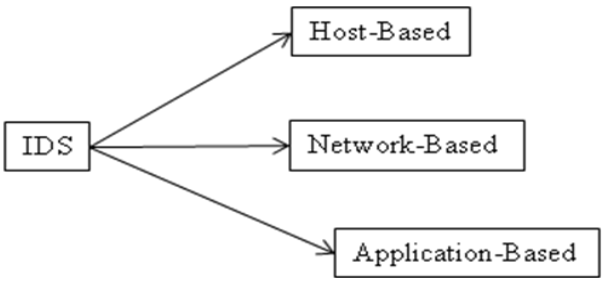


Fig 2.1. Types of Intrusion Detection System

Researchers have proposed various methods to detect abnormal operations in a system. Generally, IDS comprises of four main components: Traffic Collector, Analysis Engine, Signature Database, Management and Reporting Interface [7]. Network-based, Host-based, or Application-based intrusion detection systems use either signature mechanism to detect intrusion or anomaly approach. The signature approach uses rules for decision making on classifying intrusion based known profile of intrusion. Anomaly on the other hand classifies an operation as intrusion based on a deviation from the known normal operation of a given system.

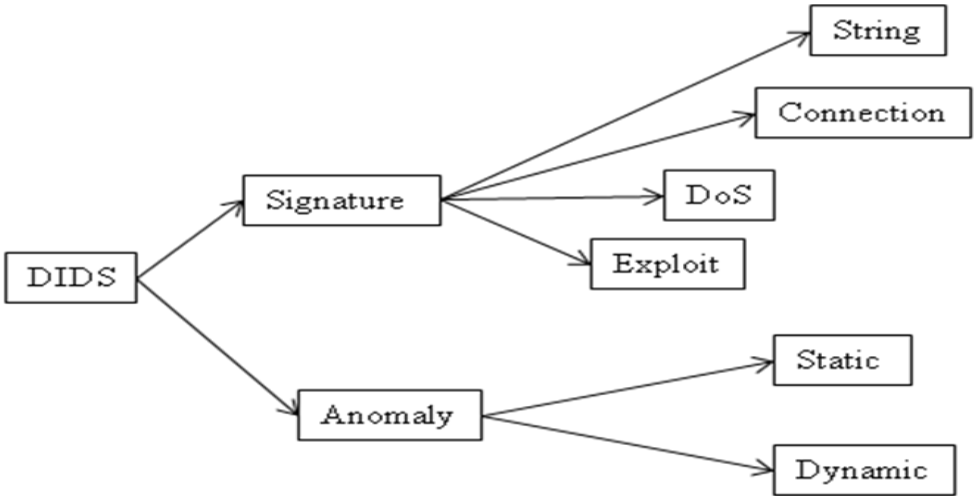


Fig 2.2. Mechanisms for intrusion detection

The work done in can be consulted for further reading [23] as it compared the various approaches using features, advantages, and disadvantages of each approach.

2.5. Review of Related Literatures

[5] proposed "an efficient hybrid system for anomaly detection in social networks". The model cascaded several machine learning algorithms that included decision tree, Support Vector Machine (SVM) and Naïve Bayesian classifier (NBC) for classifying normal and abnormal users on social networks. the anomaly detection engine uses SVM algorithm to classify social media network user as happy or disappointed, NBC algorithm is used based on a defined dictionary to classify social media users with social tendency. Unique features derived from users' profile and contents were extracted and used for training and testing of the model, performance evaluation conducted by experiment on the model using synthetic and real datasets from social network shows 98% accuracy.

(Mansoori et al. 2020) proposed "Suspicious Activity Detection of Twitter and Facebook using Sentimental Analysis". The model was designed using Natural

[24] proposed "Social Media Cyberbullying Detection using Machine Learning" the model leverage on the machine learning capability to detect the language patterns of bullies on social media network which will be used to generate a model that can automatically detect cyberbullying actions on the network. The model was evaluated using two classifier algorithms: SVM and Neural Network, TFIDF and sentiment analysis algorithms were used for features extraction. Different n-gram language models were used to evaluate the model and was found to achieved 92.8% accuracy using Neural Network with 3-grams and 90.3% accuracy using SVM with 4- grams while using both TFIDF and sentiment analysis together. NN performed better than the SVM classifier with average f-score of 91.9% while that of SVM achieved average f-score 89.8%.

(Shah, Sharma, and Bandgar 2021) proposed "Cybercrime Prevention on Social Media" in which a Facebook-like social network platform was designed as a proof-of-concept to detect cyberbullying and anormaly detection. Each user's post will be passed through "Post Classifier" to divide illegal posts from adult image post, the illegal post detection algorithm was used to check for violence-based objects. Similarly, sentiment analysis was used to check the intention of

the message to detect any harrasment words in the content of the post. A threashold was set for any of the negative words and images posted by a user, if any illegal post by a user exceed a threashold, the user will be warned, and if there is persistence in the post, the user will be banned from making a post on the platform. Convolution Neural Network (CNN) machine learning algorithm was used in the implementation

[4] "Intrusion Detection System for Data Warehouse with Second Level Authentication". This proposal was premised on the ground that earlier security mechanisms for data warehouse like "Role Based Access Controls (RBAC), Extended RBAC, Temporal RBAC (TRBAC), Risk-based access control, Intrusion Detection System (IDS) and some other customized security solutions for DWs does not detect a hacker that gain rightful access to the DW through credential theft. Therefore, a second level authentication mechanism within the IDS was proposed where a minute deviation from the user's past behavior will be detected based on providing answer to secrete question that was provided at the account setup phase.

In their research work, [12] envisioned the most important criteria of employing social network in higher education, they observed broad view of possibilities for using social networks in higher education. The conceptual framework for academic social network should have the following four main objectives: (1) To provide academic service support and academic information dissemination; (2) To enable student support and communication; (3) To exalt social and cooperate learning; (4) To provide achievement representationability.

[13] developed an "Intrusion Detection System for College (ERP) Enterprise Resource Planning System" The system used a layered approach combined with a decision tree based architecture to detect attacks, the design and simulation was carried out using Netbeans integrated development environment with MYSQL database.

[8]proposed a model for building the network intrusion detection system using a machine learning algorithm called decision tree to detect anomaly based intrusion. The system used Recursive-Feature-Elimination (RFE) to select the best features from Change Control Intrusion Detection (CCIDS) 2017dataset. The dataset was split into training and test dataset, the training dataset was fit to the classification model developed to classify the test data as malicious or benign. The precision of the proposed system indicated True-Positive-Rate (TPR) of 99.9% and the False-Positive-Rate (FPR) of 0.1%

[14]proposed a dynamic Intelligent Intrusion Detection Model based on specific AI approach for intrusion detection. It is a hybrid system that combines anomaly, misuse and host based detection. SNORT packet sniffer was used for new data collection which will be passed through the inference engine to classify the operation as host-based or anomaly. The implementation algorithm used was neural networks and fuzzy logic with network profiling. Simple data mining techniques was used to process the network data to predict anomaly detection.

[7] proposed "Network Intrusion Detection System". The system like any other intrusion detection system comprises of four major components which are: Traffic Collector for gathering activity and event data for analysis; Analysis Engine that analyzes the data that the traffic collector gathered; Signature Database which is an amalgamation of signatures known to be associated with suspicious and malicious activities; Management and Reporting Interface used by system administrators to manage the system and receive alerts when intrusions are detected. The system was implemented using Java programming language, used to detect specific attacks which are Man-in-the- Middle, DOS and ping of death.

[6] proposed "Data warehousing and data mining techniques for intrusion detection systems" the aim was to improve the performance and usability of Intrusion Detection Systems (IDS). The system model network traffic and alerts using a multi-dimensional data model and star schemas to perform network security analysis and to detect denial of service attacks. A prototype of the system was successfully tested at Army Research Labs.

[9] proposed "DBMS Application Layer Intrusion Detection for Data Warehouses (DW)". They argued that the current DIDS lack capacity to detect heterogeneous oriented DW. In the research work, specific requirements for data warehouse based IDS was defined which lead to the proposal of a conceptual approach for a real-time DIDS for DWs at the SQL command level that works transparently as an extension of the Database Management System (DBMS) between the user applications and the database server itself.

In their review, [25] looked into four different approaches to intrusion detection in a network environment. The approaches are: Artificial Neural Network (ANN) "is comprised of a collection of processing elements that are highly interconnected, and convert a set of inputs to a set of desired outputs"; Self Organizing Map (SOM), Fuzzy Logic and Support Vector Machine (SVM).

Table 1. shows the summary of related literatures

Title [Ref]	**Methodology/ Tools**	**Contribution**	**Research Gap**
An efficient hybrid system for anomaly detection in social networks [5]	R language and Python machine learning packages.	DT-SVMNB that classifies users as depressed one or suicidal one in the social network	Focus was on predicting vulnerable users on the social media not hackers
Suspicious Activity Detection of Twitter and Facebook using Sentimental Analysis [26]	Sentiment analysis, NLP based Part-Of-Speech (POS) tagging	Developed a model that can analyze the opinions posted on the internet to classify them as good, bad, or neutral.	Does not cover the hackers' intrusion detection on the social media platforms
Social Media Cyberbullying Detection using Machine Learning [24]	SVM, Neural Network, TFIDF and sentiment analysis algorithms	Achieved average f-score of 91.9% and 89.8% for NN and SVM respectively on the detection.	Handled only one aspect of social media attacks – cyberbullying. Does not cover anomaly intrusion
Cybercrime Prevention on Social Media [27]	Django, NLP, NN, CNN	Developed a customized platform and implemented anomaly post detection	The intrusion detection can only detect cyberbullying attack
DBMS Application Layer Intrusion Detection for Data Warehouses [9]	Oracle 11g DBMS, TPC-H benchmark	Proposed DBMS Application Layer Intrusion Detection for Data Warehouses	Hackers with persistent attack can evade the security mechanism
Intrusion Detection System : Overview [25]	Qualitative	highlighted four different approaches to intrusion detection in a network environment	There was no proposed design
Intrusion detection system for data	MariaDB, MYSQL, TPC-H	Developed IDS for data warehouse with	Resilient hacker can use their hacking

warehouse with second level authentication [4]		second level authentication to reinforce access control security	techniques to overcome the second level authentication
Intrusion Detection System for College ERP System [13]	JDK 1.7, MYSQL, NETBEANS	Developed IDS for college ERP system	The IDS only detect inconsistent data entry to the ERP system
Network Intrusion Detection System Using Machine Learning [8]	Python libraries, CCIDS, RFE	Achieved TPR of99.9% and FPR of 0.1% on CCIDS	Decision trees have a tendency to over-fit and can create biasness
Artificial Intelligence Techniques Applied To Intrusion Detection [14]	Data mining MySQL, SNORT, Fuzzy Logic	Used data mining techniques to process the network data used to predict anomaly detection.	Rule-based models are limited by the knowledge of the expert that developed sit
Network Intrusion Detection System [7]	Java, Traffic Sniffer	used to detect specific attacks which are Man-in-the- Middle, DOS and ping of death	Focused on NID for Man-in-the- Middle, DOS and ping of death, detection algorithm not specified
Data warehousing and data mining techniques for intrusion detection systems [6]	STAR schemas, OLAP	improved the performance and usability of Intrusion Detection Systems (IDS with the star schema	The improvement was for network intrusion, not DW intrusion detection
A Conceptual Model of Social Networking in Higher Education [12]	conceptual model	Identified 4 main functions of social network usage in higher education: academic service support; student support; social and cooperate learning; and achievement representation	Social vulnerability of the network was not explored. This will affect efficient learning
A Data Mining Approach for Attribute Selection in Intrusion Detection System [28]	WEKA	Justified that attribute selection in will improve Intrusion Detection System	IDS was not developed

2.6. Research/Knowledge Gap

There is presently no developed intrusion detection system for social media platform to curtail the activities of hackers that have turned their attention to the platform. Most of the literatures reviewed do not have the intelligence to detect anomaly usage of social media account.

Role Based Access Control (RBAC), Extended RBAC, Temporal RBAC (TRBAC), Risk-based access control, etc does not have the ability to detect an attacker who obtains access to the system using some compromised credentials. Intrusion Detection System (IDS) and some other

customized security solutions for DWs have also been proposed including second level authentication. But the same mechanism used to evade the first level authentication can still be employed to overcome the second level authentication security approach. Hence fooling the attacker with fake response will provide a better reinforcement to DW intrusion detection/prevention mechanism.

Most of the previous proposals used KDD-CUP-99and DARPA 98/99 dataset for training but these datasets have become outdated with limitations over updating of new attacks [8]. Those earlier models might not work well owing to the fact that attackers changes their signatures regularly to evade detection.

3. CONCLUSIONS

Social media network has become the nerve centre of the virtual community that connects billions of heterogeneous users for mutual interaction. Because of its dynamic nature where users can share contents freely between friends and followers, hackers are seriously exploiting this rich platform for malicious intention. Various strategies for attacking social media users have been highlighted in this paper with various preventive approaches proposed by researchers. Despite all these preventive approaches, hackers activities on the platform is on the rise hence social media intrusion detection system will be highly recommended as second line of defence against the attacks of hackers on the social media network platforms.

ACKNOWLEDGEMENTS

The authors would like to thank my lecturers for their constructive criticism and impact!

REFERENCES

[1] A. E. Omolara, A. Jantan, O. I. Abiodun, V. Dada, H. Arshad, and E. Emmanuel, "A Deception Model Robust to Eavesdropping over Communication for Social Network Systems," no. Im, pp. 1–21, 2019.
[2] K. Musial and P. Kazienko, "Social networks on the Internet," *World Wide Web*, pp. 31–72, 2012.
[3] C. Timm, *Seven Deadliest Social Networks Attacks*. USA: Elsevier Inc., 2010.
[4] A. Arora and A. Gosain, "Intrusion Detection System for Data Warehouse with Second Level Authentication," *Int. J. Inf. Technol.*, vol. 13, pp. 877–887, 2021.
[5] M. S. Rahman, S. Halder, M. A. Uddin, and U. K. Acharjee, "An efficient hybrid system for anomaly detection in social networks," *Cybersecurity*, vol. 4, no. 10, pp. 1–11, 2021.
[6] A. Singhal and S. Jajodia, "Data warehousing and data mining techniques for intrusion detection systems," *Distrib Parallel Databases*, vol. 20, pp. 149–166, 2006.
[7] G. N. Prabhu, K. Jain, N. Lawande, Y. Zutshi, R. Singh, and J. Chinchole, "Network Intrusion Detection System," *Int. J. Eng. Res. Appl.*, vol. 4, no. 4, pp. 69–72, 2014.
[8] R. A. Jamadar, "Network Intrusion Detection System Using Machine Learning," *Indian J. Sci. Technol.*, vol. 11, no. 48, pp. 1–6, 2018.
[9] R. J. Santos, J. Bernardino, and M. Vieira, "DBMS Application Layer Intrusion Detection for Data Warehouses," in *Building sus- tainable information systems.*, 2013.
[10] O. Logvinov, "Standard for an Architectural Framework for the Internet of Things (IoT)," 2021. .
[11] "Social Media Attacks," 2020.
[12] P. Jucevi and G. Valinevičienė, "A Conceptual Model of Social Networking in Higher Education," *Electron. Electr. Eng.*, vol. 6, no. (102, 2010.
[13] H. Vora, J. Kataria, D. Shah, and V. Pinjarkar, "Intrusion Detection System for College ERP System," *J. Res.*, vol. 03, no. 02, pp. 69–72, 2017.
[14] B. Shanmugam and N. B. Idris, "Artificial Intelligence Techniques Applied To Intrusion Detection," in *Proceedings of the Postgraduate Annual Research Seminar*, 2005, pp. 285–287.

[15] C. F. Noonan and A. . Piatt, *Global Social Media Directory*. USA: U.S Department of Energy, 2014.
[16] E. S. Dandaura, U. M. Mbanaso, G. N. Ezeh, and U. C. Iwuchukwu, "The Use of Social Networking Service among Nigerian Youths between Ages 16 and 25 Years," 2015.
[17] J. Bagadiya, "367 Social Media Statistics You Must Know In 2021," *Social Pilot*, 2021. [Online]. Available: https://www.socialpilot.co/blog/social-media-statistics. [Accessed: 12-Oct-2001].
[18] C. Timm, *Seven Deadliest Social Network Attacks*. USA: Syngress Publishing, Inc, 2010.
[19] C. Noonan and A. Piatt, *Global Social Media Directory*, no. October. USA: U.S. Department of Energy, 2014.
[20] H. Wilcox and M. Bhattacharya, "A Human Dimension of Hacking : Social Engineering through Social Media," in *IOP Conference Series: Materials Science and Engineering*, 2020.
[21] C. Suggs, "Hacking Social Media."
[22] J. Patterson, "Hacking: Beginner to Expert Guide to Computer Hacking, Basic Security, and Penetration Testing (Computer Science Series)."
[23] L. Wang, "Big Data in Intrusion Detection Systems and Intrusion Prevention Systems," *J. Comput. Networks*, vol. 4, no. 1, pp. 48–55, 2017.
[24] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social Media Cyberbullying Detection using Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 703–707, 2019.
[25] H. O. Alanazi, R. Noor, B. B. Zaidan, and A. A. Zaidan, "Intrusion Detection System : Overview," *J. Comput.*, vol. 2, no. 2, pp. 130–133, 2010.
[26] S. Al Mansoori, A. Almansoori, M. Alshamsi, S. A. Salloum, and K. Shaalan, "Suspicious Activity Detection of Twitter and Facebook using Sentimental Analysis," *TEM J.*, vol. 9, no. 4, pp. 1313–1319, 2020.
[27] B. Shah, N. Sharma, and S. Bandgar, "Cybercrime Prevention on Social Media," *Int. J. Eng. Res. Technol.*, vol. 10, no. 03, pp. 509–513, 2021.
[28] R. Pandey and J. Pant, "A Data Mining Approach for Attribute Selection in Intrusion Detection System," *Int. J. Comput. Appl.*, vol. 172, no. 1, pp. 11–14, 2017.

AUTHORS

Emmanuel Etuh is currently pursuing a PhD in Computer Science at the University of Nigeria, Nsukka, He obtained his first degree certificate in Computer Science from Kogi State University, Anyigba in 2009 and an MSc degree in Computer Science from Ahmadu Bello University, Zaria in 2014, His research interest include Artificial Intelligence, CyberSecurity, and Software Engineering.

Professor Bakpo, Francis Sunday received his M Eng. degree in Computer Science and Engineering from Kazakh National Technical University, Almaty (formerly USSR) in 1994 and PhD degree in Computer Engineering in 2008 from Enugu State University of Science and Technology (ESUT), Agbani. He joined the Department of Computer Science at University of Nigeria Nsukka as lecturer II in June 1996 and further progressed from the rank of lecturer II to Professor in 2010. His current research interest includes Computer architectures, computer communications networking, artificial neural network applications, and Petri net theory. He is dully registered professionally as member Nigeria Computer Society, MNCS and Computer professional of Nigeria, MCPN, respectively and has also published a number of excellent journal papers, books and conference proceeding papers in his field.

Eneh, Agozie H is a Senior Lecturer in the Department of Computer Science at the University of Nigeria, Nsukka. His Area of Specialization include: authentication protocols analysis, network security, optimisation theories, medical informatics, and assistive technologies for educating children and adolescents with learning difficulties.

AUTHOR INDEX