A publication of ISCA*:
International Society for Computers
and Their Applications

INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS

TABLE OF CONTENTS

***"International Journal of Computers and Their Applications is Peer Reviewed".**

Volume 30, No. 4, Dec 2023 **ISSN 1076-5204**

International Journal of Computers and Their Applications
A publication of the International Society for Computers and Their Applications

ISCA Headquaters • 278 Mankato Ave, #220, Winona, Minnesota 55987, USA
Phone 507-458-4517 • Email isca@isca-hq.org • URL: http://www.isca-hq.org

Guest Editorial
December 2023

This issue of the International Journal of Computers and their Applications (IJCA) has gone through the normal review process. The papers in this issue cover a broad range of research interests in the community of computers and their applications.

IJCA Contributed Papers: As mentioned earlier, this issue comprises papers that were contributed to the International Journal of Computers and their Applications (IJCA). The topics and main contributions of the papers are briefly summarized below:

CH. RAM MOHAN REDDY, D EVANGELIN GEETHA, R.V. RAGHAVENDRA RAO, and K SAILAJA KUMAR from B. M. S. College of Engineering, India & Ramaiah Institute of Technology, India present their work "Performance Prediction for Web Services Based on Dynamic Workload: A Simulation Approach." This paper addresses the challenge of predicting the performance of web services due to their variable workloads and execution environments. It proposes a methodology with a mathematical model to estimate web service response times over time. The approach considers resource configurations and workload fluctuations and includes a case study using the Travel care application. The tool SMTQA is employed for sensitivity analysis to identify bottleneck resources and enhance web service performance while establishing performance objectives.

ING. GOFUR HALMURATOV, and DOC. ING. ARNOŠT VESELÝ, CSc from Czech university of life sciences, Czech Republic presents their work "Evaluating Image Quality through Latent Space Analysis of Autoencoders – kopie." In this, the author discussed challenges in deep learning-driven image classification with limited datasets, which often require manual image quality assessment due to data augmentation and GANs. They introduced a novel method using autoencoders' latent space for automated image quality evaluation, surpassing traditional techniques in experiments.

SUSAN F. ELLAKWA, ABDELRAHMAN ELSAYED, RASHA ELNEMR, SAHEIR F. EL-LAKWAH, A. M. AZAZY from Agricultural Research Center, Egypt presents their work "A system for Identifying Entomopathogenic Nematodes." This paper addresses the need to identify Entomopathogenic Nematodes (EPNs) species as an alternative to chemical pesticides for insect control in agriculture. It proposes a system that uses Web Ontology Language (OWL) to represent EPNs species' semantic information, facilitating data exchange and interoperability. The system is implemented as a mobile application, enabling users to extract and retrieve EPNs data from the ontology, providing details on valid Heterorhabditis and Steinernema species and finding related species based on infective juveniles (IJs) features. Similarity search techniques like cosine similarity and Euclidean distance are employed, demonstrating the system's ability to recognize known EPNs species and identify similar ones.

ABDULLAH ALI BEN-NAKHI, KALIM QURESHI from Kuwait University, Kuwait and MOSTAFA ABD EL-BARR from Badr University, Egypt presents their work "Threat Modeling of IoT-based Smart Home Systems." The study explores the broad applications of the Internet of Things (IoT), with a focus on smart home automation. While IoT devices are prevalent in various sectors like healthcare, agriculture, transportation, and manufacturing, security remains a significant concern. The research employs Microsoft STRIDE, a threat modeling tool, to categorize and analyze threats in a generic smart home system design. It compares threats generated by STRIDE with those collected from a literature review, providing a comprehensive overview of system vulnerabilities beyond individual components or attack types.

ZEROUAL DJAZIA from University of Biskra, Algeria, OKBA KAZAR from United Arab Emirate University, UAE, SAAD HAROUS from University of Sharjah, and SABER BENHARZALLAH from

Batna 2 University, Algeria presents their work “Collaborative Cloud-V. Edge System for Predicting Traffic Accident Risk Using Machine Learning Based IOV.” The author discusses the profound impact of cutting-edge technologies like ICT, AI, and IoT on smart city development, particularly in the context of intelligent transportation systems (ITS). They propose an intelligent driving system that utilizes machine learning to predict road accident risks and alert drivers to avoid dangerous situations. While current Internet-of-vehicle (IOV) solutions rely on the cloud, the author introduces the concept of vehicular edge computing to address limitations like Internet disconnection and response time. The proposed framework, ICEDAS, combines cloud-based and edge-based machine learning models to enhance crash prediction and prevention, with promising results in terms of efficiency and road safety improvement.

HONG-LAM LE, THANH-TUOI LE, THI-THU-HIEN VU from Vinh University of Technology Education, Vietnam, DOAN-HIEU TRAN from Ho Chi Minh University of Banking, DINH VAN CHAU from Electric Power University, and THI-THU-TRANG NGO from Posts and Telecommunications Institute of Technology, Vietnam present their work “Evaluation of the Impact of Hyperparameters on random forest performance based on accelerometer data.” This study focuses on the impact of hyperparameters on the performance and training time of a random forest algorithm in fall detection systems, which has been relatively unexplored in previous research. The findings reveal optimal ranges for hyperparameter values to achieve high performance. Additionally, the study identifies specific combinations of hyperparameters that can either improve or diminish the random forest's effectiveness compared to default settings. Experiments were conducted using datasets from accelerometers in smartphones and wearables (MobiAct v2.0 and UP-Fall), offering valuable insights for optimizing hyperparameters in the development of more efficient fall detection systems.

MONALISA DEY, SAINIK KUMAR MAHATA from the University of Engineering and Management, India, and DIPANKAR DAS from Jadavpur University, India present their work “Exploring summarization of scientific tables: Analyzing data preparation and extractive to abstractive summary generation.” This research paper addresses the challenges of extracting meaningful information from extensive textual data, with a specific focus on tables found in scientific papers, which often contain vital but complex data. The scarcity of training data for such systems is a hurdle, prompting the creation of a high-quality corpus comprising both extractive and abstractive summaries derived from table content. Two approaches, rule-based and template-based, were used to build this dataset, validated using automated and manual metrics. Two models, T5 and Seq-to-Seq, were then trained on the dataset to generate abstractive summaries from extractive ones, with evaluation measures like BLEU, ROUGE-L, Adequacy, and Fluency indicating that the dataset can produce coherent abstractive summaries, yielding promising scores of 58.2 and 0.31 for the fine-tuned T5 model and 16.25 and 0.18 for the Seq-to-Seq model.

SIRINDA PALAHAN from University of the Thai Chamber of Commerce, Thailand presents his work “Improving Access to Trade and Investment Information in Thailand through Intelligent Document Retrieval.” This paper introduces a chatbot system that combines natural language processing and information retrieval techniques to simplify overseas investment and trade information retrieval. The system identifies the most relevant content, making it easier for users, especially beginners, to navigate this complex domain. The methodology employs the BM25 model and deep learning to rank and retrieve documents, reducing noise and improving accuracy. Experiments with Thai queries show the system's effectiveness, and a user satisfaction survey confirms its helpfulness, suggesting it as a valuable tool for Thai entrepreneurs dealing with foreign trade and investment.

AHMED AL-NAKEEB, MOUNIR EL KHATIB, ABDULRAHMAN ALHOSANI, IBRAHIM ALHOSANI from Hamdan Bin Mohammed Smart University, U.A.E, and RAED ABU ZITAR from Sorbonne University, U.A.E presents their work “Project Manager Role in Manage Project Knowledge process – An Approach to enhance project quality." This research investigates the integration of knowledge management into project management in UAE organizations, emphasizing the impact on project quality and innovation. It underscores the importance of managing project knowledge throughout the project lifecycle, with project managers leading lessons-learned sessions. A robust knowledge repository and a culture of continuous improvement are highlighted as crucial for organizational excellence. The study also explores the role of Project Management Information Systems (PMIS) in decision-making and risk management within knowledge management. The findings emphasize the link between knowledge management and project success, advocating for the integration of diverse systems and tools to enhance knowledge utilization and overall project performance.

LE VINH QUANG, TRAN HUU ĐUC, NARAYAN CHANDRA DEBNATH from Eastern International University, Vietnam and Nguyen Ngoc Long from Industrial University Ho Chi Minh City presents their work "Managing Risks in the Adoption of Cybersecurity Technology in Manufacturing Enterprises: Identification and Assessment." This research addresses the pressing need for cybersecurity technology adoption in manufacturing enterprises that increasingly rely on digitalization and connectivity. It identifies and evaluates risks associated with adopting cybersecurity solutions using the Analytic Hierarchy Process (AHP). The analysis prioritizes investment areas and mitigation strategies in manufacturing, considering risks related to strategy, organization, technology, finance, and human factors. The findings provide valuable insights for manufacturing organizations looking to enhance cybersecurity, safeguard assets, maintain production processes, and protect sensitive data, contributing to knowledge in the field of cybersecurity adoption within the manufacturing sector.

As guest editor, we would like to express our deepest appreciation to the authors and the reviewers. We hope you will enjoy this issue of the IJCA. More information about ISCA society can be found at http://www.isca-hq.org.

Guest Editor:
Ajay Bandi, Northwest Missouri State University, USA

December 2023

Performance Prediction for Web Services Based on Dynamic Workload: A Simulation Approach

Ch. Ram Mohan Reddy*, D. Evangelin Geetha*, R. V. Raghavendra Rao*, and K. Sailaja Kumar*
B. M. S. College of Engineering, Bangalore-19, INDIA
Ramaiah Institute of Technology, Bangalore - 54, INDIA

Abstract

The increase in the use of web services has rendered its workload to be highly uncertain. The workload and the execution environment of web services affect their performance since they are non-uniform and unpredictable. Hence, estimation techniques are required to predict the performance of web services considering their dynamic workload. This paper presents a methodology to estimate the response time of web services over a given time horizon by providing a mathematical model. The deployment environment is analyzed by considering the configuration of the resources and the workload that fluctuates over a period of time. A case study on Travel care application is presented to illustrate the methodology. The tool SMTQA (Simulation of Multi-Tier Queuing Applications) is used to carry out sensitivity analysis on the configuration of resources so that the behavior of the resources can be analyzed. This analysis helps to improve the performance of web services by identifying the bottleneck resources. Moreover, it provides a possibility to determine the performance objectives.

Key Words: Web services, software performance engineering, performance goal, dynamic workload, mathematical model, sensitivity analysis, SMTQA.

1 Introduction

One of the most significant aspects of Quality of Service (QoS) is performance. It is critical to evaluate the aspects that lead to client satisfaction in the efficient delivery of various services. Certainly, improving these factors is a challenge in web services since the workload from the client-side is uncertain and ever-changing. The performance of web services can be analyzed in several ways. It is necessary to analyze the workload to meet users' performance expectations.

Composite services are a typical aggregation of the complex process. In the context of web services, a composite web service is considered as a single logical unit [10]. The performance of the composite web services encompasses the effective resource usage and sharing of the workload in a service-oriented environment. It is difficult to know the workload of the composite web services. Any online service offered on the internet can be evaluated by looking at four key factors: number, service quality, complexity, and function diversity [11].

Two important performance dimensions for web applications are responsiveness and scalability [2, 18]. In the current scenario, the web users are too busy to wait for a slow responsive system; hence, responsiveness is significant. Scalability is important to maintain responsiveness, as more and more users converge on a site. There is also significant capacity planning concerns for web applications, such as the selection of the number of processing nodes, the number of processors for each node, the speed of the processors, and so on. Other significant performance tuning issues also have to be addressed for responsive web applications. One of the critical issues for the performance of web services is the issue of balancing the workload of computational tasks among the different nodes comprising the system [15].

Early in the software development process, Software Performance Engineering (SPE) provides numerous approaches for analyzing the performance of software systems. It is difficult to obtain reliable early estimations because complete information about the future system is lacking at this time. However, to establish the feasibility of a software system in terms of performance analysis, early estimates are essential [18-19].

A methodology to analyze the behavior of the hardware resources based on the dynamic workload is proposed in this paper. The methodology provides a basis for calculating the response time over a given time horizon based on the demand of the activities of the web services. Moreover, the methodology helps to define the performance goal as well.

2 Related Work

As discussed in the previous section, the non-uniform workload of web services has a strong impact on system performance. Some literature available in this context is reviewed.

The Customer Behavior Model Graph (CBMG) is a graph that describes patterns of customer behavior in e-commerce site workloads [14, 20]. The impact of a more realistic dynamic workload on online performance measures is investigated in [16]. The analysis is done by carrying out an experimental study on an e-commerce scenario with a dynamic workload. The obtained results are compared with the traditional workloads. In multi-tier web service systems, a soft resource allocation approach is proposed to handle dynamic workloads in real-time [25]. The authors have formulated the whole system by queueing the network model. To cope with dynamic workloads and to meet performance demands, an optimization approach based on sliding windows

* Emails: prof.crmr@gmail.com, degeetha@msrit.edu, rvraghavendrarao76@gmail.com, sailajakumar.k@gmail.com.

is presented. Experimental measurements are used to validate the effectiveness of the optimization method, as well as the model parameters and performance indicators. The author of [23] proposes a method for both economical and robust provisioning of resources for N-tier web applications. The authors tested the method using three different workload models in the RUBiS web application benchmark: open, closed, and semi-open. The authors claimed that their approach was flexible enough to accommodate a wide variety of workloads with varying resource demands without requiring reconfiguration.

Web service compositions are emerging to support industries by becoming more dynamic and delivering customized services to users. Dynamic web service composition approaches as the foundation of problems like transaction support, compositional correctness, etc. are compared in [9]. In [3] the performance of the clinical decision support system is shown based on a web service. The layered queueing network model is followed to design the software architecture of the system, and the systems performance goals are obtained by solving the model analytically. E-commerce applications are evaluated using the queuing network model in [8]. Through the use of techniques like Layered Queuing Network (LQN) models and SPE, performance parameters such as response time, utilization, and throughput have been verified with actual measurements. In [21] an architecture behavior designed for varying demands is discussed that are placed on the server. The paper discusses the experimental studies on an infrastructure behavior that is devised for variant loads placed on the server.

Web services performance evaluation can be done at two levels. One is at the client-side which is a direct interface with the user and the other at the server-side. The performance has to be very good at the client-side, and various factors that contribute to the effective performance of web services have to be considered as in [12]. The performance of a Web service has to be estimated properly when we integrate different web services to make a new one. Different web services can be combined based on simplicity, interoperability, flexibility, and reuse of services. Different toolkits are used to compare the performance of Integrated Web Services. The authors describe a new concept of creating web services, based on existing services to enhance the performance of a web service, also called a Hybrid approach [4].

To handle the growing client and varying workload, a session-based workload and reliability analysis is presented in [22]. The authors introduced intra-session and inter-session metrics. Describing workload and analyzing characteristics of web errors by estimating request-based and session-based web server reliability to indicate user perception is discussed. Performance analysis using a measurement-based performance analysis is explained with an example of an e-commerce application that uses a web service component in [5]. According to the setup established, higher workload intensities increase the response time of the web services in comparison with any other component. A methodology for selecting an appropriate execution environment based on dynamic workloads over a given time horizon is discussed in [7]. The authors propose a mathematical model for calculating the performance which facilitates decision-making for a given distributed application by calculating its performance. It also simulates the model proposed for the same.

In the literature, the early estimation of performance parameters for web services is not addressed. In this paper, a methodology to estimate the performance metrics, such as average response time, average service time, average waiting time, and probability of idle server and dropping of sessions is proposed considering the dynamic workload of web services. Furthermore, the methodology provides an opportunity to determine the performance goal.

3 Methodology

The performance of web services highly depends on the configuration of the resources and service requirements of various services and in particular, the demand for them. Hence, precise performance analysis is necessary to estimate the expected demand for the web services and in turn for the corresponding servers. It is acknowledged that the workload of a web service in a time horizon is particularly important for identifying the capacity of the deployment environment and the allocation of web services to web servers. Subsequently, a probabilistic technique is necessary to find out the workload in a fixed time horizon. This type of analysis in a fixed time horizon helps to find out the adequacy of web servers and application servers and also to assess the alternatives in the deployment environment so that the performance goal can be achieved.

Fluctuations in workload may have an impact on the performance due to contention for resources during the execution of the web services. Moreover, it may lead to 'drop the requests' also. Hence, an estimation mechanism has to be provided to determine the utilization of the web servers. The estimation is based on the design of the deployment environment and the workload during the time horizon. Suitable configuration for the servers can be determined by analyzing the deployment environment considering different configurations for the servers.

Considering all the above aspects, a methodology is proposed in this paper for estimating the performance of service-oriented web services over a time horizon considering aspects of WA/WS. The methodology aims to identify whether the proposed configuration of the deployment environment of web servers is adequate to satisfy the customers based on the required demand or not.

Let $i \in [i_0, I+ i_0-1]$ be the interval, where i_0 be the first interval in the time horizon that is chosen; *I* will be the number of intervals. Let A_T, be the set of activities that are expected to be processed during the 'I' intervals and each activity may trigger a particular web service of a composite service. Based on these assumptions, the methodology has been devised as follows.

- Modeling the workload of web service components using a mathematical model over a given time horizon.
- Modeling the servers and other hardware resources in the deployment environment of WS architecture in SOA.
- Formulating a procedure to calculate the response time proportional to workload.
- Prediction of performance and analyzing the behavior of

the resources across the servers of web services.

- Identifying bottleneck resources and improving the performance by carrying out sensitivity analysis on resources of composite web services.

3.1 Modeling the Workload of Composite Web Services

A workload element is a unique workload request that is generated for a given web service and must be handled by the appropriate composite web server. The workload specification covers service utilization or requests for service functions, as well as the likelihood of requests arriving and request patterns [1]. The activity diagram of Unified Modeling Language (UML) helps to model the flow of activities of the web services, and it is available during the preliminary design phase. A sequence of actions involved in the activity is the scenario of a service. The activity scenarios define the desired behavior of web service and show the execution patterns of a composite service. A scenario of service illustrates the interaction between the objects or execution of the activities at a particular time. The execution of a specific activity scenario depends on the type of user's request (event). Web services are geographically distributed, and the number of users of the system tends to vary from time to time (for example, the travel agent service, derived from [24], is a good example of how this works. A user accesses a single interface, entering the information needed to book a flight, book a hotel room, and obtain maps from the area. Each of these three obligations will be met by its web service. The problem here is that the user may submit some information that only needs to be used by certain parts of the service. For instance, a credit card is involved with airline and hotel reservations, but a website can provide maps for free. As a travel agency customer, one would want to know whether the credit card information will be secured and accessible to only necessary vendors). As a result, in a service-oriented environment, the type of requests arriving at a particular time interval is unpredictable. Hence, workload scenarios can be built based on the number and type of expected activities across a certain time horizon interval.

Let F_I be the collection of activities that represent the web service functionality that will occur throughout I time intervals, and R_i denotes the number of requests that can arrive in the interval $i \in [i_0, I + i_0 - 1]$.

Each request is characterized by:

- $p_{i,r}$– the likelihood of the request r occurring at a given time interval i
- $D^a_{i,r}$ – expected demand for the activity $a \in F_I$, if the request r arrives among those specified at the time interval i

The arrival of the request at period i and the scenario that occurred in period i-1 are used to define the workload of a certain service of a composite web service at period i. Let S_i denote the number of demand scenarios for a service that will occur in each interval i, with $p_{i,s}$ denoting the likelihood that scenario s will occur at interval *i*.

During the first interval i_0, the number of scenarios can be calculated as:

$$S_{i_0} = R_{i_0} \quad (3.1)$$

The number of scenarios in the following interval $i \in [i_0, I + i_0 - 1]$. can be computed recursively as:

$$S_i = R_i \,.\, S_i\text{-}1 \quad (3.2)$$

The probable situations at a time interval i depend on the scenarios at a time interval i-1, and the requests come at the time interval i for a certain composite web service.

The conditional probability tree is used to depict workload scenarios in the service-oriented environment because the occurrence of a service scenario at any time interval conditionally depends on the preceding scenario and the arrival of requests in that scenario. Figure 1 depicts the layout of the workload scenarios.

Let $p_{i,s}$ be the probability that the scenario of a service *s* can be utilized in the period *i*, then

$$p_{i,s} = p_{i,r} \cdot p_{i-1,v} \quad (3.3)$$

where, $p_{i,r}$ be the probability that the request r occurs in the period *i* and $p_{i-1,v}$ be the probability that the scenario *v* occurs in the period *i-1*. The demand $D^a_{i,s}$ is computed as:

$$D^a_{i,s} = p_{i,s} \forall \quad (3.4)$$

Where $D^a_{i,s}$ is the demand for activity a in time interval t respond to the workload scenario s.

Furthermore, the state (activity) at the interval *i-1* and the request arrived at *i* determine the state of the composite service in time *i*. Hence, the group of activities to be executed at time *t* are considered as state of the web service at the time *i*, and the pattern of processing the activities are modeled as the State Chart Diagram of UML. The representation of the workload scenarios is presented as a statechart diagram in Figure 2.

3.2 Modeling of Execution Environment of a Composite Web Service

The resources in the execution environment and workloads are closely related to each other. The configuration of the resources available for individual web services as well as the workload for the web services influences the performance of the composite web service.

The software architecture that includes hardware and software components becomes the base for the deployment environment of each web service. The features of the execution environment are specified by a set of attributes that categorize each resource in the environment. For analyzing the deployment environment, a different set of attributes for resources can be considered to obtain the performance metrics of those resources.

For each hardware resource, the following set of attributes are considered:

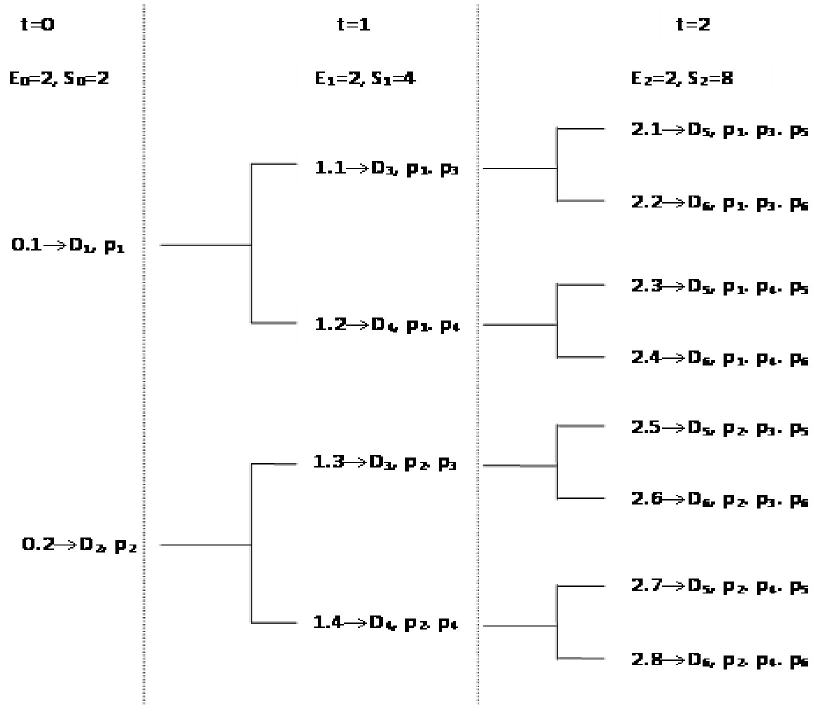


Figure 1: Workload scenarios – conditional probability

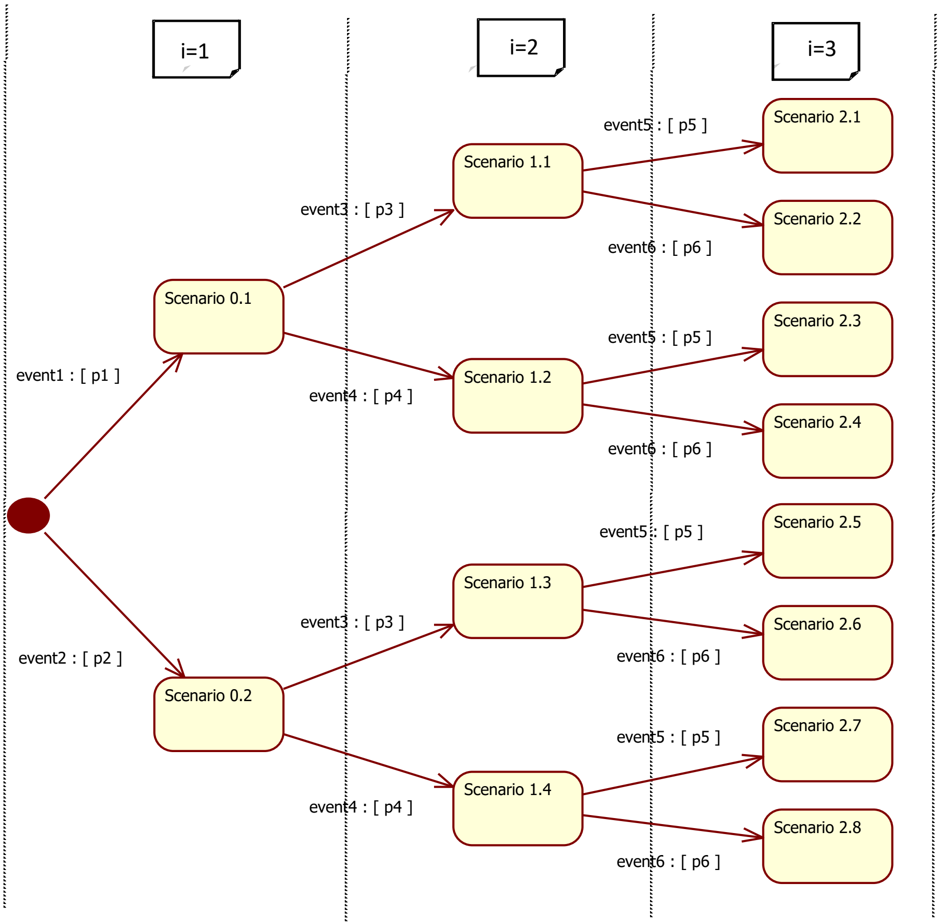


Figure 2: Workload Scenario - Statechart Diagram

- The configuration
- A collection of activities belonging to F_I that the resource can carry out
- The number of resources used by each activity

Based on the resource parameters that are to be specified, a set of values can be assigned to define the base configuration for the resources. For example, consider a resource R_1 with the configuration of 1000 KB/sec as the processing speed of a web service. Let a_1, a_2, a_3, be the activities that require the service of the resource R_1 and the resource usage for the activities be 5 seconds, 8 seconds, and 4 seconds, respectively. Another alternative type of data can be obtained by changing any of the combinations of these attributes; for example, the processing speed changing to 10000 KB/sec or the processing time for the activities to 2 seconds, 8 seconds, and 5 seconds, respectively.

3.3 Calculation of Response Time

The resource utilization of each web service must be analyzed to determine the composite web service's optimal and predefined response time over the provided time horizon. This can be made possible by calculating the resource utilization for each time interval. The configuration and the fluctuating workload scenarios lead to a difference in resource usage so that each configuration $c \in IC$ is characterized by a unique set of parameters of the resources of that web service.

Let PT_a^j be the processing time required for the activity, $a \in F_I$ for a given configuration *j*. Let $T_{i,s}^j$ be the total processing time in time interval *i* to respond to the workload scenarios *s*. Then the total processing time during the period *i* can be calculated as:

$$T_{t,s}^j = \sum_{a \in F_I} \left(D_{i,s}^a * PT_a^j \right) \qquad (3.5)$$

4 Illustration of the Methodology

For a precise illustration of the methodology, we have taken a case study on the Travel Agent web application/web service [13]. Users can use the Travel agency's website to search for available airlines, hotels, and cars that suit their search parameters, make reservations, and pay for booked services, among other things. The Travel Agency web application makes use of a local database that stores customer information as well as tourism-related data. Moreover, four external web services are used by the Travel Agency web application: the airline web service provider which provides services like flight availability, the hotel online service provider publishes a news bulletin and accepts room reservations, the vehicle web service provider conveys availability and booking information, and an independent online payment web service provider collects payments. Not only does the Travel Agency use web services, but it also exposes some of its functionality as a web service. The UML models developed for the Travel Agency web application are given in the following sections.

4.1 Modeling the Scenarios of Workload

The kind of queries that the application receives are determined by its functionalities. The overall activity of the travel care application is presented in Figure 3 with a help of a use case diagram.

Figure 4 depicts the flow of activities carried out in the application for any type of reservation made through travel care. The action begins with a login to the travel care system using the credentials created and a booking query (airline, car, and hotel). Once the type of reservation is chosen, the travel care selects the service providers and connects to the appropriate network, then completes the reservation by selecting and providing all of the necessary details, confirms the transaction by making an online payment, and ends the transaction by reconnecting the network.

The use case diagram of the Airline Reservation is shown in Figure 5 with various scenarios. The scenarios or services that are frequently used by the customers are Login for a particular travel site for reservation, Searching a flight, Seat selection, Providing booking details, Confirming reservation, Canceling a reservation, Payments, etc.

The activity diagram shown in Figure 6 describes the flow of actions carried out by a customer who reserves a flight ticket. The activity begins with login in into the travel agency by entering the credentials. Based on the destination planned, the flights are searched and selected by reserving the class, number of seats, and providing the passenger details. On confirmation of the reservation, the payment is made based on the seat count and the transaction is ended by generating the bill for the payment.

The use case diagram for Car Reservation is shown in Figure 7 that includes the following use cases: Registering and login for a particular travel reservation, selecting the location of car provider, checking the availability of car based on the particulars, selecting the car by providing details such as rent date and time, duration and number of cars, confirming and cancelling the booking, making payment on confirmation.

The activity diagram for Car Reservation is presented in Figure 8 which starts by selecting the location of the car providers nearer to the customer's location. The customer then checks for the availability of cars by providing certain particulars. If available, the customer reserves a particular type and number of cars. The customer confirms the reservation and provides personal details, makes the payment, and terminates the reservation process.

The use case diagram of the Hotel reservation is shown in Figure 9. It includes the following functionalities: Registering and login into a travel agency, selecting a location for a customer's choice, selecting the hotel in the location and checking for availability of rooms, selecting the type and number of rooms, confirming and cancelling booking, making payment. The actors include travel agencies, hotel employees, banks, and customers.

The hotel reservation process is presented in Figure 10, which starts with the selection of the location of customers' choice by viewing the map. The hotel in the selected location is checked for the availability of rooms. If available, the customer selects the room type and room count and then proceeds to book by providing check-in, check-out dates, and personal details. The booking is completed by making an online payment for the reserved room.

The conditional probability tree and the statechart diagram for Travel Care, Airline Reservation, Hotel Reservation, and

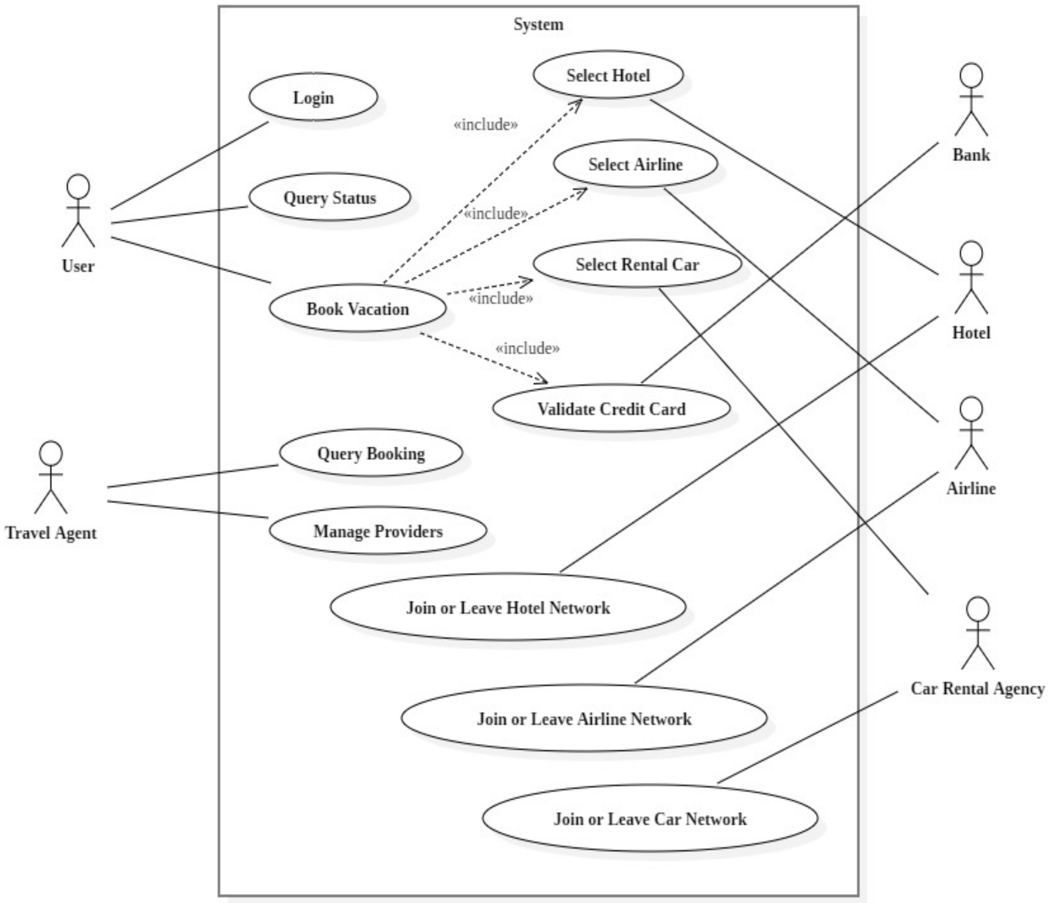


Figure 3: Use case model of travel care

Car Rental are given in Figures 11 to 15. These diagrams are used to represent the expected scenario of activities during each of the time intervals. The occurrence of these activity scenarios is dependent on customers' requests during the specified time interval and the probability of their occurrence.

4.2 Modeling the Execution Environment

The deployment environment of the Travel Care is given in Figure 16. The execution environment of the application is modeled using the deployment diagram of UML. The data for the elements of execution are given in Table 1.

4.3 Response Time Calculation

The proposed model is simulated by considering ten-time intervals, for a given time horizon t, t∈[10, 12]. To calculate the total processing time for the scenarios, the following information is needed as given in equation 3.5: i) Probability of the activity occurring in a given time interval t. ii) the processing time of each activity. Uniform distribution is used to generate the probability of occurrence of the activities. The execution time of the activities is estimated using the methodology discussed in [17]. The activity point performance prediction approach is used to calculate the activity points in this methodology. The size of the activity is calculated in terms of Lines of Code (LoC) using the gearing factor and in turn activity size in kilobytes is obtained from LoC. The processing time for the activities is calculated from the software size taking into consideration the given deployment environment. The estimated response time for the activities is calculated and tabulated in Table 2. The processing speed of the hardware resources is tabulated in Table 1.

For example, the probability of the occurrence of activities considered during the time interval t_1 is 0.9. To process the activities that are to be handled during this time interval, the time required is estimated as 0.355 seconds. Then during the time interval t_1, the time required for processing this scenario can be computed by applying Equation 3.5 as:

$$T_{t,s}^{j} = 0.9 * 0.355 = 0.32 \sec$$

The maximum, minimum, mean and execution time, variance, and standard deviation are calculated and tabulated in Table 3 for each of ten intervals (i_1, i_2, i_3 … i_{10}), as well as for the total time horizon (T* - the sum of response time acquired in intervals i_1 to i_{10}). Smith [18-19] provides the following reasons for estimating the minimum, maximum, and average response times:

i. It aids in the study of the best-average-worst case analysis.

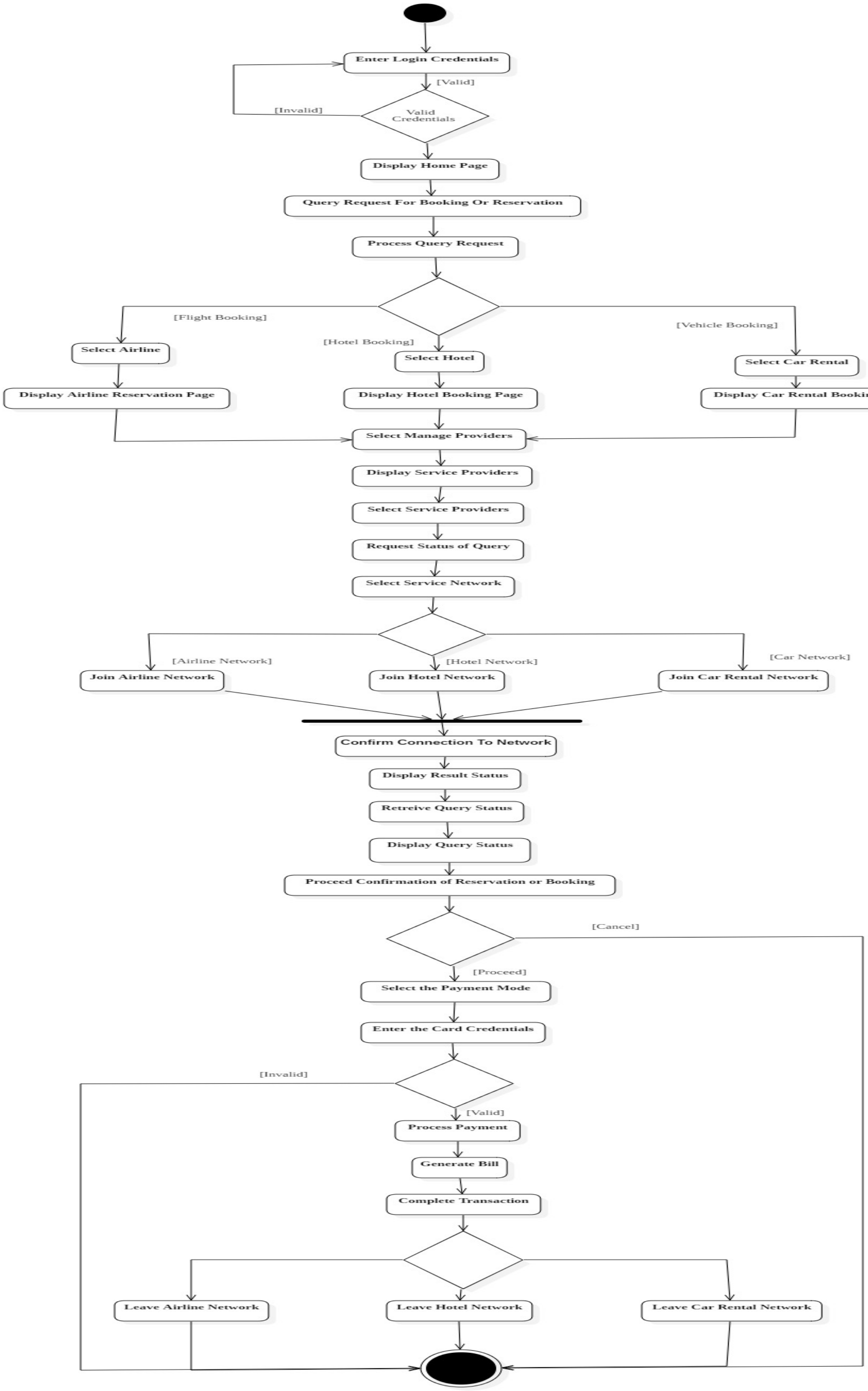


Figure 4: Activity model of travel care

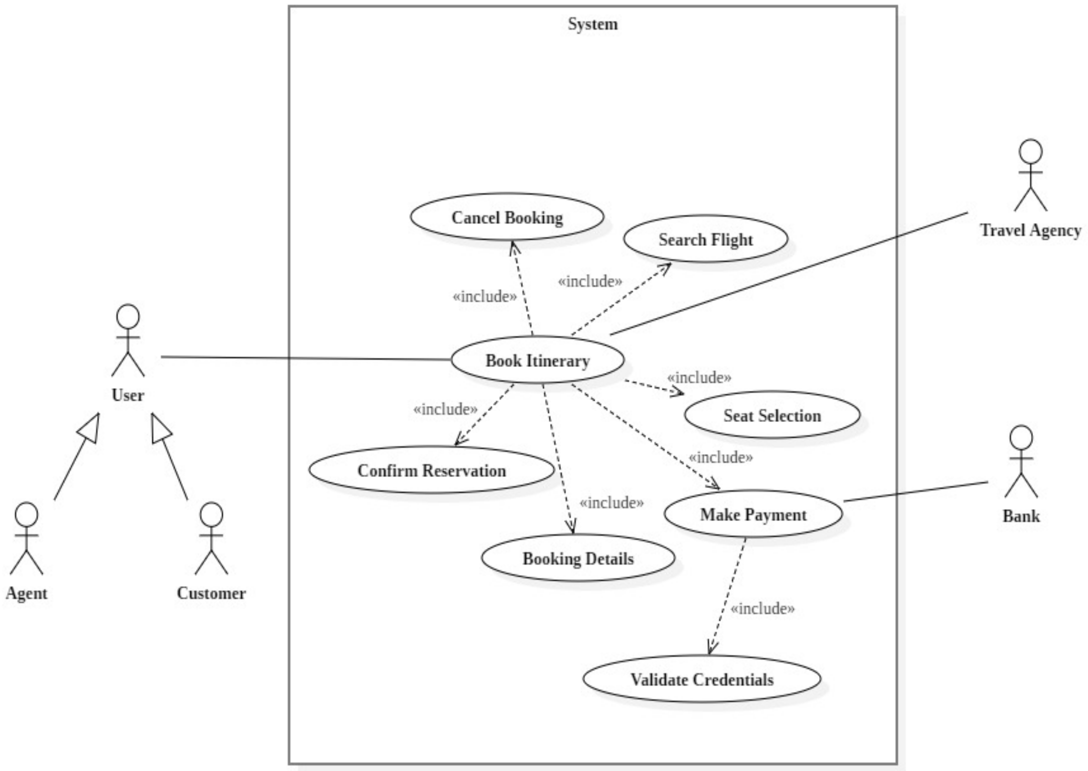


Figure 5: Use case model of airline reservation

ii. The obtained maximum response time in various trials aids in determining the application's performance goal

As shown in Figure 17, a graph is generated to display the response time acquired for ten intervals. Moreover, for the response time obtained during time intervals i_1 to i_{10}, graphs are generated and presented in the figures, from Figure 18 to Figure 27.

The observations are:

The maximum response time in interval t_1 is less related to other intervals. The only reason for this is login-related actions; user authentication can happen during this time. Furthermore, these activities are the interactions between the user and the travel reservation. The variance and the mean response time are comparably higher in the intervals t_4 to t_7, because the activities that occur during this interval need more execution time. This is due to the activities of the interactions between the travel reservation and the application servers, namely, airlines, car rental, and hotel reservation; these interactions are communicated through protocols service. The total maximum response time is obtained as 5.721 seconds, for the intervals considered.

4.4 Simulation Results

Simulation is carried out with 500 trials, where 30000 data that represent requests for web service are considered for each trial. Uniform distribution is used to generate the probability of occurrence of the requests. The maximum response time is obtained in the range of 4.73 to 5.82 seconds irrespective of the workload during the intervals t_1 to t_{10}. Moreover, a negligible difference is observed in the mean response time value. Graphs are generated for understanding the fluctuations visually. Sample graphs are presented in Figures 17, 18, and 19 for intervals 1, 3, and 4, respectively.

The preceding observations lead to a conclusion that irrespective of the workload during the given time horizon consisting of ten-time intervals, the required processing time for executing the activities is 5.82 seconds. This is the maximum time taken for the given configuration in the deployment environment of web services.

The recommendations that can be suggested are:

- The response time obtained as the maximum can be considered as the minimum required processing time to process the activities of the web service during a heavy workload since we have not considered the congestion delay. Therefore, this maximum response time value with probable congestion delay may help to define the performance goal of a composite web service.
- While a huge number of users using the system, a user cannot expect to receive the response within 5.82 seconds.
- If the performance objective is 5.82 seconds or above, then the configuration of the resources in the execution environment can be chosen as the values given in Table 1 (configuration C_1).

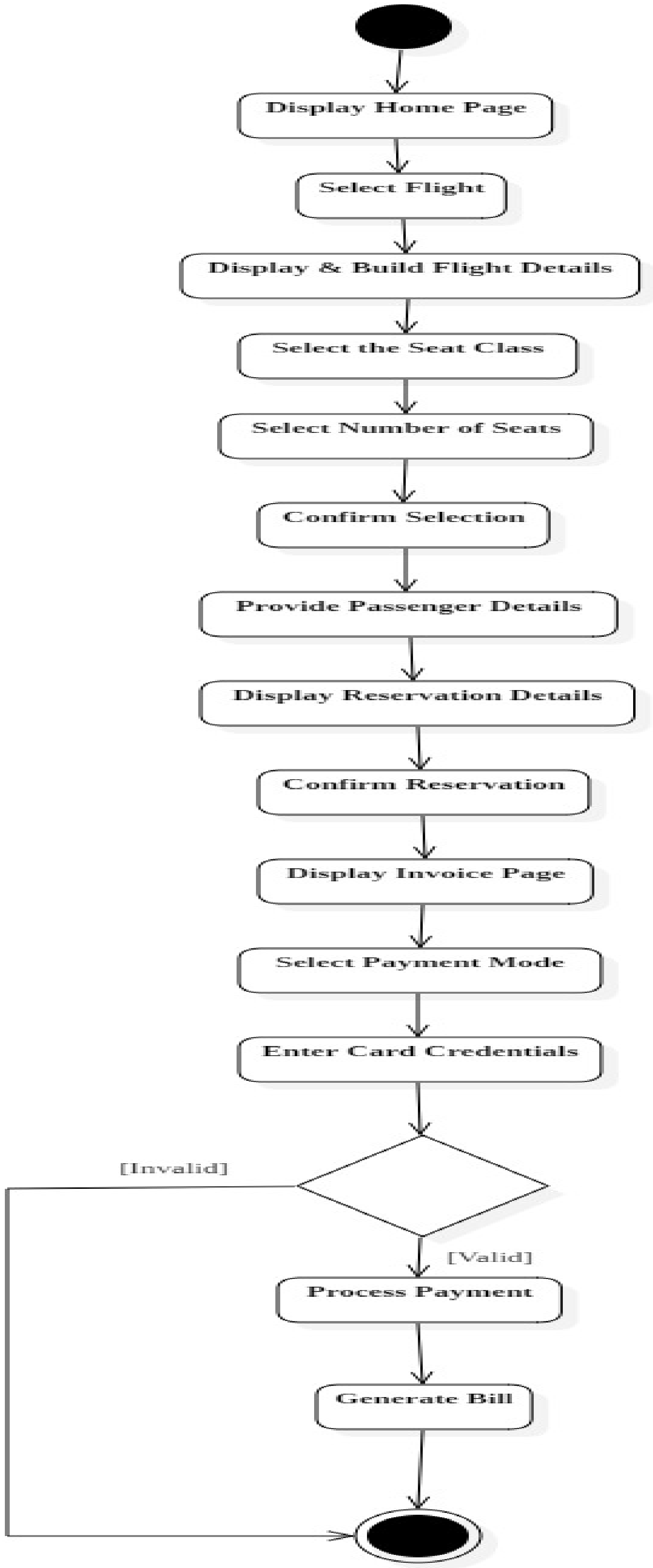


Figure 6: Activity model of airline reservation

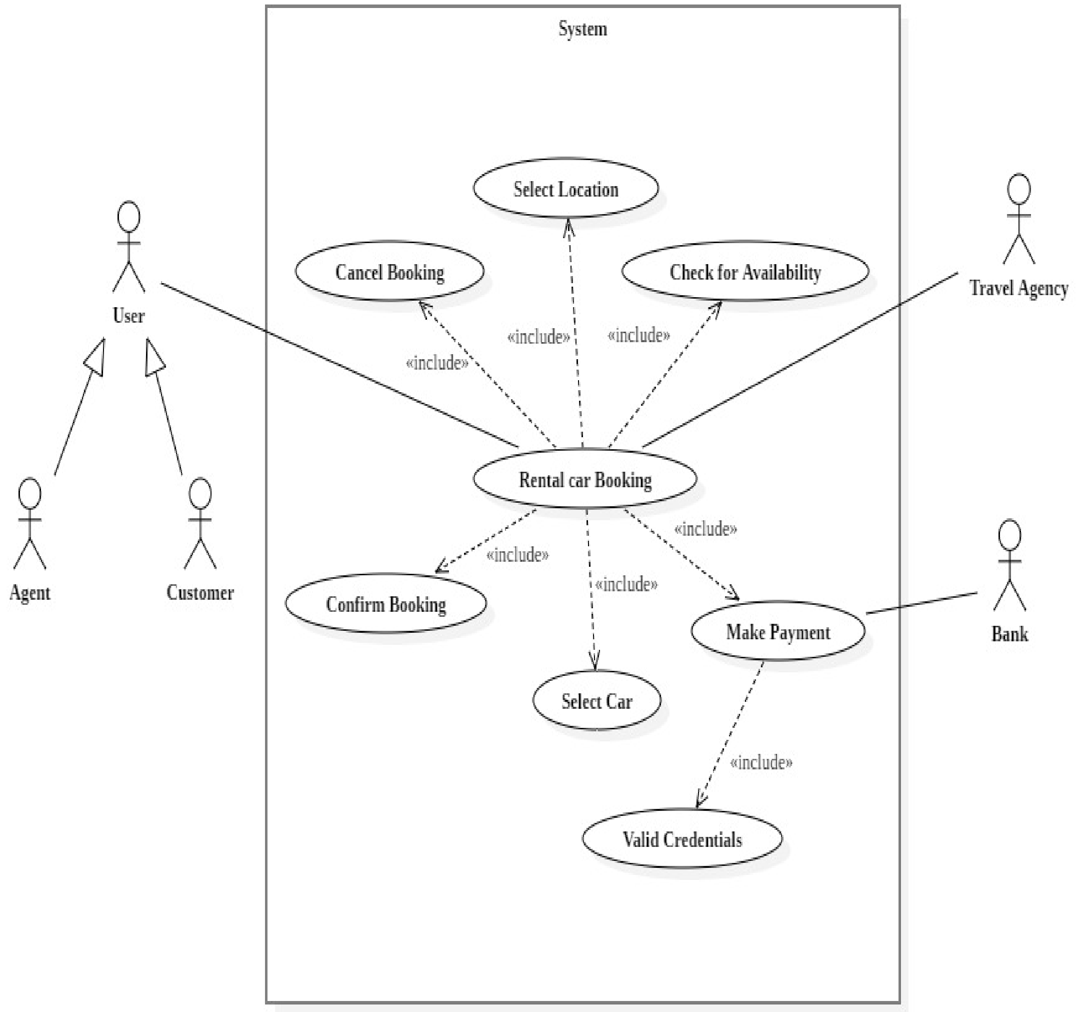


Figure 7: Use case model of car reservation

Section 7.6 discusses the analysis of changes in the configuration of hardware resources in the execution environment, as well as the identification of bottleneck resources and ideas for improvement.

4.5 Sensitivity Analysis

The environment of the composite web services for Travel Reservation is simulated using the simulation tool SMTQA (Simulation of Multi-Tier Queuing Applications) [6]. The simulation is carried out with the configuration (C1) given in Table 1, and the performance metrics are obtained. The values of the performance metrics are given in Table 4. It is observed from the table that the probability of dropping of sessions in Internet 1 and Internet 2 is 0.699 and 0.497, respectively. This is due to the low processing speed of the Internet. Hence, these are identified as bottleneck resources. To analyze the behavior of the hardware resources in the execution environment, sensitivity analysis is carried out by considering a modification in resource configuration one at a time. To improve the performance of the services, the processing speed of the Internet 1 is increased to 1050 KB. As a consequence of this, the probability of dropping requests is reduced to 0.213.

Since the number of requests processed by the Internet 1 is increased, this has an impact on the performance of the Travel Reservation server. As a consequence, the dropping of requests has happened in the Travel Reservation server, and its probability is 0.225. In turn, the probability of dropping sessions on the Internet is reduced from 0.497 to 0.297. From this observation, we could conclude that the performance of the Internet 1 and Internet 2 can be improved further by increasing their processing speed. Simultaneously, the processing speed of the Travel Reservation server also must be increased to avoid the dropping of requests.

5 Conclusions

The given methodology provides a mathematical model to estimate the response time based on the workload that fluctuates over a given time horizon. The estimation is made by i) Modeling the servers and other hardware resources in the deployment environment of WS architecture in SOA. ii) Formulating a procedure to calculate the response time pro-

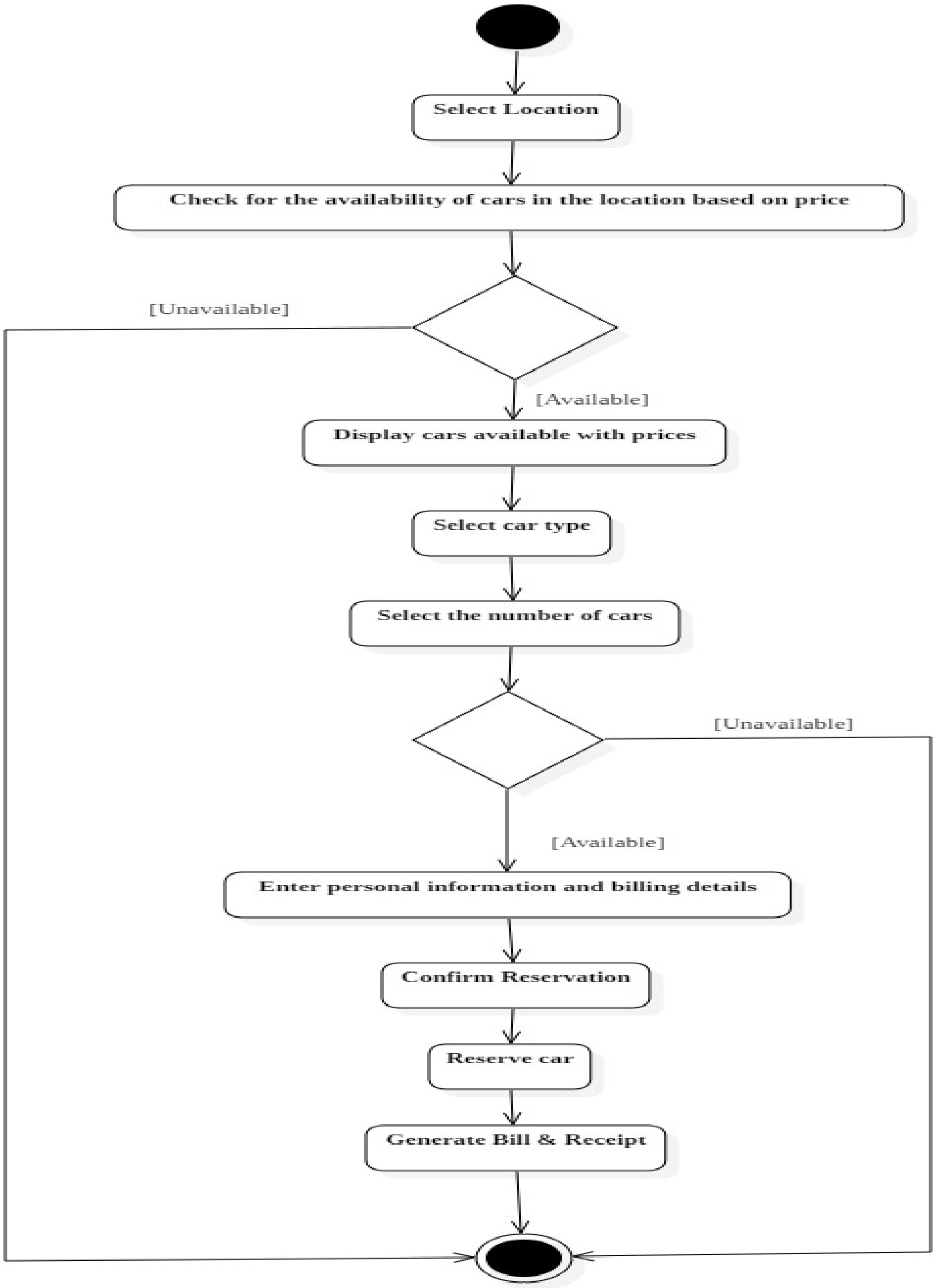


Figure 8: Activity model of car reservation

portional to workload. iii) Prediction of performance and analyzing the behavior of the resources across various web servers. iv) Identifying bottleneck resources and improving the performance of composite web service by sensitivity analysis.

Sensitivity analysis using the tool SMTQA is carried out to analyze the behavior of the hardware resources. The sensitivity analysis has indicated how changes in the base configuration of resources have an impact on the response time of the application. The methodology also helps to determine and define the performance objective of the web services by obtaining the range of values for maximum response time.

The proposed methodology can be used to predict the performance and to determine the most suitable deployment environment that can achieve the defined performance objective for web service based on non-uniform workload considering specifications of WS architecture. But, in composite web service various other services may be affecting the response time of the application; hence, the workloads of other services also need to be considered.

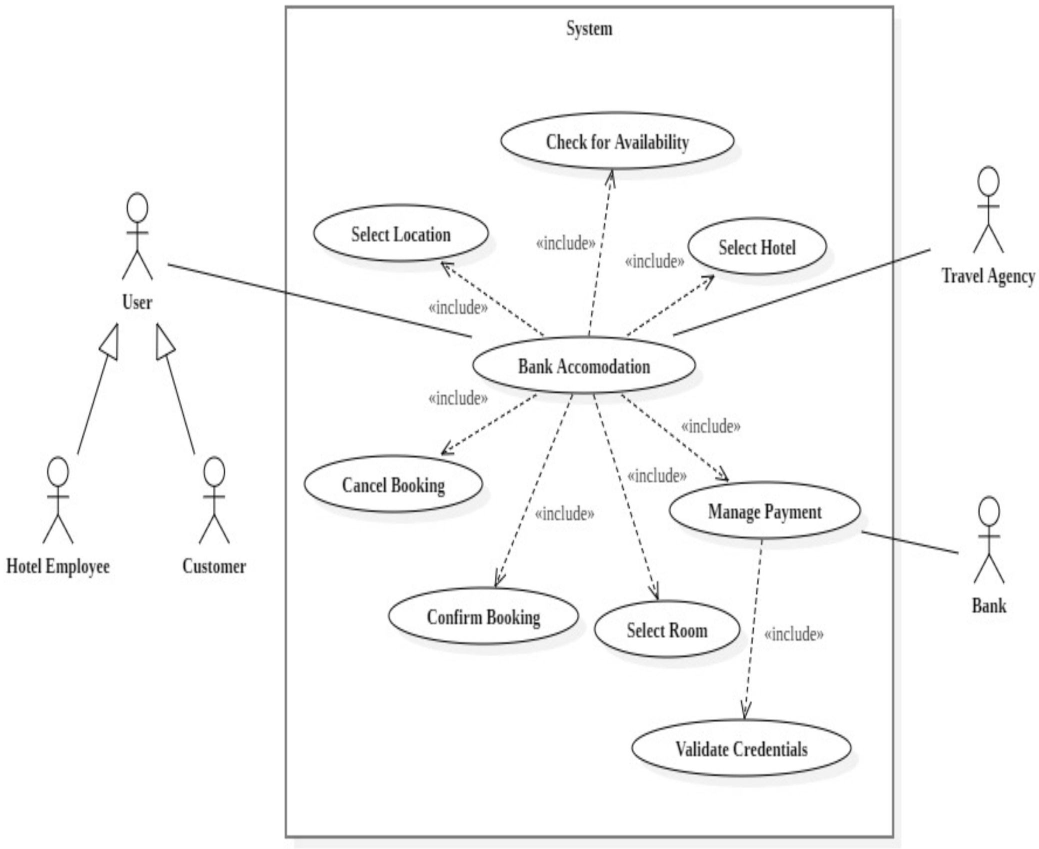


Figure 9: Use case model of hotel reservation

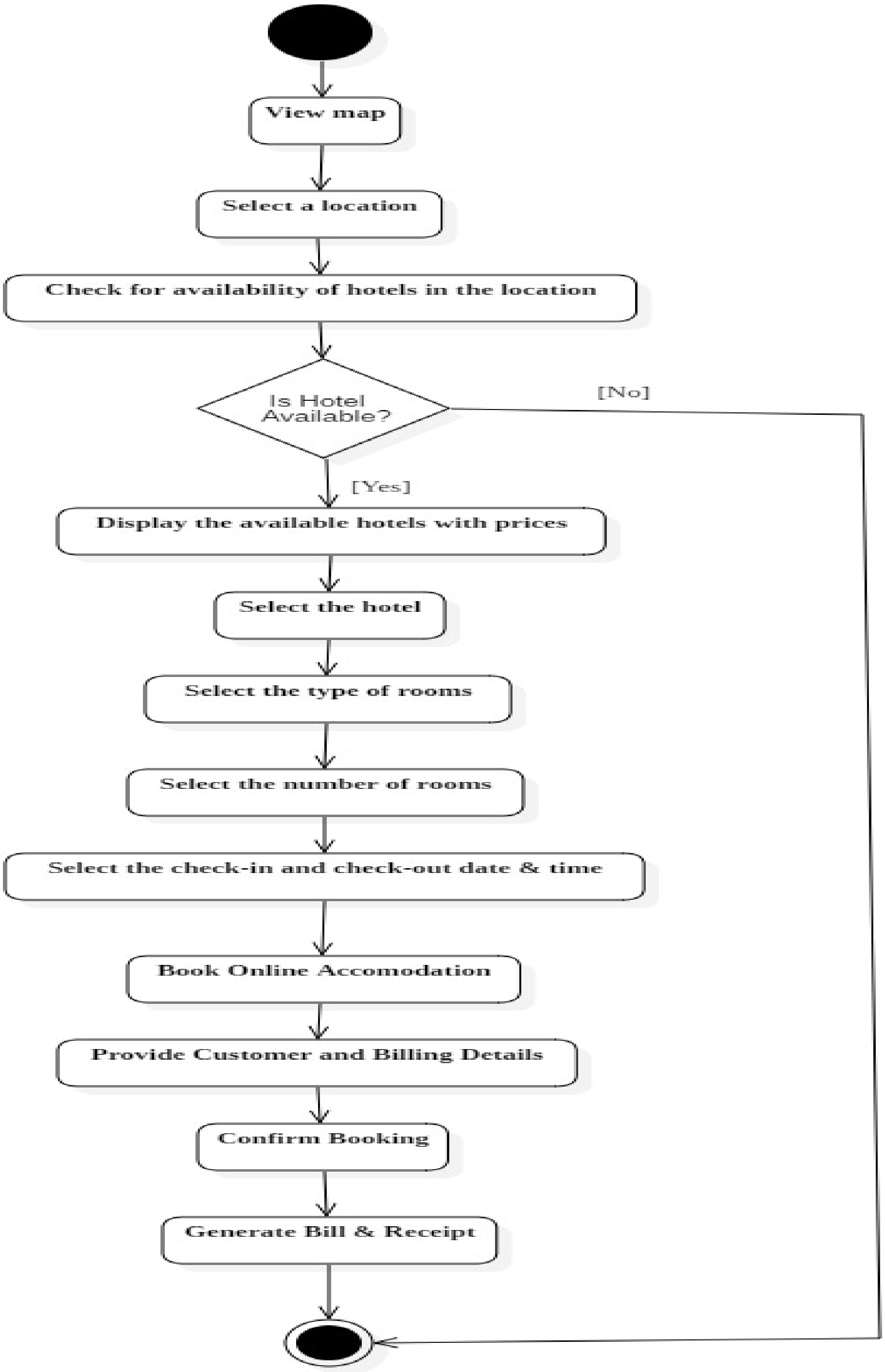


Figure 10: Activity model of hotel reservation

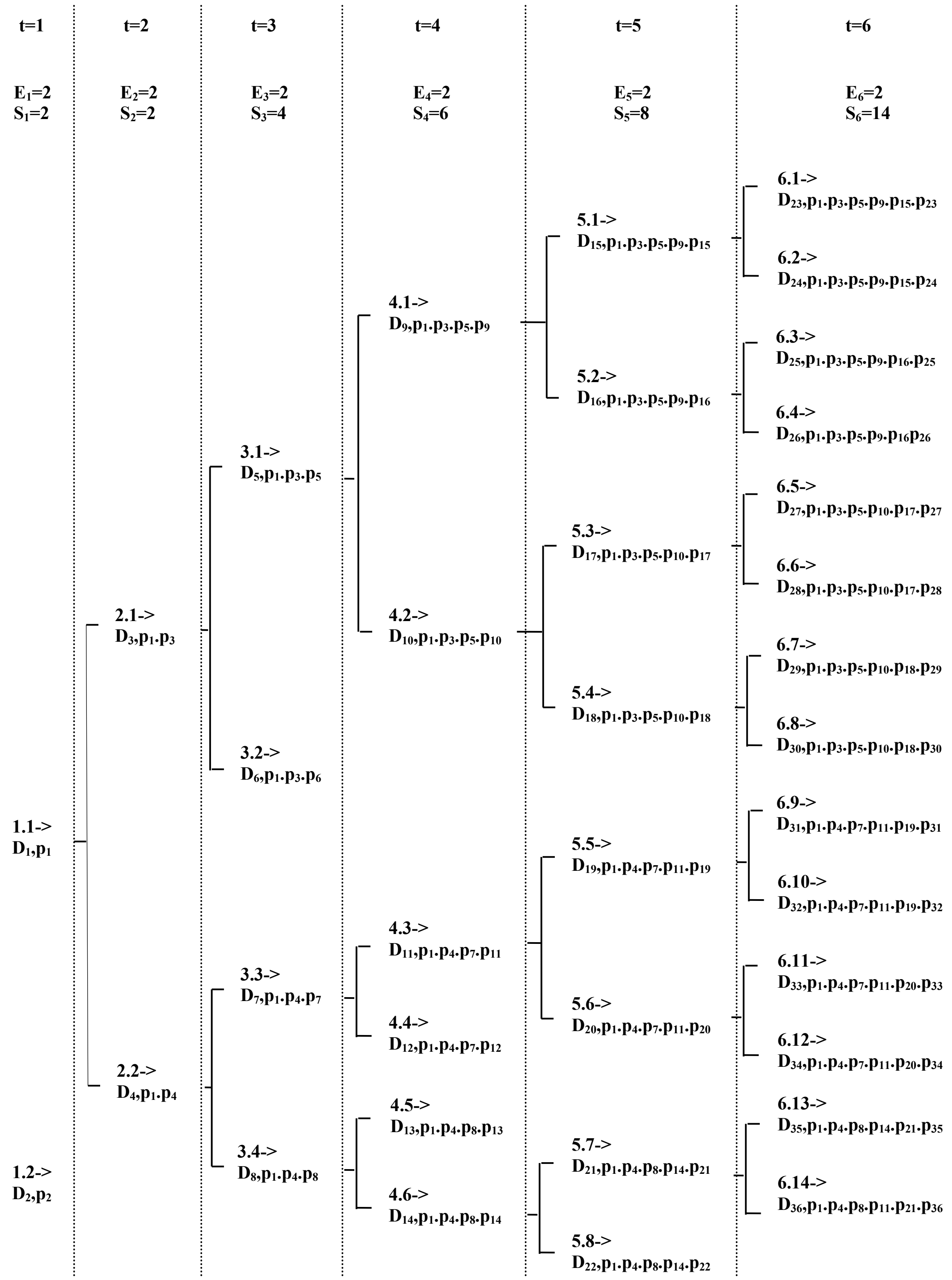


Figure 11: Conditional probability tree for the case study

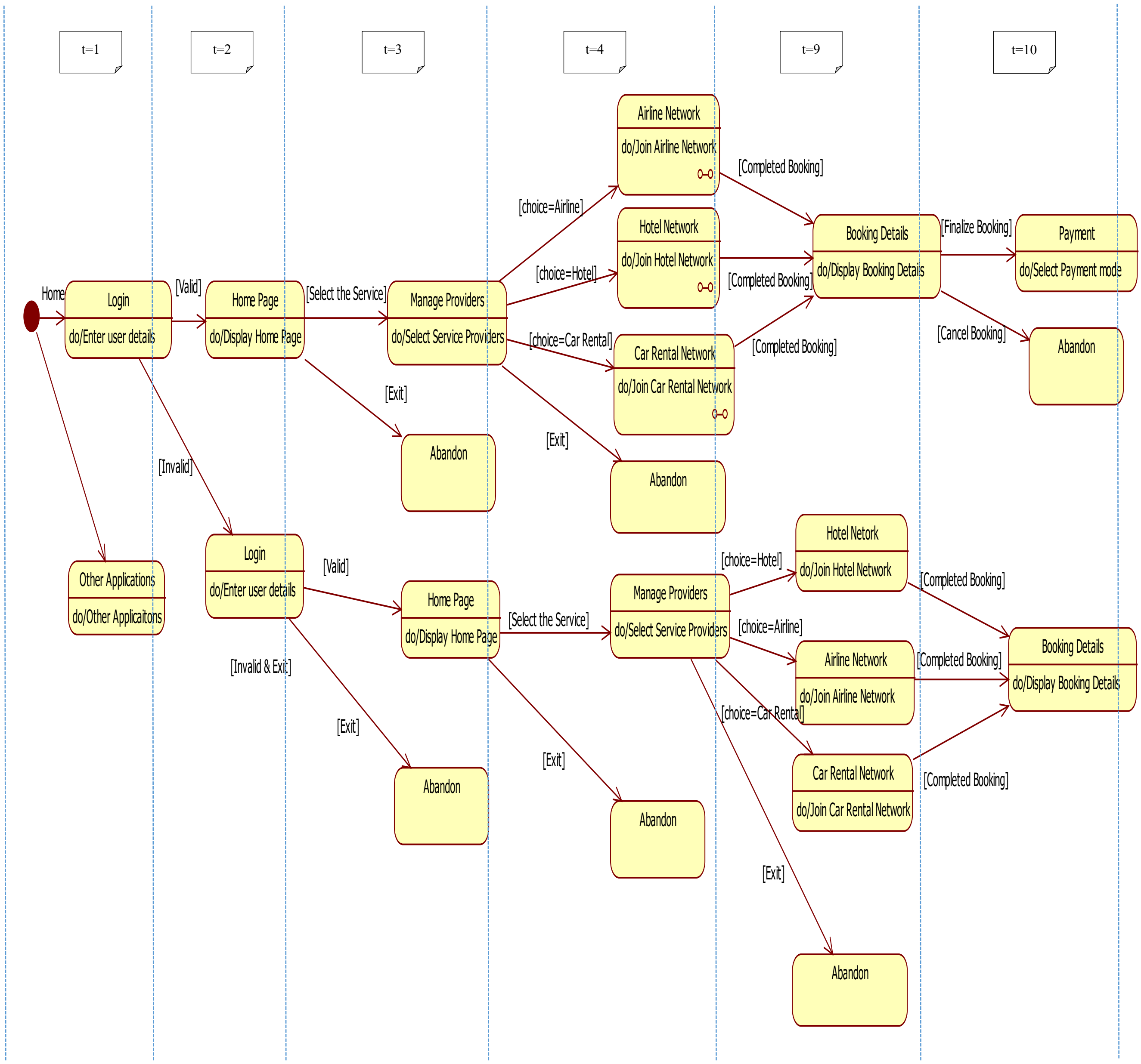


Figure 12: State chart diagram for travel care

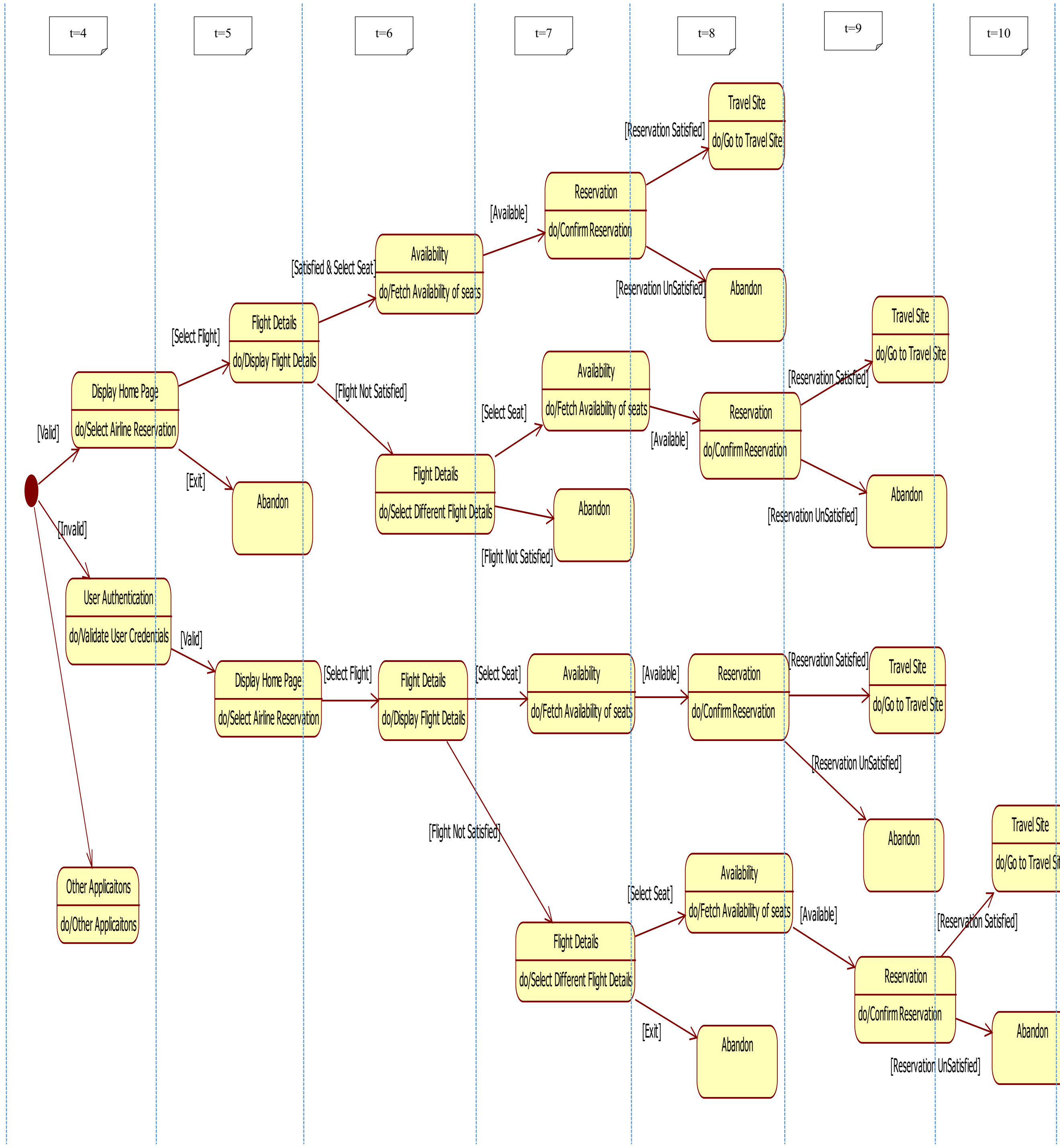


Figure 13: State chart diagram for airline reservation

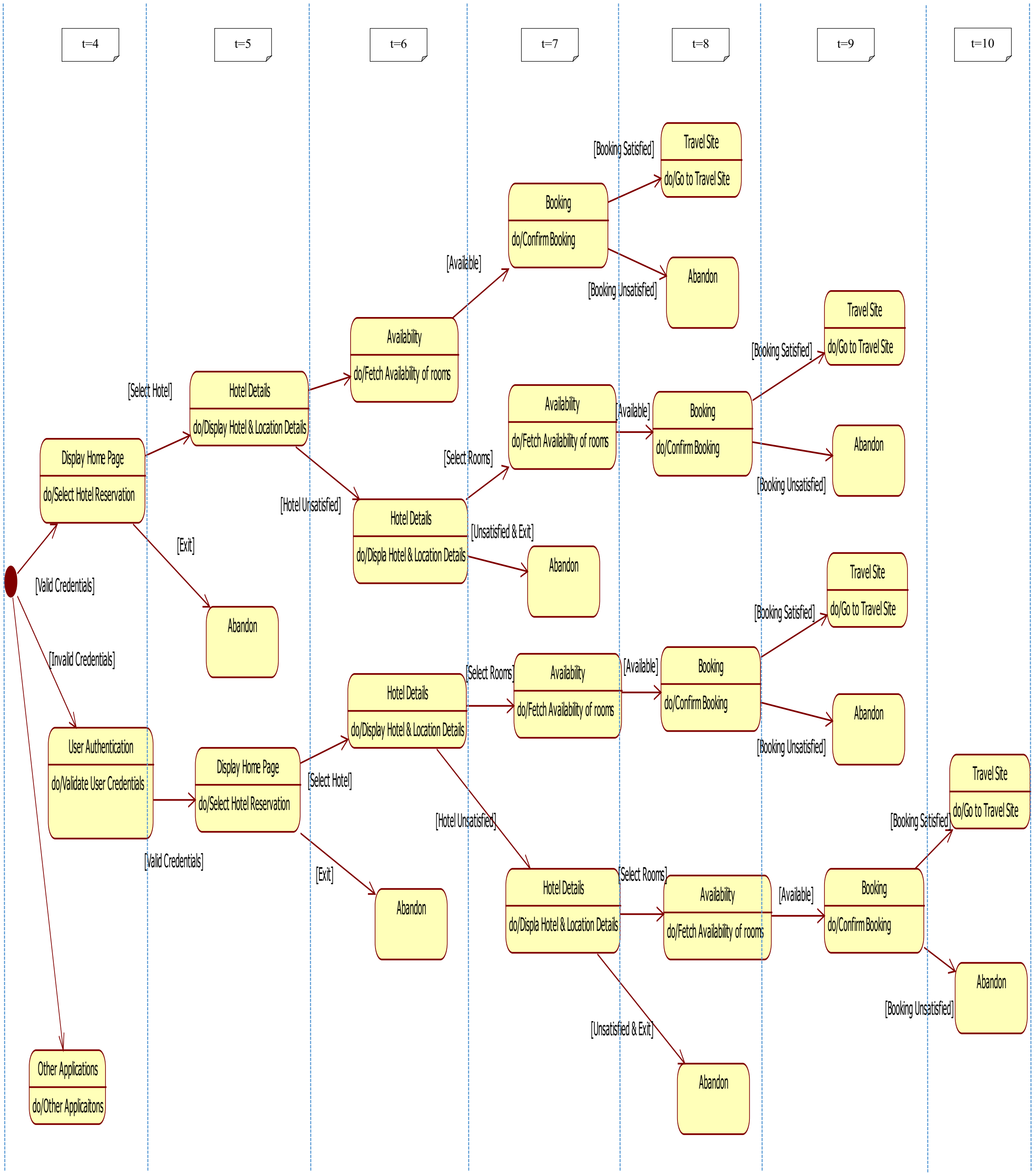


Figure 14: State chart diagram for hotel reservation

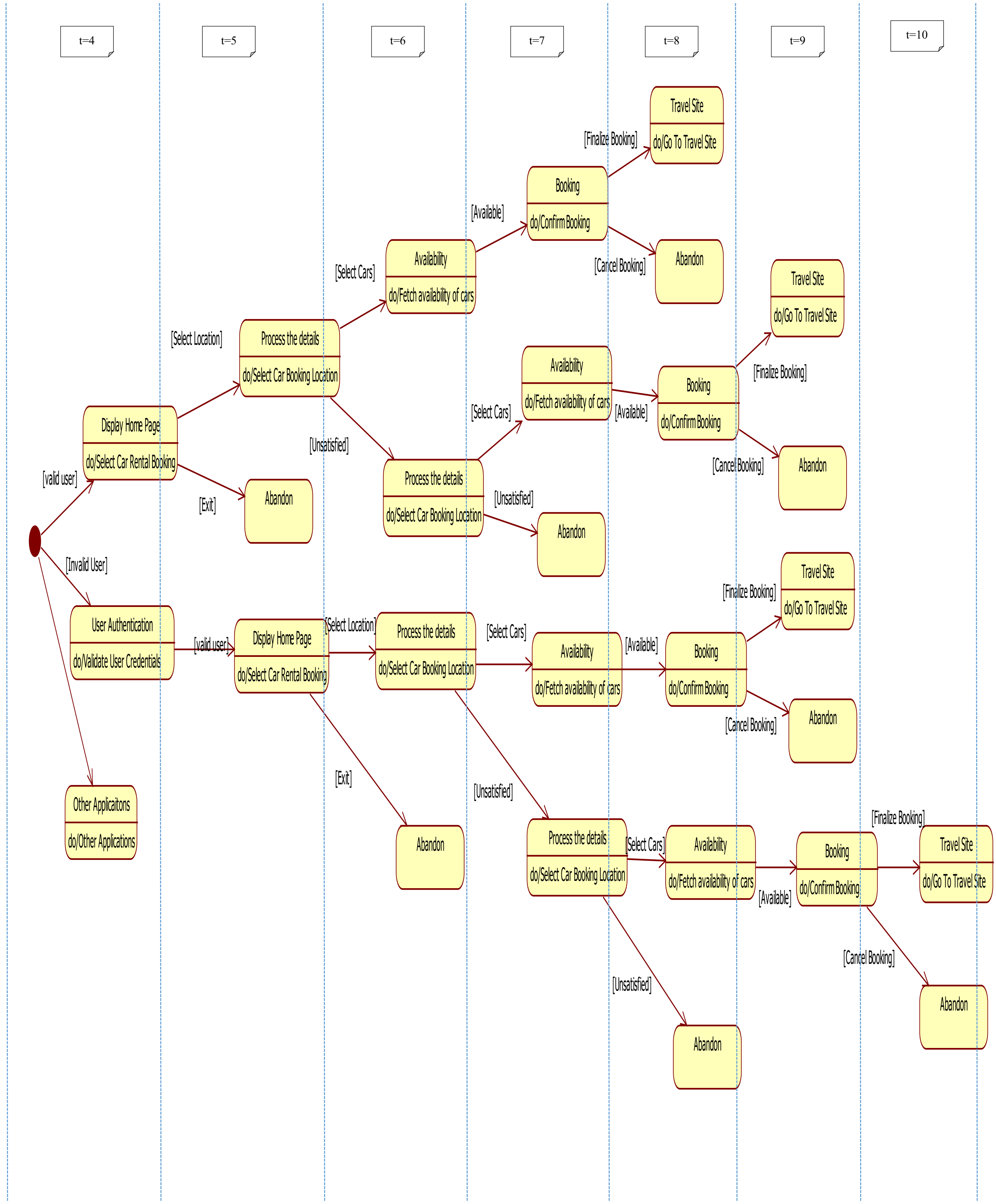


Figure 15: State chart diagram for car reservation

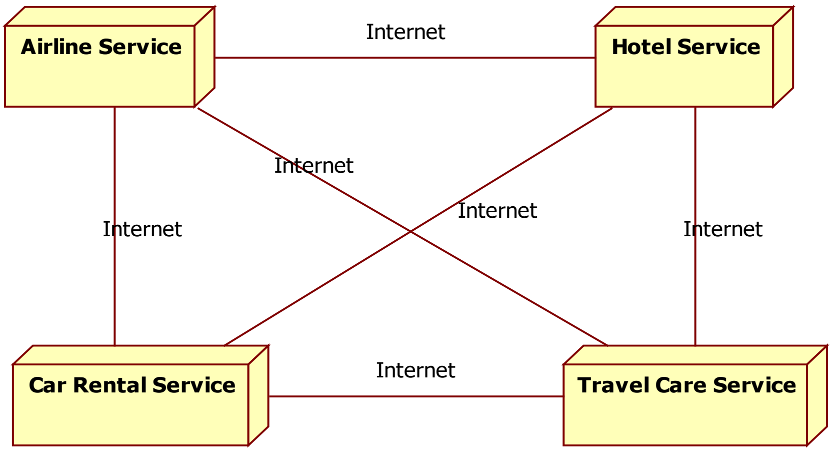


Figure 16: Exécution environment – base configuration (C_1)

Table 1: Configuration of resources in the execution environment (C_1)

Devices / Servers	*Client*	*Internet*	*Airline*	*Hotel*	*Car*	*Credit Card*	*LAN*
Processing Speed (Sec/KB)	16000	575	20000	15000	15000	20000	12500

Table 2: Size and response time of activities of travel service

Travel Care Service		
Activities	**Size (KB)**	**Response Time (Sec)**
Login	81.83094	0.355473
Query Status	55.6554378	0.472115
Select Airline	61.8863298	0.436247
Select Hotel	61.8863298	0.436247
Select Rental car	61.8863298	0.436247
Query Booking	53.8247277	0.421855
Manage Providers	78.27246088	0.465031
Join/leave airline net	95.1955894	0.543976
Join/leave hotel net	95.1955894	0.543976
Join/leave car net	95.1955894	0.543976

Table 3: Statistical values for configuration

Configuration											
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**	**Total**
Max	0.350	0.369	0.438	1.109	1.023	1.088	1.042	0.638	0.484	0.376	5.721
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005
Mean	0.174	0.099	0.066	0.170	0.093	0.077	0.169	0.086	0.041	0.019	0.993
Variance	0.010	0.006	0.005	0.028	0.014	0.010	0.015	0.006	0.002	0.001	0.429

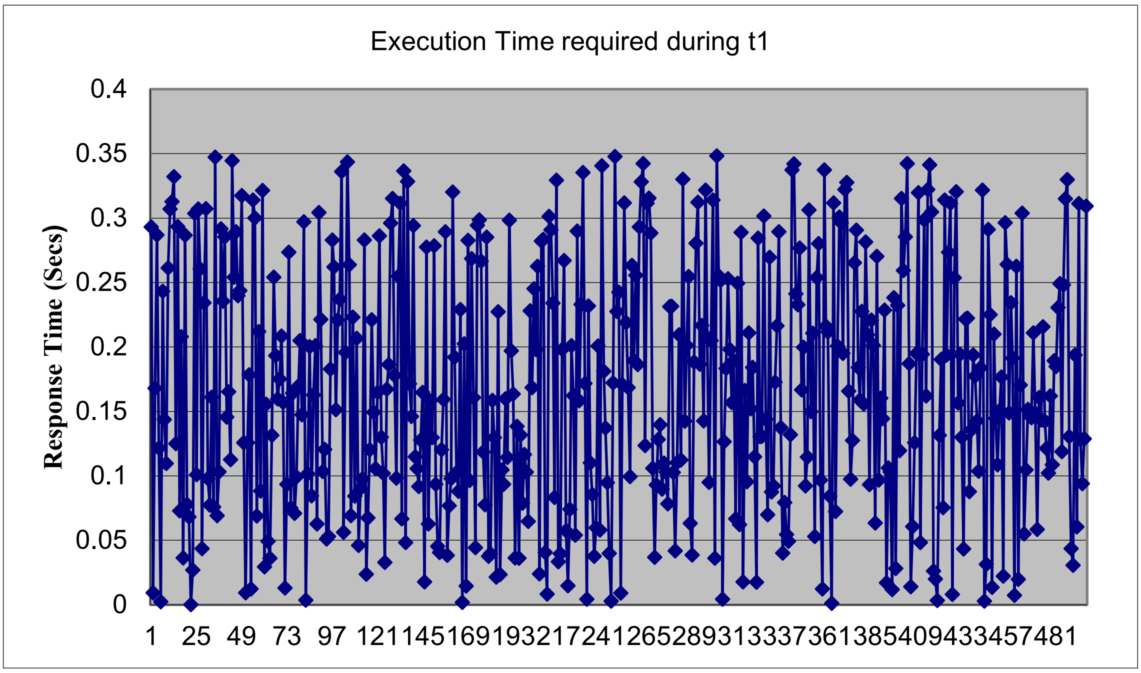


Figure 17: Response time obtained for t1

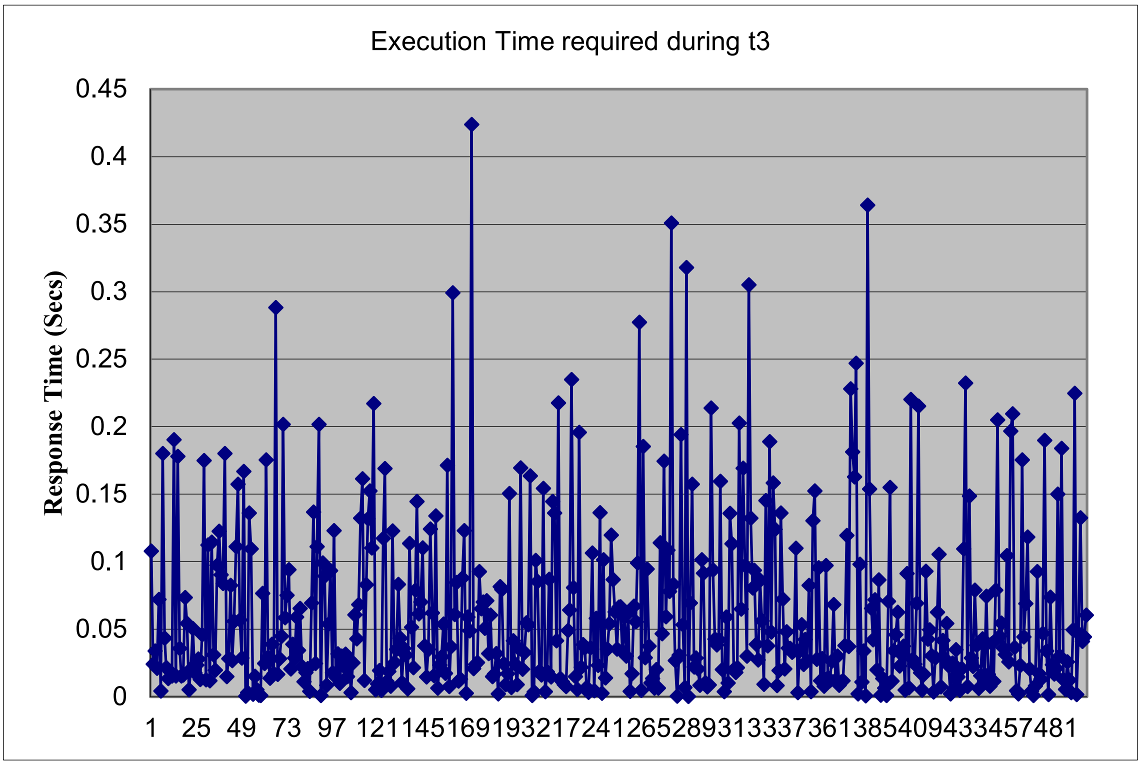


Figure 18: Response time obtained for t3

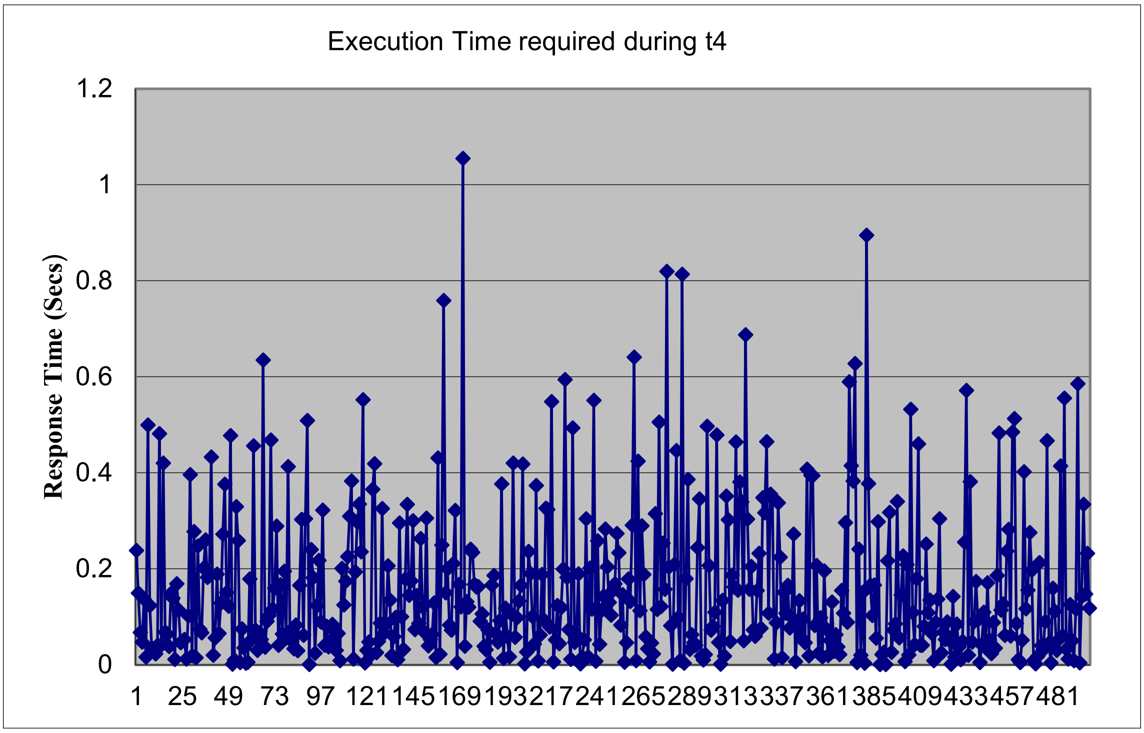


Figure 19: Response time obtained for t4

Table 4: Performance metrics of Internet speed 575 KBps with Arrival distribution is 0.01

Sl No	Layer Name	Processing Speed (Units)	Average Response time	Average service time	Average waiting time	Probability of idle server	Probability of dropping sessions
1	Client	16000	0.004	0.003	0.000	0.764	0.003
2	Internet1	575	0.144	0.052	0.091	0.001	0.699
3	Travel Reservation	20000	0.001	0.001	0.000	0.990	0.000
4	Internet2	575	0.090	0.036	0.055	0.011	0.497
5	Airline	20000	0.001	0.001	0.000	0.976	0.000
6	Hotel	15000	0.002	0.002	0.000	0.963	0.000
7	Car	15000	0.002	0.002	0.000	0.977	0.000
8	CreditCard	20000	0.004	0.004	0.000	0.999	0.000
9	LAN11	12500	0.003	0.003	0.000	0.989	0.000
10	LAn12	12500	0.003	0.003	0.000	0.984	0.000
11	LAN13	12500	0.002	0.002	0.000	0.995	0.000
12	LAN14	12500	0.004	0.004	0.000	0.997	0.000
13	WS1	10000	0.004	0.004	0.000	0.985	0.000
14	WS2	10000	0.004	0.004	0.000	0.981	0.000
15	WS3	10000	0.003	0.003	0.000	0.993	0.000
16	WS4	10000	0.005	0.005	0.000	0.997	0.000
17	LAN21	12500	0.003	0.003	0.000	0.983	0.000
18	LAN22	12500	0.002	0.002	0.000	0.978	0.000
19	LAN23	12500	0.002	0.002	0.000	0.993	0.000
20	LAN24	12500	0.003	0.003	0.000	0.998	0.000
21	DB1	20000	0.001	0.001	0.000	0.997	0.000
22	DB2	15000	0.001	0.001	0.000	0.996	0.000
23	DB3	15000	0.001	0.001	0.000	0.998	0.000
24	DB4	20000	0.004	0.004	0.000	0.999	0.000

References

[1] G. Alonso, F. Casati, H. Kuno, and V. Machiraju, *Web Services – Concepts, Architectures, and Applications*, Springer-Verlag, Berlin, Germany, pp. 123-149, 2004, doi:10.1007/978-3-662-10876-5_5.

[2] V. Almeida and D. Menasce, "Capacity Planning an Essential Tool for Managing Web Services," *IT Professional*, 4(4):33-38, July 2002, https://doi.org/10.1109/mitp.2002.1046642

[3] C. Catley, D. C. Petriu, and M. Frize, "Software Performance Engineering of a Web Service-Based Clinical Decision Support Infrastructure," *ACM Sigsoft Software Engineering Notes*, 29(1):130-138, 2004, https://doi.org/10.1145/974043.974066.

[4] L. Cheung, L. Golubchik, and F. Sha, "A Study of Web Services Performance Prediction: A Client's Perspective," 2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems, Singapore, pp. 75-84, 2011. doi: 10.1109/ MASCOTS. 2011.66.

[5] V. Datla and K. Goseva-Popstojanova, "Measurement-Based Performance Analysis of e-Commerce Applications with Web Services Components," IEEE International Conference on e-Business Engineering (ICEBE'05), Beijing, China, pp. 305-314, 2005, doi: 10.1109/ICEBE.2005.85.

[6] D Evangelin Geetha, T V Suresh Kumar, K Rajani Kanth, and Mayank P Singh, "Simulation of Multi-Tier Queuing Applications," Copyright, Diary Number: 50201/2014-CO/SW.

[7] E. Geetha D, T. S. Kumar, and K. R. Kanth, "Determining Suitable Execution Environment based on Dynamic Workload during Early Stages of Software Development Life Cycle: A Simulation Approach," *International Journal of Computational Science and Engineering*, 7(3):175, 2012, https://doi.org/10.1504/ ijcse.2012.048228.

[8] Gordon P. Gu and Dorina C. Petriu. "XSLT Transformation from UML Models to LQN Performance Models," *Proceedings of the 3rd International Workshop on Software and Performance (WOSP '02)*, Association for Computing Machinery, New York, NY, USA, pp. 227–234, 2002. https://doi.org/10.1145/584369.584402.

[9] R. Khadka, "An Evaluation of Dynamic Web Service Composition Approaches," M. J. van Sinderen, M. J. van Sinderen, B. Sapkota, and B. Sapkota (Eds.), 4th International Workshop on Architectures, Concepts and Technologies for Service Oriented Computing - ACT4SOC 2010, SCITEPRESS, pp. 67-79, 2010.

[10] F. Lecue and N. Mehandjiev, "Towards Scalability of Quality Driven Semantic Web Service Composition," IEEE International Conference on Web Services, Los Angeles, CA, USA, 2009, pp. 469-476, 2009, doi: 10.1109/ICWS.2009.88.

[11] Y. Li, Y. Liu, L. Zhang, G. Li, B. Xie, and J. Sun, "An Exploratory Study of Web Services on the Internet," IEEE International Conference on Web Services (ICWS 2007), Salt Lake City, UT, USA, pp. 380-387, 2007, doi: 10.1109/ICWS.2007.37.

[12] Ana C. C. Machado and Carlos A. G. Ferraz, "Guidelines for Performance Evaluation of Web Services," *Proceedings of the 11th Brazilian Symposium on Multimedia and the Web (WebMedia '05)*, Association for Computing Machinery, New York, NY, USA, pp. 1-10, (2005), https://doi.org/10.1145/ 1114223.1114234,

[13] Ioana Manolescu, Marco Brambilla, Stefano Ceri, Sara Comai, and Piero Fraternali, "Model-Driven Design and Deployment of Service-Enabled Web Applications," *ACM Trans. Internet Technol.,* 5(3):439-479, August 2005. https://doi.org/10.1145/1084772.1084773.

[14] D. A. Menasce, "Scaling for e-Business," *Proceedings 8th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (Cat. No. PR00728)*, San Francisco, CA, USA, pp. 511-513, 2000, doi: 10.1109/MASCOT.2000. 876578.

[15] D. Menasce and V. Almeida, "Capacity Planning an Essential Tool for Managing Web Services," *IT Professional*, 4(4):33-38, July 2002, https://doi.org/10.1109/mitp.2002.1046642.

[16] R. Peña-Ortiz, A. Gil José, Julio Sahuquillo, and Ana Pont, "Analyzing Web Server Performance under Dynamic User Workloads," *Computer Communications*, 36(4):386-395, 2013, ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2012.11.005.

[17] Ram Mohan Reddy, Evangelin Geetha D, Suresh Kumar T V., "Activity Based Performance Prediction for Software Systems," Technical Report, Department of MCA, M S Ramaiah Institute of Technology, Bangalore, 2016.

[18] C. U. Smith and L. G. Williams, "Performance Solutions: A Practical Guide to Creating Responsive, Scalable Software," October 1, 2001, https://doi.org/10.1604/9780201722291.

[19] C. U. Smith, *Performance Engineering of Software Systems*, Addison-Wesley Longman Publishing Co., Inc., ISBN:978-0-201-53769-7, January 1, 1990.

[20] J. Hyun Son and M. Ho Kim, "Improving the Performance of Time-Constrained Workflow Processing," *Journal of Systems and Software*, 58(3):211-219 September 2001, https://doi.org/ 10.1016/s0164-1212(01)00039-5.

[21] Paulo S. L. Souza mail, Regina H. C. Santana, Marcos J. Santana, Ed Zaluska, Bruno S. Faical, and Julio C. Estrella, "Load Index Metrics for an Optimized Management of Web Services: A Systematic Evaluation," July 16, 2013, DOI: 10.1371 /journal. pone.0068819.

[22] N. Thio and S. Karunasekera, "Client Profiling for QoS-Based Web Service Recommendation," 12th Asia-Pacific Software Engineering Conference (APSEC'05), Taipei, Taiwan, 8 pp. 2005, Doi: 10.1109/APSEC.2005.50.

[23] P. Xiong, Z. Wang, S. Malkowski, Q. Wang, D. Jayasinghe and C. Pu, "Economical andRobust Provisioning of N-Tier Cloud Workloads: A Multi-Level Control Approach," 31st International Conference on Distributed Computing Systems, Minneapolis, MN, USA, 2011, pp. 571-580, 2011, doi:

10.1109/ICDCS.2011.88.
[24] L. D. Xu, W. Viriyasitavat, P. Ruchikachorn, and A. Martin, "Using Propositional Logic for Requirements Verification of Service Workflow," *IEEE Transactions on Industrial Informatics*, 8(3):639-646, August 2012, https://doi.org/10.1109/tii.2012.2187908.
[25] X. Zhao, et al., "Real-Time Soft Resource Allocation in Multi-Tier Web Service Systems," 2017 IEEE International Conference on Web Services (ICWS), Honolulu, HI, USA, pp. 492-499, 2017, doi: 10.1109/ICWS.2017.57.

Ch. Ram Mohan Reddy is currently a Professor in Department of Computer Applications at the BMS College of Engineering, Karnataka, India. He received his Ph.D. Degree in Computer Applications from Visvesvaraya Technological University, Belagavi in 2017. His current research interests include software performance engineering, data analytics machine learning. He is a member of various professional bodies.

D. Evangelin Geetha is working as Associate Professor in the Department of Master of Computer Applications. Her areas of interest are software performance engineering, predictive analytics, cloud computing, object technology and distributed systems.

R. V. Raghavendra Rao received a bachelor's and Master's from S. K. University, Anantapur, India and Madras University, India, respectively. Currently pursuing Ph.D. studies at NIT Trichy. Having 20+ years of experience in teaching, research, and development. His research articles have been published in journals and conferences related to software performance, block chain technologies, and data analytics. My research is focused on using machine learning algorithms for data analytics in medical and agricultural domains.

K. Sailaja Kumar is currently working as Senior Subject Matter Expert, Data Science, UNext, Bangalore, Karnataka, India. She received her Ph.D. Degree in Computer Applications from Visvesvaraya Technological University, Belagavi. Her areas of interest are predictive analytics, software performance, machine learning. She published a large number of publications in reputed journals.

A Survey on the Impact of Hyperparameters on Random Forest Performance using Multiple Accelerometer Datasets

Hong-Lam Le*‡§
VNU Information Technology Institute, Hanoi, VIETNAM
Vinh University of Technology Education, Vinh City, VIETNAM

Thanh-Tuoi Le and Thi-Thu-Hien Vu§
Vinh University of Technology Education, Vinh City, VIETNAM

Doan-Hieu Tran§
Ho Chi Minh University of Banking, Ho Chi Minh City, VIETNAM

Dinh Van Chau§
Electric Power University, Hanoi, VIETNAM

Thi-Thu-Trang Ngo†§
Posts and Telecommunications Institute of Technology, Hanoi City, VIETNAM

Abstract

Previous studies have examined the effects of hyperparameters on random forests, but little research has been done in the context of fall detection. To address this gap, our study aimed to examine how hyperparameters influence the performance and training time of a random forest algorithm used in fall detection systems. Our findings highlighted the best range of values for each hyperparameter to achieve high performance. Moreover, we discovered that certain combinations of hyperparameters could either enhance or reduce the random forest's performance compared to the default settings. To conduct these investigations, we performed experiments using two datasets: MobiAct v2.0 and UP-Fall, which were collected from accelerometers in smartphones and wearables. These insights can contribute to the optimization of hyperparameters for more effective fall detection systems.

Key Words: Hyperparameter, random forest, max_depth, num_Tree, num_features, min_samples_leaf, min_samples_split.

* Corresponding author, created the first draft, and created the stable version of this document.
† Corresponding author, created the stable version of this document.
‡ University Information Technology Institute.
§ lehonglam@vnu.edu.vn, fredleektv@gmail.com, thuhienktv@gmail.com, hieutd@hub.edu.vn, chaudv@epu.edu.vn, trangnttl@ptit.edu.vn.

1 Introduction

Fall detection systems play a crucial role in ensuring the well-being and health of older adults and individuals with mobility impairments. These systems are designed to quickly identify falls and notify caregivers or emergency responders, allowing for prompt medical attention and reducing the risk of harm or fatality. However, current fall detection methods have drawbacks, such as low accuracy and high false alarm rates, which affect their reliability. Thus, it is vital to develop and improve fall detection systems to advance human healthcare.

Enhancing the accuracy, sensitivity, specificity, and reliability of fall detection systems is a primary goal in preventing or mitigating the negative consequences of falls. Various machine learning algorithms, such as J48, Logistic Regression (LR), K-Nearest Neighbor (kNN), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), and Random Forest (RF), have been utilized to classify different human activities, including falls. Studies have shown that RF outperforms other algorithms in fall detection using the MobiAct v2.0 and UP-Fall datasets [5, 10-11, 13]. However, the RF algorithm's performance heavily relies on the selection of hyperparameters [7, 9, 14-15], which poses a significant challenge.

The random forest algorithm is a general classification method that uses hyperparameters to adjust or optimize the loss function. These hyperparameters impact both the model's speed and accuracy, making it crucial to find the optimal values [8, 26]. However, the optimal hyperparameter values are not universal and depend on the specific problem and data. The influence of hyperparameters on random forest performance, particularly for human daily activity data crucial for fall detection systems, has not been extensively studied and requires further research.

This study determined the optimal range for each hyperparameter of the random forest algorithm to achieve high performance. The results also revealed that combining adjustments of min_samples_leaf and min_samples_split hyperparameters could decrease the random forest's performance. The hyperparameters max_depth, NumTree, and max_features had the most significant impact on performance and model-building time. Combining adjustments of NumTree and max_features, or combining max_depth with NumTree and max_features, resulted in the best performance. These findings have practical implications for constructing optimal models for fall detection systems.

The rest of this paper is structured as follows: Section 2 reviews related work and provides an analysis and summary of studies on hyperparameter tuning for random forests. Section 3 describes our research's evaluation methods, including the research process, the dataset used, and the implementation of the random forest algorithm with adjusted hyperparameters. Section 4 presents our experimental results on the effect of hyperparameters on the random forest algorithm''s performance and provides an interpretation of these results. Finally, the conclusions section summarizes our main findings, discusses their implications, suggests directions for future research, highlights study limitations, and proposes areas for further investigation.

2 Related Works

Recently, random forest algorithms have gained widespread use in machine learning systems due to their high accuracy and robustness against interference. Several studies have proposed solutions to enhance the performance of random forests by optimizing their hyperparameters. For instance, Philipp Probst et al. [15] surveyed the impact of hyperparameters on predictive performance in the field of credit risk. The research was divided into two parts. The first section provided an overview of the adjustment strategies. Later, the author team utilized the tuneRanger R package to automate the tuning of random forests with MBO (Model-Based Optimization) to demonstrate the validity of applying one of these strategies. The benchmark research was conducted on datasets from OpenML and downloaded through the OpenML R package. The authors concluded that the mtry parameter had the most impact on performance. Finally, they recommended using a high number of trees and SMBO (Sequential Model-Based Optimization) to simultaneously tune the mtry parameters, sample size, and node size to enhance the performance of the random forest algorithm.

Kavita M Kelkar et al. [9] proposed an effective learning system based on the random forest algorithm to detect learners' emotional states. This research proposed analyzing hyperparameters such as n_estimators, max_depth, min_samples_split, min_samples_leaf, and criterion to improve the system's performance. However, their paper did not provide a method for investigating the influence of hyperparameters. The experimental results showed that the optimal value for the hyperparameter max_leaf_nodes was 0, which gave the best Kappa score. Additionally, max_features did not affect the accuracy or Kappa value, while a max_depth > 10 provided more accurate results. Of the hyperparameters surveyed, n_estimators were the most important, and their value could range from 25 to 30 without affecting accuracy or Kappa values.

The research by Ningyuan Zhu et al. [27] proposed a random forest-based intrusion detection model for application in electric industrial control systems. The authors also introduced an improved mesh search algorithm (IGSA) to optimize the hyperparameters of the random forest model and improve its efficiency and accuracy. The research adjusted the seven hyperparameters of the random forest model in descending order of preference, including max_depth, min_samples_leaf, min_samples_split, criterion, n_estimators, bootstrap, and max_features. The optimization results revealed that the optimal values for the hyperparameters were max_depth = 24, min_samples_leaf = 1, min_samples_split = 2, criterion = gini, n_estimators = 240, bootstrap = default, and max_features = 9. The test results demonstrated that the intrusion detection method based on hyperparameter optimization achieved higher accuracy, resilience, F1 score, and ROC-AUC score than other methods.

Bernard et al. [1] recommend using the value mtry = sqrt(p) as it provides a reasonable error rate. They emphasize that the number of predictors influences the optimal value of mtry. When many predictors are involved, mtry should set small to both the most and least influential variables are a selection for separation. It may increase relevance but potentially reduce performance. The computation time of the model is devoted to the selected separable variables, so the computation time decreases linearly with decreasing mtry values, similar to the research of author Wright et al. [25]. Conversely, setting mtry too high may exclude less influential variables from contributing to the prediction, as more influential variables are preferred for separation, inadvertently "blurring" variables of less influence.

Probst et al. [16] proposed a general structural framework to evaluate the tunability of hyperparameters in algorithms, including Random Forest. They presented this approach through an application to 38 datasets. The results showed that the mtry parameter had the most influence on the AUC index, followed by the sample size, while the node size had less effect. Hyperparameter tuning was a crucial step, as the results showed that the hyperparameter-tuned random forest model had higher prediction accuracy than the default one. Besides, the performance of default hyperparameters was often inconsistent.

In general, authors in each field have their conclusions about how hyperparameters affect the performance of random forests. Increasing the number of trees has improved classification accuracy [17, 19], while some studies suggest that increasing the maximum depth can lead to overfitting. The computation time of the model increases with the number of trees, and with a smaller number of trees, the computation time will be faster.

However, they did recommend using a large enough number of trees to ensure a stable error rate. Some studies indicate that the best performance of random forests is typically achieved with 100 trees [6, 17], while Boehmke and Greenwell [2, Ch. 11] recommend selecting ten times the number of trees as the number of features of the dataset for forest construction. Out-of-bag error curves increase with the number of trees and can be used to test algorithm convergence [17]. Research by P. Contreras et al. [6] indicated that numTree had the most influence on model performance, particularly in the value range of 0 to 100, which is consistent with the research by Probst et al. [16]. Therefore, numTree tuning is a straightforward method to achieve optimal model performance.

Based on a search on Google Scholar, we learned about related work on hyperparameter tuning for random forest algorithms. However, we have not found studies on this topic in the range of human daily activity recognition, especially in fall detection systems. Therefore, we conducted a survey and evaluated the impact of hyperparameters on the performance of the random forest algorithm using accelerometer data for fall detection systems.

3 Evaluation Methods

3.1 Evaluation Process

In this article, the feature extraction and random forest algorithms are built based on the Java language using the Weka library version 3.9.6. Figure 1 illustrates the experimental procedure for evaluating each hyperparameter and combination of hyperparameters in the random forest.

First, two raw data sets, MobiAct v2.0 and UP-Fall, were prepared for the experiment. The training dataset was created by segmenting the raw data into windows of 256 samples with a 50% overlap, applying preprocessing techniques to the segmented data, and extracting 44 features [11]. Next, we split the dataset into a training set (80%) and a test set (20%). We independently selected the test set to represent the overall dataset.

We conducted hyperparameter selection by surveying each hyperparameter individually to assess its influence on the performance of the random forest model. Afterward, we examined combinations of two, three, four, and five hyperparameters to determine the best and worst combinations.

Finally, we analyzed the results to evaluate and draw conclusions regarding the impact of hyperparameters and their combinations on the performance and training time of the random forest model. We utilized the accuracy and F1-score measures to estimate the performance of the random forest. Accuracy and F1-score are widely recognized as performance measures of efficient models by many studies.

Accuracy (Acc) was defined as the ratio of quantity correctly classified instances to the total quantity instances in the dataset. It is a simple and popular metric for evaluating the efficiency of classification algorithms. This measure has the following definition:

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

where TP, TN, FP, and FN are symbols used in the classification problem and evaluate the accuracy of a classification algorithm, specifically:

TP (True Positive) represents the amount of data that actually belongs to a class and the algorithm correctly predicts it to be that class.

TN (True Negative) represents the amount of data that actually does not belong to a class and the algorithm correctly predicts it to not belong to that class.

FP (False Positive) represents the amount of data that actually does not belong to a class, but the algorithm predicts it does.

FN (False Negative) represents the amount of data that actually belongs to a class, but the algorithm predicts it does not belong to that class.

F1-score is a measure of the efficiency of a classification algorithm. It combines the precision and recall of the data classes in a classification model. Formally, accuracy has the following definition:

$$\text{F1-core} = 2 * \frac{(\text{Precision} * \text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})} \quad (2)$$

where Precision is the ratio between the TP and the (TP + FP), and Recall is the ratio between the TP and (TP + FN).

The F1-score ranges from 0 to 1, with a higher value indicating a better classification model. It is used in binary classification problems where the balance between precision and recall is crucial.

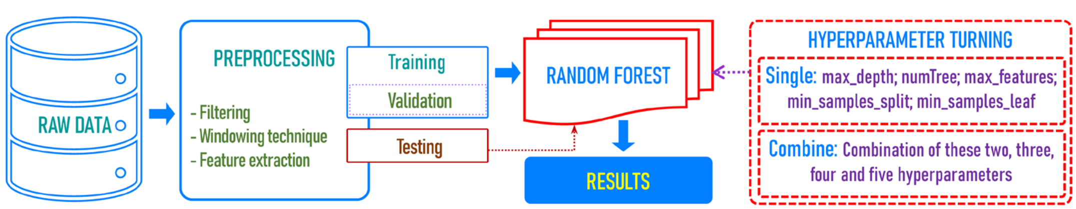


Figure 1: Hyperparameter evaluation process of random forest

3.2 Experimental Dataset

To evaluate the impact of hyperparameters, we conducted experiments using two commonly used datasets in human activity recognition and fall detection, namely MobiAct v2.0 [23] and UP-Fall detection [13]. These datasets are used widely in related research to action recognition and fall detection due to their large size and popularity. What sets these datasets apart is the diversity of the number of actions, volunteers, and collection methods. The MobiAct v2.0 dataset was collected using sensors on smartphones, whereas the UP-Fall dataset utilized wearable sensors for collection.

The MobiAct v2.0 dataset [23] was collected from the accelerometer, gyroscope, and orientation sensor of the Samsung Galaxy S3 smartphone with a sampling frequency of around 85 Hz for all activities. This dataset has 16 actions, of which 12 were daily activities (ADLs) and four fall behaviors, collected from 66 volunteers. The sampling of fall behaviors was done three times for action, with each sample lasting 10 seconds. Standing and walking activities were sampled only once for 5 minutes. Other daily activities have sampling times ranging from six seconds to 30 seconds, depending on the action. The sample rate of fall behaviors accounts for only about 8% of the total data collected from smartphone accelerometers. We only use the data collected from the accelerometer in this dataset for our research.

For the UP-Fall dataset [13], Martínez-Villaseñor and colleagues used five Mbientlab MetaSensor wearable sensors to collect raw data from the 3-axis accelerometer and gyroscope. They also used a NeuroSky MindWave headset to measure brainwave signals in the forehead, installed six infrared sensors as a grid 400mm above the lab floor to measure changes in actions, and positioned two Microsoft LifeCam Cinema cameras 1820 mm above the floor to collect images of the subject from the front and side. All types of sensors work simultaneously to collect data for each type of action. The standardized sampling frequency for all samples is 100 Hz. In this research, we only used data collected from the 3-axis accelerometer of the IMU device placed inside the right pocket of the volunteer. This part of the data includes 11 types of actions collected, of which the number of samples for five falling behaviors accounts for nearly 9% of the dataset.

The owners and authors of both datasets have permitted us to use them in community support and educational studies. These datasets have been reviewed and approved by experts in human research ethics.

3.3 Processing

The preprocessing steps applied to the raw dataset include data cleaning, noise filtering, normalization, and extraction of matching features. The data cleaning step removes missing or corrupted data, normalizes the data to keep it in a consistent range, and extracts feature that identify characteristics suitable for classifying human actions and behaviors.

The dataset used in this research includes 44 features extracted from MobiAct v2.0 and UP-Fall. We used the same preprocessing and feature extraction methods as in previous studies [11], and details of these methods have been presented in previous publications [11].

For experimentation, we segmented the data into sliding window sizes of 256 samples with an overlap rate of 50%, extracting 44 features [11]. However, the MobiAct v2.0 dataset was so large that using all the data for experimentation would have been time-consuming. Therefore, to shorten the experimental time, we used only 1/3 of the sample number of each action in the dataset to build the experiment data. The total sample quantity of the MobiAct v2.0 and UP-Fall datasets were used for experimental 15,776 and 2,025, respectively.

3.4 Random Forest Algorithm

Random Forest (RF) is a supervised learning algorithm [3] that uses multiple classifiers instead of a single one to achieve higher accuracy in predicting future cases. It is an extended version of a decision tree that uses two random steps to generate highly diverse sub-datasets, reducing variance error. Unlike traditional decision trees, each Classification and Regression Tree (CART) [4] in RF can only select a random subset of features, making the trees in the model more diverse. RF is particularly effective in handling datasets with specific issues [24] and can help resolve complex interactions between input features, enabling good over-model matching [22]. Additionally, RF can estimate the importance of each feature in the feature space [12]. Each decision tree in RF does not use all the training data or all the attributes of the data to build the tree. The information from the trees complements each other, leading to a low-bias and low-variance model with good prediction results and fast training time.

3.5 Hyperparameters in Random Forest

The random forest algorithm uses hyperparameters to control the learning process and time, which has a significant impact on its performance. Understanding the function of each hyperparameter is crucial to optimize the algorithm. These hyperparameters include the maximum depth of a tree, the number of trees in the forest, the maximum number of features for node splitting, the minimum number of samples in leaf nodes, the minimum number of samples for node separation, and the data percentage used for tree construction [3, 15]. Additionally, another set of hyperparameters focuses on dividing the nodes within each tree.

Failure to find the optimal hyperparameter value can reduce the performance of the RF algorithm. However, adjusting the hyperparameters for each dataset can improve classification performance or speed up the model's predictive ability [15]. This article investigates identifying the hyperparameter value range that yields the best RF performance and predictability. Based on relevant studies, we conclude that hyperparameters significantly affect the RF classification performance and time in various ways, such as:

(1) max_depth (M_D) is a hyperparameter that determines the

maximum depth at which a tree in a forest can grow. It is an essential hyperparameter that considerably impacts the model's accuracy [15, 21]. Increasing the tree's depth improves the model's accuracy by providing more information and data division. However, setting M_D too high will complicate the tree's structure and result in an overfit to the data. Therefore, selecting an appropriate M_D value is crucial in optimizing the model.

(2) num_tree (N_T) is the number of decision trees (DT) utilized in the forest, and it is correlative with the training dataset's size. The number of trees should be sufficient to stabilize the error rate. The higher the number of trees, the greater the classification accuracy. However, using too many trees increases the computation time. There is no rule for determining the optimal number of trees, but some sources suggest that [2, Ch. 11] the number of trees should be approximately ten times the number of dataset features. The initialization trees' number can be increased or decreased depending on hyperparameters such as max_features and min_samples_leaf.

(3) The max_features (M_F) parameter represents the maximum number of features used when the algorithm searches for node separation. It is a crucial hyperparameter that influences the model's classification performance [15, 21]. By setting M_F to a low value, the correlation between trees decreases, which enhances prediction stability. However, when M_F is too low, the selected features may not be optimal, thus affecting the performance of the forest. Conversely, when M_F is set too high, the trees become similar, which results in overfitting.

The optimal value of M_F depends on the dataset used and should be adjusted through cross-validation. For regression problems, the default value of M_F is p/3, where p is the number of features in the training dataset. For classification problems, M_F can take one of four values: "none", "sqrt", "log2", and "auto", and the default value is usually "sqrt". Let the number of features in the training dataset be p (n_features), if M_F = "none", then m_{try} = p; if M_F = "sqrt", then m_{try} = sqrt(p); if M_F = "log2", then m_{try} = log2(p)+1; and if M_F = "auto", then m_{try} = p/3. The choice of M_F value should balance the stability and accuracy of each tree in the forest.

(4) The hyperparameter min_samples_leaf (m_L) specifies the minimum number of samples required for a node to become a leaf after splitting. Changing its value can affect the depth of the tree, so allowing us to control it. A small value of m_L may lead to a deeper tree, increasing the possibility of overfitting. However, if m_L is set too high, the model may fail to learn from the data.

(5) The hyperparameter min_samples_split (m_S) represents the minimum number of samples necessary to split a node into child nodes. When the number of samples in a node exceeds m_S and is not pure, the splitting process continues until purity is attained or the sample count in the node is less than or equal to m_S. By increasing m_S, the total number of splits decreases, which reduces the number of parameters and can potentially prevent overfitting. However, increasing m_S too much can result in decreased model performance.

4 Results and Discussion

This section presents the results of experimenting with each hyperparameter to evaluate its impact on the performance of the random forest algorithm. Our program was written in Java using the library Weka 3.9.6 and running on a Dell Precision 5510 laptop which is pre-installed with Eclipse software and runs on the Windows 11 64-bit operating system. The basic configuration of this laptop includes an Intel Core i7-6820HQ CPU, 24GB RAM, and NVIDIA M1000M GPU.

To ensure the objectivity and analogy of the data, we clear the cache and restart the computer after each hyperparameter result is collected.

Five hyperparameters that were to have the most influence on classifier performance were selected to testing [15, 2, Ch. 11]. These hyperparameters included: max_depth (M_D), numTree (N_T), max_features (M_F), min_samples_split (m_S), and min_samples_leaf, (m_L). For each hyperparameter, 18 different values were selected for testing within specific ranges, as follows: $M_D \in [0, 100]$, $N_T \in [1, 100]$, $M_F \in [1, 150]$, $m_L \in [0, 150]$, $m_S \in [0.001, 150]$. We then tested combinations of two, three, four, and five hyperparameters to determine the best and worst combinations. The survey results for each hyperparameter are as follows:

4.1 Effect of Max_Depth (M_D)

This hyperparameter reflects the maximum depth a tree in the forest can grow. Conceivably, the nodes are expanding until all leaves are pure or the leaves have a sample number less than min_samples_split. The default value of this parameter in Weka is 0 [18], and in Scikit-learn, it is "none" [20]. Figure 2 shows the experimental results.

The graph in Figure 2 shows that when the tree depth is too small, the model's performance is low because the input data does not provide enough information to train the model. Although the Accuracy measure received classification results, the F1-score and MCC measures did not produce results for tree depths $M_D < 7$ for the MobiAct v2.0 dataset and $M_D < 3$ for the UP-Fall dataset, as some actions occur quickly with a small number of data samples, resulting in insufficient information for evaluation. However, as the tree depth increases, the model's performance rapidly improves, along with an increase in the time taken to build the model. Both datasets had similar results, and when the tree depth increases to a certain threshold ($M_D >$ 11 for the MobiAct v2.0 dataset and $M_D > 7$ for the UP-Fall dataset), both the performance and training time of the model reached near saturation.

4.2 Effect of numTree (N_T)

This hyperparameter represents the number of decision trees (DT) used in the forest. In toolkits such as Weka and Scikit-learn, the default number of trees is 100. In this experimental part, we evaluate the effect of the number of trees in the forest on the performance and computation time of the model. Figure 3 is the synthetic results from the experimental evaluation of the influence of numTree.

The results presented in Figure 3 demonstrate that the model's performance improves with an increase in the number of trees. To a certain threshold, adding more trees does not further enhance the model's performance. The performance of RF is better when the number of selected trees exceeds the default value. However, choosing too many trees increases the model's complexity and computational time. Therefore, increasing the number of trees to improve the model's performance is not an optimal solution.

4.3 Effect of Max_Features (M_F):

In this experiment, we evaluated the influence of the max_features hyperparameter on the performance of the random forest model. This hyperparameter controls the maximum number of features the algorithm uses when searching for node splits. In Weka, the default value for this hyperparameter is 0, while in Scikit-learn, it is 1.0. The test results for two datasets, MobiAct v2.0, and UP-Fall, are displayed in Figure 4.

The synthesized results in Figure 5 indicate that the RF achieves nearly optimal performance with the default value. Gradually increasing the value of m_L causes the model's performance and the computation time to decrease, although the reduction is not significant.

4.4 Effect of Min_Samples_Leaf (m_L)

This hyperparameter controls the minimum number of samples required for a node to be considered a leaf after a split. It is set to 1 by default in Weka and Scikit-learn. Figure 5 shows the experimental investigation of the influence of this hyperparameter.

The synthesized results in Figure 5 indicate that the RF achieves nearly optimal performance with the default value. Gradually increasing the value of m_L causes the model's performance and the computation time to decrease, although the reduction is not significant.

4.5 Effect of Min_Samples_Split (m_S)

This hyperparameter controls the minimum number of samples to split into child nodes. If the number of samples in a node > m_S, then splitting continues until £ m_S. The default values for this hyperparameter in Weka and Scikit-learn are 1e-3 and 2, respectively. Figure 6 displays the results of an investigation into the influence of this hyperparameter.

Similar, to the m_L hyperparameter, the performance of RF gradually decreases as the minimum number of samples to split into child nodes increases. However, adjusting this hyperparameter does not significantly affect the computation time.

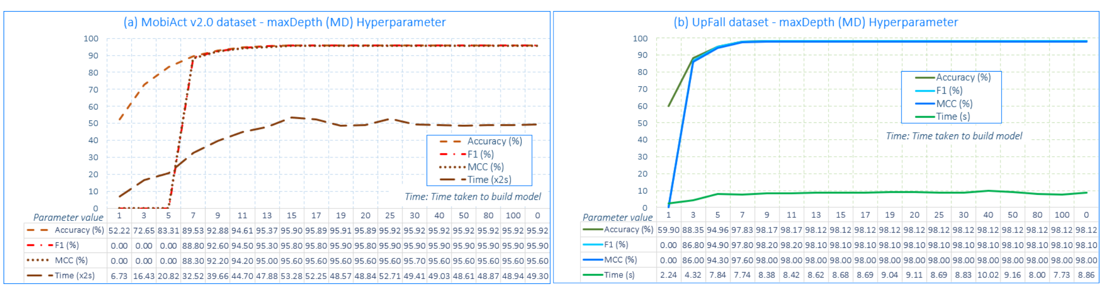


Figure 2: As the tree depth ($\boldsymbol{M_D}$) increased, the RF classification efficiency and model building time also increased on both (a) MobiAct v2.0 and (b) UP-Fall datasets

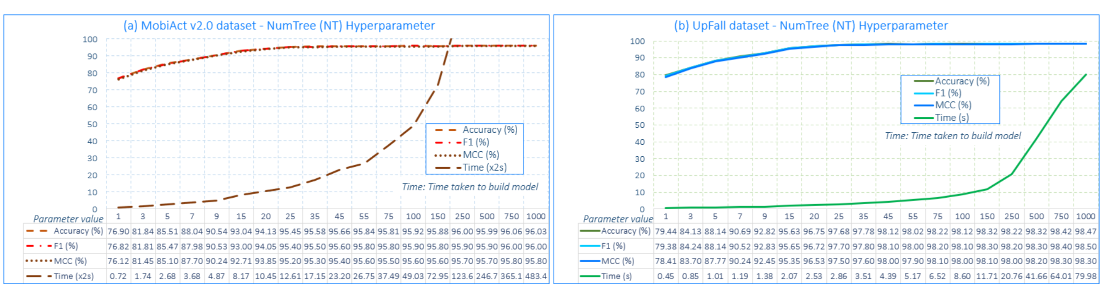


Figure 3: The accuracy in classification increases as the number of trees increases and the computation time also increases very quickly

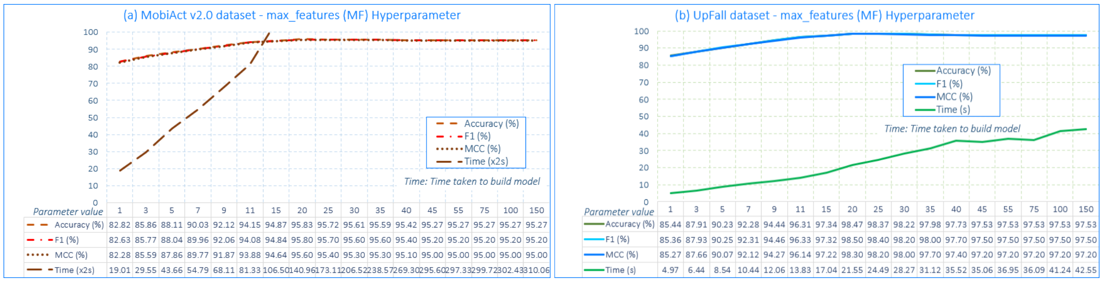


Figure 4: The performance of the model does not change much when adjusting the value of the max_features hyperparameter (MF), but the computation time increases very quickly with increasing MF

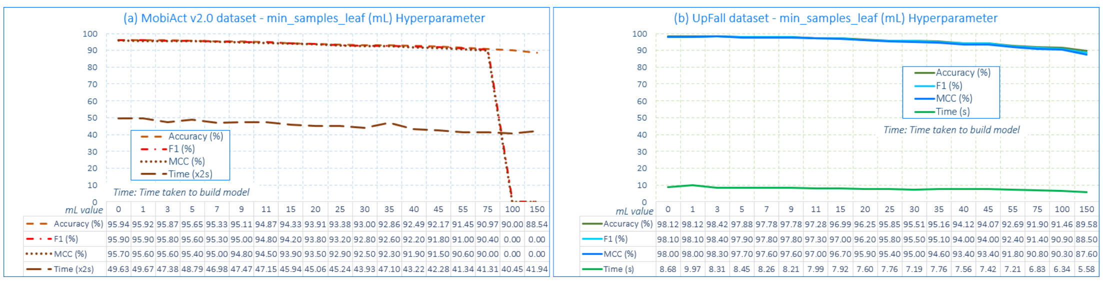


Figure 5: The model's performance decreases as the minimum number of samples in a leaf node increases

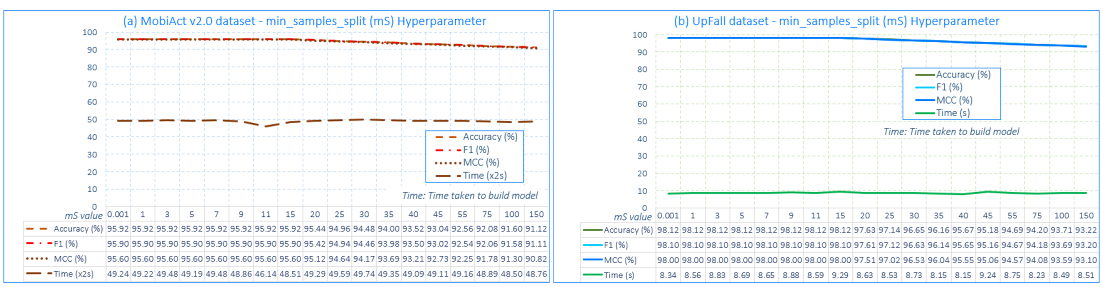


Figure 6: Increasing the minimum number of samples required to split into child nodes leads to a decrease in RF performance

4.6 Combined Tuning of Multiple Hyperparameters

In this section, we move beyond evaluating each hyperparameter independently, as described in previous sections, and investigate the simultaneous adjustment of two to five hyperparameters to assess their combined influence on the model's performance and building time. We manually select hyperparameter values based on survey results from Sections 4.1 to 4.5. The default values of hyperparameters are shown in Table 1 as cells with a dark background (blue) and light text (yellow). Specifically, we choose max_depth (M_D) = 19, NumTree (N_T) = 150, max_features (M_F) = 20, min_samples_leaf, (m_L) = 2, and min_samples_split (m_S) = 1 as they have the most positive effect on RF performance

Table 1 presents the experimental results based on the MobiAct v2.0 and UP-Fall datasets. These results indicate that combining the simultaneous adjustment of two hyperparameters, numTree, and max_features, results in the best RF performance. Additionally, combining the three hyperparameters M_D, N_T, and M_F also yields good results, similar to the combination of numTree and max_features. In contrast, combining the hyperparameters min_samples_leaf and min_samples_split yields unexpected results, with the model having the lowest performance of all combinations. Regarding time, model building is faster when hyperparameters are set to their default values than when adjusting the combination of hyperparameters.

The simultaneous adjustment of numTree and max_features is crucial in optimizing the model and often produces the most positive results. Although the combination of max_depth and

Table 1: Combined tuning of multiple hyperparameters results

No.	Code	Hyperparameters value					Metrics				Note
		MD	NT	MF	mL	mS	Accuracy (MobiAct)	F1-score (MobiAct)	Accuracy (UP-Fall)	F1-score (UP-Fall)	
		1	1	1	1	1					
1	00000	0	100	0	1	0.001	95.918	95.867	98.124	98.122	All defaults
2	11000	19	150	0	1	0.001	95.950	95.900	98.321	98.314	MD and NT are tuning
3	10100	19	100	20	1	0.001	95.931	95.893	98.469	98.465	MD and MF are tuning
4	10010	19	100	0	2	0.001	95.702	95.646	97.975	97.973	
5	10001	19	100	0	1	1	95.912	95.859	98.124	98.122	
6	01100	0	150	20	1	0.001	95.956	95.917	98.568	98.565	Best performance
7	01010	0	150	0	2	0.001	95.778	95.723	97.975	97.971	
8	01001	0	150	0	1	1	95.880	95.825	98.321	98.314	
9	00110	0	100	20	2	0.001	95.899	95.861	98.370	98.365	
10	00101	0	100	20	1	1	95.829	95.786	98.469	98.465	
11	00011	0	100	0	2	1	95.658	95.597	97.975	97.971	Worst performance
12	11100	19	150	20	1	0.001	95.924	95.885	98.568	98.565	
13	11010	19	150	0	2	0.001	95.766	95.712	97.976	97.973	
14	11001	19	150	0	1	1	95.950	95.900	98.321	98.314	
15	10110	19	100	20	2	0.001	95.855	95.814	98.370	98.365	
16	10101	19	100	20	1	1	95.931	95.893	98.469	98.465	
17	10011	19	100	0	2	1	95.702	95.646	97.975	97.973	
18	01110	0	150	20	2	0.001	95.912	95.873	98.420	98.414	
19	01101	0	150	20	1	1	95.924	95.885	98.568	98.565	
20	01011	0	150	0	2	1	95.778	95.723	97.975	97.971	
21	00111	0	100	20	2	1	95.899	95.861	98.370	98.365	
22	11110	19	150	20	2	0.001	95.816	95.775	98.420	98.414	
23	11101	19	150	20	1	1	95.956	95.917	98.568	98.565	Best performance
24	11011	19	150	0	2	1	95.766	95.712	97.975	97.971	
25	10111	19	100	20	2	1	95.855	95.814	98.370	98.365	
26	01111	0	150	20	2	1	95.912	95.873	98.420	98.414	NT, MF, mL and mS tuning
27	11111	19	150	20	2	1	95.816	95.775	98.420	98.414	Tuning of 5 hyperparameters

numTree is slightly less effective than the numTree and max_features combination, it significantly reduces model building time. It is worth noting that adjusting min_samples_leaf and min_samples_split together can reduce the model's performance. With a powerful computer system, the number of trees should be beyond 200 to keep the modelstable, which is also supported by theoretical evidence from Probst and Boulesteix [17].

5 Discussion

This research aims to identify the range of hyperparameter values that can performance enhance random forests in fall detection systems. Our findings suggest that each hyperparameter has a different impact on the performance and training time of RF, and inappropriate selection of hyperparameter values may lead to decreased model performance compared to default values.

Out of the five hyperparameters analyzed, max_depth (M_D) exerts the most substantial influence on RF classification performance. When the value of max_depth is set too small, the model performs poorly. However, optimal performance is maintained when $M_D > 15$ (for the MobiAct v2.0 dataset) and $M_D > 7$ (for the UP-Fall dataset), as shown in Figure 2.

Optimizing the tree depth can enhance the model's performance, but this comes at the expense of longer training time. Furthermore, the research underscores the significance of having adequate data samples for model training since insufficient samples may result in inadequate information for evaluation. Overall, these results indicate that a balance between tree depth and data samples is critical for attaining optimal model performance.

The default value of numTree is 100 trees that can also produce a model with good performance. However, increasing N_T beyond 100 trees can improve performance but at the cost of longer training time, which can be overcome by using powerful computers. When N_T exceeds 35 trees, RF performance becomes good and stabilizes on both datasets, as shown in Figure 3. This finding is consistent with the assessment of Probst et al. [1, 16]. Therefore, the optimal value of numTree should be chosen within the range of [35, 150] to balance performance and training time optimization.

The max_features hyperparameter plays a crucial role in the RF algorithm, especially for datasets with a large number of samples, as it significantly affects the training time. Increasing the value of max_features lead to a rapid increase in training time. In this research, with datasets consisting of 44 features, the default value of "none" outperforms the other default values of "sqrt", "log2", and "auto". As the maximum number of features is increased from 1 to 20, the performance of the random forest increases. However, once max_features exceed 20, the performance of the random forest is stable and better (Figure 4).

The simultaneous adjustment of numTree and max_features is crucial in optimizing the model and often produces the most positive results. Although the combination of max_depth and numTree is slightly less effective than the numTree and max_features combination, it significantly reduces model building time. It is worth noting that adjusting min_samples_leaf and min_samples_split together can reduce the model's performance. With a powerful computer system, the number of trees should be beyond 200 to keep the model stable, which is also supported by theoretical evidence from Probst and Boulesteix [17].

The experimental results demonstrate a correlation between the hyperparameters min_samples_leaf and min_samples_split, as both follow the same distribution and define the minimum number of samples for splitting. Figures 5 and 6 illustrate that to achieve high performance with RF, it is advisable to select small values for m_L and m_S, preferably below five.

Generally, tuning hyperparameters can have either positive or negative effects on the performance of the RF algorithm. Selecting optimal values for hyperparameters can improve RF performance and vice versa. Setting a large number of trees can enhance performance and maintain accuracy during training, but it comes at the expense of increased computational costs. Hyperparameters such as max_features, max_depth, and node size act as controls for the randomness of the RF. Of these, max_depth and numTree are the most influential hyperparameters, as per our theoretical analysis and experimental findings.

6 Conclusion

This study evaluates the impact and determines the optimal range of hyperparameters on the performance of the random forest classification algorithm in the fall detection system. Based on the test results on human activity simulation datasets collected by accelerometers, we drew the following conclusions:

(i) The number of trees (N_T) is the most crucial hyperparameter of the RF algorithm. Models with N_T < 35 yield poor classification results, while increasing N_T to 100 or higher significantly improves model performance. Therefore, to enhance the RF algorithm's performance, we should choose several trees > 100 for this hyperparameter, as supported by related literature and our experiments.

(ii) Simultaneously tuning numTree and max_features, or max_Depth and numTree, or max_Depth and numTree and max_features can improve RF performance.

(iii) The hyperparameters min_samples_leaf and min_samples_split should be selected as defaults for achieving high performance in RF. In particular, when these two hyperparameters are adjusted simultaneously, the performance of the RF is at its worst.

While our research has shown promising results in developing a random forest algorithm for fall detection, there are still certain limitations. These limitations include using only two accelerometer datasets, limiting the number of hyperparameters considered, and focusing only on the Weka toolkit without experimentation with other toolkits, such as Scikit-learn. Additionally, the hyperparameter tuning combination was limited to a manual method based on the survey results of each hyperparameter individually.

Future studies can address these limitations by expanding the range of datasets, extending the scope of surveys to multiple parameters, and utilizing various popular toolsets to evaluate the impact of hyperparameters. Also, we want to use genetic algorithms to find optimal combinations of hyperparameters for random forests to improve the performance of fall detection systems. With these lines of research in mind, we aim to develop more advanced and effective fall detection systems that enhance the safety and well-being of populations requiring special care.

Acknowledgment

The authors sincerely thank the Hellenic Mediterranean University- Department of Electrical and Computer Engineering - Biomedical Informatics and eHealth Laboratory (BMI) for sharing the MobiAct dataset. Great thanks to Martínez-Villaseñor et al. for sharing the UP-Fall dataset.

References

[1] S. Bernard, L. Heutte, and S. Adam, "Influence of Hyperparameters on Random Forest Accuracy," *Multiple Classifier Systems: 8th International Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009. Proceedings 8*, Springer, pp. 171–180, 2009.

[2] B. Boehmke and B. Greenwell, *Hands-on Machine Learning with R*. Chapman and Hall/CRC, 2019.

[3] L. Breiman, "Random Forests," *Machine Learning*, 45(1):5-32, 2001.

[4] L. Breiman, "Bagging Predictors," *Machine Learning*, 24(2):123-140, 1996.

[5] C. Chatzaki, M. Pediaditis, G. Vavoulas, and M. Tsiknakis, "Human Daily Activity and Fall Recognition using a Smartphone's Acceleration Sensor," International Conference on Information and Communication Technologies for Ageing Well and e-Health, Springer, pp. 100-118, 2016.

[6] P. Contreras, J. Orellana-Alvear, P. Muñoz, J. Bendix, and R. Célleri, "Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment," *Atmosphere*, 12(2):238, 2021.

[7] D.-M. Ge, L.-C. Zhao, and M. Esmaeili-Falak, "Estimation of Rapid Chloride Permeability of SCC using Hyperparameters Optimized Random Forest Models," *Journal of Sustainable Cement-Based Materials*, 0(0):1-19, July 2022. doi: 10.1080/21650373.2022.2093291.

[8] C. Joo, H. Park, J. Lim, H. Cho, and J. Kim, "Development of Physical Property Prediction Models

for Polypropylene Composites with Optimizing Random Forest Hyperparameters," *International Journal of Intelligent Systems*, 37(6):3625–3653, 2022.

[9] K. M. Kelkar and J. W. Bakal, "Hyper Parameter Tuning of Random Forest Algorithm for Affective Learning System," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, pp. 1192-1195, 2020.

[10] K. Lai, S. N. Yanushkevich, V. Shmerko, and M. Hou, "Capturing Causality and Bias in Human Action Recognition," *Pattern Recognition Letters*, 147:164-171, 2021.

[11] H.-L. Le, D.-N. Nguyen, T.-H. Nguyen, and H.-N. Nguyen, "A Novel Feature Set Extraction Based on Accelerometer Sensor Data for Improving the Fall Detection System," *Electronics*, 11(7):1030, 2022.

[12] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding Variable Importances in Forests of Randomized Trees," *Advances in Neural Information Processing Systems*, Vol. 26, 2013.

[13] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "UP-Fall Detection Dataset: A Multimodal Approach," *Sensors*, 19(9):1988, 2019.

[14] N. Mohapatra, K. Shreya, and A. Chinmay, "Optimization of the Random Forest Algorithm," *Advances in Data Science and Management: Proceedings of ICDSM 2019*, Springer, pp. 201-208, 2020.

[15] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and Tuning Strategies for Random Forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

[16] P. Probst, "tuneRanger: Tune Random Forest of the Ranger Package," *R Package Version*, Vol. 2, 2018.

[17] P. Probst and A.-L. Boulesteix, "To Tune or Not to Tune the Number of Trees in Random Forest," *The Journal of Machine Learning Research*, 18(1):6673-6690, 2017.

18] "RandomForest." https://weka.sourceforge.io/doc.dev/ (accessed Oct. 26, 2022).

[19] E. Scornet, "Tuning Parameters in Random Forests," *ESAIM: Proceedings and Surveys*, 60:144-162, 2017.

[20] "sklearn.ensemble.RandomForestRegressor — scikit-learn 1.1.2 Documentation." https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed Oct. 26, 2022).

[21] J. Thorn, "Random Forest: Hyperparameters and How to Fine-Tune Them," *Medium*, Sep. 26, 2021. https://towardsdatascience.com/random-forest-hyperparameters-and-how-to-fine-tune-them-17aee785ee0d (accessed Oct. 26, 2022).

[22] H. Tyralis, G. Papacharalampous, and A. Langousis, "A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources," *Water*, 11(5):910, 2019.

[23] G. Vavoulas, C. Chatzaki, T. Malliotakis, M. Pediaditis, and M. Tsiknakis, "The Mobiact Dataset: Recognition of Activities of Daily Living Using Smartphones," International Conference on Information and Communication Technologies for Ageing Well and e-Health, SCITEPRESS, pp. 143-151, 2016.

[24] L. Wang, X. Zhou, X. Zhu, Z. Dong, and W. Guo, "Estimation of Biomass in Wheat Using Random Forest Regression Algorithm and Remote Sensing Data," *The Crop Journal*, 4(3):212-219, June 2016. doi: 10.1016/j.cj.2016.01.008.

[25] M. N. Wright, T. Dankowski, and A. Ziegler, "Unbiased Split Variable Selection for Random Survival Forests using Maximally Selected Rank Statistics," *Statistics in Medicine*, 36(8):1272–1284, 2017.

[26] L. Yang and A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," *Neurocomputing*, 415:295-316, Nov. 2020. doi: 10.1016/j.neucom.2020.07.061.

[27] N. Zhu, C. Zhu, L. Zhou, Y. Zhu, and X. Zhang, "Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm," *Applied Sciences*, 12(20):10456, 2022.

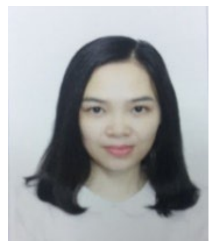
Hong-Lam Le received his B.E degree in Electrical and Electronics Engineering from Ho Chi Minh City University of Technology and Education (HCMUTE) in 2006, and M.E degree in Telecommunications Engineering from Hanoi University of Science and Technology (HUST) in 2015. He is working as a PhD student in the Information Technology Institute, Vietnam National University in Hanoi. Currently, he is a lecturer in Faculty of Electronic Engineering, Vinh University of Technology Education, Vinh city, Vietnam. His research interests include machine learning, data analysis, IoT, and sensor data processing in machine learning.

Thanh-Tuoi Le received the B.E degree in Information Technology from Vinh University (TDV) in 2003 and M.E degree in Information Technology with a specialization in Information System at the Military Technical Academy (MTA) in 2010. He has been working with the Department of Information Technology, Vinh University of Technology Education, Vinh City, Vietnam. Since 2022, he has been pursuing a Ph.D. degree in the Department of Information Technology, Hanoi National University of Education. His current research interests are in machine learning methods and their applications in molecular biology data mining.

Thi-Thu-Hien Vu received the B.E degree in Information Technology from Vinh University of Education (TDV) in 2000 and M.E the degree in Computer Science at Military Technical Academy (MTA) in 2010. Currently, she is the Vice Dean of Information Technology Department at Vinh University of Technology Education, with many years of experience in educating and researching topics such as cloud computing, embedded system, and applications.

Dinh Van Chau received his BSc from Hanoi University of Science and Technology (HUST), Vietnam, in 1998; MSc from De La Salle University, The Philippines, in 2005 and PhD at Tokyo Institute of Technology, Japan, in 2009. He is currently the Acting Rector of Electric Power University. He has many years of teaching and research experience in the fields of Engineering Physics, Nanotechnology, Energy Saving and Sustainable Development. He was principle investigator and participated in many domestic and foreign projects.

Doan-Hieu Tran received his BSc in Information Technology from the Ho Chi Minh City University of Transport in 2010 and a MSc in Information Systems from the Posts and Telecommunications Institute of Technology (PTIT) in 2014. At present, he is a Deputy Head of the Information Technology Management Department at the Ho Chi Minh City University of Banking, Vietnam. He has over 10 years of teaching and research experience in the fields of financial risk analysis, applied computer science, computer networks and communications, and Python programming for data analysis.



Ngo Thi-Thu-Trang received her B.E degree of Telecommunications and Electronics Engineering from Vietnam National University, Hanoi (VNUH) in 2002, and M.E degree of Computer and Communication Engineering from Chungbuk National University, Korea in 2005, and PhD in Communication Engineering from Posts and Telecommunications Institute of Technology (PTIT) in 2021. Now, she is a lecturer in Telecommunications Faculty 1 of PTIT. Her research interests include digital signal processing, optical communication, and broadband networks.

Collaborative Cloud-V. Edge System for Predicting Traffic Accident Risk Using Machine Learning Based IOV

Zeroual Djazia [*]
University of Biskra, ALGERIA

Okba Kazar[†]
United Arab Emirate University, UAE.

Saad Harous[‡]
University of Sharjah, UAE.

Saber Benharzallah[§]
Batna 2 University, ALGERIA

Abstract

Smart city development is profoundly impacted by cutting-edge technologies such as information and communications technology (ICT), artificial intelligence (AI), and the Internet of Things (IoT). The intelligent transportation system (ITS) is one of the main requirements of a smart city. The application of machine learning (ML) technology in the development of driver assistance systems, has improved the safety and the comfort of the experience of traveling by road. In this work, we propose an intelligent driving system for road accident risks prediction that can extract maximum required information to alert the driver in order to avoid risky situations that may cause traffic accidents. The current acceptable Internet-of-vehicle (IOV) solutions rely heavily on the cloud, as it has virtually unlimited storage and processing power. However, the Internet disconnection problem and response time are constraining its use.

In this case, the concept of vehicular edge computing (V.Edge.C) can overcome these limitations by leveraging the processing and storage capabilities of simple resources located closer to the end user, such as vehicles or roadside infrastructure. We propose an Intelligent and Collaborative Cloud-V.Edge Driver Assistance System (ICEDAS) framework based on machine learning to predict the risks of traffic accidents. The proposed framework consists of two models, CLOUD_DRL and V.Edge_DL, Each one complements the other. Together, these models work to enhance the effectiveness and accuracy of crash prediction and prevention. The obtained results show that our system is efficient and it can help to reduce road accidents and save thousands of citizens' lives.

Key Words: IOV, deep learning, deep reinforcement learning, cloud computing, V.Edge computing, cloud-V.Edge collaboration.

[*] Computer Science Department. Email: d.zeroual@univ-batna2.dz.

[†] Department of Information Systems and Security.

[‡] Computer Science Department.

[§] Computer Science Department.

1 Introduction

Every year the lives of approximately 1.3 million people are cut short, as a result, of road traffic accidents. Between 20 and 50 million people, suffer non-fatal injuries, however, many of these cases result in different kinds of disabilities. Timely and accurate prediction of traffic accidents has great potential to ensure traffic safety and reduce economic losses. To enhance this traffic safety, many studies have been conducted to help the development of an Active Traffic Management System. The main areas of interest covered by these studies are i) black-spot detection where road traffic accidents have been concentrated [16]. ii) Detection of traffic incidents in real time and alert people to reduce their effects [37]. iii) Road accidents prediction, where the prime goal of this research is to predict the road accidents before they occur [24, 17, 32]. Road accident prediction is a field of significant and contemporary scientific interest. The prediction with high spatiotemporal resolution is difficult, mainly due to the complex traffic environment, human behavior and the lack of real-time traffic-related data [29]. With the recent development of Internet-of-vehicle (IOV) technology and the advancement in wireless communications, and computational systems, new opportunities have opened-up for intelligent traffic safety, comfort, and efficient solutions.

The interest in machine learning has increased exponential due to the wide availability of parallel computing technology and a large amount of training data [5]. In particular, the success of deep learning (DL) technique led the researchers to investigate the application of machine learning for traffic accident prediction. Reinforcement learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment; its combination with DL generates a new powerful algorithm called deep reinforcement learning (DRL). These algorithms have recently been very successful in implementing road safety applications. However, the enormous resources required by these algorithms pose a great challenge to the limited computational and storage resources that are available on-board every vehicle.

The Internet-Of-Vehicle (IOV) solutions to traffic safety problems rely heavily on the cloud, as it has virtually unlimited storage and processing power; where data must be moved from the data source location (IOV sensors) to a centralized location in the cloud. However, in addition to the Internet disconnection problem, the cloud might be far from the location of sensors and devices generating these data, which will cause the response time to be slow. Therefore, this might restrict the use of a solution that is based on the cloud, for sudden car accident prediction.

The concept of V.Edge Computing is an efficient alternative to overcome the limitations of using machine-learning models in the cloud platform. Many emergency predictions take place close to the end user; therefore, they can be processed at the edge nodes. This reduces the impact of communication delay and internet disconnection. We propose an Intelligent Collaborative Cloud-V.Edge Driver Assistance System (ICEDAS) framework based on machine learning, which predicts the risks of traffic accidents.

This framework takes advantage of the strengths of the two platforms, where a Deep Q-Learning Network (DQN) algorithm is adopted in the cloud, in order to train an intelligent agent to warn the driver of any foreseeable risk of traffic accident based on the huge historical data available on the cloud. On the other hand, a deep learning algorithm can be deployed on the V.Edge platform for inference, covering potential response absences by the cloud in predicting sudden traffic risk due to the platform's proximity to the end user. The DL algorithm is trained in the cloud, taking advantage of its scalability and high-end computing resources for model training. Figure 1 illustrates the three layers of our system. The proposed (ICEDAS) aims to achieve the following main objectives:

1) The system must be able to react in a timely manner to warn the driver before entering a critical state.

2) The system must deliver adaptive messages to each driver who is at risk of a traffic accident based on their personal conditions.

3) The system must have the ability to use the cloud and V.Edge to predict the accident risk prediction in an efficient manner.

The rest of this paper is organized as follows: Section 2 presents the literature review of existing works. In Section 3, we briefly describe the proposed system architecture. In Section 4, we provide details regarding the intelligent Cloud-V.Edge system for predicting road accident risk. The experimental results are presented in Section 5, and the work is concluded in Section 6.

2 Related Work

Many researchers have investigated the use of machine learning for traffic safety and accident risk prediction during the past few decades. In this section, we review the different categories of these studies.

2.1 Traffic Accident Prediction using Classical Techniques

Numerous researchers have approached the prediction of traffic accidents by considering it as either a classification problem or a regression problem. In this section, we will explore several studies that have utilized classical machine learning techniques to address this problem. For instance, in [20] Lv et al. investigated the identification of potential traffic accidents by employing the k-nearest neighbor method with real-time traffic data. Hossain and Muromachi [12] utilized a Bayesian belief net (BBN) for real-time crash prediction on basic freeway segments of urban

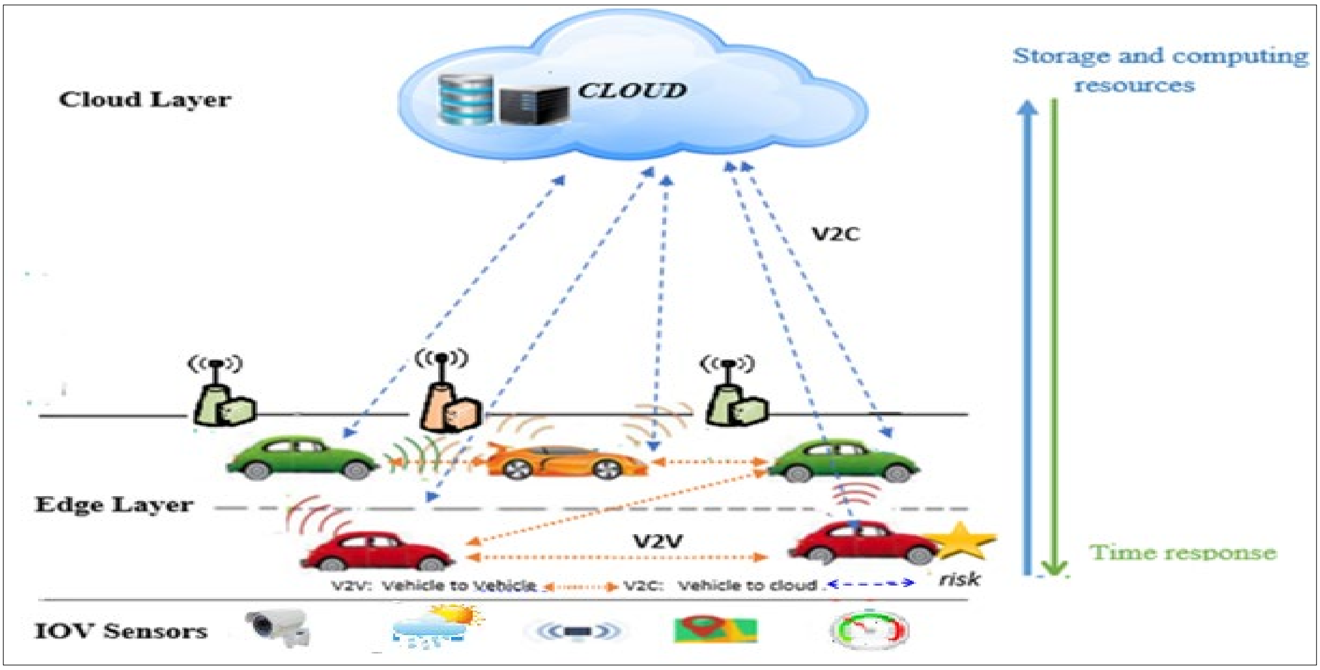


Figure 1: Vehicular cloud – edge system architecture

expressways. In another study by Lin et al in [18], a novel Frequent Pattern Tree (FP Tree) approach based on important variable selection was proposed to achieve an acceptable level of accuracy in real-time traffic accident risk prediction. Chang and Chen in [6] developed a decision tree model to build a classifier for accident prediction, achieving training and testing accuracies of 55%. In [13] Karlaftis and Vlahogianni compared statistical methods with neural networks (NN) in transportation-related research and demonstrated the promising potential of NN-based solutions. Furthermore, Zhang et al utilized in [40] a statewide live traffic database to develop real-time traffic crash prediction models. They compared Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) models. Although many previous systems have treated traffic accident risk prediction as a classification problem, their prediction accuracy has been relatively unsatisfactory.

2.2 Deep Learning for Traffic Accident Prediction

Recently, with the rapid advancements and remarkable achievements in machine learning technologies, several recent studies have embraced deep learning methods for predicting traffic accidents. In the study [28] by Ren et al., a deep learning approach based on Recurrent Neural Networks (RNN) was proposed to predict traffic accident risk, where risk was defined as the number of accidents occurring in a specific region at a given time. Another method by Chen et al called "STENN" in [7] was introduced for traffic accident prediction, incorporating multiple factors such as spatial distributions, temporal dynamics, and external factors to improve prediction accuracy. Yu et al. [38] developed an autoencoder deep architecture to examine the impact of human mobility on traffic accident risk. They utilized this approach to gain insights into how human mobility patterns influence the occurrence of traffic accidents. In the work conducted by Yuan et al. [39], a Hetero-Convolutional Long-Short Term Memory (Hetero-ConvLSTM) model was proposed to forecast the number of traffic accidents in Iowa. This model incorporated both spatial and temporal features, enhancing the accuracy of accident predictions. The focus of the research by Gutierrez et al. [10] was on developing a Deep Learning Ensemble Model that utilizes information extracted from social media to predict traffic accidents. Peng et al. [26] presented DeepRSI, a real-time road safety prediction framework that utilized mobile sensing data collected in Vehicular Ad-Hoc Networks (VANETs).

2.3 Deep Reinforcement Learning for Traffic Safety

In recent years, deep reinforcement learning (DRL), an advanced form of artificial intelligence, has gained significant importance in intelligent decision-making across various domains. DRL has found applications in robotics [35], healthcare [33], Natural Language Processing [2], and sentiment analysis [34].

In the field of transportation systems, DRL algorithms have been widely utilized, particularly in traffic control tasks. For example, DRL has emerged as the most popular machine learning methodology for traffic signal control [11]. In ramp metering control [19], a DRL-based method was proposed to leverage video traffic data for improving the efficiency of ramp metering. This approach utilized traffic video frames as inputs and learned optimal control strategies directly from visual data. In the context of intelligent transportation, an improved Deep Q-Learning Network (DQN) method has been adopted to train intelligent agents for guiding vehicles to their destinations and avoiding congestion [15]. Furthermore, in reference [5], Deep Reinforcement Learning (DRL) has been utilized to develop autonomous braking systems capable of intelligently regulating vehicle velocity to prevent collisions. For autonomous driving or self-driving cars, DRL algorithms have garnered considerable attention and have been the subject of extensive research [3, 4, 30, 41].

The DRL algorithm has proven highly efficient in solving complex decision-making problems that were previously beyond the capability of traditional machine learning techniques. However, when operating in a dynamic environment, such as in the case of traffic prediction and prevention, the algorithm requires frequent updates of the data being exploited in order to provide reliable predictions. Additionally, due to the significant storage and computing resources required, its application is best suited for deployment on a cloud platform.

3 Proposed System Architecture

Our main goal is to develop a framework that leverages machine learning techniques to help drivers in safe driving practices. We plan to achieve this by analysing large amounts of data from previous accidents. The proposed framework consists of an intelligent and collaborative driver assistance system, called 'ICEDAS' that operates between the cloud and a vehicle's edge. Figure 2 illustrates the two layers in this framework, which work together to safeguard drivers and minimize the risk of road accidents.

3.1 Cloud Layer

Cloud computing is one of the most significant trends in the information technology evolution, as it has created new opportunities that were never possible before [14]. Due to its storage capacity and computing power, we consider it the suitable location to generate the two machine learning models in our system. The first model is DRL, which is the main component in our framework. It runs on the cloud to predict accident risks. The second model is a DL, also generated in the cloud and then deploy it to the V.Edge device for inference when needed, to cover the absence of prediction by the cloud.

3.2 V.Edge Layer

Vehicular Edge Computing (VEC), based on the edge computing motivation and fundamentals, is a promising technology supporting ITS services, and smart city applications [22]. In our system, the V.Edge is used to replace the cloud in certain cases, such as internet disconnection or bandwidth overload. Vehicles equipped with cameras, radars, GPS, and other devices can sense both the internal and external environment and collect various information such as speed, road quality, position, and more. These data are either sent to the cloud in real-time for prediction by DRL, or used by the edge itself to replace cloud prediction in generating

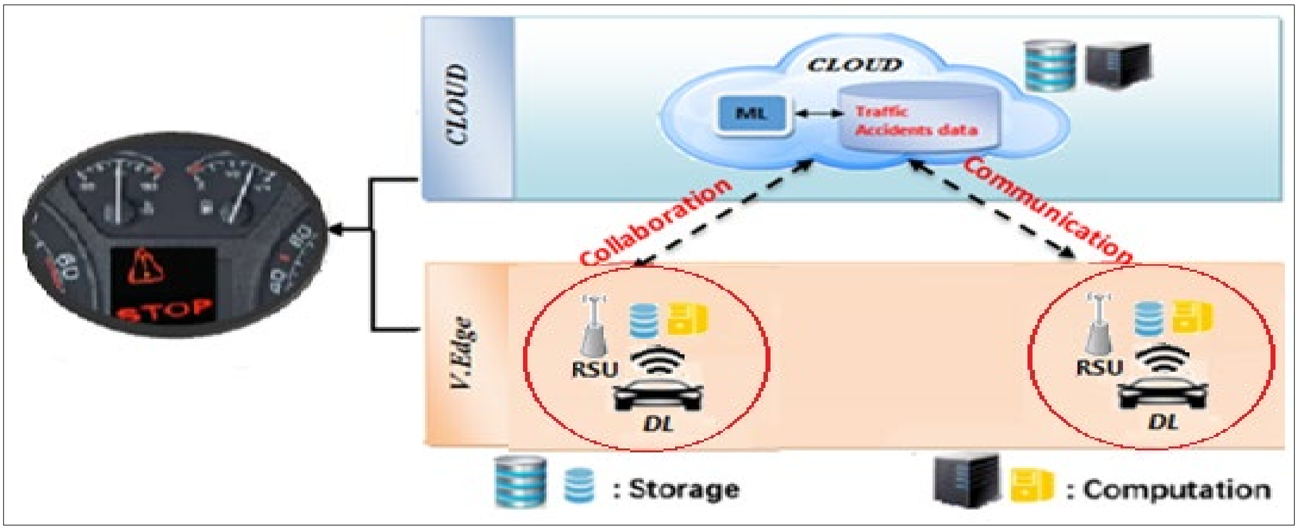


Figure 2: Cloud_DRL and V.Edge_DL system architecture

accident risk alerts using the inferred DL model.

4 Methodology

In this section, we describe the general structure of our ICEDAS. We first define possible scenarios then we discuss its operations in detail.

4.1 Scenarios

Many accidents occur when driving conditions suddenly change. ICEDAS must detect the potentially accident-causing events in advance and help the driver take the appropriate actions to avoid them. To predict a traffic accident risk, we focus on many contributing factors that frequently cause traffic accidents. They are often related to *Drivers, Roads or Vehicles* such as: *Driver's age, Driver's Sex, Driver's experience, Road condition, Light condition, Weather condition, Type of vehicle, Service year, etc.*

When a sudden change in any of the car's normal conditions is accurately detected, it may be difficult to adapt properly to this change, which may become a threat to the car. In this case, we need an intelligent risk prediction system that adapts to different situations of this risk. Markov Decision Process (MDP) is a powerful technique for modeling sequential decision-making problems. We used MDP to formulate our problem. In MDP framework, an agent interacts with a given environment state by taking actions at discrete time steps. In our system, we assume that the traffic environment follows the discrete-state. Figure 3 describes this Markov process. The state (SN-risk) implies that the system did not detect any risk. Once a risk is detected, the state (SN-risk) is changed to the state (S-risk). In practical scenarios, it is difficult to know the transition probabilities of the Markov process and the distribution of the environment states. Therefore, reinforcement-learning approach can be applied to learn the risk prediction policy through the interaction with the environment.

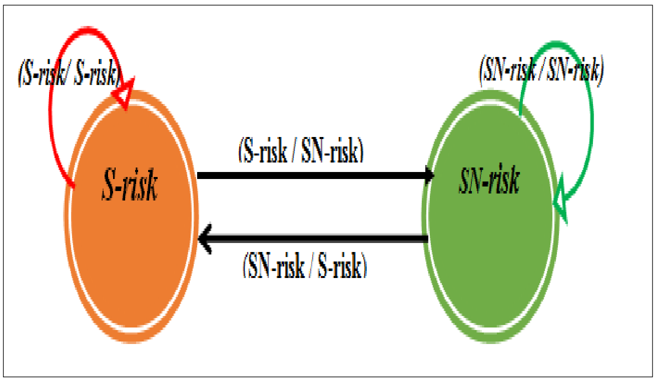


Figure 3: Traffic environment-state description by the discrete-state Markov process

4.2 Cloud Deep Reinforcement Learning for Traffic Accident Risk Prediction

In this section, we present the details of the proposed Cloud_DRL based risk prediction system. We first present the structure of DQN and explain in detail how it works to train the learning model based on accidents data available in the cloud.

4.2.1 Key Elements of Cloud _DRL. There are four key elements in this DRL system: Cloud-Agent, observation/state, action, and reward scheme.

We formulate traffic accident risk prediction problem as a reinforcement learning problem shown in Figure 4, where the Cloud-Agent interacts with the vehicle traffic environment in discrete time steps (t_0, t_1, t_2… t_N). The agent's objective is to reduce the number of accidents.

- *Cloud –Agent*: the agent observes the state of each vehicle, in its environment, defined by St_i at the beginning of time step t_i, then selects an action $At_i \in A$ to perform. The use

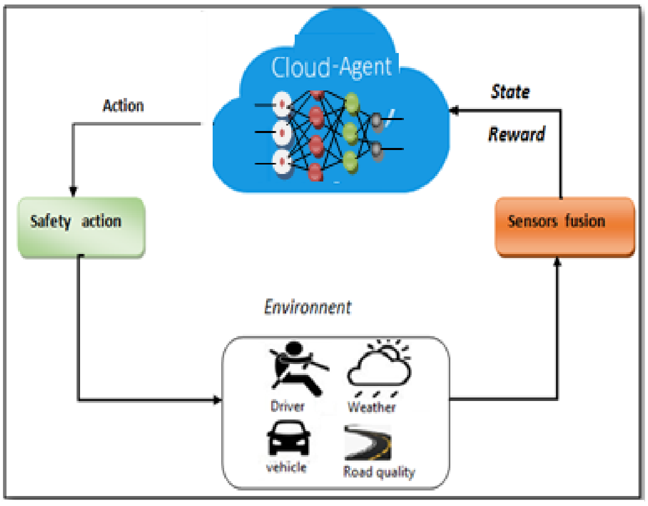


Figure 4: Proposed Cloud DRL based accident risk

of deep neural network (DNN) model, in this case, is very appropriate because of the huge number of states. The DNN take input observations about traffic accidents and produces action decisions that should be taken as its output. The DNN architecture is a multilayer-network where the Cloud–Agent explores the information (available in the Cloud) about various accidents that have occurred previously and recommend the best actions that must be applied to avoid similar accidents from happening again.

- *Action*: refers to the decision recommended by a Cloud - Agent. It is a feedback on a state of risk accident, which is one of the following actions (*Stop, Deceleration, and No-Change of lane)* as an output to avoid this risk of accident.
- *State*: is an efficient representation of current road traffic condition. The representation variables contain multiple parameters reflecting the circumstances of a specific zone of an urban transportation network to precisely describe the complexity of its dynamics. The agent learns through interacting with the environment episode by episode, where each episode ends with the prediction of accident risk for a vehicle, and the next episode starts.
- *Reward (penalty):* the agent gets a reward Rt_i at the end of time step t_i as a result of the applied action At_i. The key requirement for a successful application of reinforcement learning is to design a reward function that frames the goal of an application and guides the learning towards a desirable behavior [23]. To reduce the traffic accident risk, it is reasonable to reward the agent at each time step for choosing an action that led to the avoidance of accidents [9]. Therefore, we determine the reward (penalty) Rt_i for the agent who chooses an action At_i at time step t_i as follows:

$$Rt_i = \left\{ \begin{array}{c} 0 \quad \longrightarrow (No-Change) \\ 1 * (\mathrm{N}) \longrightarrow (Deceleration) \\ 2 * (\mathrm{N}) \longrightarrow (Stop) \end{array} \right\} \qquad (1)$$

Where *N is a negative integer,* which represents the severity of an action. The agent can perform one of these actions (*No-Change*, *Deceleration or Stop*) according to accident severity: *{(0) Negligible risk, (1) Serious risk and (2) Fatal risk}*

The goal of reinforcement learning system is to achieve a safe road traffic system with no accident risk rate during the evaluation time (T). This is represented by the Total-Reward ($T\_Rt_i \approx 0$):

$$T\_Rt_i = \sum_{i=1}^{T} |\mathrm{Rt_i}| \qquad (2)$$

4.2.2 Deep Q-Network (DQN). There are classical RL algorithms such as Q-learning, Policy Gradient (PG), Actor Critic, etc. Q-learning is one of the popular RL methods, which search for the optimal policy in an iterative fashion [5]. This algorithm is not suitable when we have a huge number of states and complex state transitions. In this work, a DQN algorithm that uses a DNN is utilized for predicting accident risks, with the aim of enhancing both the speed and accuracy of predictions. For each episode, the Cloud-Agent observes

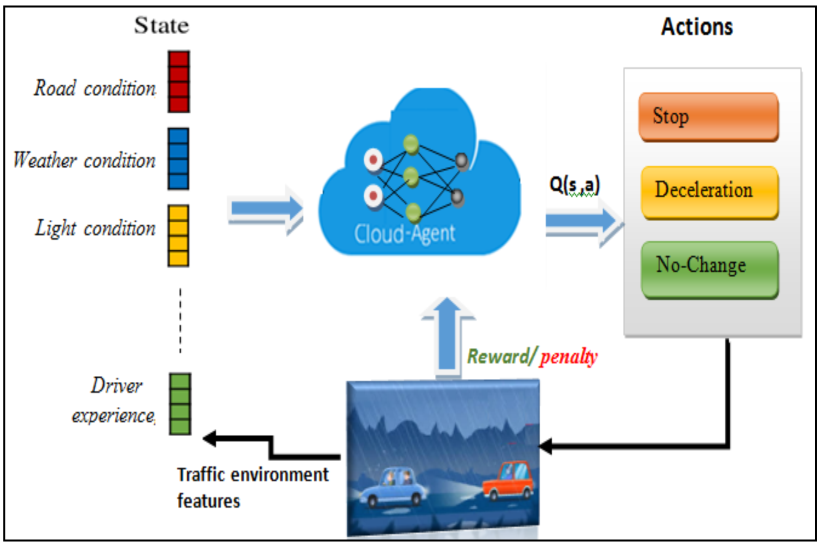


Figure 5: Cloud _DQL accident risk prediction

state St_i at the beginning of time step t_i, then makes action decision according to vehicle state, and receives a sequence of rewards (Rt_i) after time steps. If the cloud agent aims to reduce vehicle road accidents, it is sufficient to choose an action that maximizes the immediate reward Rt_i.

Since the agent aims to reduce the number of accidents in the long run, it needs to find an optimal policy noted ($\pi*$) at every possible state-action pair. To find the optimal policy $\pi*$, we need to find the optimal Q-value:

$$Q\pi*(s, a) = \max Q\pi(s, a)$$

$$= Q*(s, a) \quad (3)$$

When the state space is continuous, it is impossible to find the optimal value of the state-action pair $Q*(s, a)$ for all possible states. To deal with this problem, the DQN method was proposed, which approximates the state-action value function $Q(s, a)$ using the DNN, i.e, $Q(s, a) \approx Q\theta(s, a)$, where θ are parameters of the DNN that will be learned from raw traffic accident data.

We construct such a DNN network, where the network input is the observed traffic environment state St_i and the output is a vector of estimated Q-values Q(s, a, θ) for all actions a ∈ A under observed state St_i. Figure 5 illustrates the Cloud_DQN module for traffic accidents prediction. Real-traffic accident data was collected in a buffer called a replay buffer to train our network. We build a neural network connected to several layers so that DNN approaches the Q-value. The agent learns parameters θ by training the DNN network to minimize the following mean squared error (MSE) as the loss function. MSE can be defined as the average squared difference between the target value and the predicted value [27], as shown in Equation (4):

$$MSE = \frac{1}{K}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2 \quad (4)$$

Where y is the target value, $\hat{y}$ is the predicted value, and K is the number of training samples. Our target value should be the optimal Q value; the optimal Q value can be obtained by using the Bellman optimality Equation (5), where its Q value is just the sum of the reward (r) and the discounted maximum Q value of the next state-action pair [27]:

$$Q^*(s,a) = r + \Upsilon maxQ^*(s',a') \quad (5)$$

Therefore, we can define our loss as the difference between the target value (the optimal Q value) and the predicted value (the Q value predicted by the DQN) and express the loss function L as (6) [27]:

$$L(\Theta) = Q^*(s,a) - Q_\Theta(s,a) \quad (6)$$

Substituting Equation (5) in Equation (6), we get Equation (7).

$$L(\Theta) = r + \Upsilon maxQ(s',a') - Q_\Theta(s,a) \quad (7)$$

The Q value of the next state-action pair in the target is computed by the target network parameterized by **θ'** and the predicted Q value is computed by the main network parameterized by **θ**. The loss function is represented by Equation (8).

$$L(\Theta) = \frac{1}{K}\sum_{i=1}^{K}(r_i + \Upsilon maxQ_{\Theta'}(s',a') - Q_\Theta(s,a))^2 \quad (8)$$

The target network has the same architecture as the main network but different weights. Every N step, the weights from the main network are copied to the target network, where N is a hyperparameter that can be set by the user. Using both networks leads to more stability in the learning process and helps the algorithm to learn more effectively. To find the optimal parameter θ, we use gradient descent. We compute the gradient of our loss function$\nabla_\Theta L(\Theta)$and update the network parameter θ as:

$$\theta = \theta - \alpha\nabla_\theta L(\theta) \quad (9)$$

The algorithm for training the Cloud_DQN is defined on next page.

4.3 V.Edge Deep Learning

Deep learning is one branch among the many fields of machine learning, and it is based on artificial neural networks [21]. Since DL often requires high performance computing resources (GPUs, CPUs and storage devices) for model training and execution on massive data [36], the resources available in a vehicle may not fulfil this stringent requirement. Meanwhile, there is an imprecise trend: the more layers and parameters of a deep neural network, the more accurate the decision-making, which would undoubtedly increase the training and running cost of deep learning models (DLMs) [36]. In this case, the cloud is the best solution to handle a huge traffic accident data due to its scalability, availability of resources, and cost-effectiveness.

In some situations, such as sudden accidents, where a fast response is the most important variable in the accident risk prediction problem, it is not always effective to rely on the cloud to send risk predictions. This is because predictions sent from the cloud to the driver may be lost or delayed due to internet disconnections or bandwidth overloads.

The best solution is to build a deep learning model based on big data for traffic accident risk prediction in cloud platform, and then transfer it to the V.Edge to cover this cloud prediction absence. The V.Edge_DL can learn deep connections between traffic accidents and their spatial-temporal patterns. Deep learning is a deeper network of neurons, which consists of input layer, hidden layers, and an output layer. It aims to exploit historical traffic accident data to avoid their reoccurrence again. The input layer of the model would represent the variables that are known to influence accident severity, such as crash timing, speed limit, weather conditions, and so on. The output layer would represent the degree of risk according to the severity, which could be classified into three categories: *Negligible risk, Serious risk,* and *Fatal risk.* We construct a DL model by region, which are transferred to the *V.Edge* when it is needed. Figure 6 illustrates V.Edge_DL *construction* and *exploitation*.

Algorithm DQN algorithm

Initialize the main network with random weights θ
Initialize the target network with random weights θ' by copying the main network parameter θ
Initialize the replay memory capacity as *D;*

For *episodes = 1, M* **do**
(1) Randomly select the initial road traffic state *s*

For t = 1, T **do**
(2) Observe the state *s and select* random action *a* with probability ε otherwise select
a =argmaxQ_θ (s, a)
(3) Execute action *a and move* to the next state s'and obtain the reward *r*
(4) Store experience (s, a, r, s') in replay memory *D*
(5) Randomly sample a minibatch of *K* transitions from the replay memory *D*
(6) Compute the target value, that is, y= r + ɣ max $Q_{\theta'}$ (s', a')
(7) Compute the loss function valueL *(θ) with* (8)
(8) Compute the gradients of the loss and update the main network parameter *θ using (9)*
(9) Every N step update θ' ← *θ*
End For

End For

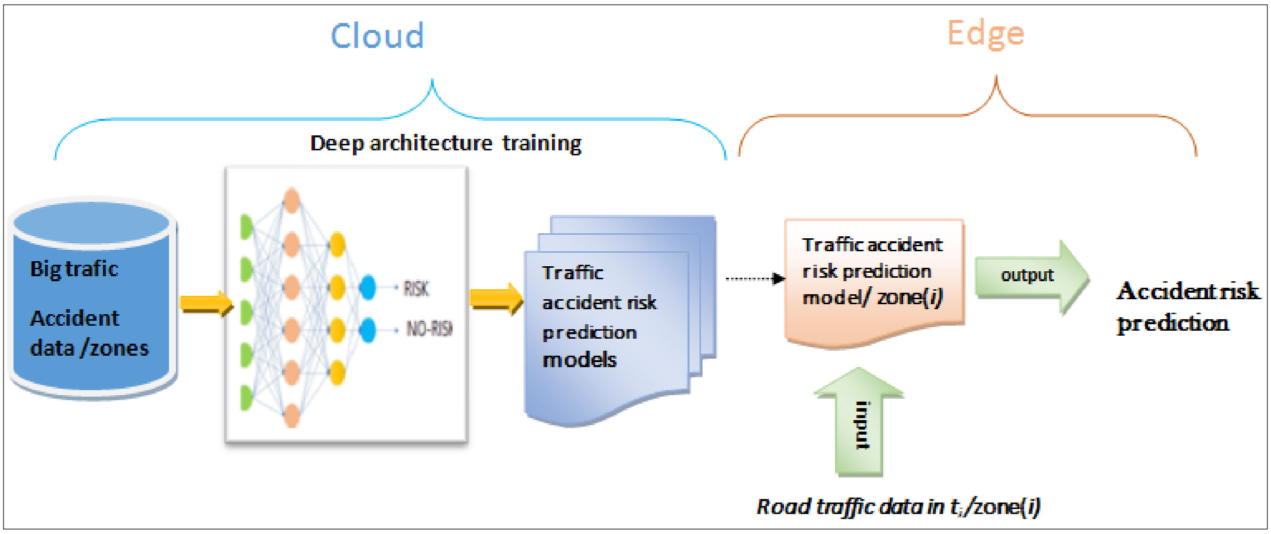


Figure 6: Regional V.Edge_DL traffic accident risk prediction

4.4 Collaboration V.Edge _DL / Cloud _DRL

The V.Edge does not have sufficient capacity to store and process a large amount of IOV data and generate DL models, so, it is not easy to ensure an absolute quality of traffic safety. Therefore, it uses models generated in cloud level, Figure 7.

We have adopted a collaborative work between Cloud platform and V.Edge platform through a distributed learning system that uses both platforms for an optimal prediction. The cooperation-communication between these platforms can have a vertical *V2C* (V.Edge- Cloud) or horizontal *V2V* (V.Edge–V.Edge) type.

4.4.1 V.Edge – Cloud.

- Communication (V.Edge - Cloud): In IOV technology, the sensors enable gathering information about the road, the vehicle and the driver, to be sent to the cloud using V2C

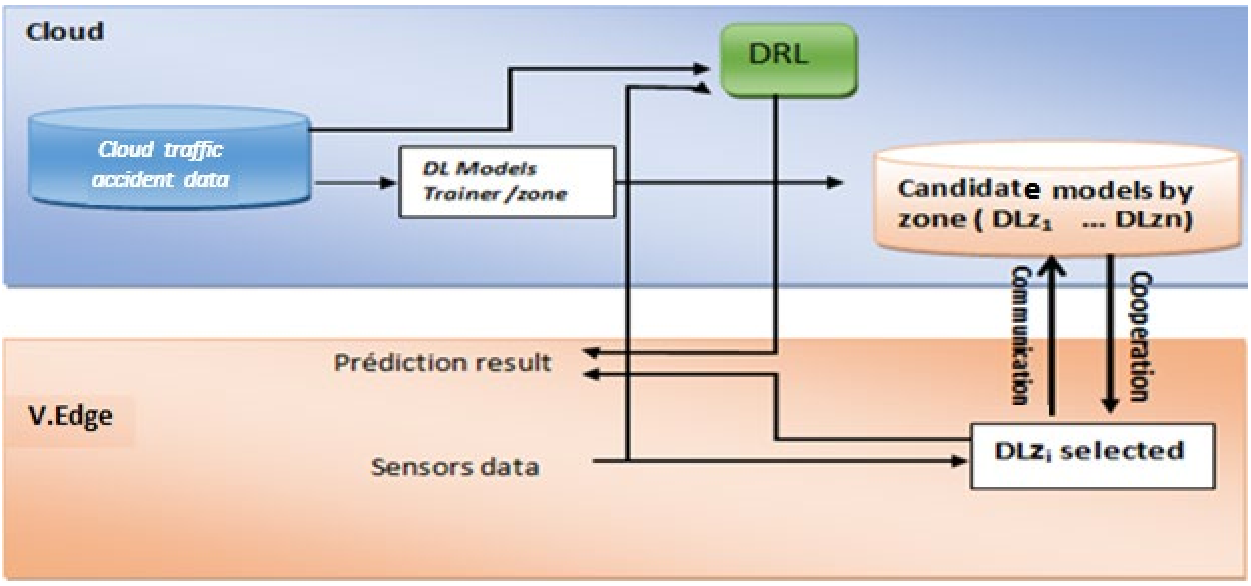


Figure 7: Collaboration / communication (V. Edge -cloud)

(vehicle to Cloud) communication and interaction. This information is used to predict if there is a risk of accident occurring and in this case, the driver is warned by an alert message sent by the Cloud_DRL entity as soon as possible.

- Collaboration (Cloud -V.Edge): Collaboration between cloud and V.Edge platforms can involve training a deep learning model in the cloud using high-end computing resources, and then deploying the model to the V.Edge device for inference when needed. This approach can help to cover potential response absences by the cloud and ensure real-time data processing and decision-making without delay.

4.4.2 V.Edge – V.Edge. Embedded a deep learning model into different vehicles enables effective collaboration and communication among them for accurate prediction of road accidents.

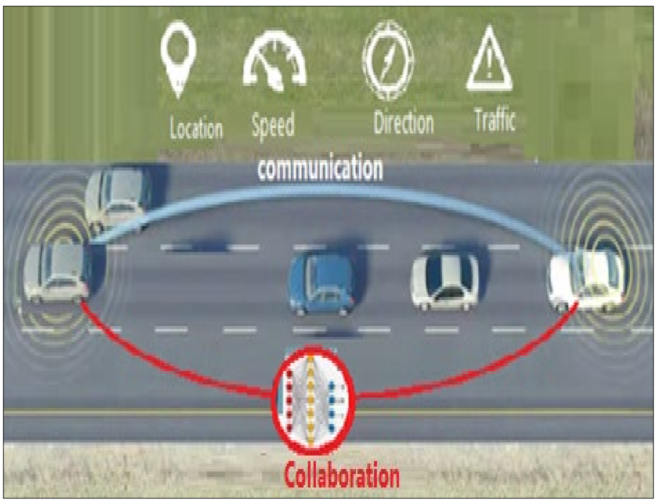


Figure 8: Collaboration / communication (V. Edge - V.Edge)

- Communication (V.Edge - V.Edge): Vehicles communicate with each other and exchange data through wireless communication protocols (V2V). The shared data may include information about the vehicle's speed, direction, location, or any other relevant data that could help the vehicles avoid potential collisions and dangerous behaviors.
- Collaboration (V.Edge – V.Edge): A vehicle can refer to another vehicle for importing the deep learning model of its new zone in case of internet problems with the cloud.

5 Experiments and Results

In this section, after describing the used data, we evaluate the effectiveness of the proposed models. Several machine learning methods are compared through a series of experiments. All implementations are in Python which utilize Tensorflow [1], Keras [8], and scikit-learn [25] libraries.

5.1 Data

To evaluate our accident risk prediction framework, we utilized road accident data from the United Kingdom, which is available on the website www.data.gov.uk. The data includes accident information ranging from the year 2005 to 2017 with 34 features, and vehicle information ranging from 2004 to 2016 with 24 features, comprising two million records. The dataset is considered a big data, which requires preprocessing to improve the performance of machine learning models and obtain accurate results. The preprocessing steps include data cleaning, data transformation, and data reduction. A machine learning feature selection method such as the Scikit-learn Random Forest library was used to identify the most relevant and correlated attributes influencing the learning process, which are depicted in Figure 9. Table 1 presents the important features description of this dataset, which will form the input vector of our ML models.

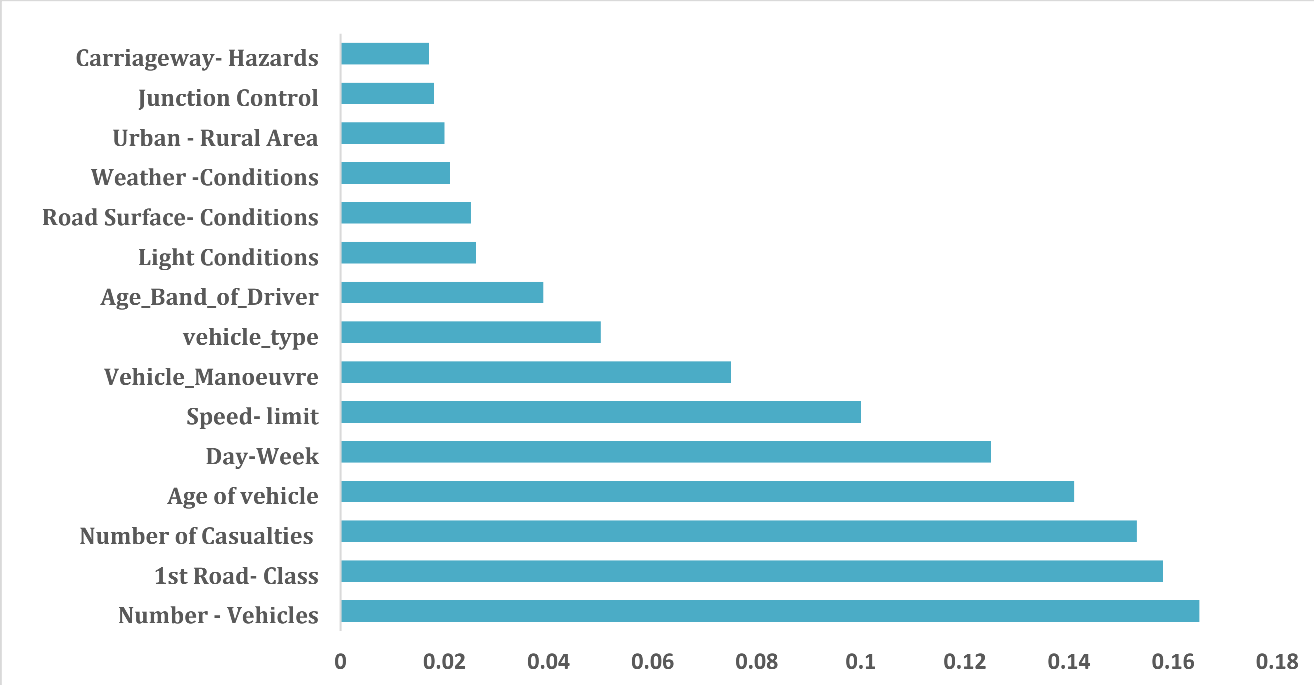


Figure 9: Attribute importance scores

Table 1: Input Factor

Variable	Label
Number – Vehicles	Vehicles involved
1st Road- Class	Motorway, A (M), A, B, C , Unclassified
Number of Casualties	Casualties involved
Age of vehicle	0-10,11-20,21-30,31-40, Above 40 years, Missing
Day-Week	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
Speed- limit	Speed limit in [mph]
Vehicle_Manoeuvre	Going ahead, Turning left/right/U, Reversing, Parked, Slowing/stopping/waiting, Overtaking, Others, Missing
vehicle_type	Pedal cycle, Motorcycle, Car, Bus, Truck, Others
Age_Band_of_Driver	<24,25–34,35–44,45–54,55–64,65–74,>75
Light Conditions	Daylight, Darkness—lights lit, Darkness—lights unlit, Darkness—no lighting, Darkness—lighting, unknown
Road Surface- Conditions	Dry, Wet, or damp, Snow, Frost or ice, Flood over 3 cm deep, Oil or diesel, Mud
Weather –Conditions	*Fine no high winds, Raining no high winds, Snowing no high winds, Fine + high winds, Raining + high winds, Snowing + high winds, Fog or mist, Other, Unknown*
Urban - Rural Area	*Urban, Rural, Unallocated*
Junction Control	*Not at junction or within 20 m, Authorized person, Auto traffic signal, Stop sign, Give way or uncontrolled*
Carriageway- Hazards	*None, Vehicle load on road, Other object on road, Previous accident, Dog on road, Other animal on road, Pedestrian in carriageway—not injured, Any animal in carriageway (except ridden horse)*

5.2 Evaluation Metrics

It is necessary to identify and estimate the efficiency and effectiveness of Cloud_DRL and V.Edge_DL in predicting traffic accidents with the data set. Our models are validated in terms of:

- Learning curves (Accuracy and Loss) for both of *Cloud_DRL and V.Edge_DL.*
- Comparison with other algorithms in terms of evaluation metrics.
- Efficiency and effectiveness in reducing the risk of road accidents with or without cooperation.

The calculation of evaluation metrics is mainly based on (N x N) confusion matrix (shown in Figure 10) that is used to display the performance of the algorithm, where N is the number of target classes. This matrix compares the actual target values with those predicted by the machine learning model. To comprehensively measure the performance of the proposed models, accuracy, sensitivity, F1 score, and other indicators are used. The concept and formula for calculating each of these indicators are shown in Table 2. Where TP denotes true positive, FP denotes false positive, TN denotes true negative, and FN denotes false negative.

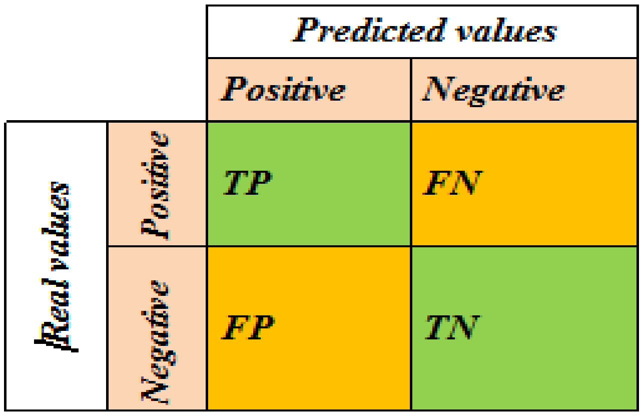


Figure 10: Confusion matrix

5.3 Results and Discussion

During the construction of our machine learning models, the dataset was divided into training dataset (80%) and test dataset (20%).

Table 2: Main metrics for classification

Metric	Formula	Interpretation
Accuracy (Acc)	$\frac{(TP+TN)}{TP+TN+FP+FN} * 100\%$	Gives the proportion of the total number of predictions that were correct
Precision (Pre)	$\frac{TP}{(TP+FP)} * 100\%$	How accurate the positive predictions are
Recall (Sensitivity)	$\frac{TP}{(TP+FN)} * 100\%$	Gives information about the True Positives that are correctly classified during the test.
Specificity	$\frac{TN}{(TN+FP)} * 100\%$	Gives information about of True Negatives that are correctly classified during the test.
F1-score	$\frac{2*TP}{(2*TP+FN+FP)} * 100\%$	Hybrid metric useful for unbalanced classes

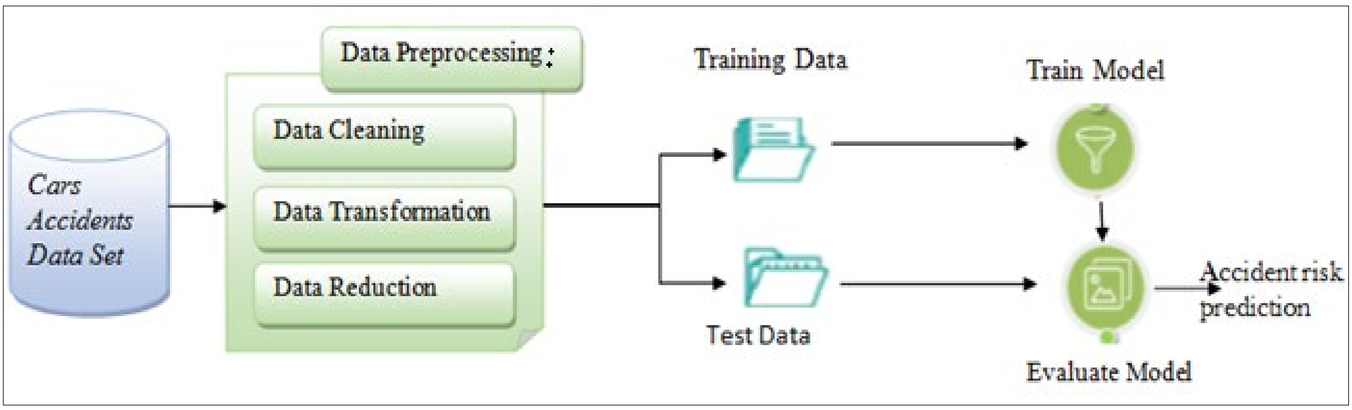


Figure 11: Experiment procedure

5.3.1 Cloud_DRL Vs V.Edge_Dl Learning Curves. To build the best traffic accident predictive framework, we used a Convolution Neural Network (CNN), which is one of the best classification algorithms based on artificial neural networks, for both the Cloud_DRL and V.Edge_DL models. CNN is designed to learn automatically and adaptively using multiple building blocks such as convolution layers, pooling layers, and fully connected layers.

For the V.Edge_DL model, the CNN algorithm consists of five convolutional layers with 32 filters of size 3, and five max pooling layers. The output of these layers was then flattened and passed through two fully connected layers before being processed by a softmax activation function to produce three output predictions. In the case of Cloud_DRL model, DRL was integrated into the same Convolutional Neural Network (CNN) architecture used in the first model, to produce the same number of outputs, each representing an action to be performed. The results obtained in terms of accuracy and loss for both models are displayed in Figure 12.

5.3.2 Performance Comparison. The proposed models Cloud_DRL and V.Edge_DL are compared to other well-known algorithms [31], such as: *Logistic Regression* (*LR*), *support vector machine (SVM), decision trees (DT), Random forests (RF)*, and *XGBoost* in terms of Accuracy, Sensitivity, Specificity, Precision, and F1-score measures. The experiment results are summarized in Table 3. Figure 13 visualized the results in Table 3.

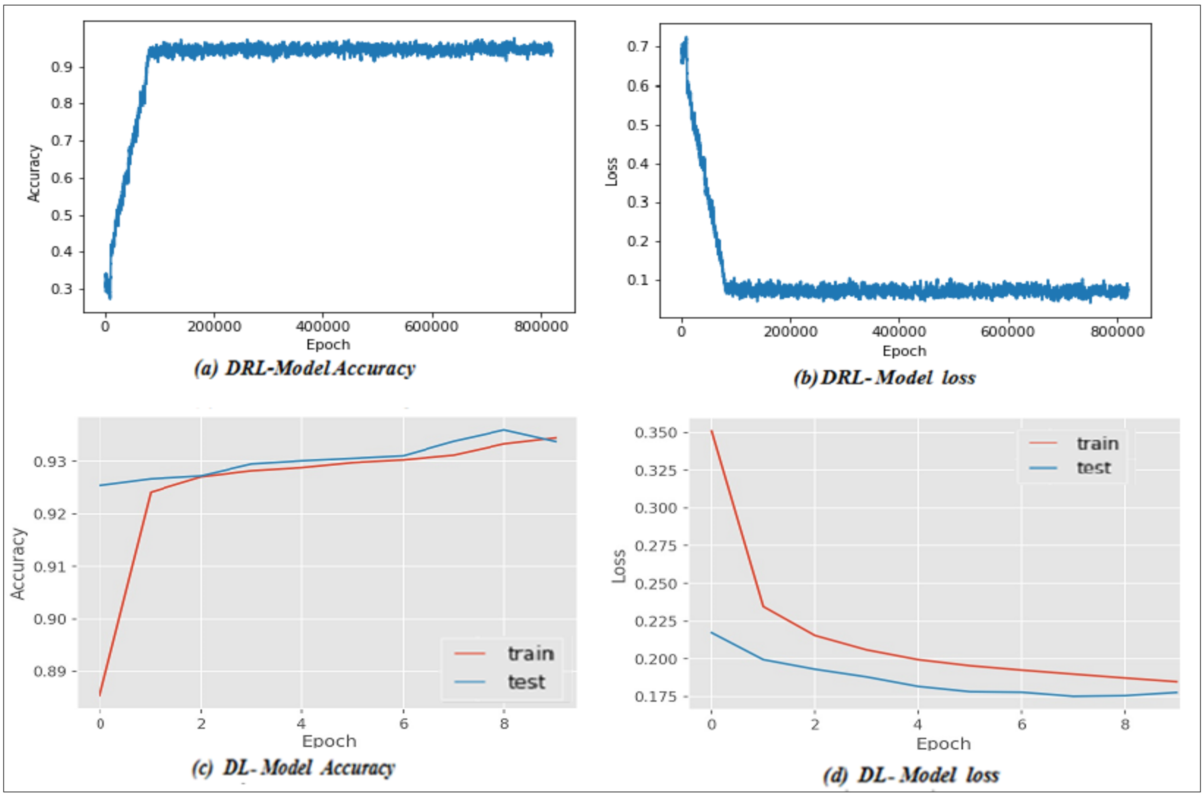


Figure 12: Learning curves for Cloud_DRL and V.Edge_Dl models

Table 3: Comparison of (Cloud_DRL, V.Edge_DL) with baseline

Classification Techniques	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1 score*
Cloud_DRL	**0.94**	**0.98**	**0.92**	**0.84**	**0.91**
V.Edge_DL	**0.93**	**0.98**	**0.90**	**0.83**	**0.90**
LR	0.76	0.28	0.97	0.83	0.42
SVM	0.89	0.82	0.95	0.93	0.87
DT	0.91	0.89	0.90	0.92	0.91
RF	0.92	0.89	0.91	0.94	0.92
XGBoost	0.93	0.99	0.96	0.66	0.80

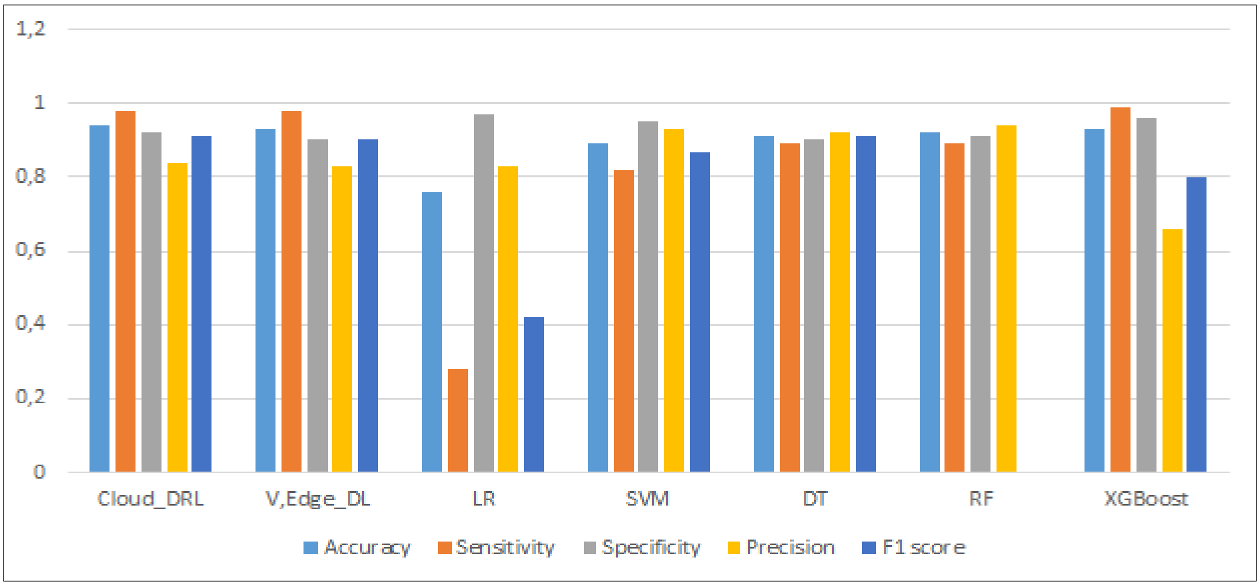


Figure 13: Visual comparison with baselines

5.3.2 Cloud_DRL – V.Edge_DL Collaboration. *Cloud_DRL***,** *V.Edge_DL* collaboration leads to efficient and effective prediction of traffic accident risk. The results obtained by each model individually and then together are shown in Figure 14.

5.3.4 Discussion. Figure 12 represents the accuracy and the loss of both models *Cloud _DRL* and *V.Edge _DL*. Figure 12 (a) plots the increment of *Cloud_DRL* accuracy in function of epoch's number; its accuracy starts very low and ends very high. The main reason of this distinction is due to a balance between the two explorations and exploitation strategies. At the beginning of the algorithm, each action is performed randomly, which is useful for helping the agent learn more about its environment. Whenever the agent takes more steps, the exploration decreases, and the agent starts to exploit more of the good actions that it has detected. Towards the end of the training process, the search space becomes very limited. Therefore, the agent concentrates more on the exploitation step. This leads to a significant increase in accuracy. It is the same similar justification for the loss curve (Figure 12 (b)), which reduces the error to a minimum.

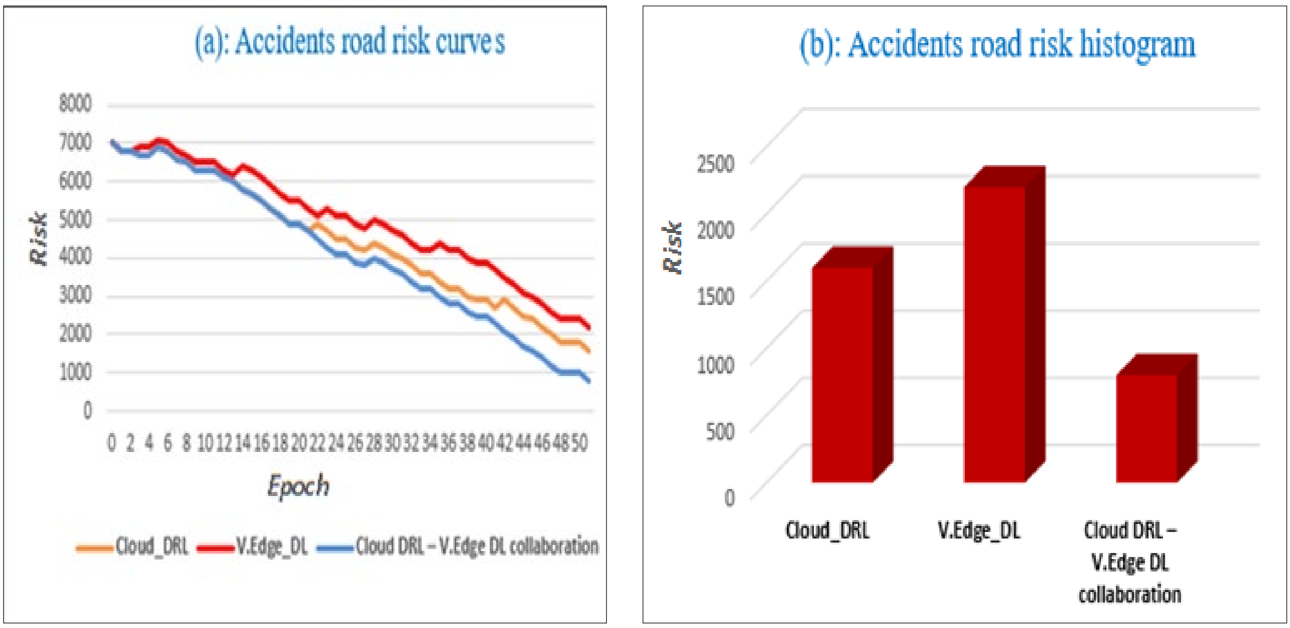


Figure 14: Cloud _DRL - V.Edge_DL Collaboration

DL has also shown better performance (Figure 12(c, d)), but DRL remains the strongest in solving dynamic problems where the environment changes over time and the optimal decision-making strategy may vary depending on the state of the environment. This poses challenges for traditional DL algorithm that lack the ability to adapt to changing conditions.

To present how well our accident risk prediction models are performing, we compared them with other algorithms that use the same performance measures. Table 3 summarizes the obtained results when applying these machine learning algorithms including *LR*, *SVM*, *DT*, *RF* and XGBoost. We note that *Cloud_DRL* and *V.Edge_DL* give a high performance in term of Accuracy, Sensitivity, Specificity, Precision, F1-score measures. We can see that the *Cloud _DRL* achieved the highest degrees of accuracy 94%, Sensitivity 98%, Specificity 92%, Precision 84% and F1-score 91%. After *Cloud_DRL* algorithm, the *V.Edge_DL* classifier generates a good result with 93 percent accuracy, 98 percent Sensitivity, 90 percent Specificity, and 83 percent Precision and 90 percent F1-score ; where all the implemented ML methods also perform excellently. Only LR performs relatively poorly with accuracy of less than 80%.

In Figure14, we tested our framework on a sample of past road accidents to evaluate its effectiveness in reducing the risk of traffic accidents by using equations (1) and (2) with (N= -100), running it through 50 epochs. We started by evaluating each model individually and then combined the two models to demonstrate the importance of their collaboration. As shown in Figure 14 (a). The red curve in the graph shows the decrease in the risk rate of road accidents when only *V.Edge_DL* was applied. The risk value decreased from 7000 to 2200 over time; in contrast, when *Cloud_DRL* was used, the risk value decreased further to 1600, as shown by the orange curve. However, the best solution for reducing traffic accident risk was achieved by combining the two models, as demonstrated by the blue curve. With their collaboration, the risk level decreased to almost zero (800). The same objective is represented by the histogram in Figure 14 (b), which shows the level of traffic safety that has been achieved by each model individually and by their collaboration.

Our research project faced several limitations, particularly regarding the quality of historical data and the complexity of machine learning techniques. The availability of high-quality historical data was a significant challenge, as it affected the accuracy and reliability of our results. Additionally, the complexity of implementing machine learning algorithms posed difficulties in achieving optimal performance. Despite these hurdles, we were able to achieve commendable results and contribute positively to the field of road safety.

6 Conclusion

Smart Cities provide a range of capabilities that can enhance the daily lives of residents. One crucial application of Intelligent Transportation Systems (ITS) is the improvement of road safety. The prediction of traffic accident risk plays a vital role in achieving this objective, a collaborative Cloud-V.Edge driver assistance system (ICEDAS) that utilizes machine learning based IOV that has been proposed to address this challenge. To leverage the advantages and mitigate the drawbacks of both platforms, the proposed framework includes two models. The first, Cloud_DRL with accuracy of 94%, utilizes a substantial amount of crash data stored in the cloud. It also suggests various preventive actions, including stopping, decelerating, or not changing lanes in cases of negligible risk. The second model, V.Edge_DL with an accuracy of 93%, functions as an assistant, compensating for Cloud_DRL's lack of prediction due to issues such as internet disconnection or bandwidth overload. To evaluate the effectiveness of the framework to reduce the risk of accidents, we conducted tests on a randomly selected sample of past road accidents running it through 50 epochs. The results obtained indicate that the collaboration between the two models significantly reduces the risk (from 7000 to less than 800), about 90% for the selected sample, surpassing the performance of either model alone.

The future work would involve integrating computer vision into the current project, using machine learning and Cloud-Edge Computing. This approach will develop advanced systems to prevent accidents and enhance road safety by analysing images and videos to detect hazardous situations, identify risky behaviors, and assess real-time road conditions. This holds great promise for further development, particularly in the field of self-driving vehicles. Additionally, given the positive results achieved in road safety, the system can find applications in other areas like fire detection and industrial risk management.

.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, and X. Zheng, "Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," ArXiv preprintarXiv: 1603.04467, 2016.

[2] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh, and B. Gupta, "Deep Learning for Arabic NLP: A Survey," *Journal of Computational Science*, 26:522-531, 2018, https://doi.org/10.1016/j.jocs.2017.11.011.

[3] S. Aradi, "Survey of Deep Reinforcement Learning for Motion Planning of Autonomous Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 23(2):740-759, 2020, Doi: 10.1109/TITS.2020. 3024655.

[4] N. M. Ashraf, R. R. Mostafa, R. H. Sakr, and M. Z. Rashad, "Optimizing Hyperparameters of Deep Reinforcement Learning for Autonomous Driving Based on Whale Optimization Algorithm," Plos one, 16(6):e0252754, 2021, https://doi.org/10.1371/ journal.pone.0252754.

[5] H. Chae, C. M. Kang, B. Kim, J. Kim, C. C. Chung, and J. W. Choi, "Autonomous Braking System via Deep Reinforcement Learning," 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp. 1-6, October 2017, Doi: 10.1109/ITSC.2017.8317839

[6] L. Y. Chang and W. C. Chen, "Data Mining of Tree-Based Models to Analyse Freeway Accident Frequency," *Journal of Safety Research*, 36(4):365-375, 2005. https://doi.org/10.1016/j.jsr.2005.06.013.

[7] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, "Learning Deep Representation from Big and

Heterogeneous Data for Traffic Accident Inference" Proceedings of the AAAI Conference on Artificial Intelligence, 30(1):338-344, February 2016, https://doi.org/10.1609/aaai.v30i1.10011.

[8] François Chollet, Keras, 2015. https://github.com/fchollet/keras.

[9] J. Gao, Y. Shen, J. Liu, M. Ito, and N. Shiratori, "Adaptive Traffic Signal Control: Deep Reinforcement. Learning Algorithm with Experience Replay and Target Network," arXiv preprint arXiv:1705.02755, 2017, https://doi.org/10.48550/arXiv.1705.02755.

[10] F. A. González and C. A. Pedraza, "Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data," *Computers*, 11(9):126, 2022, https://doi.org/10.3390/computers11090126.

[11] M. Gregurić, M. Vujić, C. Alexopoulos, and M. Miletić, "Application of Deep Reinforcement Learning in Traffic Signal Control: An Overview and Impact of Open Traffic Data," Applied Sciences, 10(11):4011, 2020, https://doi.org/10.3390/app10114011.

[12] M. Hossain, and Y. A. Muromachi, "Bayesian Network Based Framework for Real-Time Crash Prediction on the Basic Freeway Segments of Urban Expressways," *Accident Analysis & Prevention*, 45:373-381, 2012, https://doi.org/10.1016/j.aap.2011.08.004.

[13] M. G. Karlaftis and E. I. Vlahogianni, "Statistical Methods Versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights," Transportation Research Part C: Emerging Technologies, 19(3):387-399, 2011, https://doi.org/10.1016/j.trc.2010.10.004.

[14] J. H. Kiswani, S. M. Dascalu, and F. C. Harris Jr, "Cloud Computing and Its Applications: A Comprehensive Survey," *International Journal of Computer Applications*, 28(1):3-24, 2021.

[15] S. Koh, B. Zhou, H. Fang, P. Yang, Z. Yang, Q, Yang, and Z. Ji, "Real-Time Deep Reinforcement Learning based Vehicle Navigation," *Applied Soft Computing*, 96:106694,.2020, https://doi.org/10.1016/j.asoc.2020.106694, Get rights and content.

[16] R. Kumar Thangavel, S. Athithan, S. Sarumathi, M. Aruna, and B. Nithila, "Blackspot Alert and Accident Prevention System." 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1-6, 2019, Doi: 10.1109/ICCCNT45670.2019.8944412.

[17] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das, F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by using Machine Learning in Bangladesh," 2019 7th International Conference on Smart Computing & Communications (ICSCC), IEEE, pp. 1-5, 2019, Doi: 10.1109/ICSCC.2019.8843640.

[18] L. Lin, Q. Wang, and A. W. Sadek, "A Novel Variable Selection Method based on Frequent Pattern Tree for Real-Time Traffic Accident Risk Prediction," Transportation Research Part C: Emerging Technologies, 55:444-459, 2015, https://doi.org/10.1016/j.trc.2015.03.015.

[19] B. Liu, Y. Tang, Y. Ji, Y. Shen, YA Du, Deep Reinforcement Learning Approach for Ramp Metering based on Traffic Video Data," *Journal of Advanced Transportation*, 1-13, 2021, https://doi.org/10.1155/2021/6669028.

[20] Y. Lv, S. Tang, and H. Zhao, "Real-Time Highway Traffic Accident Prediction based on the k-Nearest Neighbor Method" 2009 International Conference on Measuring Technology and Mechatronics Automation, IEEE, 3:547-550, April 2009, Doi : 10.1109/ICMTMA.2009.657.

[21] Y. Ma, Z. Liu, and C. McAllister, "Deep Reinforcement Learning for Portfolio Management," *International Journal of Computers and Their Applications*, 30(2):93-106, June 2023.

[22] R. Meneguette, R. De Grande, J. Ueyama, G. P. R. Filho, and E. Madeira, "Vehicular Edge Computing: Architecture, Resource Management, Security, and Challenges," *ACM Computing Surveys (CSUR)*, 55(1):1-46, 2021, https://doi.org/10.1145/3485129.

[23] A. Murad, F. A. Kraemer, K. Bach, and G. Taylor, "Information-Driven Adaptive Sensing based on Deep Reinforcement Learning," *Proceedings of the 10th International Conference on the Internet of Things*, pp. 1-8, October 2020, https://doi.org/10.1145/3410992.3411001.

[24] J. Paul, Z. Jahan, K. F. Lateef, M. R. Islam, and S. C. Bakchy, "Prediction of Road Accident and Severity of Bangladesh Applying Machine Learning Techniques," 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), IEEE, pp. 1-6, December 2020, Doi: 10.1109/R10-HTC49770.2020.9356987.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and E. Duchesnay, "Scikit-Learn: Machine learning in Python," *Journal of Machine Learning Research*, 12:2825-2830, 2011.

[26] Z. Peng, S. Gao, Z. Li, B. Xiao, and Y. Qian, "Vehicle Safety Improvement Through Deep Learning and Mobile Sensing," *IEEE Network*, 32(4):28-33, 2018, Doi: 10.1109/MNET.2018.1700389.

[27] S. Ravichandiran, "Deep Reinforcement Learning with Python: Master Classic RL, Deep RL, Distributional RL, Inverse RL, and more with OpenAI Gym and TensorFlow," Packt Publishing Ltd, 2020.

[28] H. Ren, Y. Song, J. Liu, Y. Hu, and J. Lei, "A Deep Learning Approach to the Prediction of Short-Term Traffic Accident Risk," arXiv preprint arXiv:1710.09543, 2017.

[29] H. Ren, Y. Song, J., Wang, Y. Hu, J. Lei, "A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp. 3346-3351, November 2018, Doi: 10.1109/ITSC.2018.8569437.

[30] C. Spatharis and K. Blekas, "Multiagent Reinforcement Learning for Autonomous Driving in Traffic Zones with Unsignalized Intersections," *Journal of Intelligent Transportation Systems*, pp. 1-17, 2022 https://doi.org/10.1080/15472450.2022.2109416.

[31] H. Tatsat, S. Puri, and B. Lookabaugh, "Machine Learning and Data Science Blueprints for Finance," O'Reilly Media, 2020.

[32] A. Thaduri, V. Polepally, and S. Vodithala, "Traffic Accident Prediction Based on CNN Model," 2021 5th

International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, pp. 1590-1594, 2021, Doi: 10.1109/ICICCS51141.2021.9432224.

[33] A. Tolba, O. Said, and Z. Al-Makhadmeh, "MDS: Multi-Level Decision System for Patient Behavior Analysis based on Wearable Device Information," *Computer Communications*, 147:180-187, 2019, https://doi.org/10.1016/j.comcom.2019.08.022.

[34] M. Usama, B. Ahmad, J. Yang, S. Qamar, P. Ahmad, Y. Zhang, and J. Guna, "REMOVED: Equipping Recurrent Neural Network with CNN-Style Attention Mechanisms for Sentiment Analysis of Network Reviews," *Computer Communications*, 148:98-106, 2019, https://doi.org/10.1016/j.comcom.2019.08.002.

[35] S. Wan, Z. Gu, and Q. Ni, "Cognitive Computing and Wireless Communications on the Edge for Healthcare Service Robots," *Computer Communications*, 149:99-106, 2020. https://doi.org/10.1016/j.comcom.2019.10.012.

[36] C. Yang, Y. Wang, S. Lan, L. Wang, W. Shen, and G. Q. Huang, "Cloud-Edge-Device Collaboration Mechanisms of Deep Learning Models for Smart Robots in Mass Personalization," *Robotics and Computer-Integrated Manufacturing*, 77:102351, 2022, https://doi.org/10.1016/j.rcim.2022.102351.

[37] H. Yang, Y. Wang, H. Zhao, J. Zhu, and D. Wang, "Real-Time Traffic Incident Detection using an Autoencoder Model," 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp. 1-6, 2020, Doi: 10.1109/ITSC45102.2020.9294455.

[38] L. Yu, B. Du, X. Hu, L. Sun, W. Lv, and R. Huang, "Traffic Accident Prediction Based on Deep Spatio-Temporal Analysis," 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, IEEE, pp. 995-1002, August 2019, Doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00195.

[39] Z. Yuan, X. Zhou, and T. Yang, "Hetero-convlstm: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 984-992, July 2018. https://doi.org/10.1145/3219819.3219922.

[40] Z. Zhang, Q. Nie, J. Liu, A. Hainen, N. Islam, and C. Yang, "Machine Learning based Real-Time Prediction of Freeway Crash Risk using Crowdsourced Probe Vehicle Data," *Journal of Intelligent Transportation Systems*, pp. 1-19, 2022, https://doi.org/10.1080/15472450.2022.2106564.

[41] Z. Zhu and H. A. Zhao, "Survey of Deep RL and IL for Autonomous Driving Policy Learning," *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14043-14065, 2021, Doi: 10.1109/TITS.2021.3134702.

Zeroual Djazia (photo not available) is a researcher in the Computer Science Department at Biskra University in Algeria. She received her master's degree from Batna University, also in Algeria, with a specialization in Image Parallel Processing. Currently, she holds the position of assistant professor. Her research interests primarily revolve around artificial intelligence, with a specialization in machine learning, internet of things, cloud computing, data mining, and data analytics.

Okba Kazar obtained his master's degree in 1997 from Constantine University in Algeria, specializing in the field of artificial intelligence. He later earned his PhD from the same university in 2005. He served as a Professor in the Computer Science Department at Biskra University and played a pivotal role in the establishment of the LINFI laboratory at the same institution. He is a prolific author with numerous publications to his name, and serves as a member of the editorial board for several journals. His research interests encompass a wide range of areas, including multi-agent systems and their applications, ERP, advanced information systems, web services, semantic Web, big data, robotics, internet of things, and cloud computing. Currently, he holds the position of professor at the Information Systems and Security Department at the United Arab Emirates University in the UAE.

Saad Harous Senior Member, IEEE received the Ph.D. degree in computer science from Case Western Reserve University, Cleveland, OH, USA, in 1991. He is currently the Chair and a Professor with the College of Computing and Informatics, University of Sharjah, Sharjah, UAE. He has more than 30 years of experience in three different countries (USA, Oman, and UAE). He has published more than 200 journal and conference papers. His research interests are in distributed computing, technology-based learning, fault tolerant computing, and communication networks.

Saber Benharzallah is a professor and researcher in the Computer Science Department of Batna 2 University (Algeria). He received his Ph.D. degree in 2010 from the Biskra University (Algeria). Prof. Benharzallah is currently director of laboratory LAMIE (Batna 2 University). His research interests include internet of things, service-oriented architecture, and context aware systems.

Threat Modeling of IoT-based Smart Home Systems

Abdullah Ali Ben-Nakhi*
Kuwait University, KUWAIT

Mostafa Abd El-Barr†
Badr University in Cairo, EGYPT

Kalim Qureshi‡§
Kuwait University, KUWAIT

Abstract

Internet of things has a wide range of applications such as healthcare, agriculture, transportation, and industrial manufacturing. Smart homes automation occupies a large segment of applications. Although these devices are improving constantly, security is still a challenge for them. In this study, Microsoft STRIDE is considered. STRIDE threat modeling is a tool used to simplify threats categorization. STRIDE modeling tool used to design a generic smart home system and analyze the design in terms of vulnerabilities in the design. The tool generated threats compared to the collected threats from the systematic literature review. Comparing STRIDE generated threats with the collected threats is to examine the system from a complete overview, not just examine a single component or attack type.

Key Words: Smart homes, security, threat modeling, IoT, systematic literature review, STRIDE modeling.

1 Introduction

Internet of Things is a broad term that covers a wide range of applications. Examples include smart homes, wearable devices, smart grids, and connected cars [11]. But, IoT-enabled smart homes application stands out as the most prominent application under IoT. By 2022 smart home market share is expected to reach 53.45 billion dollars [22], and by 2023 the number of smart homes is expected to exceed 300 million houses [23]. Smart homes are described as the ability to control and monitor different home devices and appliances remotely via the internet. Smart home applications encompass the needs of the residents to provide comfort, safety, security, and energy-saving for the inhabitants [70]. Although IoT-enabled smart homes devices are used for a wide range of applications, security and surveillance are the most dominant application requirements. According to the statistics, the amount of smart home security and surveillance devices sold is expected to triple between 2017 and 2022 [7].

The reason for the enormous demand for a home security system is that these devices should give a sense of safety and security to homeowners. Another critical application for smart home automation is healthcare automation at home for the elderly and people with disabilities [18]. From the existing statistics, we infer that the demand and the usage of smart home devices are growing over time. Hence, this growth motivates us to:

- Analyze the risk resulting from vulnerability to figure the severity of the vulnerability to households.
- Map the proper mitigation method using threat modeling to simplify the risk overview to the user-level.

Therefore, mitigating the vulnerabilities encountering smart homes is becoming difficult over time and makes smart homes exposed to a wide range of attacks.

The main aim of this study is to find out the current state of the art in smart home devices in terms of their security by simplifying the overall view of these threats to facilitate vulnerabilities mitigation.

The security challenges can be grouped under three categories as follows [1]:

- Data security: IoT devices rely heavily on exchanging data among them. These data can have personal or confidential nature which makes them a valuable target for attackers, therefore, authenticity and confidentiality of data must be guaranteed.
- Communication security: due to the heterogenicity of smart home devices the medium to exchange is diverse, in this case, integrity and access control are required.
- Application security: The collected data from devices finally pour into the application end to be processed, in this case, the application security itself is a concern and the contained data privacy.

*,Department of Information Science, College of Life Sciences. Email: bennakhi1@gmail.com.
†Department of Electrical Engineering, College of Engineering. Email: mostafaabdelbarr@gmail.com.
§Corresponding author. Tel: +96597639430, ‡Department of Information Science, College of Life Sciences. Email: kalimuddin.qureshi@ku.edu.kw.

Although the security of smart home device is a critical part of the design of the product, device manufacturers do not take it seriously for many reasons. Manufacturers overlook security features to compete in a very competitive market, which allows them to release products faster and prevent the product from be obsolete in a short period [9].

The multilayered architecture of smart home systems due to hardware architecture design differences according to manufacturer uses different communication standards and protocols in their products. Also, in the application layer, a variety of operating systems and firmware are used to operate these devices [6]. The heterogeneous nature of smart home systems imposes several security challenges [8]. Moreover, IoT-enabled smart home devices are known for their CPU power, memory, storage, and power limitation due to their constraints of the design [14]. These limitations cause difficulties in implementing advanced security mechanisms. Another major security challenge is the human factor. The limited technical knowledge usually led to misconfiguration and misuse of these devices makes them susceptible to social engineering attacks [4]. Unlike traditional information systems, IoT-enabled smart home devices increase security challenges because the cyber and physical threats converge as the application makes demands [25].

The research in this work includes the following. We start by examining related work to our topic to find the shortcomings as well as the strength in the articles. We then consider the STRIDE threat modeling in categorizing extracted threats. We follow that with mapping the proper countermeasures. In addition, the STRIDE threat modeling tool is used to model generic smart homes to generate the potential threat from the STRIDE perspective. The resulting threats from the STRIDE tool are compared to the excreted threats.

2 Literature Review

This section studies the related works that appear in the literature. In [10] the researchers classified smart home attacks into main categories according to IoT layers. Each layer is examined separately for security issues and solutions proposed. There was no methodology proposed to elicit attacks. The attacks on each layer were obtained from previous articles in the field of IoT security. A survey of a recent solution is conducted to present state-of-the-art risk mitigations.

The research in [2] states that the increased number of IoT-based devices connected to the internet used in smart homes widen the attack vectors. In this research, they collected 12 common attacks on smart homes presented in related works and explained the possible impact of these threats. They stated that although there are several solutions to improve smart home security, it is difficult to achieve comprehensive security. This is because of the increased number of technologies used in IoT devices.

The researchers in [14] stated that since IoT-based smart homes are connected to the internet they directly or indirectly expose the inhabitant's privacy and data to vulnerabilities. Although inhabitants' physical security is important, they only take into consideration cyber security. They classify attacks into two classes internal and external. Internal attacks are possible when attackers are nearby, and external attacks are conducted through the internet. In the article, they only identified five common attacks on smart homes and no countermeasure was suggested.

In [12] threats are classified into two categories: passive and active attacks. Passive attacks are where the attacks are eavesdropping the information with no intention to manipulate the information. Whereas, in active attacks, the attacker intends to cause harm to the system such as modifying the messages between equipment or interrupting the system functionality. Due to the popularity of smart home devices, house holders are more concerned about security and privacy risks in smart homes. They summarized the most vulnerable attack in three sets: connected objects, cloud, and applications. Within these three categories, security risks are classified into three levels: low, intermediate, and high-security issues. For each of these levels, the proposed countermeasures are derived from other researchers' works.

In [5] a comprehensive survey is conducted in IoT in general and focuses on smart home security issues. With taking into consideration user awareness, this paper analyzes the technical perspective of IoT security risk. Based on the current known IoT attacks, they state that smart home systems are vulnerable to many threats because of a lack of security in the design phase of these systems. Although they identified many attacks in their work, there is no clear taxonomy for categorizing threats. In addition, the suggested solution solely concentrates on protocols used in smart home automation devices.

In [26] a survey on security challenges was taken in a physical/network layer in smart home automation and the exciting proposed solutions for these challenges. Additionally, security attacks on the voice control interface are also examined. Since voice control is one of the most used smart home features, they proposed a two-factor voice command validation framework to improve the voice control interface. However, they did not show how the survey was conducted to summarize the collected attacks.

In [13] the researchers simplify the threat identification by using the attack trees methodology. The attack trees proposed are based on the smart home model with different configurations to identify the different attack's surfaces. The goal of this work is to improve the security of smart homes in terms of hardware and software. The root of the tree represents the attacker's goal, the branches represent the attacker's method to achieve the goal, and the leaves represent the final step to achieve the attacker's goal. However, the attack tree clearly explained that no countermeasures were suggested to mitigate the identified attacks.

In [24] a survey is conducted on IoT-enabled cyber-attacks since 2010 in all IoT application domains. Two types of attacks were examined in the article. They were either based on real-life scenarios or produced in controlled environment published attacks by other researchers, and they exclude attacks that cannot be verified practically. They study these attacks on IoT-enabled devices to find not only the effect on the device itself, but also on the target system of which the device is part of. In addition, they categorized attacks based on the attacker's goals. Then they assess the threat severity based on risk assessment standards such as ISO 27005 and NIST SP800-30. Therefore, security controls are proposed to mitigate the threats in the short term and long term. The proposed security controls are grouped into six categories: physical access, logical access, hardware, software, network, and procedures. They conclude improving smart home systems needs a comprehensive analysis. Table 1 summarizes this type of work.

3 Threat Modeling

Data was collected in the previous section from studies using systematic literature review and conducting a survey with experts in the field. In this section, threat modeling is conducted to simplify attacks categorization. According to [17] threat modeling applies to a wide range of applications such as software, systems, networks, and things in the Internet of Things. In this section, a threat model is designed to facilitate the mitigation of risk in smart homes from the user perspective to understand how the system works. Threat modeling in our case was conducted to identify and understand the threats in smart home systems and then decide on the proper mitigation method in a simple manner. Threat modeling is useful in clarifying at which point the element in the system is vulnerable to attacks and analyzing the type of attack that could be conducted. Also, threat modeling helps to anticipate potential threats that might be missed by users and understand the risks from the attacker's perspective. Threat modeling is important in the system design phase. There are many threat modeling frameworks such as PASTA, OCTAVE, and LINDDUN. PASTA is designed specifically for organizations planning to merge their strategic objectives with the anticipated risks, and OCTAVE is a complex threat modeling tool that requires great effort and dedicated time to examine the information assets [15]. Whereas the LINDDUN framework focuses solely on privacy risks in the system [20].

3.1 Microsoft STRIDE Threat Modeling

We adopted the Microsoft STRIDE framework since it applies to IoT threats and provides an easier overview of risks [21]. STRIDE is an acronym that stands for Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of privilege. STRIDE utilizes data-flow diagrams to illustrate the flow of data between elements in the system and interaction with external entities. STRIDE simplifies interactions between elements for easier to understand threat modeling.

This data flow diagram identifies threats encountered during transitions and interactions. STRIDE framework maps the element to the potential threat category. Also, it rates threat severity to provide risk level [3]. In our case, when a house owner purchases a system or devices, threat modeling provides an overview about what the households are facing, and the proper mitigations needed to take into consideration. Table 2 is based on Table 2 from [21] STRIDE threat categories presented and defined with which security property was violated.

Threat modeling in our case includes: a layout design of a typical smart home system, threat list, countermeasures list and the action undertaken to combat each threat. Conducting threat modeling using STRIDE framework must consider these four steps [16]:

a) Identify the assets of the system
b) Identification of threats
c) Rating of the threats
d) Propose countermeasures

In this section, steps a-d are implemented using STRIDE threat modeling to simplify threats extracted in Section 4 categorization, risk rating, and countermeasure mapping. After that, a generic smart home system data flow diagram was designed using the STRIDE modeling tool to analyze potential threats, understand the data transactions and countermeasures suggested. In the end, the extracted threats from Section 4 are

Table 1: Security challenges reviews and surveys

Article	Type	Threats Number	Threat Type	The Future
Smart Home Is Not Smart Enough to Protect You - Protocols, Challenges and Open Issues [10].	Review	5	Perception layer Network layer Application layer	Investigate smart homes components
A review on smart home present state and challenges: linked to context-awareness internet of things (IoT) [2].	Review	12	Physical attacks Malicious code Eavesdropping Personal information abuse	Develop smart home solution
IoT-based smart homes: A review of system architecture, software, communications, privacy, and security [14].	Review	3	Data breaches Authorization User privacy	Computer engineers and specialist involves in smart homes development

Table 2: STRIDE threats [22]

Threat	Property Violated	Definition
Spoofing	Authentication	Pretending to be something or someone other than yourself
Tampering	Integrity	Modifying something on disk, ona network, or in memory
Repudiation	Non-Repudiation	Claiming that you did not do something or were not responsible.
Information Disclosure	Confidentiality	Providing information to someone not authorized to see it
Denial of Service	Availability	Absorbing resources needed to provide service
Elevation of Privilege	Authorization	Allowing someone to do something they are not authorized to do

compared to threats generated from STRIDE tool and discussed.

3.2 Identify Assets of the System

There are several implementations of smart homes. This implementation depends on the needs of the household. But there is a basic infrastructure that is mutual among various implementations. The common assets are central hub or gateway, devices such as surveillance cameras, cloud service, and control software (Figure 1). According to Microsoft IoT security architecture [19] the best practice for STRIDE threat modeling is to divide IoT layout into zones: devices, field gateway, cloud gateway, and services. Zone separations provide data boundaries that facilitate threat detection through data transition. Table 3 maps the generic smart home IoT layout into STRIDE model layout.

Table 3: IoT layout conversion to STRIDE model

IoT layout	STRIDE layout
Gateway or Hub	Field gateway zone
IoT devices	Device zone
Cloud service	Cloud gateway zone
Control software	Services zone

STRIDE framework streamlines the threat categories for easier user understanding. In Table 4, 23 extracted threats classified into STRIDE categories according to definition and property were violated. This conversion facilitates understanding of the risk from the non-security expert

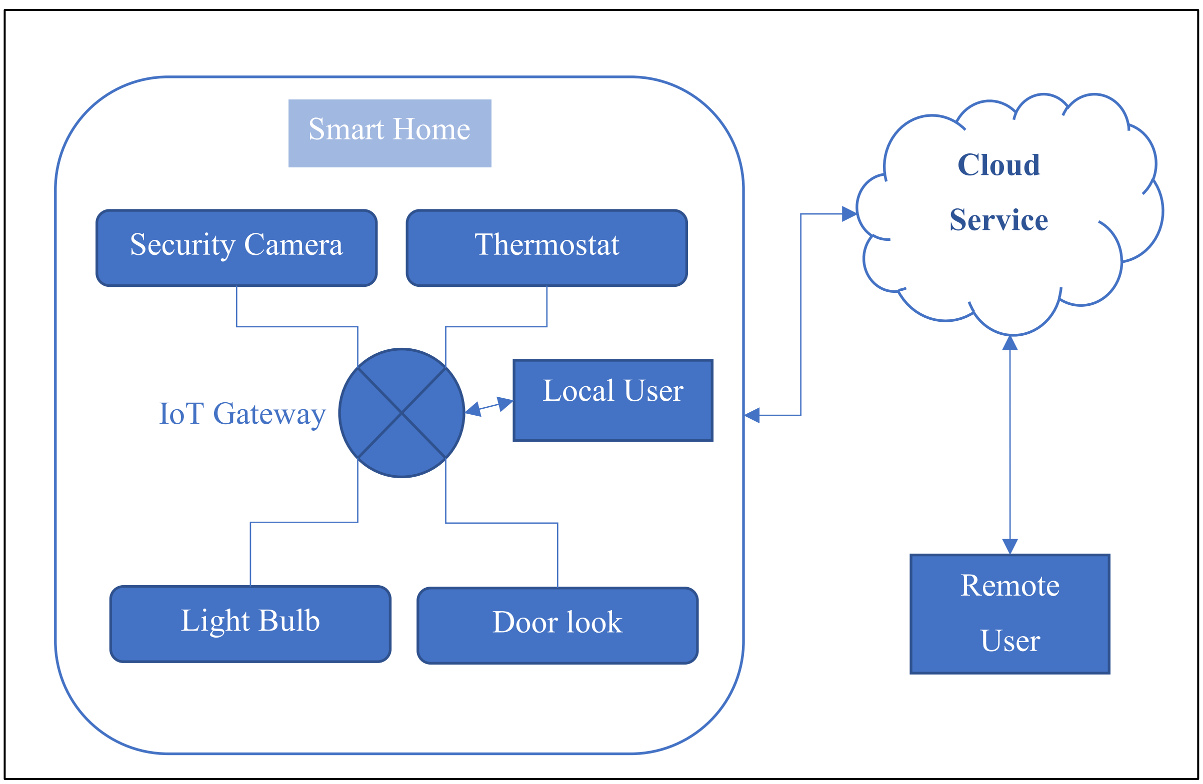


Figure 1: Smart home layou

Table 4: Threats STRIDE conversion

STRIDE Categories	Property Violated	Extracted Threats
Spoofing	Authentication	Spoofing, impersonation, brute force attack, MITM, masquerade attack, unauthorized access
Tampering	Integrity	Forgery, malicious code injection, physical attack, unsecured interfaces, gain initial access
Repudiation	Non-Repudiation	No repudiation threats were mentioned in the selected papers
Information Disclosure	Confidentiality	Data leakage, eavesdropping, insecure communication, abuse attack, open ports, replay attack
Denial of Service	Availability	DoS, DDoS, jamming, or interruption attacks
Elevation of Privilege	Authorization	Over privileged, lack of authentication,

background. Also, a threat with similar methods is grouped into the same category. For example, man-in-the-middle (MITM) and spoofing attacks both intercept messages between two parties and alter them, which is called active eavesdropping. Unlike, replay attack which is passive eavesdropping with no modification done to the transmitted messages. The extracted threats are classified according to the attacker's goal, needing a medium to reach the final goal. However, in the non-repudiation category, there are no threats mentioned from the extracted threat.

3.3 Rating of the Threats

After the risk is identified, threat assessment is conducted to rate the threats and prioritize the mitigation process. In some cases, not all threats are feasible to mitigate. Some of the threats can be ignored. Threat assessment conducted using Microsoft DREAD framework. DREAD acronym for Damage potential, Reproducibility, Exploitability, Affected users, Discoverability. In Table 5 DREAD risk factor was explained in [14], each question answered with a range of severity from 1-3. Then after scoring all the risk factors, calculate the total of each threat. If scores are 5-7 the risk is low, 8-11 risk is medium, 12-15 risk is high.

Table 6 shows the threats rating for the extracted 23 threats.

4 Proposed Countermeasures

After classifying risk into SRTIDE classes we select the proper countermeasures from Table 6 to threats in Table 7.

5 Implementing STRIDE Threat Model

In this section the STRIDE threat modeling tool is used to implement smart home generic data flow architecture. Designing a home system using a threat modeling tool is important to understand the potential threats encountering households. Furthermore, the modeling aims to understand the vulnerabilities from the attacker's perspective. Threat modeling is conducted before deploying the system to identify potential threats. When modeling the system these elements are taken into consideration: processes, data stores, data flow, and external entities. Based on Figure 2 and Table 7 common IoT-enabled smart home design uses Microsoft STRIDE threat modeling tool as shown in Table 8.

According to [19] it is recommended that IoT architecture be divided into zones, where each zone has its data and authentication method needs. Then zones are separated by trust boundaries (dotted lines) to represent data transition from one source to another. As shown in Figure 2 device zone and local user zone interact with the home gateway. Then the data is transferred to the cloud gateway through an internet connection to the service provider cloud to store, analyze data and allow the remote user to interact with the home system.

Using a threat modeling tool, a report is generated based on the designed system. The report states that this design posed 97 threats and examples depicted on how the attacker conducts the attack are included. These threats are summarized into spoofing, forgery, DoS, data leakage, data repudiation, sniffing, interruption, impersonation, code injection, lack of authentication, lack of authorization. The modeling tools do not only depict potential threats but also often suggest countermeasures as shown in Table 8. All the identified threats generated from the modeling tool are included in the threats list extracted from selected prime studies in Section 4, except for repudiation which was not mentioned as a threat before in 35 selected prime studies.

6 Results and Analysis

The motivation behind this research was the growth of the use of IoT devices specifically in smart home applications. Such systems give the households a sense of security and control over the house. The smart home system consists of newly emerged IoT-enabled devices known for their limitations. IoT device limitations have imposed security concerns because security control mechanisms require major computational power. For this reason, we conducted a rigorous systematic literature review to identify the recent vulnerabilities threatening smart home systems and security control proposed to mitigate these vulnerabilities.

6.1 Threat Modeling Findings

As stated previously, there is heterogeneity in threats categorizations in different articles proposed. Different

Table 5: DREAD risk factors

Risk Factor	Meaning
Damage potential	How great is the damage if the vulnerability is exploited?
Reproducibility	How easy is it to reproduce the attack?
Exploitability	How easy is it to launch an attack?
Affected users	As a rough percentage, how many users are affected?
Discoverability	How easy is it to find the vulnerability?

Table 6: DREAD threat rating

Vulnerability	D	R	E	A	D	Total	Priority
DDoS or DoS	1	3	3	3	3	13	High
Data leakage	3	2	3	2	1	11	Medium
Eavesdropping	2	2	3	2	1	10	Medium
Forgery	2	2	1	2	2	9	Medium
MITM	2	3	2	1	2	10	Medium
Lack of authentication	3	3	2	3	1	12	High
Unauthorized access	3	3	2	3	1	12	High
Malicious code injection	3	1	1	3	1	9	Medium
Over privileged	2	2	2	1	2	9	Medium
Replay attack	3	2	2	3	1	11	Medium
Physical attack	3	1	1	3	1	9	Medium
Impersonation	2	3	3	2	3	13	High
Spoofing	3	3	2	3	3	14	High
Brute force attack	3	1	2	2	3	11	Medium
Insecure communication	3	3	2	3	3	14	High
Abuse attack	2	3	1	2	2	10	Medium
Jamming or interruption attacks	3	2	3	3	3	14	High
Lack of encryption	3	2	3	3	2	13	High
Open ports	2	2	2	2	2	10	Medium
Masquerade attack	3	3	2	3	1	12	High
Gain initial access	3	3	1	3	1	11	Medium
Unsecured interfaces	2	2	2	2	2	10	Medium

Table 7: Map countermeasures to STRIDE

STRIDE Categories	Extracted Threats	Countermeasures
Spoofing	Spoofing, impersonation, brute force attack, MITM, masquerade attack, unauthorized access	Authentication solutions
Tampering	Forgery, malicious code injection, physical attack, unsecured interfaces, gain initial access	Detection and identification and authentication solutions
Repudiation	No repudiation threats were mentioned in the selected papers	No counter measures introduced
Information Disclosure	Data leakage, eavesdropping, insecure communication, abuse attack, open ports, replay attack	Secure communication and Blockchain-based solutions
Denial of Service	DoS, DDoS, jamming, or interruption attacks	Detection and identification solutions
Elevation of Privilege	Over privileged, lack of authentication	Authentication solutions

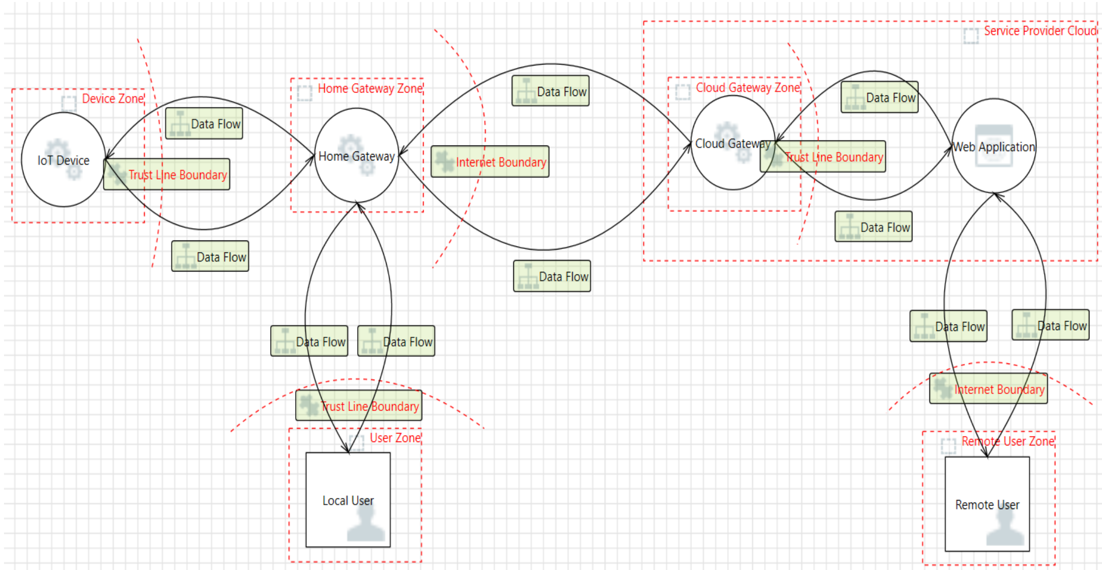


Figure 2: Smart home model

Table 8: STRDE generated threats

Category	Threat	Countermeasure
Spoofing	Spoofing	Authentication mechanism
Tampering	Forgery	Input validation mechanism
Repudiation	Data repudiation	Logging or auditing to record
Information Disclosure	Sniffing	Encrypting the data flow (secure communication)
Denial Of Service	DoS attack	Input validation mechanism
Elevation Of Privilege	Impersonation	No mitigation provided
Elevation Of Privilege	Code injection	No mitigation provided
Denial Of Service	Interruption of service	No mitigation provided
Elevation Of Privilege	Lack of authorization	Authenticated state-changing requests mechanism

articles have different threat analysis methodologies, and these methodologies are based on threats detected after incidents occurred. These approaches lack identifying vulnerabilities in the system overall when all elements are connected. Using threat modeling in our research simplified threat categorizations and identified the vulnerabilities in the system.

Microsoft STRIDE threat modeling gives threat classification with a simplified viewpoint. Using STRIDE threat modeling simplifies classifying the collected threats into six categories according to the violated security property to cope with heterogeneity in classification presented in the collected studies. Threat modeling identifies vulnerabilities in early stages such as design or deployment, in addition to mapping the countermeasures in a simplified view from the user perspective. It provides a threat rating framework that helps users to stress the highest risks in smart home systems when deployed. The researchers encompassed all threats in the STRIDE model except for repudiation. Although repudiation is important to audit activities in the system, it was overlooked by researchers. Repudiation could be at risk of data forgery. Smart home devices work as a system. It needs a basic element to operate as a communication medium and gateway hub. For this reason, we took a step forward to design a generic smart home system in the STRIDE modeling tool to find vulnerabilities as a whole system. The designed data flow diagram shows not only the risk at elements of the system, but it shows the risk data exposed to when in transition from element to others in the system. The tool generated 9 unique threats across the system. These generated threats are included in the collected threats. Repudiation was recognized as a threat even though it was neglected by studies collected.

7 Limitation of Research Work

The lack of security standardization and hardware limitations resulted in a slow or lack of security practices in these devices.

In this study, we conduct a systematic literature review to identify the exposed threats in the last five years in these devices and introduce novel countermeasures to mitigate the security issues. IEEE, ACM, Scopus, Science Direct, Springer, and MDPI databases were selected for the systematic review but we have to add more database to get more stable results.

The security of smart home devices is a critical part of the design of the product, but device manufacturers do not take it seriously for many reasons. Manufacturers overlook security features to compete in a very competitive market, which allows them to release products faster to prevent the product from being obsolete in a short period.

8 Conclusion and Future Works

In conclusion, this study discussed the security challenges encountered by IoT-enabled smart homes. A survey was conducted to elicit the latest challenges from experts in the field. The result was 22 unique threats that were identified. Because of variations in threats classification STRIDE threat model is used to simplify categorization from the house owner viewpoint. Furthermore, the STRIDE tool is used to design a generic smart home system layout using a dataflow diagram. The design helps with finding threats in the system as a whole, unlike the collected studies. Then a comparison is made between the generated threats from the tool with the extracted threats from the studies. Threat modeling plays a significant role when a house owner decides to design a smart home system. As future work, we are planning to create a framework that enables the user to design a smart home system by selecting the components of the smart home system from a predefined list such as smart lock, smoke detectors, or surveillance camera. This tool identifies the threats for each component by checking the CVE database for vulnerabilities and how to mitigate them. If a threat exists for such a component, the framework enables the user to download the patch to mitigate the issue. Otherwise, the framework suggests a proper mitigation procedure to combat the threat. Also, the framework guides the user on how to configure the available features in devices to enhance security. This method adds an active layer of protection for the smart home system and enhances the overall security of the smart home system.

References

[1] H. Abbas, W. Iqbal and M. Daneshmand, "An In-Depth Analysis of IoT Security Requirements, Challenges, and Their Countermeasures via Software-Defined Security," *IEEE Internet of Things Journal*, 7(10):10250-10276 2020. DOI:10.1109/JIOT.2020.2997651.

[2] Zahrah A. Almusaylim and Noor Zaman, "A Review on Smart Home Present State and Challenges: Linked to Context-Awareness Internet of Things (IoT)," *Wireless Networks*, 25:3193-3204, 2019.

[3] P. Aufner, "The IoT Security Gap: A Look Down into the Valley Between Threat Models and Their Implementation," *International Journal of Information Security*, 19(3):3-14, 2020.

[4] A. Awad and B. Ali, "Cyber and Physical Security Vulnerability Assessment for IoT-Based Smart Homes," *Sensors*, 18(3):817, 2018.

[5] B. Bastos, S. Shackleton, and M. El-Mosssa, "Internet of Things: A Survey of Technologies and Security Risks in Smart Home and City Environments," IET: Living in the Internet of Things: Cybersecurity of the IoT - 2018, 7 pp. 2018.

[6] J. M. Batalla and F. Gonciarz, "Deployment of Smart Home Management System at the Edge: Mechanisms and Protocols," Neural Computing and Applications, pp 1301-1315, 2019.

[7] M. Burhan, R. A. Rehman, B, Khan, and B.-S. Kim, "IoT Elements, Layered Architectures and Security Issues: A Comprehensive Survey," *Sensors*, 18(9):2796-2833, 2018.

[8] Resul Das, Muhammed Zakeriya Gunduz, Analysis of Cyber-Attacks in IoT-based Critical Infrastructure, International Journal of Information Security Science, 8(4):122-133, 2019.

[9] M. Frustaci, P. Pace, G. Aloi, and G. Fortino, "Evaluating Critical Security Issues of the IoT World: Present and Future Challenges," *IEEE Internet of Things Journal*, 5(4):2483-2495, Aug. 2018. doi: 10.1109/JIOT.2017. 2767291.

[10] I. Hmeidi, F. Shatnawi, W. Mardini, and Y. Khamayseh, "Smart Home Is Not Smart Enough to Protect You - Protocols, Challenges and Open Issues," *Procedia Computer Science*, 160:134-141, 2019.

[11] K. Karimi and G. Atkinson, "What the Internet of Things (IoT) Needs to Become a Reality," Retrieved June 3, 2020, from https://www.mouser.cn/pdfdocs/ INTOTHNGSWP.PDF.

[12] K. Khawla and T. Tomader, "A Survey on the Security of Smart Homes: Issues and Solutions," *ICSDE'18: Proceedings of the 2nd International Conference on Smart Digital Environment*, pp. 81-87, October 2018. https://doi.org/10.1145/3289100.3289114.

[13] D. Meyer, J. Haase, M. Eckert, and B. Klauer, "A Threat-Model for Building and Home Automation," IEEE 14th International Conference on Industrial Informatics (INDIN), Poitiers, France, pp. 860-866, 2016. doi: 10.1109/INDIN.2016.7819280.

[14] D. Mocrii and Y. Chen, "IoT-Based Smart Homes: A Review of System Architecture, Software, Communications, Privacy and Security," *Internet of Things*, 1(2):81-98, 2018.

[15] L. O. Nweke and S. Wolthusen, "A Review of Asset-Centric Threat Modelling Approaches," *International Journal of Advanced Computer Science and Applications*, 11(2):1-6, 2020.

[16] B. A. Omotosho, "Threat Modeling of Internet of Things Health Devices" *Journal of Applied Security Research*, 14(1):1-16, 2019.

[17] OWASP. (n.d.). "Application Threat Modeling," (OWASP) Retrieved July 2, 2021, from https://owasp.org/www-community/Application_Threat_Modeling#.

[18] S. U. Rehman and S. Manickam, "A Study of Smart Home Environment and its Security Threats" *International Journal of Reliability*, Quality and Safety Engineering, 23(3):1640005-1640014, 2015.

[19] R. Shahan and P. Meadows, "Internet of Things (IoT) Security Architecture," (Microsoft), Sept 10, 2018. Retrieved July 4, 2021, from https://docs.microsoft.com/en-us/azure/iot-fundamentals/iot-security-architecture.

[20] N. Shevchenko, "Threat Modeling: 12 Available Methods," Retrieved from SEI Blog - Carnegie Mellon University, December 3, 2018. https://insights.sei.cmu.edu/blog/threat-modeling-12-available-methods/

[21] A. Shostack, *Threat Modeling: Designing for Security*, Wiley, 2014.

[22] "Smart Home Market Size & Share will hit $53.45 Billion by 2022," April 12, 2017. (Zion Market Research) Retrieved June 3, 2020, from https://www.globenewswire.com/news-release/2017/04/12/959610/0/en/Smart-Home-Market-Size-Share-will-hit-53-45-Billion-by-2022.html.

[23] "Smart home - Statistics & Facts," March 3, 2020). (Statista) Retrieved June 3, 2020, from https://www.statista.com/topics/2430/smart-homes/.

[24] I. Stellios, P. Kotzanikolaou, and M. Psarakis, "A Survey of IoT-Enabled Cyberattacks: Assessing Attack Paths to Critical Infrastructures and Services," *IEEE Communications Surveys & Tutorials*, 20(4):3453-3495, 2018.

[25] K. Taghizad-Tavana, M. Mohsen Ghanbari-Ghalehjoughi, N. Razzaghi-Asl, and S. Sayyad Nojavan, "An Overview of the Architecture of Home Energy Management System as Microgrids, Automation Systems, Communication Protocols, Security, and Cyber Challenges," *Sustainability*, 14(23):159382022. https://doi.org/10.3390/su142315938.

[26] Y. Yang, L. Wu, G. Yin, and L. Li, "A Survey on Security and Privacy Issues in Internet-of-Things," *IEEE Internet of Things Journal*, 4(5):1250-1258, 2017.

Abdullah Ali Ben-Nakhi is an Associate Computer Engineer at State Audit Bureau of Kuwait, September 2015 – Present where he completed computing security projects; Servers and virtual machine administrator, Disasters recovery planning, Access control and managing users' permissions on different level of systems, Harden operating systems security, Harden operating webservers security, Prototyping and designing mobile application, Maintaining electronic archiving system and, etc. Mr. Adullah graduated in 2014 from Portland State University, Portland, Oregon, USA. He also completed his master in Computing Information Systems in 2022 and his email: bennakhi1@gmail.com.

Mostafa Abd-El-Barr received his PhD degree from the Department of Electrical and Computer Engineering, University of Toronto, Canada in 1986. He was with the Department of Information Science, College of Computing Sciences and Engineering (CCSE), Kuwait University 2003-2020. He was also an Adjunct Professor with the ECE Department, University of Victoria (UVic), BC, Canada 2009-2020. He is now the Chairman of the Electrical Engineering Department, Badr University in Egypt. His research interests include Information Security, Design and Analysis of Reliable & Fault-Tolerant Computer Systems, Computer Networks Optimization, Parallel Processing/Algorithms, Multiple-Valued Logic (MVL) Design & Analysis, VLSI System Design, and Digital Systems Testing.

He is the author and/or co-author of more than 185 scientific papers published in journals and conference proceedings/symposia. He has three books published (two are translated to the Chinese Language). Professor Abd-El-Barr is a Senior IEEE Member and a member of the International Association of Engineers (IAENG). He is a Senior International Associate Editor of the International Journal of Information Technology & Web Engineering and a member of the Editorial of the International Journal of Computer Science and Security (IJCSS). He is also an official IEEE/EAC/ABET evaluator. Dr. Abd-El-Barr is a Certified Professional Engineer in Ontario, Canada.

Kalim Qureshi is an Associate Professor of Information Science Department, Kuwait University, Kuwait. His research interests include network parallel distributed computing, thread programming, concurrent algorithms designing, task scheduling, performance measurement and medical imaging. Dr. Qureshi receive his Ph.D. and MS degrees from Muroran Institute of Technology, Hokkaido, Japan in (2000, 1997). He published more than 70 journal papers in reputed journals. His email address: kalimuddin.qureshi@ku.edu.kw.

A System for Identifying Entomopathogenic Nematodes

Susan F. Ellakwa*, Abdelrahman Elsayed*, Rasha Elnemr*,
Saheir F. El-lakwah†, and A. M. Azazy†
Agricultural Research Center, Giza, EGYPT

Abstract

High agricultural production is essential for food security. Hence, it is important to increase crop yields while minimizing losses. Insects cause considerable economic losses. Entomopathogenic Nematodes (EPNs) are alternatives to chemical pesticides for controlling insects. For researchers, determining the type of EPNs is not an easy task. Thus, a tool for identifying EPNs species is needed. In this paper, we introduce a method for developing a system to identify EPNs species according to their morphometric traits. We used Web Ontology Language (OWL) to build the ontology of EPNs species and represent their semantic information. Ontology helps in data representation, exchange, and interoperability. The proposed system was implemented as a mobile application that extracts and retrieves EPNs data from ontology. It displays the details of valid Heterorhabditis and Steinernema species. Also, it enables us to find species that are related to the given infective juveniles (IJs) features. We used techniques of similarity search such as cosine similarity and Euclidean distance to compare different EPNs species and identify similar species based on appearance features. The results indicate that the system can recognize the known EPNs species and it helps to identify similar species.

Key Words: Ontology, knowledge representation, entomopathogenic nematodes (EPNs), computer applications.

1 Introduction

Nematodes are one of the most diverse and numerous organisms that live in aquatic and terrestrial environments [5, 15]. Nematodes are the most plentiful animals on the planet, and over 25,000 species have been identified [8]. They are divided into two types; the first is phytoparasitic nematodes (or phytonematodes) which are severely harming crops and incurring massive economic losses globally [1]. The second type is Entomopathogenic Nematodes (EPNs), which are parasitic nematodes that can destroy harmful insects. They are used as an alternative to chemical pesticides. Both genera of EPNs, either Heterorhabditis or Steinernema, are regarded as superior biocontrol agents against a variety of insect pests globally. Hence, they contribute to a clean environment [4, 14]. EPNs are incredibly small, with body widths measured in microns. They have developed the ability to live in different habitats and coexist with various creatures in a symbiotic way. Their life cycle includes the egg phase, four larval phases, and the adult phase. The third larval phase called infective juvenile (IJ) is infectious. In actuality, IJ is only a free-living phase; all subsequent developmental phases are only found inside infected hosts. EPNs have a life cycle that spans between 5 and 10 days, according to the temperature, the existence of a bacterial symbiont, and their ability to lower the resistance of their insect host. EPNs are widely dispersed, however, the species differ according to environments and geographic locations [11, 14].

In order to make it possible for nematologists to evaluate and analyze the nematode data, numerous efforts have been made to manage it. For example, an association of scientists in the domains of biology and computer called Wormbase [10] developed an ontology for the Caenorhabditis elegans nematode in order to meet information accessibility requirements. The constructed ontology has a great focus on cells and anatomy and seeks to provide data that can handle sophisticated retrieval system queries. Nemys [12] is another web-based system for databases on nematodes. It avails information about the taxonomies of nematodes such as classification, parent, child, environments, and taxonomic citation.

Researchers also try to use machine learning techniques to automatically identify different nematodes based on microscopic images. For nematodes that parasitize plants, in [17] they developed an image dataset. They present a deep learning model for identifying plant-parasitic nematodes and confirm that machine learning can assist in the identification of pests in Indonesian soil. Further, in [1] (Andre da Silva Abade et al) applied a convolutional neural network(CNN) to identify phytonematodes in soybean crops in Brazil. They proposed a publicly available, open-source dataset called NemaDataset that contains the five major phytonematode species that harm soybean crops. Furthermore, in [18] they investigated the potential of creating a CNN appropriate for categorizing microscope images of three species EPNs (H. bacteriophora, S. carpocapsae, and S. feltiae). For IJs, they reached an average validation accuracy of 88.28%.

* Climate Change Information Center & Renewable Energy & Expert Systems.
† Plant Protection Research Institute.

The cooperation between nematologists and computer scientists helps in advancing the field of nematodes. Ontology is a way to organize and standardize the used concepts in a specific domain. With the use of ontologies, people and various application systems can communicate about a given topic in a common and shared way [6]. Ontologies provide a generally accepted understanding of a subject and capture domain knowledge generically. An ontology can explain the types of entities that exist in a domain, how they might be related to one another, and what they imply. Ontologies are employed to support communication and understanding among the various parties. By addressing the issue of semantic heterogeneity, ontologies enable semantic interoperability between various online applications and services. With the help of ontologies, the web can become a self-navigating and self-understood environment where search engines return results tailored to the user's needs [7].

Empowering nematologists with a facility to identify the morphometric characteristic of EPNs advances the field of nematology. So we present a tool to identify EPNs species. First, we built an ontology for the species of EPNs Steinernema and Heterorhabditis. This ontology facilitates and accelerates EPNs research by integrating EPNs data. We used the Web Ontology Language (OWL) [2]. OWL is one of the semantic web languages made by The World Wide Web Consortium to represent detailed information about various things, and representing the relation between them. The ontology allows users to easily query these data. Second, we developed

Table 1: Entomopathogenic nematodes species

Heterorhabditis	Steinernema				
H. amazonensi	*S. abbasi*	*S. brazilense*	*S. huense sp. n*	*S. ohioense*	*S. scarabaei*
H. atacamensis	*S. aciari*	*S. cameroonense*	*S. ichnusae*	*S. oregonense*	*S. schliemanni*
H. bacteriophora	*S. affine*	*S. carpocapsae*	*S. innovation*	*S. pakistanense*	*S. siamkayai*
H. baujardi	*S. africanum*	*S. caudatum*	*S. intermedium*	*S. papillatum*	*S. sichuanense*
H. beicherriana	*S. akhursti*	*S. caudatum*	*S. jeffreyense*	*S. phyllophagae*	*S. silvaticum*
H. brevicaudis	*S. anatoliense*	*S. ceratophorum*	*S. jollieti*	*S. poinari*	*S. surkhetense*
H. downesi	*S. anomalae*	*S. changbaiense n. sp.*	*S. karii*	*S. populi*	*S. taiwanensis*
H. floridensis	*S. apuliae*	*S. cholashanense*	*S. khoisanae*	*S. portoricense*	*S. tami*
H. georgiana	*S. arasbaranense*	*S. citrae*	*S. khuongi*	*S. puertoricense*	*S. texanum*
H. hawaiiensis	*S. arenarium*	*S. colombiense*	*S. kraussei*	*S. puertoricense*	*S. thanhi*
H. indica	*S. ashiuense*	*S. costaricense*	*S. kraussei*	*S. pui*	*S. thermophilum*
H. mexicana	*S. asiaticum*	*S. cubanum*	*S. kushidai*	*S. puntauvense*	*S. tielingense*
H. noenieputensis	*S. australe*	*S. cumgarense*	*S. lamjungense*	*S. pwaniensis*	*S. tophus*
H. safricana	*S. backanense*	*S. diaprepesi*	*S. leizhouense*	*S. ralatorei*	*S. unicornum*
H. taysearae	*S. bakwenae*	*S. eapokense*	*S. litchii*	*S. rarum*	*S. vulcanicum*
H. argentinensis	*S. balochiense*	*S. ethiopiense*	*S. litorale*	*S. riobrave*	*S. websteri*
H. marelatus	*S. bedding*	*S. everestense*	*S. loci*	*S. riojaense*	*S. weiseri*
H. megidis	*S. beitlechemi*	*S. fabii*	*S. longicaudum*	*S. ritteri*	*S. xinbinense*
H. zealandica	*S. bertusi*	*S. feltiae*	*S. minutum*	*S. robustispiculum*	*S. xueshanense*
	S. bicornutum	*S. glaseri*	*S. monticolum*	*S. sacchari*	*S. yirgalemense*
	S. biddulphi	*S. goweni*	*S. neocurtillae*	*S. sandneri*	*S. scarabaei*
	S. bifurcatum	*S. guangdongense*	*S. nepalense*	*S. sangi*	*S. schliemanni*
	S. boemarei	*S. hebeiense*	*S. nguyeni*	*S. sasonense*	*S. siamkayai*
	S. borjomiense	*S. hermaphroditum*	*S. nyetense*	*S. scapterisci*	*S. sichuanense*

a mobile application that reads and queries EPNs ontology and retrieves species details. The app users can get the details of different species and search for specific species based on their morphometric characteristics.

The paper is organized as follows: Section 2 lists the materials and methods. Section 3 presents a validation test for the developed application. Section 4 describes the results and discussion. Finally, Section 5 is the main conclusion.

2 Materials and Methods

2.1 Data Collection

The EPNs data was obtained through conducting a meeting with a domain expert who guided us with a reference book [9]. This book covers the most valid EPNs species till 2016. Hence, we used the most recent research papers to collect EPNs data which are available up to the paper writing date. We collected morphometric data of about 120 species of Steinernema and 19 species of Heterorhabditis. Table 1 shows a list of EPNs species that have been included in the developed EPNs ontology.

2.2 Ontology Building

We used OWL language to build EPNs. Also, we used a web-based version of Protégé [13] to build and edit the ontology. Protégé is available in desktop and web-based versions. We used the web-based version to enable our team to access and edit the ontology. The parent class of the EPNs ontology is Nematode. It has a child which is Entomopathogenic-nematodes. Steinernema, and Heterorhabditis are represented as subclasses of Entomopathogenic-nematodes.

As shown in Figure 1, S. abbasi and S. affine are subclasses of Steinernema. Subclasses in the ontology represent species. Each class or subclass has properties, these properties (annotation properties) represent morphometric characteristics as shown on the right side of the Figure 1. These properties have values that are used in the class description, for example, a class S. abbasi has L (body length) = 541 μm, W (greatest body width) = 29 μm, EP (distance from anterior end to excretory pore) = 48 μm, NR (distance from anterior end to nerve ring) = 68 μm, ES (esophagus length) = 89 μm.

2.3 Similarity Search

We used a vector space model to represent species in EPNs. We handle each species as a vector and each characteristic as a feature or a dimension of this vector, such as body length, pharynx length, and tail length. So, we can apply distance measures such as Euclidean distance and cosine similarity. First, the Euclidean distance (Ed) between two vectors a={x1, x2, x3, ………, xn} and b={y1,y2,y3,………., yn} is calculated as in Equation (1)

$$Ed\,(a,b) = \sqrt[2]{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2 + \cdots + (x_n-y_n)^2} \quad (1)$$

After calculating the Euclidean distance, we normalize the results using the Min-Max technique in Equation (2) to be in

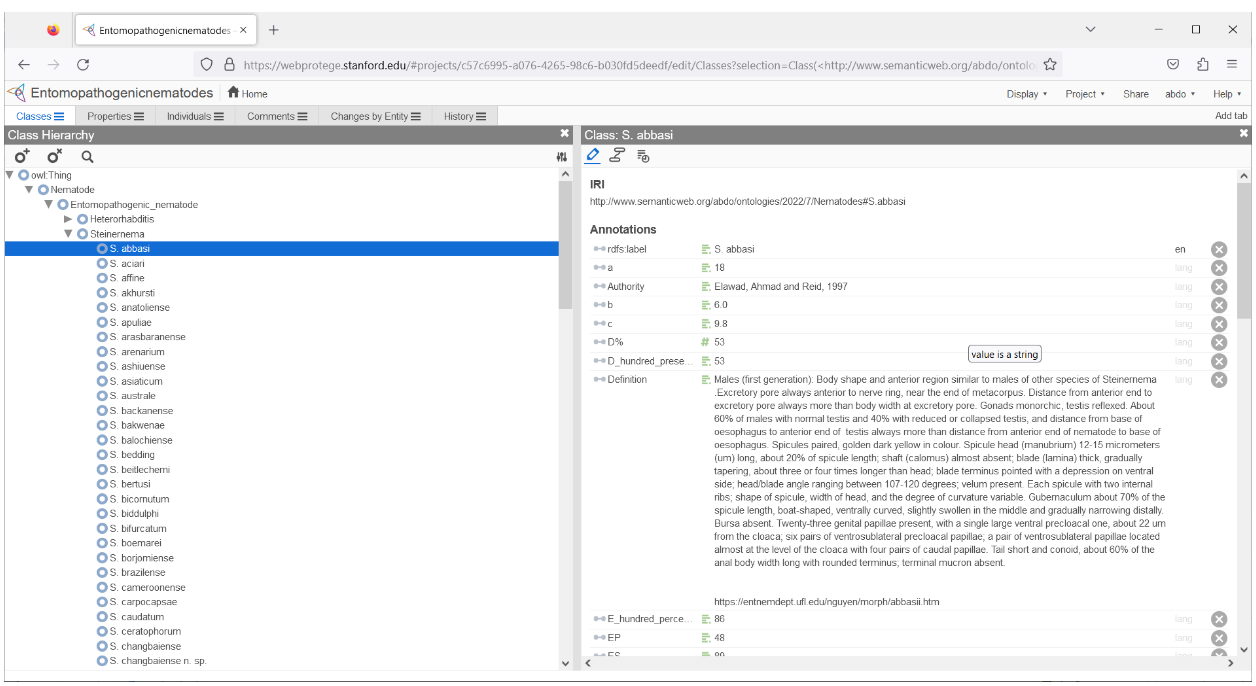


Figure 1: Entomopathogenic nematode ontology

intervals from zero to one [0, 1] typically known as feature scaling [3].

$$\text{Resultscaled} = (\text{Result} - \text{Resultmin})/(\text{Result}_{\text{max}} - \text{Result}_{\text{min}}) \quad (2)$$

In the proposed system, the Min-Max technique is modified by not subtracting the Min-value (Resultmin) from the value (Result).

$$\text{Resultscaled} = (\text{Result})/(\text{Result}_{\text{max}} - \text{Result}_{\text{min}}) \quad (3)$$

Then we subtracted the result from 1 and multiplied it by 100 to convert the distance to similarity. If we subtracted the minimum result from the result, it would give the first nearest similar species 100 percent, so we removed this step from the Min-Max normalization technique.

Second, the cosine similarity (Cs) between two vectors a and b is calculated as in Equation (4).

$$\text{CS}(a, b) = \frac{a.b}{||a||*||b||} \quad (4)$$

The outcome of the cosine similarity falls between (0, 1). Its higher value indicates that the vectors are more similar. Additionally, we multiplied the cosine similarity by 100 to present the results in a way that is easy for system users to understand. The system displays the results in two tables, one using cosine similarity and the other using Euclidean distance.

2.4 Developing a Mobile Application

In this phase, a mobile app was built using ".NET MAUI". It is a development platform introduced by Microsoft. Through .Net MAUI, the developers can develop apps which are compatible with various operating systems, including Android, iOS, Windows, and macOS by writing a single code. First, we downloaded the developed EPNs ontology from the Protégé in RDF/XML format as an OWL file. After that, we wrote a C# programming language code to develop the mobile app. We used XDocument and XPath to parse the OWL file.

We designed a page to display the Heterorhabditis and Steinernema species. It enabled the end user to get details about specific species. Figure 2 displays screenshots of the main page of the developed application and the species details for H. mexicana. Further, we developed a page that allows the end-user to do an advanced search. The end-user finds similar species based on the input characteristics of infective juveniles.

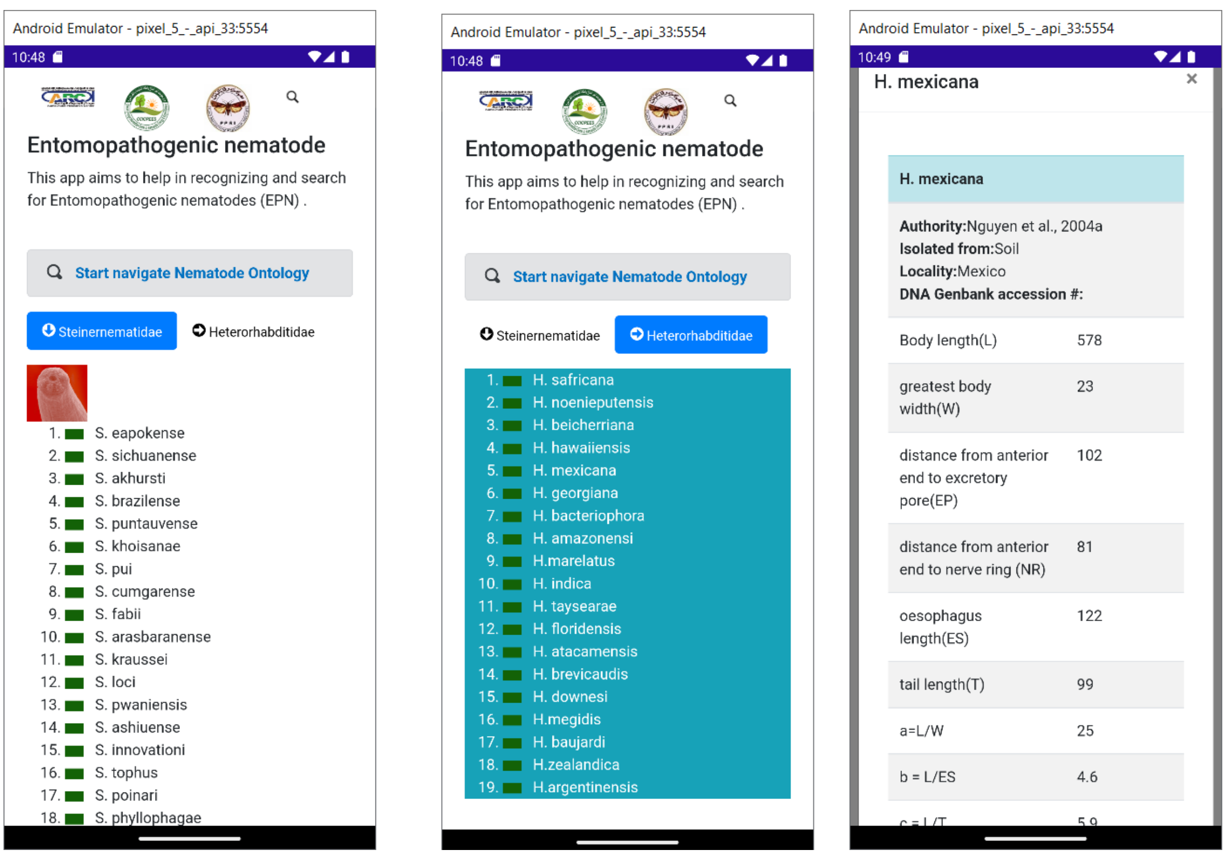


Figure 2: Mani page of the application and the details screen of species

3 Validation Test

To validate the proposed system a group of test cases has been validated by the domain experts and the system gave valid results. Furthermore, a questionnaire was conducted. A group of researchers evaluated the developed applications and answered the following questions.

1. Are you satisfied with the application?
2. Is the information about each type of entomopathogenic nematode adequate?
3. Does the application help in identifying the entomopathogenic nematode and the similarities between them?
4. Add suggestions to improve the application.

Number of researchers who responded to the questionnaire is 15. Figure 3 was their response. All researchers are satisfied with the application and recommend new enhancements.

The researchers suggested the following ideas to enhance the proposed application

- Is it possible to add identification by the standard specifications for males and females?
- The application should not be limited to nematodes, their types, and classification, but should include their relationship to other fields to interest all researchers who are not specialized in nematodes, such as the effect of natural and chemical compounds on them and the degree of toxicity and their impact.
- It is recommended to enter similar strains that were isolated from other places with the genetic number on the GenBank.
- Is it possible to add a picture of each type?
- Can we apply it to the iPhone as well?
- The application is part of an integrated pest control program recommended. Is it possible to add some research on nematode?
- The data should be up to date.

4 Results and Discussion

To evaluate the proposed system, we conducted many experiments. We used morphometric characters of infective juveniles data of some known species. We started with three species of the genus Heterorhabditis (H. megidis, H. bacteriophora, and H. indica). Then we did many experiments for species of the genus Steinernema. The proposed system will find all species that are similar to the given infective juvenile characteristics, and it will display the results in descending order based on the more common similarities.

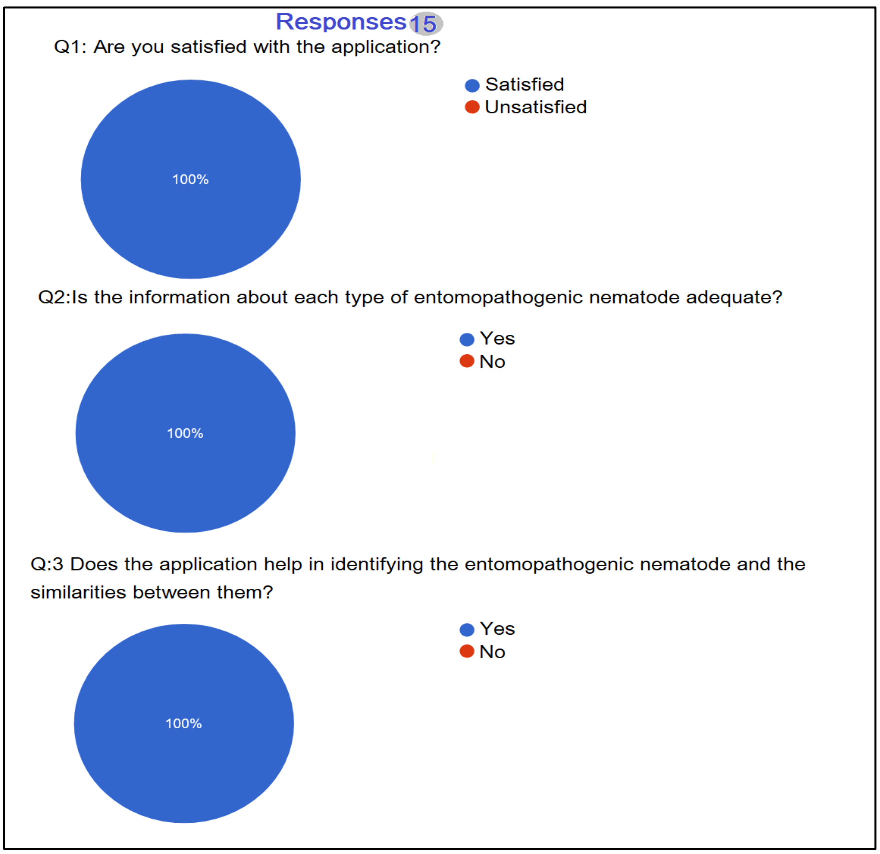


Figure 3: Responses to the evaluation questionnaire

4.1 Heterorhabditis Results

The system was used to compare the infective juvenile properties of (H. megidis, H. bacteriophora, and H. indica) with those of other species. Table 2 shows the top five results of the system using cosine similarity. The system results indicate that the system can identify the given species with 100% of known species. All the remaining five nearest species are of type Heterorhabditis, for example (H. amazonensi) is 99.88% similar to (H. megidis).

On the other hand, Table 3 shows the top five results of the system using Euclidean distance, the top five similar species of (H. bacteriophora and H. indica) are of Heterorhabditis, while in the case of (H. megidis), the system can identify H. megidis is similar to itself with 100 percentage, but the nearest five species are of type Steinernema, Since the body length of (H. megidis) is the tallest one of Heterorhabditis, this property is close to body lengths of these results species of Steinernema (S. ashiuense, S. sangi, S. everestense, S. robustispiculum, and S. unicornum), but the similar ratio is low between (H. megidis) and those results, for example (S. unicornum) is 77.56% similar to (H. megidis) because (H. megidis) belongs to Heterorhabditis, the system results agree with the assumption that (H. megidis) is a distinct species [16].

4.1.1 Identifying *H. atacamensis* Similar Species. We selected (H.atacamensis) to be tested on the system since (Hunt, D. J. and Nguyen, K. B) mention that "H. atacamensis is most similar to H. marelatus, H. downesi, H. amazonensis, H. bacteriophora, H. georgiana and H. safricana," [7]. The system's outcome is based on data results from H. atacamensis as is shown in Figure 4.

The system results state that (H. atacamensis) is 100% similar to itself. Further, the results of the top ten systems when using Euclidean distance are similar to those in the reference book [7] (H. downesi, H. marelatus, H. amazonensis, H. bacteriophora, and H. safricana). On the other hand, the results of the system when using cosine similarity agree with [7] (H. downesi, H. bacteriophora, and H. safricana). According to the findings, Euclidean distance appears to be more effective than cosine similarity in identifying the most closely related species of (H. atacamensis). The lack of recognition of (H. georgiana) using Euclidean distance and cosine similarity may be due to incomplete data on (H. georgiana).

Table 2: The results of running the system on some species of heterorhabditis using cosine similarity

Species	Similar species	Cosine similarity * 100
H. megidis	*H. megidis*	100%
	H. amazonensi	99.88009 %
	H. marelatus	99.8461 %
	H. mexicana	99.82926 %
	H. beicherriana	99.77374 %
	H. floridensis	99.74422 %
H. bacteriophora	*H. bacteriophora*	100%
	H. argentinensis	99.99906 %
	H. safricana	99.89509 %
	H. beicherriana	99.89402 %
	H. zealandica	99.8653 %
	H. baujardi	99.85868 %
H. indica	*H. indica*	100%
	H. baujardi	99.94233 %
	H. mexicana	99.92343 %
	H. beicherriana	99.91491 %
	H. safricana	99.91303 %
	H. amazonensi	99.85803 %

Table 3: The results of running the system on some species of heterorhabditis using euclidean distance

Species	Similar species	Euclidean distance
H. megidis	*H. megidis*	100%
	S. ashiuense	80.01814 %
	S. sangi	78.93036 %
	S. everestense	78.87545 %
	S. robustispiculum	78.53144 %
	S. unicornum	77.56857 %
H. bacteriophora	*H. bacteriophora*	100%
	H. argentinensis	99.61745 %
	H. safricana	94.79906 %
	H. mexicana	94.04926 %
	H. beicherriana	93.65662 %
	H. floridensis	91.99536 %
H. indica	*H. indica*	100%
	H. baujardi	96.65402 %
	H. noenieputensis	95.74544 %
	H. mexicana	93.57463 %
	H. floridensis	91.44147 %
	H. amazonensi	91.41016 %

4.2 Steinernema Results

Tables 4 and 5 display the results of the system by comparing the infective juvenile characteristics of (S. eapokense, S. apuliae, and S. glaseri) with other species using cosine similarity and Euclidean distance, respectively. The system results indicate that the system can identify the given species with 100% of known species. Further, results show the top five similar species for each species.

The system's results show that some results agree with both cosine similarity and Euclidian distance. For instance, common species that are similar to (S. eapokense) are (S. minutum, S. surkhetense, and S. backanense). We observe that Vietnam is the location of (S. eapokense), and the locations of the top species that are comparably shown in Table 6 are located in Asia.

4.2.1 Identifying *S. bicornutum* Similar Species. We selected (S. bicornutum) as a species of Steinernema to evaluate and test the system results. Figure 5 shows screenshots of the search screen and the result of the system on data of (S. bicornutum). It shows how to use the proposed system to locate the closest species that are comparable to one of Steinernema. The results of the system allow nematologists to compare any species to those that already exist, and to identify those that are similar to one another in terms of physical characteristics. The cosine similarity and Euclidean distance agree on similar species for most of the top ten (S. affine, S. monticolum, S. intermedium, S. biddulphi and S. lamjungense).

5 Conclusion

EPNs are one of the biological control agents. They are used in integrated pest management (IPM) programs. We aimed with this paper to help researchers identify the correct EPNs species. We utilized semantic web technologies such as Web Ontology Language (OWL) to build an ontology and provide semantic information for the nematode community. Further, we introduced a method for identifying EPNs by adopting morphometric characteristics. We used cosine similarity and Euclidean distance to compare different EPNs species and identify similar species.

Euclidean distance appears to be more effective than cosine similarity in identifying the most closely related species. We built a mobile application reads and parses a developed ontology to display EPNs species details and enrich researchers with a

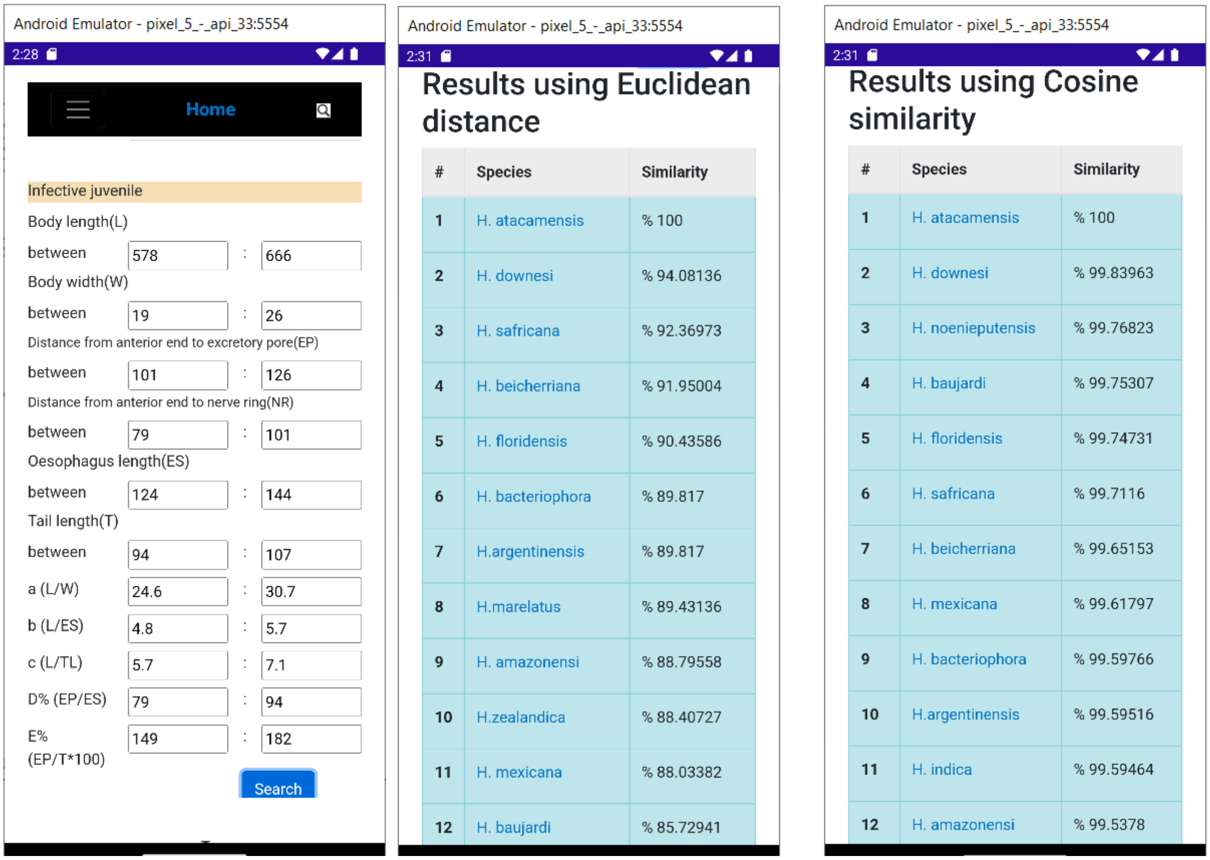


Figure 4: System results for *H.atacements* nearest species

Table 4: The results of running the system on some species of Steinernema using Cosine similarity

Species	Similar species	Cosine similarity * 100
S. eapokense	*S. eapokense*	100 %
	S. backanense	99.97615 %
	S. surkhetense	99.94342 %
	S. balochiense	99.90489 %
	S. minutum	99.89507 %
	S. sasonense	99.89184 %
S. apuliae	*S. apuliae*	100 %
	S. pui	99.98583 %
	S. aciari	99.98195 %
	S. ethiopiense	99.95362 %
	S. thanhi	99.94365 %
	S. brazilense	99.92453 %
S. glaseri	*S. glaseri*	100 %
	S. arenarium	99.77076 %
	S. affine	99.65601 %

	S. pwaniensis	99.57247 %
	S. lamjungense	99.48499 %
	S. sandneri	99.43186 %

Table 5: The results of running the system on some species of Steinernema using Euclidian distance

Species	**Similar species**	**Euclidean distance**
S. eapokense	*S. eapokense*	100 %
	S. minutum	97.40327 %
	S. surkhetense	97.26441 %
	S. cumgarense	96.99397 %
	S. asiaticum	96.85899 %
	S. backanense	96.4893 %
S. apuliae	*S. apuliae*	100 %
	S. aciari	94.41532 %
	S. khoisanae	91.4565 %
	S. bakwenae	89.82047 %
	S. pui	89.36111 %
	S. guangdongense	88.29194 %
S. glaseri	*S. glaseri*	100 %
	S. arenarium	86.28402 %
	S. aciari	79.97972 %
	S. brazilense	79.60566 %
	S. apuliae	78.06147 %
	S. vulcanicum	77.63134 %

Table 6: Location of species

Species	Location
S. minutum	*Thailand*
S. surkhetense	*Nepal*
S. balochiense	*Pakistan*
S. sasonense	*Vietnam*
S. cumgarense	*Vietnam*
S. asiaticum	*Pakistan*

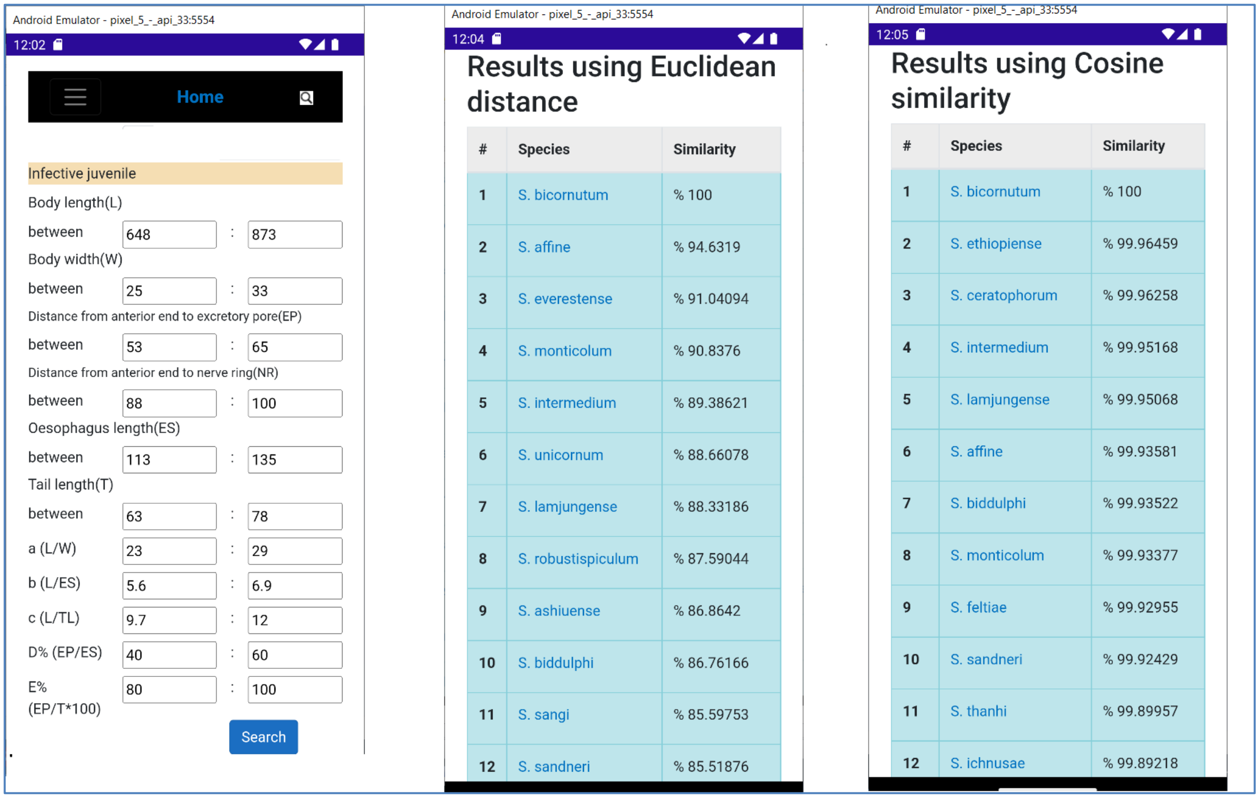


Figure 5: System results for *S. bicornutum* nearest spec

tool to search and find similar species. The result explained the strength of the proposed method in finding similarities between different species. One of the limitations of this research is that it doesn't include male morphometric characteristics in identifying similar species due to lack of time and system simplification. For future work, we can apply morphometric characteristics of males in comparison between species and automatically identify the most similar species for known and newly discovered EPNs species. Furthermore, integrating image identification techniques with similarity may enhance sytem performance and results.

References

[1] A. D. S. Abade, L. F. Porto, P. A. Ferreira, and F. D. B. Vidal, "Nemanet: A Convolutional Neural Network Model for Identification of Nematodes Soybean Crop in Brazi,l" arXiv preprint arXiv:2103.03717, 2021.

[2] S. Bechhofer, "OWL: Web Ontology Language," Encyclopedia of Database Systems, Springer, Boston, pp. 2008-2009, 2009.

[3] G. Ciaburro, V. K. Ayyadevara, and A. Perrier, "Hands-on Machine Learning on Google Cloud Platform: Implementing Smart and Efficient Analytics using Cloud ML Engine," Packt Publishing Ltd, 2018.

[4] A. R. Dillman P. W. Sternberg, "Entomopathogenic Nematodes," *Current Biology*, 22(11):R430-R431, 2012.

[5] W. K. Dodds and M. R. Whiles, "Chapter 10—Multicellular Animals", W. K. Dodds & M. R. Whiles (Eds.), Freshwater Ecology (Second Edition), Academic Press, pp. 221-257, 2010.

[6] Susan F. Ellakwa, E. El-Azhary, and P. Passent El-Kafrawy, "Integrated Ontology for Agricultural Domain," *International Journal of Computer Applications*, 54(2):46-53, 2012.

[7] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The Role of Ontologies in Biological and Biomedical Research: a Functional Perspective," *Brief. Bioinform.*, 16:1069-1080, 2015. https://doi.org/10.1093/bib/bbv011

[8] K. L. Howe, B. J. Bolt, S. Cain, J. Chan, W. J. Chen, P. Davis, ... and P. W. Sternberg, "WormBase 2016: Expanding to Enable Helminth Genomic Research," Nucleic Acids Research. 44(D1):D774-D780, 2016.

[9] D. J. Hunt and K. B. Nguyen, "Advances in Entomopathogenic Nematode Taxonomy and Phylogeny," Brill, 2016.

[10] R. Y. Lee and P. W. Sternberg, "Building a Cell and Anatomy Ontology of Caenorhabditis Elegans," *Comparative and Functional Genomics*, 4(1):121-126, 2003.

[11] S. Morand, "Life-History Traits in Parasitic Nematodes: A Comparative Approach for the Search of Invariants," Functional Ecology, pp.210-218, 1996.

[12] Nemys eds. Nemys: World Database of Nematodes. Accessed at https://nemys.ugent.be on 2023-03-02. doi:10.14284/366, 2023.

[13] N. F. Noy, M. Crubézy, R. W. Fergerson, H. Knublauch, S. Tu, J. Vendetti, and M. Musen, "Protégé-2000: An Open-Source Ontology Development and Knowledge-Acquisition Environment," AMIA 2003 Open Source Expo, 2003.

[14] W. Nurashikin-Khairuddin, S. N. A. Abdul-Hamid, M. S. Mansor, I. Bharudin, Z. Othman, and J. Jalinas, "A Review of Entomopathogenic Nematodes as a Biological Control Agent for Red Palm Weevil, *Rhynchophorus ferrugineus* (Coleoptera: Curculionidae)". *Insects*, 13(3):245, 2022. Doi: 10.3390/insects13030245. PMID: 35323543.

[15] G. O. Poinar, "Chapter 9—Phylum Nemata," J. H. Thorp & D. C. Rogers (Eds.), Thorp and Covich's Freshwater Invertebrates (Fourth Edition), Academic Press, pp. 169-180, 2016.

[16] G. O. Poinar Jr., Trevor Jackson, and Michael Klein, "Heterorhabditis *megidis* sp. n. (*Heterorhabditidae*: Rhabditida), Parasitic in the Japanese Beetle, *Popillia japonica* (Scarabaeidae: Coleoptera)," *Proceedings of the Helminthological Society of Washington*, 54(1):53-59, 1987.

[17] Nabila Husna Shabrina, Ryukin Aranta Lika, and Siwi Indarti,. "Deep Learning Models for Automatic Identification of Plant-Parasitic Nematode," *Artificial Intelligence in Agriculture*, 7:1-12, 2023.

[18] J. Uhlemann, O. Cawley, and T. Kakouli-Duarte, "Nematode Identification using Artificial Neural Networks," DeLTA, pp. 13-22, July 2020.

Susan F. Ellakwa is a researcher at the Climate Change Information Center and Renewable Energy and Expert Systems, Agricultural Research Center (ARC), Giza, Egypt. She received her PhD degree in Computer Science from the Faculty of Science, Menoufia University, Menoufia, Egypt, in 2013. From 2016 to 2019 she was head of the Knowledge Engineering and Expert System Tools Department at the Central Laboratory for Agricultural Expert Systems (CLAES), Agricultural Research Center (ARC), Giza, Egypt. Her research interests include Artificial Intelligence, Machine learning, and Expert systems.

Abdelrahman Elsayed is a researcher at Climate Change Information Center and Renewable Energy and Expert Systems, Agricultural Research Center, Giza, Egypt. He received his Ph.D. in 2017 from the faculty of Computers and Information, at Cairo University, Egypt. His research focuses on artificial intelligence applications and data-intensive algorithms.

Rasha Elnemr (photo not available) is a Statistics and Mathematics Specialist in the Climate Change Information Center and Renewable Energy and Expert Systems, Agricultural Research Center (ARC), Giza, Egypt since 2014. She received her Ph.D. in Computer Science from the Faculty of Science, Menoufia University, Menoufia, Egypt, in 2019. She is a Lecturer at the Higher Institute for Computer and Administrative Science. Her research interests are in artificial intelligence, machine learning, expert systems, and bioinformatics.

Saheir F. El-lakwah is a Professor in Pest Physiology Dept., Plant Protection Research Institute, Agricultural Research Center (ARC), Egypt. She received her Ph.D. in "Comparative physiology" at the Department of Zoology, Faculty of Science, Cairo University, Egypt (2008). Her research interests and publications are in integrated pest management, biological control, enhance efficacy of entomopathogenic nematoda, and studying the effect of plant extracts on insect physiology.

A. M. Azazy a professor in Pest Physiology Dept. Plant Protection Research Institute, Agricultural Research Center (ARC), Egypt. He received his Ph.D. in Zoology (Entomopathogenic Nematodes), Faculty of Moshtohor, Plant Protection Department, Zagazig University branch Benha (2001). His research interests and publications are in isolation entomopathogens (nematodes, fungi, and bacteria) from the soil, mass rearing of entomopathogenic nematodes, control of insect pests of the field crops, evaluation of pesticides against pests, ability to make the IPM programs for pests under open field and greenhouse conditions

Evaluating Image Quality through Latent Space Analysis of Autoencoders

Gofur Halmuratov* and Arnošt Veselý†
Czech University of Life Sciences, Prague, CZECH REPUBLIC

Abstract

In the domain of deep learning-driven image classification, the underpinning algorithms often grapple with reduced performance efficacy when working with constrained datasets. Such algorithms typically thrive in scenarios with a considerable volume of data at their disposal, but they falter when their operational spectrum narrows down to a handful of images. This necessitates data augmentation or the synthesis of new data instances via Generative Adversarial Networks (GANs). However, these methodologies do not always yield images that align with the desired quality criteria. As such, the onus of evaluating and sieving out low-quality images falls on the researchers, who must conduct a meticulous manual review of each image. This approach, albeit thorough, is riddled with the challenges of being highly time-consuming and resource-exhaustive.

In this study, we introduce a novel methodology that leverages the latent space of autoencoders for image quality assessment. This unique approach bypasses the need for manual review, allowing us to infer image quality by analyzing the latent space representation. We furnish empirical evidence of our methodology's efficacy through extensive experimentation, which unveils its superior performance over conventional image quality evaluation techniques.

Key Words: Autoencoder, latent space analysis, image quality, perceptual metrics, regularization, deep learning.

1 Introduction

Deep learning algorithms have shown remarkable performance in image classification and recognition tasks. However, their effectiveness is often hindered when working with a small number of images [2]. When the training dataset consists of only a few images, researchers usually augment them or synthesize new images using generative adversarial networks (GANs) [5]. Nevertheless, these methods may not always produce images of the required quality. As a result, manual review of individual images is necessary to remove low-quality ones from the dataset, which is a time-consuming and resource-intensive process [10]. One solution to this issue is to evaluate the quality of images through the analysis of the latent space of autoencoders. An autoencoder is a type of neural network that learns to encode and decode data, reducing the dimensions of the input data into a latent space [6]. The latent space is a compressed representation of the input data, where each dimension represents a learned feature of the input [1]. Analyzing the latent space of autoencoders can provide a more reliable measure of image quality than manual review [8]. Several studies have investigated the use of autoencoders and their latent space for image quality evaluation. In [7], a novel method was proposed for evaluating image quality using the variance of the latent representation. In [12], a perceptual loss function was introduced to improve the quality of the generated images using GANs. A regularization term was added to the loss function to encourage the generated images to be more similar to the real images in the latent space. In [4], the latent space was used to evaluate the quality of face images, and the results were compared with the human perceptual evaluation. Other studies have used autoencoders and their latent space for image generation tasks. In [3], an adversarial autoencoder was introduced, which combines the advantages of both autoencoders and GANs. In [9], a conditional autoencoder was proposed, where the input data and a condition vector are combined to generate high-quality images. In this paper, we propose a methodology for evaluating the quality of images by analyzing the autoencoder latent space. We demonstrate the effectiveness of our approach in comparison to conventional methods for determining image quality [11]. Our method can be used in various image-related tasks, including image restoration, synthesis, and recognition.

2 Methodology

We start with a small number of training datasets and our goal is to improve classification accuracy. First, we attempt to classify the existing datasets using deep learning algorithms. If the results are not satisfactory, we try image augmentation or generative adversary networks to improve the datasets.

However, if these methods do not produce the desired results, we turn to quality determination using the latent space of an autoencoder. Here are the steps involved:

* ORCiD ID: Ing. Gofur Halmuratov https://orcid.org/0000-0003-4403-8046. Email: halmuratov@pef.czu.cz
† ORCiD ID: doc. Ing. Arnošt Veselý, CSc https://orcid.org/0000-0001-8979-1336. Email: vesely@pef.czu.cz

1. We train an autoencoder for all the input training datasets and encode all available images. By decoding them, we obtain the latent space of the autoencoder.
2. We represent the first class of images in latent space.
3. Points of training datasets are generated in the latent space, and we hypothesize a sphere that filters out 95% of the points. We find the radius of this sphere.
4. We synthesize the required number of images using image augmentation or generative adversarial networks.
5. The newly synthesized images are imaged again in latent space. Images outside the sphere are considered low-quality and are excluded from the deep learning models. The ones inside are included in our training dataset. This sphere serves as a qualifier to divide the quality of images into good or bad.
6. Starting from step 2, we repeat the process for the rest of the classes.

By the end of this process, we have high-quality images to train deep learning models, and our training dataset is now larger. In summary, this methodology improves classification accuracy by using deep learning algorithms, image augmentation, generative adversary networks, and quality determination using the latent space of an autoencoder. This methodology aims to improve classification accuracy despite having a small number of training datasets. It does this by using deep learning algorithms and image augmentation or generative adversary networks to improve existing datasets. If these steps do not lead to significant improvement, quality determination is performed using the latent space of an autoencoder. Finally, the high-quality images obtained from this process are used to train deep learning models.

3 Results

We evaluated our proposed methodology on several open access datasets, including MNIST, CIFAR-10 and CIFAR-100. Our approach is designed to improve classification accuracy on datasets with very few images, so we intentionally used a small number of images from each class (5, 10, 20, and 100) in each of the test datasets. We then compared the classification results of these images with and without our proposed methodology to assess its effectiveness. All models used the same autoencoder, which is presented in Table 1: Model 1 and Table 2: Model 2.

Table 3 shows the datasets, the number of images in each

Table 1: Model 1. VAE encoder model

Layer(type)	**Output Shape**	**Param#**	**Connected to**
encoder_input(InputLayer)	[(None,28,28,1)]	0	[]
conv2d(Conv2D)	(None,14,14,512)	5120	['encoder_input[0][0]']
conv2d_1(Conv2D)	(None,7,7,1024)	4719616	['conv2d[0][0]']
flatten(Flatten)	(None,50176)	0	['conv2d_1[0][0]']
dense(Dense)	(None,400)	20070800	['flatten[0][0]']
z_mean(Dense)	(None,100)	40100	['dense[0][0]']
z_log_var(Dense)	(None,100)	40100	['dense[0][0]']
z(Lambda)	(None,100)	0	['z_mean[0][0]','z_log_var[0][0]']

Table 2: Model 2. VAE decoder model

Layer (type)	**Output Shape**	**Param #**
z_sampling(InputLayer)	[(None,100)]	0
dense_1(Dense)	(None,50176)	5067776
reshape(Reshape)	(None,7,7,1024)	0
conv2d_transpose(Conv2DTranspose)	(None,14,14,1024)	9438208
conv2d_transpose_1(Conv2DTranspose)l	(None,28,28,512)	4719104

Table 3: Results of experiments

Dataset	Number of images per category	Data augmentation	Number of qualified images	Accuracy before using proposed method	Accuracy after using proposed method
MNIST	5	500	425	20.6%	25.8 %
MNIST	10	1000	895	25.8%	30.2 %
MNIST	20	2000	1578	27.1%	30.9 %
MNIST	100	5000	3685	25.4%	24.8 %
CIFAR-10	5	500	458	18.9%	21.6 %
CIFAR-10	10	1000	901	23.4%	24.2 %
CIFAR-10	20	2000	1570	28.2%	31.2 %
CIFAR-10	100	5000	4001	34.3%	40.9%
CIFAR-100	5	2500	1978	14.6%	18.2 %
CIFAR-100	10	5000	3951	16.4%	29.5 %
CIFAR-100	20	10000	8012	18.2%	21.8 %
CIFAR-100	100	50000	38417	24.5%	27.6 %

dataset, the number of images after applying our methodology, the overall classification accuracy, and the classification accuracy of the qualified images.

The assessment of the proposed methodology was carried out across three standard benchmarks: MNIST, CIFAR-10, and CIFAR-100. The critical premise of this analysis was to gauge the effectiveness of the classification model when restricted to datasets with a limited number of images per category.

For the MNIST dataset, the proposed methodology showcased marked improvements. With only 5 images per category, the model's accuracy experienced an augmentation of 25.24%, increasing from 20.6% to 25.8%. When the model was supplied with 10 images per category, the observed accuracy escalated from 25.8% to 30.2%, marking an improvement of approximately 17.05%. Furthermore, with 20 images per category, the model's accuracy ascended from 27.1% to 30.9%, signifying a growth of about 14.02%. Intriguingly, a marginal degradation of 2.36% in accuracy, from 25.4% to 24.8%, was observed when the model was tested with 100 images per category.

The application of the methodology on the CIFAR-10 dataset also yielded enhanced accuracy. With 5 images per category, the accuracy was amplified from 18.9% to 21.6%, denoting a relative improvement of 14.29%. For 10 images per category, the accuracy ascended from 23.4% to 24.2%, marking a modest growth of 3.42%. When provided with 20 images per category, the model accuracy improved from 28.2% to 31.2%, constituting a relative gain of 10.64%. Moreover, the accuracy exhibited a substantial surge of 19.24%, improving from 34.3% to 40.9% when tested with 100 images per category.

The CIFAR-100 dataset, although more complex, still witnessed improvements in accuracy with our methodology. For 5 images per category, accuracy augmented from 14.6% to 18.2%, marking an improvement of 24.66%. When 10 images per category were used, a significant leap in accuracy was observed, from 16.4% to 29.5%, translating to an outstanding improvement of 79.88%. For 20 images per category, accuracy rose from 18.2% to 21.8%, implying a relative growth of 19.78%. Lastly, for 100 images per category, the model's accuracy improved from 24.5% to 27.6%, denoting an enhancement of approximately 12.65%.

The results emanating from this study underscore the effectiveness of our proposed methodology. The results reveal consistent improvements in classification accuracy across MNIST and CIFAR-10 datasets. Even in the more complex CIFAR-100 dataset, our methodology continued to show efficacy, indicating its robustness in diverse and challenging contexts. The observed anomalous reduction in accuracy for MNIST with 100 images per category, however, suggests the need for an in-depth investigation into the intricate dynamics at play and potential refinement of the method for better performance.

In summary, our methodology exhibits a commendable potential to enhance the performance of image classification models, particularly when dealing with datasets that offer limited images per category. Future directions of this research will focus on dissecting the unexpected result observed with MNIST (100 images per category), and broadening the applicability of our methodology across other datasets of varying complexity.

4 Discussion and Conclusions

The proposed methodology, conceived to address the challenge of improving classification accuracy with a restricted number of training datasets, was rigorously evaluated on a series of established open-access datasets, namely MNIST, CIFAR-10, and CIFAR-100. Our methodology employs an amalgamation of deep learning algorithms, image augmentation or Generative Adversarial Networks (GANs), and if the results from these steps are insufficient, quality determination is performed via the latent space of an autoencoder.

An inherent limitation of this approach is its dependence on the hypothesis that the autoencoder can accurately capture the true distribution of images within the dataset. This assumption may not stand in scenarios involving complex datasets or datasets containing outliers not well represented by the autoencoder. Additionally, the quality determination phase operates under a predetermined threshold to segregate high-quality from low-quality images, which may not universally hold for all datasets.

The empirical results from our methodology, although tested on smaller datasets, have been promising. The most substantial improvements were observed for datasets with smaller quantities of training images. For instance, with just five images per category in the CIFAR-100 dataset, we observed an improvement of approximately 24.66%, increasing the accuracy from 14.6% to 18.2%. When the model was trained with 10 images per category, the accuracy escalated from 16.4% to a notable 29.5%, translating to an outstanding improvement of nearly 80%. This highlights the efficacy of our methodology in scenarios with limited training data.

However, the observed anomalous reduction in accuracy for the MNIST dataset with 100 images per category, dropping by 2.36% from 25.4% to 24.8%, prompts further investigation. This indicates that while our methodology is potent in scenarios with limited data, its performance may vary in cases where a larger pool of training images is available.

Further exploration into the effectiveness of our methodology on larger, more complex datasets is necessary. As the greatest improvements were observed for datasets with smaller numbers of training images, it remains to be seen how the methodology would perform when the number of training images increases significantly. This highlights the need for further research into the methodology's scalability and its performance with larger, more complex datasets.

In conclusion, the proposed methodology provides a robust and efficacious strategy for bolstering classification accuracy in scenarios where the number of training images is limited. Despite its reliance on certain assumptions and pre-set thresholds, it shows a consistent trend of enhancing classification accuracy across various datasets. Further studies are warranted to explore its effectiveness with larger datasets and more complex image distributions, which could potentially unlock its full potential and extend its applicability.

References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798-1828, 2013.

[2] Y. Bengio and Y. LeCun, *Scaling Learning Algorithms Towards AI,* Large-scale Kernel Machines, MIT Press, Cambridge, ;MA 02142, USA, 1(1):1-41, (2007).

[3] Y. Blau and T. Michaeli, "The Perception-Distortion Tradeoff," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228-6237, 2018.

[4] A. Dosovitskiy and T. Brox, "Inverting Visual Representations with Convolutional Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4829-4837, 2016.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S., Ozair, .and Y. Bengio, "Generative Adversarial Nets," Advances in Neural Information Processing Systems, pp. 2672-2680, 2014.

[6] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, 313(5786):504-507, 2006.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708, 2017.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.

[9] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," *Proceedings of the International Conference on Machine Learning*, pp. 1060-1069, 2016.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, 13(4):600-612, 2004.

[11] Y. Zhang, X. Song, H. Qi, J. Liu, and H. Zhang, "An End-to-End Framework for Image Quality Assessment based on Conditional Generative Adversarial Networks," *IEEE Transactions on Multimedia*, 20(10):2678-2690, 2018.

[12] J. Y. Zhu, T. Park, P, Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223-2232, 2017.

Gofur Halmuratov (photo not available) emerges as a dynamic force at the heart of the Czech University of Life Sciences (ČZU), distinguishing himself as an Engineer (Ing.) and a dedicated Ph.D. student. Situated within the Provozně Ekonomická Fakulta (Faculty of Economics and Management) and the Katedra Informačního Inženýrství (Department of Information Engineering), Gofur is a pivotal contributor to the academic landscape.

In his capacity as a Ph.D. student, Gofur is actively engaged in pushing the frontiers of knowledge, embodying a spirit of curiosity and innovation. His role transcends mere research, as he generously offers his time for consultations and official hours, illustrating a steadfast commitment to the academic community and the development of future information engineers.

Communication with Engineer Gofur Halmuratov is facilitated through email, providing a direct conduit to his profound insights and collaborative potential. Within the vibrant academic milieu of ČZU, Gofur's contributions are instrumental in shaping the future of information engineering, rendering him an invaluable asset within the Provozně Ekonomická Fakulta.

Arnošt Veselý (photo not available)stands as a luminary figure within the academic community of the Czech University of Life Sciences (ČZU). Possessing the esteemed title of Associate Professor and having attained a Doctorate in Sciences (CSc.), he is an eminent presence within the Provozně Ekonomická Fakulta (Faculty of Economics and Management). His academic journey is intricately intertwined with the Katedra Informačního Inženýrství (Department of Information Engineering), where he imparts profound expertise in the realm of information engineering.

As a teacher Associate Professor Veselý's dedication to education is discernible, manifesting in his pivotal role as a guiding force for students. His influence extends beyond the confines of the classroom, encompassing consultations and administrative responsibilities, with his office at PEF/E550 serving as a focal point for academic discourse. Beyond his educational endeavors, Associate Professor Arnošt Veselý is a prolific author, leaving an indelible mark on the scholarly landscape associated with ČZU.

For those seeking his erudition or collaboration, Associate Professor Veselý is easily accessible through email or the ČZU hotline. His multifaceted role as an educator, consultant, and researcher underscores his commitment to the comprehensive development of students and the advancement of knowledge in the dynamic field of information engineering.

Improving Access to Trade and Investment Information in Thailand through Intelligent Document Retrieval

Sirinda Palahan*
University of the Thai Chamber of Commerce, Bangkok 10400, THAILAND

Abstract

Overseas investment and trade can be daunting for beginners due to the vast amount of complex information. This paper presents a chatbot system that integrates natural language processing and information retrieval techniques, aiming to simplify the document retrieval process. The proposed system identifies the most relevant content, enabling users to navigate the intricate landscape of foreign trade and invest more efficiently. Our methodology combines the BM25 model and a deep learning model to rank and retrieve documents aiming to reduce noise in the document content, thereby enhancing the accuracy of the results. Experiments with Thai natural language queries have proven the effectiveness of our system in retrieving pertinent documents. A user satisfaction survey further validated the system's effectiveness. Most respondents found the system helpful and agreed with the suggested documents, indicating its potential as a beneficial tool for Thai entrepreneurs navigating foreign trade and investment.

Key Words: Document retrieval, Thai natural language processing, deep learning, Chatbot system, ranking approach.

1 Introduction

Foreign trade and investment can be a daunting experience for beginners due to the many rules, regulations, and customs involved. The internet has become a valuable resource for finding relevant documents, such as articles and guides that cover all aspects of foreign trade and investment. However, with the growing numberof documents available, it is becoming increasingly difficult to find useful information, and beginners can feel overwhelmed and intimidated by the comprehensive nature of the materials.

To address this issue, traditional information retrieval methods such as TF-IDF, a term-based approach, have been utilized to rank documents based on their relevance to a user's query. Two components of documents that can be considered for retrieval are the document content and headings. The majority of research studies (Wu, Luk, Wong, & Kwok, [24]; Guo, Alamudun, & Hammond, [4]; Zaragoza, Craswell, Taylor, Saria, &Robertson, [26]; Robertson, Zaragoza, [20]; Robertson, Zaragoza, & Taylor, [21]; Zhai & Lafferty, [27]; Hiemstra, [5]; Amati & Van Rijsbergen, [1]) employ the content of documents in term-based approaches. Retrieving documents based on their entire content can yield more comprehensive outcomes, as it considers all the information encompassed within the document, including details that might not be expressed in the headings. However, the full content of a document may contain a significant amount of irrelevant or extraneous information, which can make it difficult to accurately retrieve relevant results. Moreover, the accuracy of such an approach can be limited by the quality of the tokenization process. If the tokenizer fails to accurately represent the terms in the document, the results may be inaccurate, leading to less relevant search results. The document headings can provide valuable information for accurate document retrieval. Retrieving documents based on the headings can avoid noise in the content that might be irrelevant to the query. It can be faster than retrieving documents based on their full content. This is because headings are often shorter and more focused, making it easier to quickly determine their relevance to the query. Our proposed method addresses this issue by considering both the document content and headings for retrieval, resulting in more accurate and efficient search results.

To enhance document retrieval accuracy, deep learning models have been integrated into the system to consider not only the terms but also the semantic aspects of documents. These models employ techniques suchas neural embeddings and deep neural networks to learn distributed representations of text, which can capture the underlying semantic meaning of words in a document. However, utilizing deep learning models for low-resource languages like Thai can be particularly challenging due to the complexity of the language. Thai lacks clear word boundaries, exhibits sentence pattern variations, and contains compound word ambiguity, all of which contribute to its complexity and make the development of effective IR techniques difficult. Nevertheless, we believe that leveraging natural language processing techniques is critical to retrieving relevant documents in Thai to address entrepreneurs' queries. This paper offers two primary contributions. First, we propose a deep-learning-based model specifically designed to interpret user queries in Thai and retrieve the most pertinent documents. Second, we integrate this model into a chatbot, providing a comprehensive end-to-end solution tailored to the business and economic sectors. By doing so, the model is tested in a practical environment, ensuring its utility and effectiveness. The ultimate goal of this system is to deliver succinct and relevant document content to aid Thai entrepreneurs in maneuvering through foreign trade and investment processes. This paper is

* Email: sirinda_pal@utcc.ac.th.

organized as follows: Section 2 provides an overview of related works, while Section 3 discusses the methodology of the proposed system. Section 4 presents the experimental results, and Section 5 provides conclusions and future directions for this work.

2 Literature Review

2.1 Information Retrieval

Several studies have been conducted in the Information Retrieval community on ad-hoc document retrieval tasks. In the ad hoc retrieval task, a system is given a user query, and its task is to return the top most relevant documents in a corpus, according to some ranking functions. The traditional Information Retrieval approaches evaluate users' queries primarily based on a word match at a syntactic level (Wu, Luk, Wong, & Kwok, [24]; Guo, Alamudun, & Hammond, [4]; Zaragoza, Craswell, Taylor, Saria, and Robertson., [26]; Robertson, Zaragoza, & Taylor, [21], [20]; Zhai & Lafferty, [27]; Hiemstra, [5]; Amati & Van Rijsbergen, [1]). The TF-IDF approach considers the frequency of words within a document, i.e., term-frequency (TF), and the inverse document frequency (IDF) of the same words across the entire collection of documents. While the TF-IDF approaches have been widely used in information retrieval, they have limitations in accurately ranking documents based on their relevance to a user's query. Because they rely solely on the occurrence of query terms within the document, without considering other factors that may affect the relevance of the document to the user's query, such as document length or the proximity of query terms within the document. To address these limitations, the BM25 approach was developed, which takes into account not only the frequency of terms in a document but also the length of the document and the average length of documents in the collection. By incorporating these additional factors, BM25 has been shown to outperform TF-IDF and language modeling approaches in various retrieval tasks (Naseri, Dalton, Yates, & Allan, [14]; Soni & Roberts, [22]; Lan, Ge, & Kong, [8]).

However, both approaches do not consider the order and semantics of words within the document. The advancement of deep learning methods has led to an improvement in IR tasks and ad-hoc information retrieval tasks. The Deep Learning (DL) methods have been employed in IR to add semantic aspects to the retrieval tasks. There are three broad categories of DL approaches to IR based on their influence on the query representation, the document representation, or ranking models (Mitra, Craswell, [11]). Mitra, Nalisnick, Craswell, & Caruana, [13] proposed the Dual Embedding Space Model (DESM) to map the query words into the input space and the document words into the output space and calculate a relevance score by combining the cosine similarities across all pairs of the query-document words. The experiments showed that the proposed model could rank top documents better than a traditional ranking method based on TF-IDF. Yang et al. (Yang, Zhang, & Lin, [25]) applied BERT, a deep learning model, to improve the document representation where documents are longer than the length of input BERT. The authors overcome the problem by applying inference over individual sentences and combining sentence scores into document scores. The approach was tested on TREC microblog and newswire test collections, and the test results showed their approach was effective. One of the first ranking models is the Deep Structured Semantic Model (DSSM) (Huang, He, Gao, Deng, Acero, & Heck, [6]), a neural ranking model for the ad-hoc retrieval task. In this approach, queries and documents are mapped into a low-dimensional space, and relevance is calculated by determining the distance between them. Finally, Lu & Li, [10] applied a deep learning model to matching tasks, such as finding relevant answers to a given question. The authors compared their deep matching model to the inner-product-based model. The empirical results showed that the proposed model outperformed inner-product-based matching models on real-world datasets.

There has been an increasing interest in combining traditional term-based approaches with semantic approachesin information retrieval, aiming to capture the benefits of both approaches (Galke, Saleh, & Scherp, [3]; Noraset, Lowphansirikul, & Tuarob, [16]). This hybrid approach leverages the strengths of term-based approaches in capturing local and discrete representations of text, as well as the strengths of deep learning-based approaches in capturing the semantic meaning of words and phrases. Mitra et al. (Mitra, Diaz, & Craswell, [12]) proposed a document ranking model that combines traditional methods relying on terms in the body text and newer models relying on distributed representations using two separate deep neural networks. One network matches the query and document using local representation, and the other matches them using distributed representation. The two networks are trained together as a single neural network, and the authors demonstrated that this approach outperforms traditional methods and other models based solely on neural networks. Galke et al. (Galke, Salleh, & Scherp, [3]) evaluated different techniques that use word embedding for information retrieval in short query scenarios. The techniques include word centroid similarity, paragraph vectors, Word Mover's distance, and IDF re-weighted word centroid similarity. The results show that word centroid similarity is the best technique, especially when the word frequencies are re-weighted using IDF before aggregating the respective word vectors of the embedding. In some cases, the proposed cosine similarityof IDF re-weighted word vectors outperforms the traditional TF-IDF baseline by 15%.

2.2 Natural Language Processing for the Thai Language

Natural language processing (NLP) is a branch of computer science, specifically artificial intelligence, which aims to train computers to understand spoken words and written text, much like a human would. As NLPresearch has expanded into more languages over the past decade, the number of languages covered has also increased. Similarly, NLP development for the Thai language has been growing continuously. The combination of natural language processing (NLP) and information retrieval has enabled researchers to develop systems capable of retrieving articles and answering questions automatically. Limkonchotiwat, Phatthiyaphaibun, Sarwar, Chuangsuwanich, & Nutanong [9] proposed a novel system called WabiQA, which leverages NLP and information retrieval techniques to answer questions in the Thai language. Specifically, the system uses Thai Wikipedia articles as the knowledge source and retrieves the article most likely to contain the answer. A bi-directional LSTM model is then employed to locate candidate answers, which are ranked and presented to theuser. The system was evaluated in the National Software Contest 2019 and outperformed competitors' systems by a significant margin. The research findings suggest the potential to develop intelligent NLP applications for low-resource languages such as Thai using existing tools and resources, with implications for a wide range of NLP tasks.

In Kawtrakul, Andres, Ono, Thumbanon, et el. [7]), a framework was introduced to develop a Thai document retrieval system using NLP techniques. The researchers employed a rule-based strategy to extract phrases and identify the relations between the terms in the phrases. These phrases were then used to create multilevel indexes, which were used to retrieve documents with inverted indexes that match the input query and belong to the same category. The experimental findings indicated that using NLP techniques to create an index set was more effective than not using them. In Sukhahuta Smith (Sukhahuta & Smith, [23]), an approach was proposed to enhance the presentation of Thai language documents by utilizing syntactic analysis. The approach involved analyzing documents and transforming them into a syntactic tree structure. The construction of the syntactic tree followed a specific procedure: first, the document was tokenized using predefined rules for a particular lexicon; next, each word was assigned a part-of-speech tag. Then, the context-free phrase structure grammar was applied to identify the syntactic surface structure, and finally, a syntactic tree was generated from the tagged words. The resulting tree was used to extract a list of concept definitions for each document. The evaluation of the proposed approach was based on the accuracy of the extracted concepts. The precision and recall were measured, and the average values obtained were 42% and 70%, respectively.

3 Methodology

This section describes the architecture of our system and its process for suggesting relevant documentsto users based on their queries. As depicted in Figure 1, the system is composed of four key modules: a conversational interface, a document processing module, a query processing module, and a ranking module. The document processing module initially preprocesses the documents and stores document representations in a database for efficient retrieval. When a user submits a query via the conversational interface, the query processing module steps in to extract the keywords from the query, subsequently utilizing these keywords to represent the query. Following this, the ranking module evaluates the documents based on their relevance to the user's query. The document earning the highest relevance score is selected and recommended as the most suitable response to the user's query. It's important to note that the document processing, query processing, and ranking modules were implemented on a cloud service, ensuring scalability and accessibility. Additionally, an API gateway is employed to connect the system to the conversational interface, allowing for seamless interaction between the user and the system. The subsequent subsections provide a detailed breakdown of eachcomponent and its role in the overall operation of the system.

3.1 The Conversational Interface

The first component of the system is a conversational interface acting as a chatbot to interact with users. In this work, the LINE messenger app was chosen as the conversational interface because it is Thailand's most-used messaging app (NAVER Corporation, [15]). This choice ensures that the system is easily accessible to Thai users, leveraging the popularity and familiarity of the LINE app in the country.

The conversational interface operates within the LINE messenger app and is showcased in Figure 2: (Left) Conversational interface showing query response; (Right) Document view when opened. When a user submits a query, such as "ขอข้อมูลเกี่ยวกับการนำเข้าสินค้าในเมียนมาร์หน่อยค่ะ" (Please

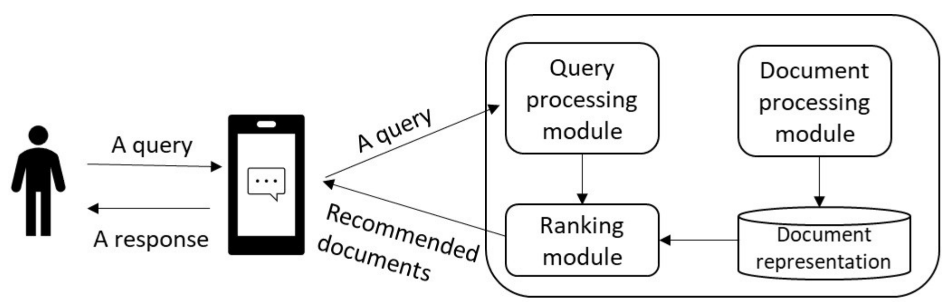


Figure 1: The architecture of the system

provide information about importing goods into Myanmar), the system initiates the document retrieval process. It then identifies and delivers the top three most relevant documents directly to the user through the LINE messenger interface, in the form of summaries and direct links to the full documents, as shown in Figure 2 (Left). Users can then click on these links to read the full documents, as shown in Figure 2 (Right). This approach ensures a quick and streamlined response to the user's information needs.

3.2 Document Processing Module

This study leverages guidebooks on foreign investment and trade as a rich source of information to answerentrepreneurs' queries. These guidebooks, however, encompass various topics, some of which may not be relevant to a particular query. To mitigate this issue, we divided the guidebooks into smaller, more manageable sections based on their headings. This approach ensures that each returned document contains focused and concise information. In our methodology, headings are treated as summaries of their respective sections and are processed as follows. Initially, we analyzed the guidebooks, which were downloaded in PDF format. We employed regular expressions to identify the start and end points of the table of contents, from which we extracted the headings and their corresponding page numbers. Each heading's content was then isolated into individual PDF files by extracting the specific pages under each heading. The headings were further utilized as file names to facilitate document retrieval. Subsequently, the system processes the headings, tokenizes sentences into words and extracting keywords.

In the keyword extraction process, our system leverages the ORCHID Part-of-Speech (POS) tagging system, a specialized resource designed for processing the Thai language. The ORCHID POS tagging system classifies individual words into their corresponding grammatical categories such as nouns, verbs, and adjectives, and annotates these tags to the words. This annotation allows for a more accurate and contextual interpretation of Thai natural language queries. The tagged and annotated words are essential for identifying relevant keywords with greater precision. However, not all tags are considered for keyword extraction. We focus on a select set of

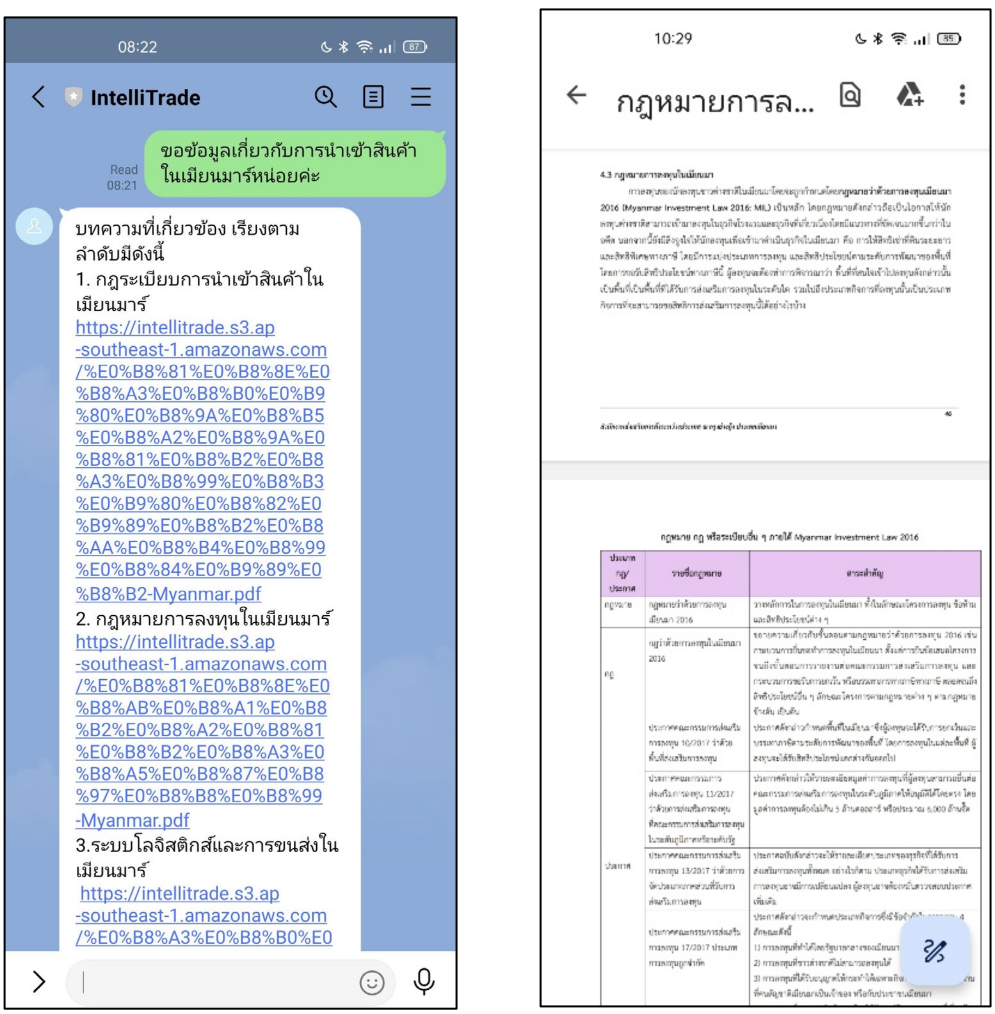


Figure 2: (Left) Conversational interface showing query response; (Right) Document view when opened

tags that are most likely to include meaningful keywords. The selected tags are as follows:

- CMTR: Measurement classifier - This tag is used for words that classify measurements, a commonfeature in the Thai language that often carries significant semantic weight.
- NPRP: Proper noun - Proper nouns often include crucial information such as names of places, organizations, or individuals.
- NCMN: Common noun - Common nouns are often key to understanding the main topics covered in a document
- NTTL: Title noun - Title nouns, which are used to denote titles or positions, can provide valuable context.
- VACT: Active verb - Active verbs describe actions and are often crucial for understanding the activities discussed in a document.
- VSTA: Stative verb - Stative verbs express states or conditions and can contribute valuable information about the circumstances or characteristics described in a document.

These specific tags are selected based on their likelihood to contain meaningful and relevant information. Once the keywords are extracted, they are then mapped to word vectors using the Large Thai Word2Vec model (Phatthiyaphaibun, [18]). This pre-trained deep learning model, specifically designed for the Thai language, employs the Word2Vec architecture. It has been trained on an extensive corpus of Thai text, capturing the semantic nuances of Thai words and phrases. The model creates a vector representation of each word by analyzing its contextual relationship with surrounding words in the corpus. To create a vector representation for each document, the system calculates the mean of all the keyword vectors. This strategy provides a condensed yet comprehensive representation of each document, which effectively accounts for the complexities and nuances of the Thai language.

To illustrate, Table 1 shows the original headings and their corresponding keywords in Thai, along with English translations in parentheses.

Table 1: The original headings and corresponding keywords

Heading	Keywords
ลักษณะภูมิประเทศ (Characteristic of terrain)	['ลักษณะ', 'ภูมิประเทศ'] [‘characteristic’, ‘terrain’]
สภาพภูมิอากาศ (Characteristic of climate)	['ลักษณะ', 'ภูมิอากาศ'] [‘characteristic’, ‘climate[
วัฒนธรรมและมารยาททางธุรกิจ (Business culture and etiquette)	['วัฒนธรรม', 'มารยาท', 'ธุรกิจ'] [‘culture’, ‘etiquette’, ‘business’]

Moreover, to represent each document more comprehensively, our approach incorporates the content of each document alongside the extracted heading keywords. This involves tokenizing the content into sentences and subsequently into words. Given that the conversion from PDF to text files can sometimes result in incorrect words, especially in Thai due to its complex script and lack of spaces between words, our system employs Peter Norvig’s spelling correction algorithm (Norvig, [17]) to maintain the accuracy of the content. Next, stop words are removed, an essential step in reducing noise and focusing on meaningful words, especially important given the high context nature of the Thai language. The resulting tokens, along with the extracted keywords, are then used to represent the document.

In summary, the document representation includes two vectors: one containing the heading keywords and the other containing the content tokens. The system preprocesses and stores heading keywords and content tokens in a database, which are then retrieved for ranking when a query is received. This enables efficient and quick retrieval of relevant documents based on the input query.

3.3 Query Processing Module

Processing user queries indeed presents a greater complexity compared to handling document headings and content. This complexity primarily arises due to the conversational nature of queries, which can often include stop words, misspelled words, or unnecessary characters. Furthermore, nuances inherent to the Thai language—such as its lack of space between words, multiple valid spellings for the same word, and the use of tone marks—add layers of difficulty to the processing of user queries. In response to these challenges, our system executes several steps to process user queries effectively. The first step is tokenization, which splits the user’s query into individual words. Tokenization is particularly challenging in Thai due to the absence of spaces between words, which is a standard feature in many other languages. To address this, our system employs advanced tokenization techniques that are specifically designed for the Thai language.

The second step involves removing stop words and question marks from the queries. Stop words, which are commonly used words such as ’and’, ’is’, ’at’, etc., typically provide little value for document retrieval and can unnecessarily complicate the process. Similarly, question marks are also eliminated as they do not contribute to identifying relevant documents.

Next, we apply normalization methods to the queries. These methods are crucial in dealing with the Thai language’s characteristics, such as multiple valid spellings for the same word and the use of tone marks. Our normalization process eliminates duplicate spaces, removes dangling characters, and consolidates repeating vowels. These actions help standardize the queries and make them more amenable to the subsequent document retrieval process.

Finally, after these pre-processing steps, keywords are extracted from the user queries. These keywords serve as the cornerstone of the document retrieval process, helping to identify and rank the most relevant documents in response to the user's query. Overall, despite the complexities presented by user queries and the Thai language, our system is designed to effectively process and handle these challenges, thereby ensuring accurate and efficient document retrieval. The final representation of each user query is a set of keywords derived from the original query. To provide a clear illustration of this process, Table 2 presents examples of original user queries in Thai and their corresponding extracted keywords, with English translations provided in parentheses.

3.4 Ranking Module

The ranking module in this system is responsible for ranking the retrieved documents based on their relevance to the user's query, considering both the document heading and content for ranking purposes. To rank documents based on their headings, each keyword is first mapped to a pre-trained word embedding vector using a deep learning model. This enables the model to capture the semantic and contextual meaning of each keyword in the document representation. The final word embedding vector is then obtained by calculating the mean of all the keyword vectors. This approach allows the model to capture the overall meaning and context of the document's heading, providing a more comprehensive representation for accurate ranking. Using the same approach, the user query is also converted into a word embedding vector. For each document, the system calculates the cosine similarity score between the word embedding vectors of its heading and the user's query. These scores are then used to create a heading ranking array, which ranks the documents based on their similarity to the query.

For content ranking, the BM25 algorithm is employed, taking into account the frequency and importance of query terms in the document to rank the documents based on their relevance to the user's query. The BM25 algorithm considers both term frequency (TF) and inverse document frequency (IDF) to calculate the relevance score of a document to a query. The TF component measures the frequency of the query terms within a document, while the IDF component measures the rarity of the query terms in the document collection. By combining these two components, the BM25 algorithm can capture both the specificity of the query and the importance of the terms in the document.

The rankings obtained from both approaches are combined using the Borda count method. This method produces a final rank by assigning weights to the ranks of each document in both arrays. The document with the highest final rank is recommended as the most relevant document to the user's query. The Borda count method is used because it provides a more robust and accurate ranking by taking into account the rankings of all documents, rather than just focusing on the top-ranked documents. It is also known for its fairness, simplicity, and robustness, making it a reliable and widely used approach for rank aggregation in various applications (Dwork, Kumar, Naor, & Sivakumar,[2]; Renda & Straccia, [19]).

Algorithm 1 outlines the procedure for ranking and retrieving documents that are most relevant to a user's query. The algorithm starts by extracting a set of keywords, denoted as K_{query} from the user's query. For each document d in the database, the algorithm retrieves a set of heading keywords $K_{document}$ and a set of content tokens $C_{document}$ that have been preprocessed and stored.

Two metrics are calculated for ranking:

1. The BM25 score between K_{query} and $K_{document}$, quantifies the relevance of the document's headings to the query.
2. The cosine similarity score between K_{query} and $C_{document}$, which measures the angular distance between the keyword vectors, thereby indicating content relevance.

The Borda count method is then used to aggregate these scores into a final ranking score for each document. Once all the documents have been scored, they are sorted in descending order based on their final ranking scores. The top-ranked document is then returned to the user, completing the retrieval process.

Table 2: The original queries and corresponding keywords

Query	Keywords
ขอข้อมูลนิคมอุตสาหกรรมในเมียนมาร์หน่อยค่ะ (May I get information on industrial estates in Myanmar?)	[‘นิคมอุตสาหกรรม’, ‘ข้อมูล’, ‘เมียนมาร์’] ['industrial estates', 'information', 'Myanmar']
ประชากรในเมียนมาร์มีเท่าไรคะ (What is the population in Myanmar?)	[‘ประชากร’, ‘เมียนมาร์’] [‘population’, ‘Myanmar’]
เมืองหลวงของกัมพูชาชื่ออะไรเหรอ (What is the name of the capital city of Cambodia?)	[‘เมืองหลวง’, ‘ชื่อ’, ‘กัมพูชา’] ['capital city', 'name', 'Cambodia']

Algorithm 1: The proposed ranking algorithm

Require: User query q
Ensure: The top-ranked document
Steps:
Extract the set of keywords K_{query} from q
for each document d in the database **do**
  Retrieve the set of heading keywords K_{document}
  Retrieve the set of content tokens C_{document}
  Calculate the BM25 score between K_{query} and K_{document}
  Calculate the cosine similarity score between K_{query} and C_{document}
  Use the Borda count method to aggregate the BM25 score and cosine similarity scores arrays andcalculate the final ranking score
end for
Sort the documents in descending order based on their final ranking scores
return The top-ranked document to the user

4 Results and Discussions

This section presents an evaluation of the proposed system's performance, including a comparison of different ranking approaches and an assessment of the system's effectiveness based on user survey feedback.

4.1 Performance Evaluation

This section embarks on a comprehensive evaluation and comparison of the proposed system's performance. Utilizing a test dataset of 406 matched query-document pairs, carefully constructed by Thai entrepreneurs and senior students in international business management, the system's effectiveness is thoroughly assessed. The documents, sourced from the Department of International Trade Promotion website, pertain to foreign trade and investment in four countries: Vietnam, Laos, Cambodia, and Myanmar. This collection of high-quality, up-to-date guidebooks is aimed at enhancing the competitiveness of Thai entrepreneurs in global markets

To validate the effectiveness of our proposed system, we undertook a detailed comparative analysis against two distinct ranking methods. The first of these methods based its ranking solely on documentcontent, employing the well-established BM25 method. Conversely, the second method leveraged only document headings, using the Word2Vec algorithm to calculate cosine similarity between query keywords and document headings.

The performance was evaluated with two primary metrics - accuracy rate and processing time in seconds. These measurements are outlined in Table 3. Through this analysis, our proposed methodology demonstrated superior performance, outstripping the other approaches in terms of accuracy. Our system attained an impressive accuracy rate of 65.76%. In comparison, the BM25 method, which strictly considers document content, yielded an accuracy of 57.88%. The Word2Vec-based method, focusing solely on document headings, achieved an accuracy of 62.81%. This clear advantage validates the effectiveness of our system's comprehensive approach, which blends the consideration of document content and headings to deliver more precise and relevant document suggestions.

However, it is worth noting that the proposed method took longer to execute, with an elapsed time of 0.034 seconds compared to 0.010 seconds for the BM25 method and 0.028 seconds for the Word2Vec-based method. Therefore, there is a trade-off between accuracy and speed. In real-world applications, the choice of ranking method will depend on the specific needs of the user. If accuracy is the top priority, then the proposed method may be the best choice, despite its longer execution time. It is also worth noting that the speed of the proposed method can be improved by adding more computing resources. This would allow for faster execution times without sacrificing accuracy. However, if speed is more important than accuracy, then the BM25 method or the Word2Vec-based method may be a better option, as they are faster but less accurate.

Table 3: Performance comparison of ranking approaches

Approach	Accuracy rate (%)	Computational time (seconds)
BM25	57.8818	0.0105
Word2Vec-based	62.8079	0.0279
Proposed method	65.7635	0.0344

4.2 User Experience Evaluation

We utilized a user survey to ascertain the system's effectiveness from the user's perspective. Initially, participants were asked to submit several queries, with the system configured to suggest the top three most relevant documents in response to each query. Subsequently, participants were asked to respond to two key questions:

- On a scale of 1 - 5, with 1 signifying ’highly irrelevant’ and 5 indicating ’highly relevant’, how would you gauge the relevance of the documents suggested by the system in correspondence to your query?
- On a scale of 1 - 5, with 1 representing ‘not useful at all’ and 5 signifying’extremely useful’, how would you appraise the overall utility of the system in aiding with your query?

The intent behind the first question is to evaluate the user-perceived relevance of the suggested documents concerning their specific queries. Each participant was requested to rate the overall relevance of the recommended documents, with each rating corresponding to each query. The second question aims to measure the overall practical usefulness of the system from the users' perspective. This question is assessed on a per-user basis, providing an aggregate view of the system's utility across multiple queries made by each user. The user satisfaction survey involved 15 participants, composed of 10 entrepreneurs and five senior students specializing in international business management. Each participant submitted 3-4 queries, resulting in a total of 50 queries assessed by the system. The survey responses indicated a moderate level of satisfaction with the system's performance. The average rating for the first question, concerning the relevance of the documents suggested by the system, was 3.18 out of 5. The second question, evaluating the overall usefulness of the system, received a slightly higher average rating of 3.70 out of 5.

The survey results suggest that users generally found the system beneficial and moderately effective in assisting with their queries. An average rating of 3.70 out of 5 for the second question indicates that users see value in the system and deem it fairly useful in their query process. However, the average rating of 3.18 for the first question suggests that the system could potentially improve its document suggestion relevance. While this score is above average, it indicates some room for improvement. Users might have received some document recommendations that were not as relevant as they expected, or the relevance of the documents could have varied across different queries.

For example, one user queried "ขอข้อมูลเกี่ยวกับการนำเข้าสินค้าในเมียนมาร์หน่อยค่ะ" (Please provide information about importing goods into Myanmar). The relevant documents returned were as follows:

- กฎระเบียบการนำเข้าสินค้าในเมียนมาร์ (The regulations for importing goods in Myanmar)
- กฎหมายการลงทุนในเมียนมาร์ (Investment Laws in Myanmar)
- ระบบโลจิสติกส์และการขนส่งในเมียนมาร์ (Logistics and Transportation System in Myanmar)

The first result, "The regulations for importing goods in Myanmar," is the most relevant, directly addressing the specifics of import regulations. The second result, "Investment Laws in Myanmar," could provide a broader context about the legal environment for business and trade in Myanmar, which would also be useful for an entrepreneur considering importing goods into the country. The final result, "Logistics and Transportation System in Myanmar," could provide practical insights into the logistical aspects of importing goods, making it a valuable resource for understanding the transportation infrastructure and potential challenges.

While the search results for the query about importing goods into Myanmar are generally accurate and relevant, user satisfaction may vary depending on their specific information needs. The first result, "The regulations for importing goods in Myanmar," directly addresses the query, providing the most pertinent information about import regulations. However, the second and third results, though related to the broader context of conducting business in Myanmar, may not align precisely with the user's immediate information requirement about importing goods. Therefore, some users might feel these documents are not exactly what they were looking for, emphasizing the need for continual refinement in document suggestion relevance.

In conclusion, the results are promising, but there is room for improvement, particularly in enhancing the relevancy of the suggested documents. It might be worth considering refining the natural language processing techniques or information retrieval methods used in the system. Potential areas of focus could include improving keyword extraction, the ranking algorithm, or the approach to understanding the user's queries more accurately.

4.3 Discussion

Our proposed method's superior accuracy rate, surpassing both the BM25 and Word2Vec-based methods, can be attributed to its comprehensive approach that combines both content and heading-based strategies. By leveraging both these elements, our system is equipped to provide a more holistic understanding of the document's context, leading to more accurate document ranking. Additionally, our approach includes the novel feature of identifying country names in the queries, further refining search results by providing location-specific responses. This is a significant factor contributing to the enhanced accuracy of our method. However, this increased accuracy comes at the cost of execution speed. The comprehensive nature of our approach, integrating multiple techniques and additional features, inherently requires more computational time compared to simpler methods. This trade-off between accuracy and speed is a common challenge in information retrieval and natural language processing tasks.

In contrast, the BM25 method, which solely focuses on document content, performed less accurately but demonstrated a faster execution time. The BM25 algorithm, being a probabilistic approach, relies heavily on term frequency and inverse document frequency. While it is efficient, the method might overlook nuanced contextual elements present in the document headings, which could explain its lower accuracy rate. Similarly, the Word2Vec-based method, which only considers document headings, also exhibits a lower accuracy rate. Despite using a sophisticated technique like Word2Vec, which understands semantic relationships between words, limiting the analysis to only headings might result in missing vital information from the main content of the document.

4.4 Limitations

While our system has shown effectiveness in assisting Thai entrepreneurs, there are some limitations. The system is primarily designed for Thai natural language queries, limiting its applicability to a more diverse, multilingual audience. Additionally, the system relies heavily on guidebooks for its source material. These guidebooks cover a wide array of topics that may not all be relevant to a specific query. To address this, we segmented the guidebooks into smaller, topic-specific sections based on their headings. However, this segmentation itself could be a limitation, as it assumes that the headings accurately summarize the content, potentially leading to information being overlooked. Lastly, while the guidebooks are officially endorsed by Thai authorities, the system's reliance on them means that the timeliness of the information is contingent on these pre-existing resources. As such, regular updates are essential to maintain the system's relevance and effectiveness.

5 Conclusions

This paper presents a methodology that utilizes a ranking approach to provide a more accurate and relevant ranking of retrieved documents by considering both the document headings and content. The system utilizes a deep learning model to measure the semantic similarity between a user's query and a document's heading. It also leverages the BM25 algorithm to rank documents based on their term frequencies in the content. To retrieve the best-matched document, the system first preprocesses the documents and precomputes keyword vectors of all documents for later similarity calculation. Upon receiving a query, the system extracts a set of keywords from the query and compares it to heading keywords and content tokens of documents in the database. The system calculates the BM25 and cosine similarity scores, which are then aggregated using the Borda count method to produce a final ranking score for each document. The top-ranked document is returned as the most relevant document to the user.

The system was evaluated using a manually constructed dataset containing 406 queries. According to the experiment result, the proposed method achieved an accuracy rate of 65.76% with an elapsed time of 0.034 seconds. In comparison, the BM25 method achieved an accuracy rate of 57.88% with an elapsed time of 0.010 seconds, and the Word2Vec-based method achieved an accuracy rate of 62.81% with an elapsed time of 0.028 seconds. The proposed method outperformed both the BM25 method and the Word2Vec-based method in terms of accuracy rate.

The user satisfaction survey, involving 15 participants, provided insightful feedback on our system. Each participant submitted 3-4 queries, leading to a total of 50 queries assessed. The results suggested a moderate satisfaction level, with an average rating of 3.18 out of 5 for document relevance and a slightly higher average of 3.70 out of 5 for overall system usefulness. These ratings indicate that users found value in the system, deeming it beneficial for their query process. However, the feedback also highlights room for improvement, specifically in enhancing the relevance of suggested documents.

There are opportunities to enhance the matching module in the future to further improve the accuracy of the retrieved documents. Another potential area of improvement is to further enhance the document processing module to better capture the important information within the document content and headings. Additionally, exploring the use of more advanced deep learning models and incorporating automatic summarization techniques could also improve the accuracy of our approach.

References

[1] G. Amati and C. J. Van Rijsbergen, "Probabilistic Models of Information Retrieval based on Measuring the Divergence from Randomness, *ACM Transactions on Information Systems (TOIS)*, *20*(4):357-389, 2002.

[2] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, " Rank Aggregation Methods for the Web," *Proceedings of the 10th International Conference on World Wide Web*, pp. 613-622, 2001.

[3] L. Galke, A. Saleh, and A. Scherp, "Word Embeddings for Practical Information Retrieval," *Informatik 2017*, pp. 2155-2167, 2017.

[4] S. Guo, F. Alamudun, and T. Hammond, "Résumatcher: A Personalized Résumé-Job Matching System," *Expert Systems with Applications*, *60*:169-182, 2016.

[5] D. Hiemstra, *Using Language Models for Information Retrieval*, Citeseer, 2001.

[6] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data," *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 2333-2338, 2013.

[7] A. Kawtrakul, F. Andres, K. Ono, C. Thumkanon, T. Thanyasiri, N. Khantonthong, N. Buranapraphanont, "The Implementation of VLSHDS Project for Thai Document Retrieval," *Proc. First International Symposium on Advance Informatics,* Tokyo, Japan, 2000.

[8] G. Lan, Y. Ge, and J. Kong, "Research on Scoring Mechanism based on BM25f Model," *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (itaic)*, pp. 1782-1786, 2019.

[9] P. Limkonchotiwat, W. Phatthiyaphaibun, R. Sarwar, E. Chuangsuwanich, and S. Nutanong, "Handling Cross and Out-of-Domain Samples in Thai Word Segmentation," 2021.

[10] Z. Lu and H. Li, "A Deep Architecture for Matching Short Texts," *Advances in Neural Information Processing Systems*, 26:1367-1375, 2013.

[11] Bhaskar Mitra and Nick Craswell, "An Introduction to Neural Information Retrieval," *Foundations and Trends® in Information Retrieval*, 13(1):1-126, 2018.

[12] B. Mitra, F. Diaz, and N. Craswell, " Learning to Match using Local and Distributed Representations of Text for Web Search" *Proceedings of the 26th International Conference on world wide web,* pp. 1291-1299, 2017.

[13] B. Mitra, E. Nalisnick, N. Craswell, and R. Caruana, "A Dual Embedding Space Model for Document Ranking," *arXiv preprint arXiv:1602.01137*, 2016.

[14] S. Naseri, J. Dalton, A. Yates, and J. Allan, "Ceqe: Contextualized Embeddings for Query Expansion," *Advances in Information Retrieval: 43rd European*

Conference on IR Research, ECIR 2021, Virtual Event, March 28-April 1, 2021, Proceedings, Part i 43, 43:467-482, 2021.
[15] NAVER Corporation, *2020 NAVER Annual Report,* Retrieved from https://www.navercorp.com/en/investment/annualReport, 2020.
[16] T. Noraset, L. Lowphansirikul, and S. Tuarob, "Wabiqa: A Wikipedia-Based Thai Question-AnsweringSystem," *Information Processing and Management*, 58(1):102431, 2021.
[17] P. Norvig, *How to Write a Spelling Corrector,* http://norvig.com/spell-correct.html, 2016.
[18] W. Phatthiyaphaibun, *Ltw2v: The Large Thai Word2vec (1.0),* Retrieved from https://doi.org/10.5281/zenodo.7280277, 2022.
[19] M. E. Renda, and U. Straccia, "Web Metasearch: Rank vs. Score based Rank Aggregation Methods," *Proceedings of the 2003 ACM Symposium on Applied Computing,* pp. 841-846, 2003.
[20] Stephen E. Robertson and Hugo Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends® in Information Retrieval*, *3*(4):333-389, 2009.
[21] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 Extension to Multiple Weighted Fields," *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management,* pp. 42-49, 2004.
[22] S. Soni and K. Roberts, "Patient Cohort Retrieval using Transformer Language Models," *Amia Annual Symposium Proceedings*, 2020:1150, 2020.
[23] R. Sukhahuta and D. Smith, "Information Extraction for Thai Documents," *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, pp. 103-110, 2000.
[24] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF Term Weights as Making Relevance Decisions," *ACM Transactions on Information Systems (TOIS)*, 26(3):1-37, 2008.
[25] W. Yang, H. Zhang, and J. Lin, "Simple Applications of Bert for Ad Hoc Document Retrieval, *arXiv preprint arXiv:1903.10972,* 2019.
[26] H. Zaragoza, N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson, "Microsoft Cambridge at Trec 13: Web and Hard Tracks," *Trec*, 4:1-1, 2004.
[27] C. Zhai and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval," *ACM Sigir Forum*, 51:268-276, 2017.

Sirinda Palahan is a faculty member in the Department of ICT at the School of Science and Technology, University of the Thai Chamber of Commerce, Thailand. With a Ph.D.in Computer Science and Engineering from Penn State University, USA. Dr. Sirinda Palahan possesses an academic background in the field. Her research interests primarily revolve around data analytics and natural language processing.

Exploring Summarization of Scientific Tables: Analysing Data Preparation and Extractive to Abstractive Summary Generation

Monalisa Dey*
Jadavpur University, Kolkata, INDIA
Sainik Kumar Mahata†
Institute of Engineering and Management, Kolkata, INDIA
Dipankar Das‡
Jadavpur University, Kolkata, INDIA

Abstract

This research paper focuses on the challenges involved in extracting relevant information from the vast amount of textual data available in digital resources. Specifically, we address the complexities associated with tables in scientific papers, which contain crucial data but can be difficult to understand and summarize. In fact, the data required for training such systems is scarce. To overcome this challenge, we propose the development of a high-quality corpus consisting of summaries, both extractive and abstractive, from table content. To develop such a corpus containing pertinent extractive-abstractive summary pair, we used two approaches, namely, rule-based and template-based. The developed dataset was then validated using various automated and manual metrics. Subsequently, two models, viz., T5 and Seq-to-Seq, were trained on this dataset, to generate abstractive summaries from extractive ones. By using extractive summaries as a starting point, the system can produce new sentences that capture the essence of the tables rather than all the specifics. We have used standard evaluation measures like BLEU, ROUGE-L, Adequacy, and Fluency to evaluate both systems. BLEU and ROUGE-L scores of 58.2 and 0.31 for fine-tuned T5 model and 16.25 and 0.18 for the Seq-to-Seq model suggest that our developed dataset can produce coherent abstractive summaries.

Key Words: Summarization; extractive; abstractive; rule-based; evaluation metrics; template-based; T5; Seq-to-Seq

1 Introduction

In this digital age, cloud resources like websites, social media platforms, news outlets, and communication channels, are seeing an unprecedented increase in information. A huge amount of textual data, generated from different places like academic books, research papers, legal proceedings, health sector, etc., presents a huge challenge when you need to have access to relevant data [1]. Due to this, the development of automatic techniques for extraction as well as summarization of textual resources has assumed significant importance in our lives. Automatic text summarization can be categorized into two types; extractive and abstractive. The extractive approach consists of selecting and extracting crucial sentences/paragraphs from the original document, to create a coherent summary. In contrast, abstractive summarization aims to develop a summary that incorporates novel lexemes and phrases.

In our work, we primarily concentrate on scientific papers. Multiple research papers and articles inundating our lives pose difficulties in swiftly and effectively extracting pertinent information from non-textual elements such as tables, graphs, figures, and flowcharts [2]. Tables, an integral part of scientific papers, provide an efficient medium to present intricate information concisely and in an organized manner. Nevertheless, tables present a challenge to conventional retrieval methods due to the fusion of content and presentation they embody. Therefore, the provision of a succinct summary of table data assumes a paramount role. This concise summary grants researchers the ability to comprehend the information without undertaking the task of reading the entire paper, thereby saving valuable time and effort. By offering a bird's-eye view of the data, users can swiftly grasp the essence of the table content without delving into complexities, empowering them to go through more results.

Table content summarization systems are therefore needed to extract table summaries effectively. However, the development of such systems faces a major challenge: the lack of appropriate corpora for training and evaluating summarization algorithms. A dataset suitable for summarizing patient information may prove futile when attempting to summarize weather reports or cricket scores.

To address the challenge posed by the scarcity of suitable training and testing corpus for scientific table summarization systems, we decided to develop a comprehensive corpus containing table-content summaries. Our study aims to generate two distinct types of summaries, extractive and abstractive. Our first contribution is constructing the table-content summary corpus. To do this we employed multiple techniques, encompassing information extraction, and extensive validation. Our approach entailed the creation of two types of summaries: extractive and abstractive. In extractive summaries, we carefully selected pertinent phrases from the text of the paper that conveyed information about the table,

*Email: monalisa.dey.21@gmail.com
†Email: sainik.mahata@gmail.com
‡Email: dipankar.dipnil2005@gmail.com

while in abstractive summaries, we extracted the accompanying captions for each table. These methodologies collectively furnished us with a robust and comprehensive corpus of table-content summaries, tailor-made for training and testing table summarization systems.

Our next contribution is to develop methods for selecting pertinent extractive-abstractive summary pair. To accomplish this, we used two approaches, namely, rule-based and template-based. The most crucial task was to recognize and separate the captions from the body content for each table in the rule-based strategy. Again, correctly extracting table mentions and citations from the body text was also very challenging. To overcome these obstacles, methods that could precisely parse the document and extract the relevant data had to be developed.

In the template-based approach, we also developed template summaries using several techniques like TF-IDF and Transition based approaches. These techniques identify a set of terms or a template, that is significant in representing the summary of that table. Both automatic and manual evaluation techniques are utilized for evaluating the quality of this dataset.

Finally, we compared both methods and chose the better-performing method for our next tasks. We have used the most famous ROUGE [16] and BLEU [25] tools to evaluate our results.

Subsequently, after solving every challenge, we successfully developed the corpus, which will serve as an important resource for any researcher interested in table summarization.

The next contribution in this paper is the development of an extractive-to-abstractive summary generation system, which would take as input the chosen extractive summary for each table and would generate an abstract representation of the same. For this, we employed two models; a transformer-based T5 model [1] that was fine-tuned using our own data as well as a seq-to-seq model. Figure 1 shows the flow diagram for the whole process.

The rest of the paper is structured as follows. The most recent developments in this area of research are covered in Section 2. The procedure for developing the dataset is described in Section 3. The methods and models for extractive to abstractive summary creation are described in Section 4, and the conclusion is presented in Section 5.

2 Related Work

The effort to summarize a huge body of information has given rise to two main techniques: extractive and abstractive summarization. Relevant sentences are carefully selected from the source document and presented together in extractive summarization. Abstractive methods, however, follow a different route. Here, new phrases are developed. This section includes a review of literature on extractive to abstractive summarizing, as well as table-based summarization that highlights significant contributions to the field of study.

[1]https://huggingface.co/docs/transformers/model_doc/t5

Nallapati et al. [24] introduced an abstractive text summarization approach using seq-to-seq RNNs and explored techniques to improve the quality of generated summaries. Gehrmann et al. [8] proposed a bottom-up abstractive summarization approach that incrementally constructs a summary by predicting key content selection and generating natural language phrases. Paulus et al. [27] presented a deep reinforced model for abstractive summarization that incorporates a reinforcement learning framework to train the model to generate high-quality summaries.

Pasunuru et al. [26] focused on generating video captions but leveraged entailment rewards to reinforce abstractive summarization, improving the quality and informativeness of the generated captions.

Zhou et al. [34] outlines a neural document summary system that efficiently combines extractive and abstractive approaches to provide clear and detailed summaries. The algorithm jointly learns to score and pick phrases. See et al. [28] in their paper, explored a network-based pointer generator for summarization, which combines extractive and abstractive methods to generate summaries by copying content from the original data.

Liu et al. [19] studied how using BERT, a pre-trained encoder, for text summarization, achieves improved performance by leveraging the rich contextual representations. Liu et al. [18] introduced PreSumm, a neural model for abstractive text summarization that incorporates transformer-based architectures and achieves state-of-the-art performance on various summarization datasets. Ma et al. [21] explored the integration of external knowledge sources into pre-trained transformers for abstractive summarization, aiming to enhance the summary generation process with additional information.

Li et al. [15] proposed DRGAT, a dual-reading graph attention network, which leverages graph attention mechanisms to capture global and local dependencies for abstractive summarization.

Table-based summarization has also been addressed in many research papers. Arvind et al. [3] proposes a structure-aware sequence-to-sequence learning model for generating natural language text from tables. Lebret et al. [13] presents a neural text generation model that converts structured data, including tables, into natural language text, with a specific focus on generating biographical summaries.

Li et al. [14] proposes structured attention networks for table-to-text generation, which effectively capture dependencies between table elements and generate coherent and informative summaries. Krishrnamurty et al. [12] presents a neural semantic parsing model for semi-structured tables, enabling the conversion of table content into a structured representation that can be used for generating natural language summaries.

Dong et al. [5] provides an overview of neural text generation techniques in structured data-to-text applications, including table-based summarization, discussing challenges and potential solutions. Agarwal et al. [9] proposes an effective hierarchical encoder that leverages structural information in tables for table-to-text generation, improving the quality and coherence of

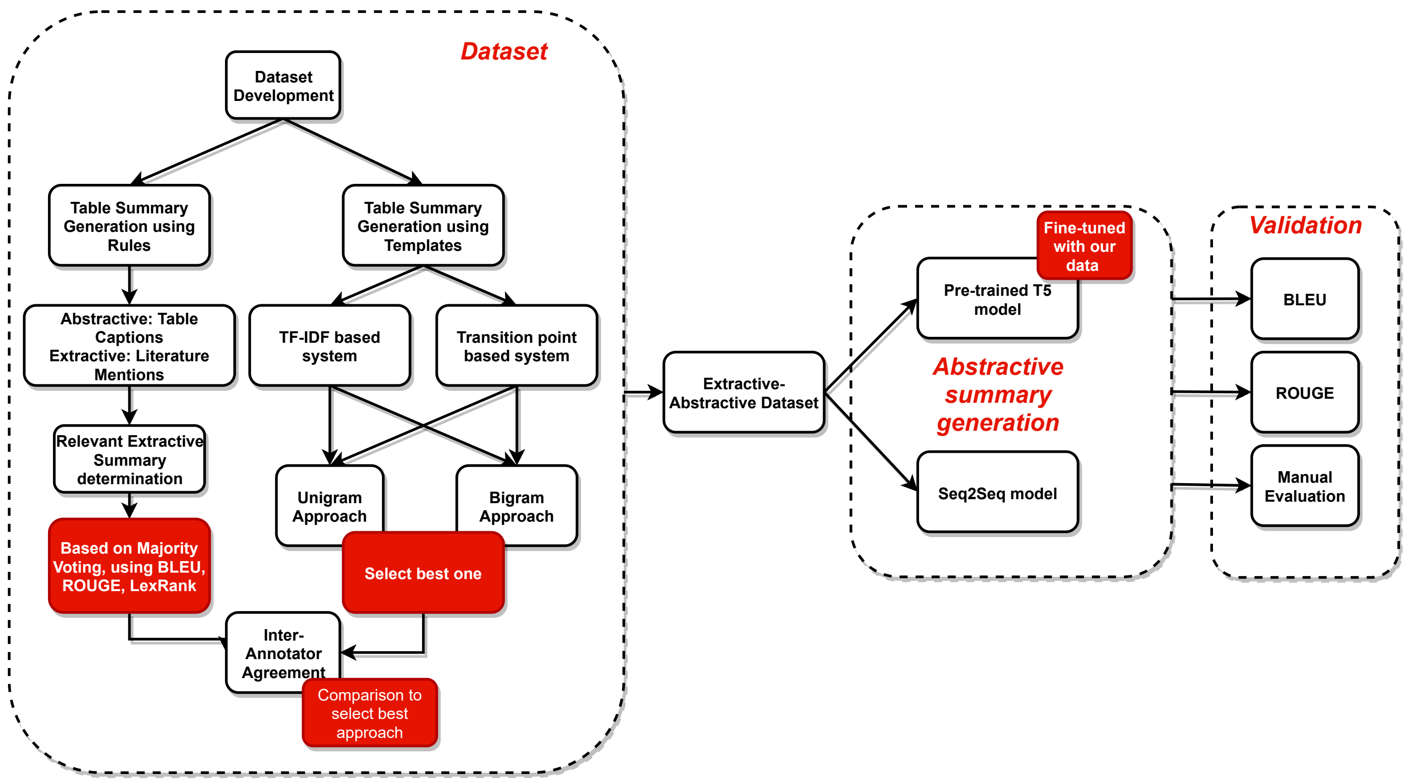


Figure 1: Flow of the data extraction, summary generation and testing process

generated summaries.

Gao et al. [7] introduces Table2Seq, a neural sequence generation model for table-to-text tasks, employing a novel column-aware attention mechanism to capture important table structure information. Lin et al. [17] presents hybrid pointer generator networks for table-to-text generation, incorporating both extractive and abstractive methods to generate coherent and accurate summaries. Cui et al. [4] proposes a descriptive sentence generation model for table structure, enabling the generation of natural language descriptions that capture the essence of tabular data.

Ma et al. [22] provides a structure-aware convolutional seq-to-seq network to efficiently capture the hierarchical structure of tables to produce consistent and useful summaries. Zhang et al. [32] introduces a table-based neural text generation model that incorporates semantic constraints to guide the generation process and improve the coherence and fidelity of generated summaries. Jiang et al. [11] proposes a structure-aware transformer model for table-to-text generation, effectively capturing table structure information to improve the quality and coherence of generated summaries.

Zheng et al. [33] proposes a multimodal framework for table-to-text generation, which leverages both extractive and abstractive methods to generate summaries, taking into account both table content and associated textual descriptions.

As discussed above a lot of work has been done in the abstractive summarization field. However, none of these works addresses the challenge of developing scientific table summarization systems. In our work, we have addressed the challenge of developing a corpus containing extractive and abstractive summaries of scientific tables. We have also proposed systems for dataset development as well as extractive to abstractive summary generation for tables.

3 Dataset Development

As mentioned the absence of a dataset made it necessary for us to first develop a dataset. To construct this corpus, we obtained scientific articles from digital libraries as they typically include tables that present important information about research findings. We gathered 1,500 papers across 20 distinct domains in computer science, such as Automatic Summary, Machine Learning, and Machine Translation. Each article has an average of approximately 250 sentences, not including titles, author names, and section headings. We then employed two approaches to prepare our dataset as described in the following sections.

Dataset Preprocessing: Preprocessing is an important step in any data analysis task, as it can improve the quality and usability of the data, and make it easier to extract insights and information from it. This involves cleaning the data, transforming it into a suitable format, and extracting the relevant information without any errors or inconsistencies. Next features were extracted from the baseline format and this information was organized in a way

Table 1: Dataset statistics

Paper Type	#Tables	Type: Text	Type: Numeric	ESummary		ASummary
				#Eavg	Elen_avg	Alen_avg
Automatic Summary	510	130	370	3	16	11
Machine Learning	700	373	327	4	18	12
Machine Translation	420	150	280	3	16	10
Named Entity Recognition	789	553	236	2	16	14
Question Answering	553	120	433	3	15	13
Sentiment Analysis	421	125	296	2	14	14
Speech Recognition	700	432	286	5	13	13
Text Classification	567	265	302	3	15	15
Text Segmentation	700	432	268	2	13	13
Word Sense Disambiguation	650	324	326	1	11	13
Total no. of papers	1,500					

that makes it easier to work with in the upcoming sections.

3.1 *Table Summary Generation Using Rules*

Caption Identification: To provide a clear representation of the data presented in a table, a well-written caption is crucial. Captions may vary depending on the domain and writing style, so we developed a method to distinguish caption sentences from other sentences in the document. We discovered that captions in various papers consist of four parts: ⟨TABLE⟩, which is the word "Table", followed by ⟨INTEGER⟩, an integer indicating the table number in the paper. The integer is then followed by a ⟨DELIMETER ⟩, which is a delimiter such as a period or a colon. Finally, we have ⟨TEXT⟩, which is a description of the content of the table. If a sentence follows this structure, we label it as a caption sentence and consider it as a summary of the table's contents. This approach enables us to handle the diversity of caption formats, resulting in coherent and informative summaries for each table.

Relevant Sentence Extraction: Although captions can effectively describe the contents of a table, studies have demonstrated that captions alone may not provide readers with a complete understanding of the information presented in a table. To address this limitation, we have observed that tables are referenced at least once in the corresponding scientific document. Therefore, we have developed a method to extract the reference text for a table to obtain a more comprehensive understanding of it.

The initial step was to segment the document text into sentences. To identify relevant sentences, we followed the same approach as for caption extraction, but we omitted the delimiter part. Additionally, we noted that sentences located near the reference sentence are useful in providing context for the table. We assigned scores to each sentence based on its proximity and distance to the reference sentence. If the distance was within a certain threshold length (+/-) 1, we considered the sentence important and included it in the summary.

By including relevant sentences in the summary, we can provide readers with a more complete understanding of the table and its context. This approach complements the captions and addresses the issue of insufficient information presented by captions alone.

As understood, a table can have multiple extractive summaries but only one abstractive summary. This process is described in Figure 2.

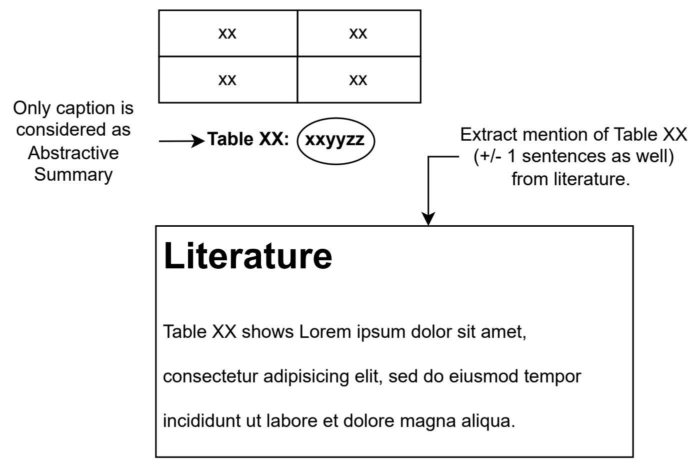


Figure 2: Process of caption identification and relevant sentence extraction

Annotation: After generating the abstractive and extractive summaries, we developed an annotated dataset to build a well-structured and easy-to-use corpus that can be automatically evaluated. We employed two separate approaches to evaluate the standard of our system-generated output as extractive and abstractive summaries which are discussed in the following subsections.

We also developed an output feature file using our system, which included additional features besides the summaries. These features included the paper ID, table ID, number of rows in a table, row attributes, column attributes, and the type of the table (numeric, text, or hybrid). By including these features, we aimed to provide more context about the tables and enable better analysis of the data.

Overall, the annotated dataset we constructed provides a comprehensive resource for researchers and practitioners in the field of computer science. The abstractive and extractive summaries and the additional features of the dataset make it an ideal resource for training and testing Natural Language Processing (NLP) models and for conducting further research in the field. Table 1 displays the characteristics of the corpus we developed.

Table 1 denotes the total number of tables in the dataset. Table types can be text or numeric. ESummary denotes extractive summary and ASummary denotes abstractive summary. $\#Eavg$ is the average number of extractive summaries present per table per paper in a particular paper type.

Whereas, $Elen\_avg$ is the average length (in words) of an extractive summary in a particular paper type. $Alen\_avg$ on the other hand denotes the average length (in words) of an abstractive summary in a particular paper type.

3.1.1 Relevant Extractive Summary Selection. Extractive Summary Selection (ESS) is the process of selecting the most relevant extractive summary. It is crucial as it ensures that the model is trained on the best possible data and is more likely to produce accurate and informative summaries. Since our main aim is to ensure that the best quality extractive summary is selected for further works, we have used standard quality assessment tools like ROUGE (ESS_R), BLEU (ESS_B) and LEXRANK (ESS_L) and used a majority voting technique between them for selecting the most relevant extractive summaries.

It must be remembered that in the upcoming sections, the abstractive and the extractive summaries are considered as the reference and generated summaries respectively.

ESS_B: The BLEU score evaluates the accuracy of translations or summaries produced by computers in comparison to one or more references produced by humans. For every table i, the BLEU score between the abstractive summary and the extractive summaries for table i is calculated.

ESS_R: The effectiveness of automated summarization is evaluated using the ROUGE method. It determines how comparable the produced summary and the reference summary are based on the overlap of n-grams and their respective frequencies. The difference between the abstractive summary and the relevant extractive summaries for each table i is measured by the ROUGE score.

ESS_L: LexRank [6] is a graph-based algorithm for ranking sentences in a document based on their similarity to each other. It uses the concept of eigenvector centrality to score sentences based on their similarity to other sentences in the document. Sentences with high LexRank scores are considered to be the most important and relevant to the document. For every table i, the LEXRANK score of every $extractive_{ij}$ for table i is calculated.

Majority Voting Technique: Once the BLEU, ROUGE, and LEXRANK scores for each extractive summary $extractive_{ij}$ were obtained, we then wanted to select the most relevant and highly scored extractive summary. However, since the three metrics are different and have different ways of calculation, it was necessary to normalize the values first.

After normalizing, the majority voted summary by all the metrics was finally selected as the most relevant extractive summary as shown in the Equation (1), where Metric denotes either BLEU, ROUGE, or LEXRANK.

$$Relevant\_ES = MaxVoted(MAX(Metric(abstractive_i, extractive_{ij})) \quad (1)$$

3.1.2 Dataset Quality Evaluation. Initially in the dataset, an abstractive summary AB_1 had multiple extractive summaries E_1, E_2 mappings denoted by $AB_i \rightarrow E_j$, where i is the total number of abstractive summaries and j is the total number of extractive summaries for each i. However, after selecting the most significant extractive summary for each table as discussed in the previous sections, we have made the dataset more relevant and compact.

We have employed two methods for validating and evaluating the quality of the corpus namely, Inter Annotator agreement-based validation and Automatic Evaluation.

The following subsections provide a succinct overview of the evaluation methodology of the corpus.

Inter Annotator Agreement-based Validation: In order to validate this dataset, we employed two human annotators, A_1 and A_2, who were tasked with evaluating the mapping between an abstractive summary and the selected extractive summary for a particular table.

Each annotator was tasked to identify whether the mappings were valid according to their opinion. A valid mapping was given a score of "1" and an invalid mapping was given a score of "0". The dataset had 6,010 tables so the annotators were asked to validate a total of 6010 $AB_i \rightarrow E_j$ mappings.

Table 2 presents the confusion matrix constructed using the two annotators provided agreement-based scores for both of the labels (Valid - "1" and Invalid - "0").

With the help of these scores, we then calculate the agreement between annotators A_1 and A_2 using Cohen's Kappa[2] agreement analysis approach.

Cohen's Kappa coefficient score κ, which is defined in Equation (2) [31], which is used to illustrate the degree of agreement.

$$\kappa = \frac{Pr_a - Pr_e}{1 - Pr_e} \quad (2)$$

where Pr_a is the observed proportion of full agreement between two annotators. In addition, Pr_e is the proportion expected by a chance and so indicates a kind of random agreement between the annotators.

The final value of κ ranges from -1 to 1, with 1 denoting total agreement, -1 denoting complete disagreement, and 0 denoting agreement by chance.

The analysis of agreement using Cohen's Kappa, in this case, shows that for the abstractive to extractive mappings, the value

[2] https://en.wikipedia.org/wiki/Cohen's_kappa

Table 2: An inter annotator agreement analysis to validate the dataset

No. of Mappings ($AB_i \rightarrow E_j$) : 6,010		Annotator 1	
		Valid (Score =1)	Invalid (Score=0)
Annotator 2	Valid (Score =1)	5175	40
	Invalid (Score=0)	45	210
Kappa Score	0.824		

of κ is 0.824 with an agreement of 95% confidence interval. A higher κ value indicates a stronger agreement.

The purpose of this experiment was to assess the effectiveness of the proposed method in accurately identifying the summary of table content in a given document.

Automatic Evaluation: We employed two evaluation metrics, BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation), to further confirm the results of the external annotators. Based on n-gram matching, BLEU calculates the degree of similarity between a machine-generated summary and one or more reference summaries. The ROUGE family of assessment measures, on the other hand, focuses on the recall of significant data from the produced summary.

To do this, we selected all the 6010 $AB_i \rightarrow E_j$ mappings and calculated the BLEU and ROUGE scores of the extractive summary with respect to its abstractive summary. For this calculation, we used the AB_i as the reference summary and the most relevant extractive summary E_j as the candidate summary.

Table 3 reports the average BLEU and ROUGE-L (F1) scores for all combinations, while Figure 3 shows the BLEU scores that were obtained for the summary mappings for all possible combinations as mentioned above.

Similarly, Figure 4 depicts the Rouge scores that were obtained for the summary mappings for all possible combinations, viz. (i) *Both annotators agree*, (ii) *A1 agrees, A2 disagrees*, (iii) *A1 disagrees, A2 agrees* and (iv) *Both annotators disagree*. We have taken 40 summary mappings as its the least number of mappings in the confusion matrix above.

After analyzing the charts, we can come to the conclusion that the BLEU and ROUGE scores of the sample mappings that were agreed as VALID by both the annotators have higher values than the other combinations.

This essentially supports our theory that the summary samples serve as the best ones when both expert annotators are in agreement, demonstrating the datasets' quality.

3.2 Summary Generation Using Templates

Along with the dataset corpus using rules as discussed in the previous section, we have also proposed the development of two models that develop summary templates representing the extractive summary of a table. These templates can be utilized by any NLP researcher to study and develop more accurate and coherent summaries and also to generate abstractive summaries.

To do this, our system first identifies a set of significant terms

Table 3: An inter annotator agreement analysis to validate the dataset

No. of Mappings : 6,010			
Both Agree		A2 Agree	
Avg_BLEU	Avg_ROUGE-L	Avg_BLEU	Avg_ROUGE-L
94.1	0.79	49.5	0.40
A1 Agree		Both Disagree	
Avg_BLEU	Avg_ROUGE-L	Avg_BLEU	Avg_ROUGE-L
48.25	0.41	5.12	0.05

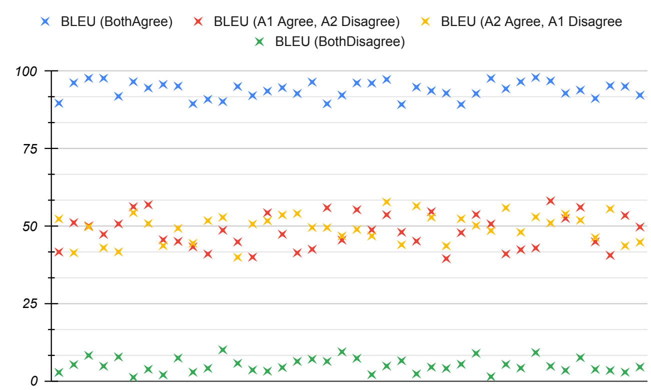


Figure 3: Inter-annotator BLEU scores for rule-based approach

from each scientific paper that is downloaded. These terms are then utilized to produce an extractive summary of the tables contained in the paper.

To assess the quality of the summaries in the dataset, we employ automatic as well as manual evaluation techniques.

The two systems that we have developed are explained in the subsection below.

3.2.1 TF-IDF.

Under this system, we have proposed two approaches namely Unigram and Bigram approach to extract templates for extractive summary.

Unigram Approach: The corpus has a total of 1,500 papers, and for each table in each paper, we calculate the TF-IDF score of all the terms, excluding stop words, non-alphanumeric characters, and unnecessary punctuations.

We only consider terms that are both within the set of unique words and belong to the highest-scored terms for the template. This set of terms is referred to as the Template for Match (TS). While each table can have multiple extractive summaries, there is only one TS for all the summaries of a particular table. Therefore, we rank the summaries to determine which extractive summary matches best with the TS.

Bigram Approach: Another approach called the Bigram approach is developed, which considers a pair of consecutive words instead of single words. In this approach, the TF-IDF score is calculated for each bigram in the document. The

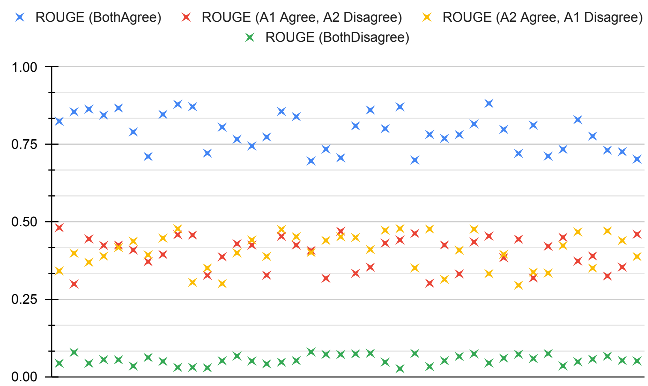


Figure 4: Inter-annotator ROUGE scores for rule-based approach

metrics used for selecting terms in the Template for match (TS) are BLEU and ROUGE. All of these scores are considered background knowledge in the textual entailment method, which is used for ranking the extractive summaries.

3.2.2 Transition Point Based System. Transition Point (TP) is a frequency number that separates the vocabulary of a text into two categories, low frequency, and high-frequency phrases. We have used TP to show how effective it is in indexing text because the mid-frequency phrases are so closely related to a document's conceptual substance. The Transition point system uses two methods: unigram and bigram.

Unigram Approach: A $document_i$ and its vocabulary $V_i = \{(w_j, tf_i(w_j)) | w_j \in D_i\}$, where $tf_i(w_j) = tf_{ij}$, and TP_i be the transition point of D_i. A collection of significant keywords that effectively represent the content of the document D_i using the following method.

$$R_i = w_j \mid ((w_j, \mathrm{tf}ij) \in V_i), (TPi \cdot (1-u) \le \mathrm{tf}ij \le TPi \cdot (1+u)) \tag{3}$$

The value of u lies between 0 and 1. Empirical evidence from experiments presented by Urbizagástegui [29] suggests that a threshold value of u=0.4 is effective. The transition point, TP, is calculated using the following formula:

$$TP = \frac{(-1 + \sqrt{8 \times I_i})}{2} \tag{4}$$

The value of I_1 denotes the count of words that occur only once in the document. The terms whose frequencies lie in proximity to the transition point (TP) are considered significant and are given a higher weight for summarization purposes, while the remaining terms are assigned a weight close to zero.

Bigram Approach: The method described above for selecting important terms was enhanced by incorporating words with similar characteristics. This was achieved through the use of a co-occurrence bigram formula described in the work by López et al. [20]. The system based on this bigram approach was divided into three subsystems, namely Left Approach, Right Approach and Left-Right approach. Specifically, given a document D_i consisting only of terms selected using the TP unigram approach (c), the new important terms for D_i were obtained in different ways for the three subsystems. To achieve this, the bigram of each document was taken, and the TF score was calculated as the number of times the bigram occurred in the document. This approach enabled the identification of important bigrams that provided additional context and meaning to the document.

Left Approach: This approach examines the bigram's TF score, selecting only those with a value greater than one. For each of these bigrams, if the term from R_i appears in the rightmost position, the left term is selected as the new term to be added to D_i.

Right Approach: This approach focuses on the right-most term of the bigram when the terms in R_i appear in the left-most position.

Left-Right Approach: The final approach is a combination of the first and second approaches, where both left and right terms are considered if they appear in the bigram and satisfy the condition of having a minimum frequency of two.

3.2.3 Dataset Quality Evaluation. ***Automatic Evaluation:*** The experimentation results after applying automatic evaluation methods like BLEU and ROUGE to the output of the above-mentioned approaches are reported in Table 4.

After developing template-based summaries, it was noticed in the experimentation that by varying the number of terms for Template for Matching (TS), we get different BLEU and ROUGE scores for different numbers of terms. It was further noticed that BLEU scores increased with a smaller number of terms however, there was a decrease in the ROUGE score as the number of terms increase as seen in Table 4.

If the TF-IDF Unigram and Bigram approaches are compared, it is very clear that TF-IDF Unigram approach has better performance in extracting the summary templates.

Moreover, if the Transition Point Approach is considered, the scores are extremely lower than the other approaches.

The results of the automatic evaluation are depicted in Figures 5 , 6, 7 and 8.

If we notice Figures 5, 6, and 7, the observed disparity in the chart emphasizes the importance of agreement between annotators, as it directly impacts the quality and accuracy of the summaries. Thus, when both annotators reach a consensus on the validity of a summary, it can be seen that the summary in question exhibits a stronger resemblance to the summaries, as evidenced by its elevated ROUGE and BLEU scores.

Thus after using automatic evaluation measures, it can be

Table 4: Automatic evaluation scores for summary templates

TF-IDF Unigram Approach				
Terms	**BLEU**	**ROUGE-L**		
		Precison	**Recall**	**F-Measure**
10	46	0.26	0.71	0.34
20	40	0.53	0.60	0.51
30	36	0.72	0.50	0.53
TF-IDF - Bigram Approach				
Terms	**BLEU**	**ROUGE-L**		
		Precison	**Recall**	**F-Measure**
10	0	0.003	0.009	0.004
20	0	0.003	0.009	0.004
30	0	0.002	0.001	0.003
Transition Point System				
Bigram-Approaches	**BLEU**	**ROUGE-L**		
		Precison	**Recall**	**F-Measure**
	0.044	0.14	0.17	0.08
Left Approach	0.08	0.11	0.16	0.21
Right Approach	0.11	0.19	0.21	0.22
Left-Right Approach	0.13	0.14	0.02	0.12

concluded that the TF-IDF unigram approach performed better overall.

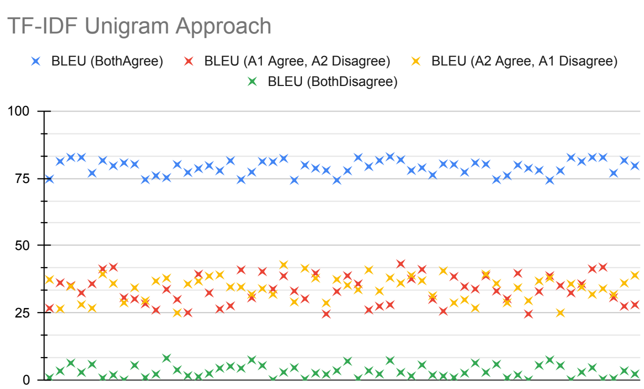


Figure 5: Inter-annotator BLEU scores for TF-IDF unigram Approach

Inter Annotator Agreement-based Validation: In order to validate further, we employed two human annotators, A_1 and A_2, who were asked to evaluate the quality of the summary templates obtained by the TF-IDF unigram approach. We selected only the TF-IDF unigram approach as it was clearly noted that it outperforms all the other approaches.

Each annotator was tasked to identify whether the mappings were valid according to their opinion. A valid mapping was given a score of "1" and an invalid mapping was given a score of "0". The dataset had 6,010 tables so the annotators were asked to validate a total of 6010 $AB_i \rightarrow E_j$ mappings.

Table 5 presents the confusion matrix constructed using the two annotators provided agreement-based scores for both of the labels (Valid - "1" and Invalid - "0").

With the help of these scores, we then calculate the agreement

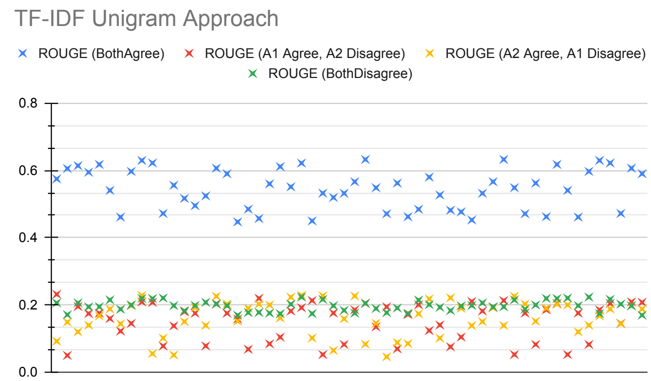


Figure 6: Inter-annotator ROUGE scores for TF-IDF unigram Approach

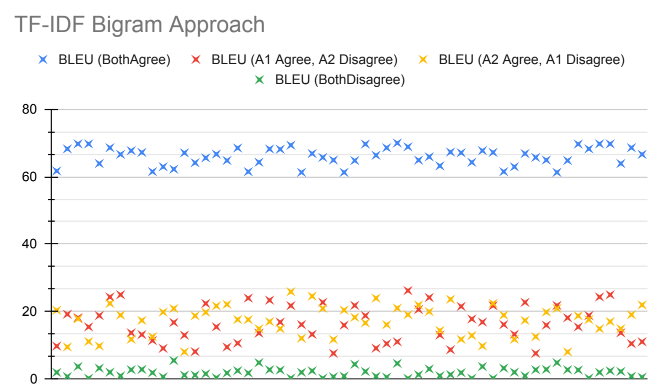


Figure 7: Inter-annotator BLEU scores for TF-IDF bigram Approach

between annotators A_1 and A_2 using Cohen's Kappa agreement analysis approach.

The analysis of agreement using Cohen's Kappa, in this case, shows that for the abstractive to extractive mappings, the value of κ is 0.568.

Comparison: Table 6 shows the comparison between the rule-based approach and template-based TF-IDF approach. In both cases, we have only taken those summary pairs which have been marked as valid summaries by both the external annotators. By observing the BLEU, ROUGE-L, and agreement scores we

Table 5: An inter annotator agreement analysis to validate the summary templates

No. of Mappings ($AB_i \rightarrow E_j$) **: 6,010**		Annotator 1	
		Valid (score = 1)	**Invalid (score = 0)**
Annotator 2	Valid (score = 1)	4,010	467
	Invalid (score = 0)	510	1,023
Kappa score		0.568	

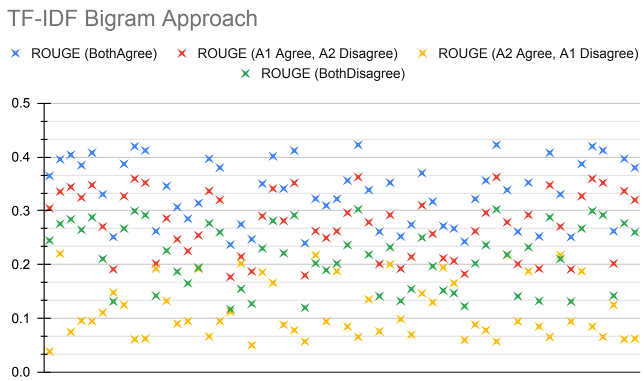


Figure 8: Inter-annotator ROUGE scores for TF-IDF bigram Approach

Table 6: Comparison between rule-based and TF-IDF-Unigram approach where both annotators have agreed

Rule Based Summary_Both Agree		
Avg_BLEU	**Avg_ROUGE-L**	**Kappa_Score**
94.1	0.79	0.824
Template Based_TF-IDF_Unigram_Both Agree		
Avg_BLEU	**Avg_ROUGE-L**	**Kappa_Score**
78.7	0.54	0.568

can easily conclude that the rule-based approach of dataset development is better than the template-based approaches. For this reason, we will be considering the summary generated by the rule-based approach in our next tasks.

4 Extractive to Abstractive Summary Generation

In this section, we propose models to develop abstractive summaries from relevant extractive summaries from the dataset.

Abstractive summarization has the potential to generate summaries that capture the underlying meaning of the text and can provide a more coherent and readable summary. One approach to generating abstractive summaries is to use extractive summaries as a starting point.

Using extractive summaries as a starting point for generating abstractive summaries has several benefits. Firstly, it can reduce the complexity of the summarization task. By using extractive summaries, the abstractive summarization system can focus on generating new sentences that capture the essence of the original document, rather than trying to capture all the details in the original document. Secondly, it can improve the quality of the generated summary. Extractive summaries provide a summary of the main ideas and key concepts in the original text, which can guide the generation of new sentences. This can result in a more coherent and informative summary than generating the summary from scratch. Finally, using extractive summaries can also improve the efficiency of the summarization process. Extractive summarization is a less computationally intensive task than abstractive summarization. By using extractive summaries as a starting point, the abstractive summarization system can reduce the amount of processing required to generate the final summary.

4.1 Abstractive Summary Generation

In general, the process of extractive text summarization involves two steps: sentence scoring and sentence selection. The first step involves determining the importance or relevance of each sentence, while the second step involves ranking the sentences based on their scores and selecting the most important ones to form the summary.

However, in this work, the process has been simplified even further. We have developed a dataset consisting of tables, along with their extractive and abstractive summaries. This means that instead of having to score and select individual sentences, we can focus on developing models to select the most relevant extractive summary for each table.

By using this approach, the researchers are able to leverage the information contained in the tables themselves, as well as the extractive summaries that have already been developed. This makes the process of summarization more efficient and accurate, as the models are able to focus on identifying the most important information in the tables and selecting the extractive summary that best captures this information.

Thus by incorporating additional sources of information and using pre-existing summaries as a starting point, we are able to develop more accurate and efficient summarization models.

In this section, we propose two separate tasks. The first task is the selection of the most relevant extractive summary and the second task is generating an abstractive summary by taking the selected extractive summary as the starting point.

In this task, our input is the carefully selected extractive summary from the previous task and the output is an abstractive summary. We have employed two approaches namely, the T5 model and Seq to Seq Model.

4.1.1 T5 Small Model. T5, which stands for "Text-to-Text Transfer Transformer", is a powerful NLP model developed by Google. It is based on the Transformer architecture, specifically designed for various text-based tasks, including text summarization.

The T5 model can take input in the form of text and generate output in the form of text as well. It follows a "text-to-text" approach, where the input and output are both represented as text strings.

In our work, we utilized T5 small model for generating an abstractive summary. The inputs given to the model are the selected extractive summary for each table.

The steps followed in utilizing this model are as follows:

Pre-processing the Text: In this step, we went through the input text and ensured that it is in a suitable format for the T5 model. This involves removing unnecessary details, formatting the text, and performing any other necessary adjustments.

Fine-tuning the Model: Despite the fact that this is a pre-trained model, we have to fine tune it using our training dataset to make sure the output is relevant to the task at hand. For this purpose, we utilized 4,800 data samples ($AB_i \rightarrow E_i$), where AB_i is the abstractive summary and E_i is the selected relevant extractive summary using the majority voting technique as discussed in the previous sections. The training input was the extractive summary and the output was the related abstractive summary. We have used a dropout rate of 0.1 and a constant learning rate of 0.001. In this task, we experiment with input lengths of 4096 and 8192 and output lengths of 512.

Summary Generation: Once the model was trained we tested the model using the remaining 1,210 selected extractive summary samples (AB_i) and generated an output summary.

It must be noted that in order to instruct T5 to perform summarization, we had to prepend the input text with a specific task description. For example, we added "summarize:" to indicate that the model should generate a summary. The formatted input would then be "summarize: [input text]."

4.1.2 Seq to Seq Model. We decided to employ the sequence-to-sequence (seq2seq) model, which builds on the notion of sequence learning using neural networks, to produce abstractive summaries from extractive ones. Basically, the model accepts a sequence

$$X = \{x_1, x_2, ..., x_n\} \tag{5}$$

as input and attempts to produce the target sequence

$$Y = \{y_1, y_2, ..., y_m\} \tag{6}$$

as output.

where the input and target symbols, respectively, are x_i and y_i. The encoder and the decoder are the two components that make up the seq2seq model's architecture. We experimented with word-level embedding in a manner similar to the work of [23], and our model utilized the seq2seq architecture. Figure 9 illustrates how the seq2seq architecture operates at the word level. We used the Keras library for implementing the model.

Encoder: We utilized LSTM cells in the encoder design. One hot tensor of word-level embedded extractive summaries served as the cell's input. The outputs from the encoder were deleted but the internal states of each cell were kept. This is done in order to maintain context-level information. The decoder cell was then given these states as beginning states.

Decoder: An LSTM cell was once more used for building the decoder, with initial states as the encoder's hidden states. Sequences and states can both be returned by it. Williams' (1989) theory of "teacher forcing" learning was applied in this instance. One hot tensor of abstractive summaries (embedded at the word level) served as the decoder's input, while the target data was the same but with an offset of one-time step ahead of the input. From the initial states that the encoder passed on, the information for generation is acquired. As a result, the decoder has the ability to produce target data [t+1,...] conditioned on the input sequence given targets [..., t]. One word is predicted for each output time step, thus predicting the output sequence.

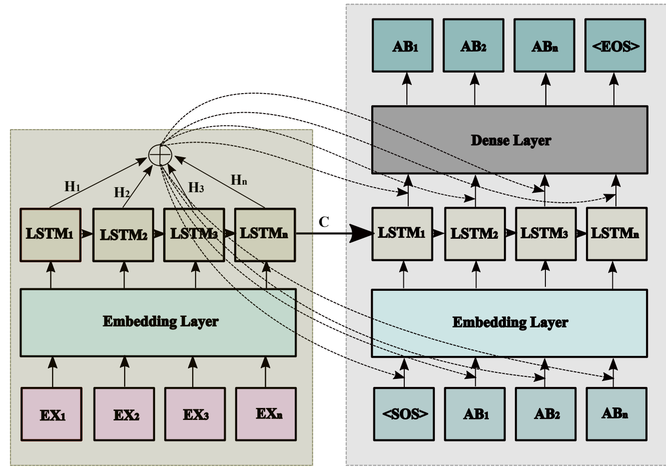


Figure 9: Architecture of the seq2seq model used to generate abstractive summaries from extractive summaries

Attention: The majority of computational neuroscience's work has focused on brain functions like attention [30]. This concept is based only very loosely on how people concentrate their visual attention. It is no longer necessary to encode the complete source sentence into a fixed-length vector thanks to the attention mechanism. Instead, we allowed the decoder to concentrate on various aspects of the source text at each stage of output production. In essence, we let the model decide what to *attend* based on the input sequence and the prediction's current status.

The context vector c_t is calculated mathematically at each time step t as a weighted sum of the source hidden states,

$$c_t = \sum_{t=1}^{Tx} \alpha_t h_t \tag{7}$$

Each attention weight α_t represents how much relevant the t^{th} source token x_t is to the t^{th} target token y_t and is computed as :

$$\alpha_t = \frac{1}{Z} exp(score(E_y(y_t - 1), s_{t\text{-}1}, h_t)) \tag{8}$$

where

$$Z = \sum_{k=1}^{Tx} exp(score(E_y(y_t - 1), s_{t\text{-}1}, h_k)) \tag{9}$$

The normalization constant is called Z. The function score() calculates the degree to which the source symbol T_x and the target symbol y_t match using a feed-forward neural network with a single hidden layer. The target hidden state is represented by s_t and the target embedding lookup table by E_y.

For training the model, *batch size* was set to 64, *number of epochs* was set to 100, *activation function* was softmax, *optimizer* chosen was rmsprop and *loss function* used was sparse categorical cross-entropy. The learning rate was set to 0.001.

4.2 *Evaluation*

We tested the performance of the T5 model using ROUGE and BLEU metrics as they are the most standard summarization metrics available. Additionally, we have also evaluated the quality of the dataset by two other metrics namely adequacy and fluency [10], with the help of two linguists, familiar with the English language. The average values of the metrics are also reported. Adequacy shows how much of the meaning of the source summary is expressed in the generated abstractive summary. Whereas, fluency depicts how well-formed the generated summary sentence is grammatically and can be easily understood by a native speaker. Fluency and adequacy are measured in the range of 1 to 5, where 1 represents the lowest value while 5 represents the highest.

Table 7 shows the results. As observed from the results, we can see that the pre-trained fine-tuned T5 model has performed better than the Seq-to-Seq model. This is because the size of the training and testing dataset was less and hence the model could not be trained adequately.

Table 7: Comparison between T5 and seq-to-seq model

Metrics	Fine-tuned T5 Model	Seq-to-Seq Model
Avg_BLEU	58.2	16.25
Avg_ROUGE-1	0.36	0.21
Avg_ROUGE-L	0.31	0.18
Avg_Adequacy	4.02	2.07
Avg_Fluency	3.91	1.83

5 Discussion and Future Scope

From the results in Table 4, we have clearly seen that the TF-IDF unigram approach outperforms the other two approaches by a substantial margin which paved the way for us to proceed the next tasks with the TF-IDF unigram approach outputs. As a future scope, we can experiment with the outputs of the other approaches as well and make a comparison after all tasks are done.

In addition, from the results shown in Table 7, we can easily see that the Fine tuned T5 model outperforms the Seq-to-Seq model by quite a margin. Though the margin looks small for some parameters like Rouge-1 it is quite substantial concerning the abstractive summary output. This is mainly due to the fact that the dataset that we developed is not sufficient enough for a deep learning model. Thus, as a future prospect, we can aim to increase the size of the summary corpus so that the models can be trained more efficiently to generate coherent summaries.

Furthermore we can aim to introduce a new graph based neural network model, as graph summaries may provide a more succinct representation of an input sentence. Since, a graph-based network's main goal is to maintain predefined properties that are important for particular tasks, including queries on the output summary, it can be highly beneficial for our task as well.

6 Conclusion

In this work, we have addressed the challenges that occur while trying to summarize tables in scientific papers. The data required for training such scientific table summarization systems are very scarce and hence we have proposed the development of a high-quality corpus consisting of both extractive and abstractive summaries. We have proposed two approaches for the same; rule-based and template based. In the rule-based approach we have taken the hypothesis that the caption of the table is its abstractive summary and the citation sentences of the table in the paper are its extractive summaries. The dataset evaluation results clearly depict that the rule-based approach generates much better summaries than the template-based approach. Subsequently, two models, fine-tuned T5-small and seq-to-seq models were also proposed to generate abstractive summaries from extractive ones. The results clearly show that the pre-trained T5 model performs better than the seq-to-seq model.

References

[1] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. Text Summarization Techniques: A Brief Survey. *arXiv preprint arXiv:1707.02268*, 2017.

[2] C. An, M. Zhong, Y. Chen, D. Wang, X. Qiu, and X. Huang. Enhancing Scientific Papers Summarization with Citation Graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:12498–12506, 2021.

[3] V. Arvind and B. Ravindran. Table-to-text Generation by Structure-aware Seq2Seq Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4881–4888, 2019.

[4] Y. Cui, Z. Chen, W. Che, and T. Liu. Table2Text: Descriptive Sentence Generation for Table Structure. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1421–1431, 2020.

[5] L. Dong, F. Wei, and C. Tan. Neural Text Generation in Structured Data-to-Text Applications. In *(EMNLP-IJCNLP)*, pages 1321–1331, 2019.

[6] G. Erkan and D. R. Radev. Lexrank: Graph-Based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[7] T. Gao, C. Zhu, K. Zhang, and X. Ren. Table2Seq: Neural Sequence Generation for Table-to-Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1323–1324, 2019.

[8] S. Gehrmann, K. Deng, and A. M. Rush. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4098–4109, 2018.

[9] F. X. Q. B. Gong, Heng and T. Liu. Table-to-Text Generation with Effective Hierarchical Encoder on Three Dimentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3143–3152, 2019.

[10] M. Hearne and A. Way. Statistical Machine Translation: A Guide For Linguists and Translators. *Language and Linguistics Compass*, 5(5):205–226, 2011.

[11] L. Jiang, M. Yu, M. Zhou, and G. Neubig. Better Structure-aware Transformer for Table-to-Text Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 45–67, 2021.

[12] J. Krishnamurthy, P. Dasigi, and M. Gardner. Neural Semantic Parsing with Type Constraints for Semi-Structured Tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1516–1526, 2017.

[13] R. Lebret, D. Grangier, and M. Auli. Neural Text Generation from Structured Data with Application to the Biography Domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1203–1213, 2016.

[14] C. Li and W. Lam. Structured Attention Networks for Table-to-Text Generation. *ACM Transactions on Information Systems (TOIS)*, 38(2):1–27, 2020.

[15] Z. Li, X. Liu, Y. Chen, and M. Sun. DRGAT: A Dual-Reading Graph Attention Network for Abstractive Summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 142–156, 2021.

[16] C.-Y. Lin. Rouge: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[17] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Hybrid Pointer Generator Networks for Table-to-Text Generation. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–27, 2020.

[18] Y. Liu and M. Lapata. PreSumm: A Neural Model for Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3142–3152, 2020.

[19] Y. Liu, M. Lapata, and F. Li. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3730–3740, 2019.

[20] F. R. López, H. Jiménez-Salazar, and D. Pinto. A Competitive Term Selection Method for Information Retrieval. In *International Conference on Intelligent Text Processing and Computational Linguistics, Springer*, pages 468–475, 2007.

[21] Q. Ma and S. Sakti. Incorporating External Knowledge into Pre-trained Transformers for Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–13, 2020.

[22] Z. Ma, J. Li, G. Zhou, and J. Su. Structure-Aware Convolutional Sequence-to-Sequence Model for Table-to-Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 112–125, 2020.

[23] S. K. Mahata, A. Garain, A. Rayala, D. Das, and S. Bandyopadhyay. A Hybrid Approach To Machine Translation For Lithuanian To English. *2019 Fourth Conference on Machine Translation*, pages 283–286, 2019.

[24] R. Nallapati, F. Zhai, and B. Zhou. Abstractive Text Summarization using Sequence-to-Sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 280–290, 2016.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[26] R. Pasunuru and M. Bansal. Reinforced Video Captioning with Entailment Rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 979–985, 2017.

[27] R. Paulus, C. Xiong, and R. Socher. A Deep Reinforced Model for Abstractive Summarization. In

Proceedings of the 6th International Conference on Learning Representations (ICLR), pages 1–13, 2018.

[28] A. See, P. J. Liu, and C. D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083, 2017.

[29] R. Urbizagástegui. Las Posibilidades de la Ley de Zipf en la IndizaciON AutomATica. *Reporte de la Universidad de California Riverside*, 1999.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[31] A. J. Viera and J. M. Garrett. Understanding Interobserver Agreement: The Kappa Statistic. *Fam med*, 37(5):360–363, 2005.

[32] Y. Zhang, Y. Zhang, Z. Wang, and L. Huang. Table-based Neural Text Generation with Semantic Constraints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 102–113, 2021.

[33] Y. Zheng, X. Ye, Z. Lin, and Z. He. Extractive or Abstractive? A Multimodal Framework for Table-to-Text Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–20, 2021.

[34] Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 654–663, 2018.

Monalisa Dey is an Assistant Professor at the Institute of Engineering and Management, Kolkata (IEM) in the Computer Science and Engineering Department. She is currently pursuing her Ph.D. from Jadavpur University. In her previous roles, she was associated with JIS College of Engineering, as an Assistant Professor. She has completed her B.Tech and M.Tech, in Computer Science and Engineering, from West Bengal University of Technology and NIT, Durgapur. She has more than 20 research papers in reputed journals and conferences.

Sainik Kumar Mahata is an Assistant Professor at the Institute of Engineering and Management, Kolkata (IEM) in the Computer Science and Engineering Department. He also shoulders administrative responsibilities in the capacity of Assistant Head of Department and Placement Coordinator. He has completed his Ph.D. from Jadavpur University. He was awarded the prestigious Visvesvarya Ph.D Fellowship during his Ph.D studies. He has strong proficiency in Python, Natural Language Processing and Machine Translation. In his previous roles, he was associated with JIS College of Engineering, and Narula Institute of Technology in the role of an Assistant Professor. He has more than 30 research papers in reputed journals and conferences.

Dipankar Das is an Assistant Professor in Computer Science and Engineering Department at Jadavpur University, Kolkata, India. He was a Young Faculty Research Fellow, Visvesvaraya PhD Scheme for Electronics IT, Media Lab Asia, Ministry of Electronics and Information Technology, India. He completed his Ph.D from Jadavpur Unievrsity. His areas of interest are Natural Language Processing, Social Networks, Sentiment/Emotion Analysis, Information Extraction, Ontology Engineering, Psycho linguistics, Machine Learning, Code-Mixing, Dialougue Management, Fake News etc. He has more than 90 research papers published in reputed conferences and journals.

Project Manager's Role in Manage Project Knowledge Process: An Approach to Enhance Project Quality

Ahmed Al-Nakeeb*, Mounir El Khatib
Hamdan Bin Mohammed Smart University, Dubai, U.A.E

Raed Abu Zitar
Sorbonne University, Abu Dhabi, U.A.E.

Abdulrahman Alhosani, Ibrahim Alhosani
Hamdan Bin Mohammed Smart University, Dubai, U.A.E

Abstract

This research investigates the intricate integration of knowledge management in project management, focusing on the quality of explicit and tacit knowledge in generating innovative insights. Through a cross-sectional and case study design, the research analyzes the interplay of project knowledge management in UAE organizations, emphasizing the sources of input for lessons learned and their impact on the quality of project knowledge. Project management, recognized as a knowledge-intensive activity, requires a robust knowledge management framework to improve project success rates. This paper highlights the significance of managing project knowledge from project initiation to completion, emphasizing the critical role of project managers in conducting lessons-learned sessions. It also discusses the importance of a strong knowledge repository and a culture of continuous improvement in achieving organizational excellence. Furthermore, the research explores the utilization of Project Management Information Systems (PMIS) as effective tools for decision-making and risk management in the context of knowledge management. The findings underscore the intrinsic relationship between knowledge management and project success, calling for integrating diverse systems and tools to maximize knowledge utilization and overall project performance.

Key Words: Knowledge management; project management; project knowledge; PMIS; lessons learned.

1 Introduction

Every project creates new knowledge. Knowledge refers to the understanding of the subject acquired from learning or experience, which helps in drawing conclusions, and it is different from information, which refers to the refined form of data. We may say that information is free, but knowledge is not [1-2]. Knowledge can be either explicit or tacit; both are important for lessons learned. Explicit refers to what can be expressed in a word, number, or letter so it can be codified, documented, and stored easily, while tacit refers to the knowledge mind of experts, which depends on experience or belief and cannot be codified. It is worth mentioning that both are important as each complements the other and supports the knowledge for contexts. Additionally, tacit and explicit knowledge are harmonized, and (new) knowledge is created through a continuous dialogue between the two types [3].

Project Management, according to the Project Management Institute PMI, is "…the application of knowledge, skills, tools, and techniques to project activities to meet the project requirements" [4]. Yet, Project Management is essentially a knowledge-intensive activity, and there is a growing interest in how knowledge management integrates with Project Management practices [5]. Nowadays, many organizations consider knowledge management and organizational learning as core aspects of their processes to improve project management's success rate. Therefore, PM will handle Manage Project knowledge process before the project start and throughout the project life cycle, which is essential to accomplishing project objectives, and new knowledge will be recorded as input within OPA, which will be helpful for upcoming project plans. Traditional knowledge management systems are focused on the ability to capture knowledge in centralized systems and make it available at a later date, as per [6].

There are several benefits of such a process, such as to produce or improve project outcomes and knowledge created to support organizational operations for further project implementation. This is essential as success factors differ from one project to another. As concluded by [7], the path to achieving success in project management remains elusive, with no definitive formula in sight. It is likely that there will not be a one-size-fits-all solution, as success hinges on numerous variables that can vary from one project to another and from one organization to another. Also [6], argued that by participating in a collaborative ecosystem of knowledge, organizations, for the first time, can accelerate their learning skills and efficiency, innovation, and agility.

Organizational learning requires the provision of new working models. Such models will either support overcoming concerns that occur as a thread or encourage opportunities. On the other hand, the process is not useful if it is not performed properly since only allocating the budget for the database and assigning a PM time, the organization will be partially engaged and therefore will not get huge benefits, which will be only either limited or gain no visible

* Email: a.alnakeeb@hbmsu.ac.ae.

benefits, in fact, they need to invest in cultural and processual aspects to gain massive benefits and proper implementation. If your project team lacks efficient knowledge transfer, this situation leads to wasted activity and poor project performance, as argued by [8]. He added that in many cases, the team will fail to perform roles and assignments as no information management practices are implemented to share and distribute project data.

Project managers play a critical role in managing projects using project management knowledge. Project knowledge consists of new project details required for the project manager and members. Before sharing details with members, a PM must verify Project knowledge inputs like project management plan, project documents, OPA, EEF, deliverables, and others. Moreover, as determined by [9], one of the PM roles is learning from their own or others' experiences, like subject matter expert SMEs, utilizing lessons learned for early warning signs (EWS) and pretested remedies. Therefore, it is evident that the PM is responsible for the success of the project implementation when supported by proper knowledge management practices.

2 Research Methodology

This research utilized two main designs as a framework for collecting & analyzing the data and to provide a plan for the study: cross-sectional & case studies. Cross-sectional design is a research methodology commonly used in various fields, including social sciences, epidemiology, and market research. It involves collecting data from a diverse group of participants or subjects at a single point in time to examine and analyze variables of interest. This approach aims to provide a snapshot or a "cross-section" of the population at that specific moment, allowing researchers to draw inferences about the relationships between variables without the need for longitudinal data collection over time. A cross-sectional design was selected as interviews with six subject matter experts in the project management field were conducted. At the same time, in the case study, we verified the approach used in managing project knowledge in organizations in the UAE. In addition, we benefited from the latest PMBOK 6th edition [10], which has a new chapter to address managing project knowledge, as this helped to enrich our understanding. In this research, the following important questions are addressed;

1. What are the sources of input for lessons learned & what do you expect to know from lessons learned?
2. How do lessons learned enhance the managing project knowledge?
3. What is expected as the outcome & benefit of lessons learned for Managing Project Knowledge?
4. What is the lessons-learned process used, and how can it be improved regarding technology related? Is the process unified in your company?
5. What are the tools and techniques used in the lessons learned process, and how to improve it?
6. Do companies use Project Management Information Systems (PMIS) for Knowledge Management Systems? How do they benefit project managers?
7. How can Managing Project Knowledge support decision-making & risk management?

3 Literature Review

3.1 Manage Project Knowledge

Manage Project Knowledge process is defined as the process of using existing knowledge and creating new knowledge to achieve the project's objectives and contribute to organizational learning, as per the Project Management Body of Knowledge [10]. Manage Project knowledge process performed throughout the project life cycle helps to use existing knowledge, which is required to accomplish project objectives and to create new knowledge that is recorded and will be helpful for upcoming project plans by improving project outcomes and knowledge that are created to support organizational operations for further project implementation.

Organizational learning, or in other words, the lessons learned process, plays an important role in providing new models, which require either overcoming concerns as threads or encouraging opportunities to occur. On the other hand, if an organization partially engages in the lessons learned process by allocating a budget for the database and assigning a PM time only, it will not get huge benefits, which will be either limited or no visible benefits, since they need to invest in cultural and processual aspects to gain massive benefits and proper implementation.

One of the project manager's roles is to manage a project using project knowledge, which is very useful; hence, it has new project details required to implement the project for both the project manager and project members. Project managers are responsible for verifying Manage Project Knowledge MPK inputs like project management plan, project documents, Organizational Process Assets OPA, Enterprise Environmental Factors EEF, and deliverables to take action to ensure the project is completed with respect to project constraints. Moreover, as a project manager is responsible for project implementation and for proper knowledge management practice, one of the project manager roles, as determined by [9], is learning from their own or others' experience, like subject matter expert SMEs. Additionally, utilizing lessons learned for early warning signs (EWS) and pretested remedies. Utilization of this is supported by knowledge sharing using a mixture of knowledge types (explicit & tacit) along with different tools and techniques to support retrieving the data and getting help for making conclusions.

3.2 Lessons Learned and Why We Need Them

Lessons learned are the documented information that contains both positive and negative experiences of a project as learning comes from project failures and successes; thus, lessons learned must be captured as an ongoing effort throughout the project's life [12]. They also argued that learning from failure will help avoid repeating previous similar situations, and learning from success will maximize the opportunity for good processes or practices to complete existing and future work successfully. Thus, a project manager should strongly encourage the team to have this mindset as this also shows a commitment to project management excellence where well-defined processes to capture or utilize the lessons learned help to benefit.

Lessons learned are essential for data documentation, and it is considered as one of the main inputs of the knowledge management process. Recording the previous faults and defects would assist the project team in examining and analyzing the root causes of each error, which, in turn, will enrich current information databases. [13] stated that when knowledge is shared through social communication channels, some data remains unwritten and unspoken as tacit knowledge. The purpose of the lessons learned process is to ensure that the tacit knowledge is converted into explicit knowledge and stays available and accessible for other people as a reference so they can use it whenever they need it. Moreover, [14] added that lessons learned show both the successful and faulty sides of each project, and it is documented in the managerial summaries and detailed reports. The lessons learned (LL) session is a very important part of the lessons learned process. The project management institute published that the lessons learned are an essential approach to reduce the number of errors in the future, and LL databases can be utilized to store information about project cost, time, and scope for similar projects [13].

According to [15], recording the lessons learned can lead to better planning in the initial stage of the coming projects and can be used to support the analysis of risks. Thus, the project team can rely on the previous documents to define all possible risks and reduce the budget for unforeseen risks. Furthermore, [15] added that the recorded lessons are essential to spread additional information among project team members and other project managers. In addition, he argued that lesson learning could be used to evaluate the performance of top management and project managers as well. Sharing knowledge effectively reduces the main causes of previous schedules and cost overruns; furthermore, lessons learned can be used as evidence to clarify the recorded issues from a managerial and technical aspect.

According to [16], the analysis carried out for their paper has shown that incorporating insights gained from previously collected lessons learned in past project experiences is viewed as a crucial component of project management methodologies.

Lessons learned (part of OPA) and other project details are entered as project knowledge for new project implementation, so it is essential to be managed perfectly by the project manager to have positive project progress. Before the project starts, the project manager can have a session with SME and other project managers with experience in the same project nature to capture and get the latest and useful lessons learned. [12] highlighted that a project kick-off meeting is the best time for discussing lessons learned. Therefore, PMs communicate using different methods, such as storing files in share driver, communicating by email, or meeting with concerned stakeholders about organizational or individual knowledge with three classifications: work, method, and result. During project implementation, areas for improvements obtained from lessons can be found for both organizations as a routine (policies, regulations, procedures, etc.) or individual as skills, experience, etc., and (working practice) knowledge. Hence, lessons are identified and then analyzed by team members; after that, they will be stored as a new or updated idea as part of new knowledge generation practice, which finally goes to knowledge storage as stored updated lessons learned and to be retrieved for upcoming projects.

Moreover, a Lessons Learned Knowledge bank could be established as recommended by professional institutes to centralize all information gained by each project manager, which can provide benefits for the same category of data for all project managers. Otherwise, an initial database is to be created with the help of experts and advisors, which will include lessons gathered from previous projects. Such a database will support eliminating duplication of the same lessons from different projects. Noting that having a number of occurrences of an issue can lead to defining common problems. Additionally, an organizational database enables the project manager and project team member to retrieve relevant lessons where lesson learned knowledge bank could be maintained by a three stages process:

1. Assessing previous project lessons learned during the planning and delivery phases of a new project
2. Recording lessons throughout the project in project logs
3. Writing lessons learned report not only during the project implementation but also during the project closure phase

Lessons Learned Register details would have new updates once the project is completed, and it will be an input to OPA that can be used for upcoming projects. Such a register will include a category for easy access to certain knowledge, description, impact, proposed actions and recommendations, challenges, risks, and opportunities. Documentation of LL purpose is to share and use knowledge to either:

- Encourage opportunities for desired outcomes.
- Eliminate the occurrence of non-desired outcomes.

As a result, organizations can make significant savings for upcoming projects in terms of time, cost, and quality; the same can be achieved by preventing the repetition of problems and encouraging opportunities. Furthermore, records on this database need to be updated and integrated with new knowledge and remove obsolete knowledge [17]. It is acknowledged that both organizations and individuals tend to learn more from failures than from success [18], and that failures contain valuable information; however, organizations vary in their ability to learn from them [19].

3.3 Lessons Learned Process, Tools, and Techniques

According to [12], there are three levels under lessons learned for organizations that need action and intervention from the project manager. These levels are: Level-1 is the lessons learned process, level-2 is the Evaluation of the Lessons Learned Repository, and level-3 is related to lessons learned metrics.

Source: Rowe, S. F. & Sikes, S. (2006). Lessons learned:

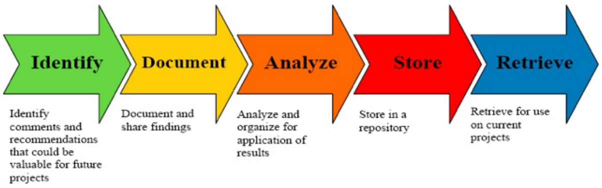


Figure 1: Lessons learned process

taking it to the next level. Paper presented at PMI® Global Congress 2006—North America, Seattle, WA: Newtown Square, PA: Project Management Institute.

Level-1 lessons learned process:

It consists of five steps: identify, document, analyze, store, and retrieve. These steps are required in order to define activities needed to capture and use lessons learned successfully. Therefore, the project manager needs to encourage the participation of team members to gain higher results.

The first step is to Identify Lessons Learned. In this step, concerns are identified as experiences or recommendations of opportunities that occurred during the existing project. This step consists of two activities: preparing the session and conducting the session. There are three key questions for this step: what went right, what went wrong and what needs improvement; it recommends that the facilitator share what would be discussed earlier and guide the discussion.

The second step is to Document Lessons Learned, where the project manager should report captured lessons learned and inputs from participants from the conducted session, then share it with stakeholders in detail, and submit a summarized report to management while the final report should be included with the other project documentation.

The third step is about analyzing lessons learned, where analyzing and organizing of captured lessons learned would take place as this includes needed training or improvement in the project management process.

The fourth step is to store lessons learned. In this step, the Lessons Learned Register could be updated with the outcome from the above-mentioned step. The same can be included in project documents or project share folder/drive to have feasible access to all team members. Also, update or store to be with an appropriate keyword search capability to ease retrieving appropriate lessons.

The fifth step is to retrieve lessons learned. The lessons learned process is to retrieve lessons to be used for either running projects or upcoming ones to overcome issues or encourage opportunities. Retrieve should be for most appropriate and updated data.

Level-2 Evaluation of Lessons Learned Repository:

Effective tools and starting analysis of stored lessons learned are what organizations need at level 2. Where the process became part of organizational culture. Although organizations consistently capture lessons learned, they are not fully utilizing them, as noted by [12].

Based on the lessons learned process, identified, and gathered data are collected in the organization's repository, which is the project manager's primary responsibility. In level 2, the project manager and others involved will start the analysis of available documents in order to provide action to raise project knowledge and implementation. For a successful analysis, it is important to empower the person doing the analysis to enable him or her to implement the approved solution. Therefore, there is a need to enhance such tasks by enhancing the project manager and member's knowledge with specific training. Additionally, existing processes or procedures may require changes to adapt to new setups, and it is worth mentioning that there should be a unified template called lessons learned inputs forms as a key tool to have consistency of input that plays an important role in identifying recurring issues and providing proactive solutions, see Figure 2.

Project Name:						Project Manager:		
When	What			Who	How	When	What	Other
Date	Lesson Description	Category	Priority	Owner	Lesson Action	Date	Status	Comments

Figure 2: Lessons learned template

Source: https://vgpblog.files.wordpress.com/2015/03/lessons- learned-plan-value-generation-partners.png.

Such templates should be shared with project members during team sessions that include previously defined fields like category, lessons learned, action taken, results after action taken, root cause, and keywords. Such identification needs to be added to the lessons learned repository.

Based on gathered data, different types of reports are classified as levels 1 and 2. Lessons learned detailed and summary reports obtained from the session that could be shared immediately with stakeholders fall under level 1. On the other hand, a level 2 report can be found as a detailed report, having responses gathered during sessions for each lesson along with findings as recommendations. Then, approved actions need documentation in detail while implemented.

Figure 2 shows a sample of how the template looks, and it could be customized based on need. A real example captured from the interview can be found in the appendix.

Level 3: Lessons Learned Metrics

The executive level is mainly interested in metric data for further approval and decision. Therefore, it is essential to convert obtained data from the completed analysis to metrics since executives seek the "Evelyn Woods Speed Reading Technique" or equivalent due to their busy schedules [12]. Subsequently, an executive lessons learned report requires clear data and information supporting expediting approval or decision-making. Such a report can contain the first page as an overview of analyzed data and recommended next step, and the following two pages can have graphic presentations such as a pie or bar chart with a clear legend in order to reflect data registered on the first page, providing a clear picture of what was wrong and how it changed after correction action. Therefore, effective metrics reports will be generated and available as key achievements that capture consistent and maintained lessons learned from a centralized repository.

Generally, in order to achieve higher outcomes, a group of team members, SMEs, and other project managers must work with project managers to document the lessons learned to ensure that all elements and issues are recorded. In addition, the assigned group will help project managers revise the important files and key findings, which will facilitate the analysis of captured data, and they will be able to distinguish between several types of faults related to similar projects. In addition, sharing the lessons learned and discussing the previous defects is essential to strengthen the social relations between project managers and all stakeholders and enhance their ability of thinking and negotiation skills.

Overall, achieving an effective metric report depends mainly on the quality and type of data captured in lessons learned, which must be consistent and maintained in a centralized repository.

3.4 Lessons Learned Tools and Techniques

There are several tools and techniques that the project manager or facilitator can depend on to gather data needed for knowledge management, which contain both explicit and tacit knowledge [10].

1. Subject matter expert or expert judgment guidance and feedback
2. Knowledge management
3. Information management
4. Gather feedback from project team members by way of formal and informal interviews
5. Use closed focus groups to develop lessons learned
6. Keep an active register for the lessons learned
7. Share lessons learned with peers in other projects
8. Continuous research into new methods for achieving similar projects
9. Emails and meetings
10. Collaborative tools such as portals, surveys, and data capture forms to gather lessons
11. Relevant information from other projects
12. Organizational learning

3.5 Improvement of Lessons Learned Process, Tools and Techniques

Deming's Cycle, or PDCA, is a simple, straightforward process that is recommended to be used for improvement concepts. From PDCA we can have and improve lessons learned and use lessons learned in a continuous life cycle. Indeed, to increase the project success rate and develop organizational learning, the lessons learned process needs to be improved.

One of the most important steps for improvement is to capture LL throughout the project life cycle and not only at the end. So, the PM should encourage all members to participate and record such learning (best practice and area for improvement) from failure or success; the same is applicable and useful to improve the process, which is considered part of procedures and expected deliverables from project management. Additionally, recommendations for best practices are to be included in the summary and in the first paragraph of any executive report to enhance organization processes and procedures.

Through a cycle of continuous learning, a culture of successful projects can be created as this helps to avoid making mistakes or reinventing the wheel at each project start, [13]. Hence, the 5-step continuous process for capturing knowledge is recommended to be used by project managers and Project Management Offices PMO, as shown in Figure 3.

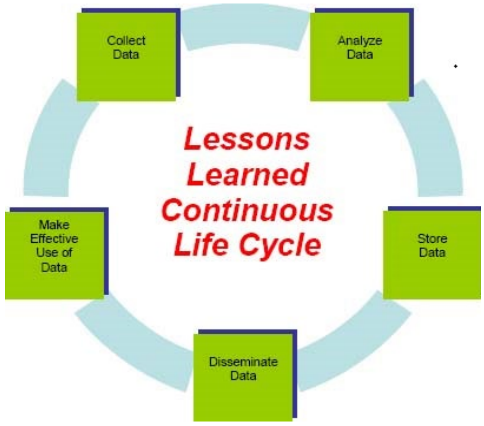


Figure 3: Lessons learned continuous life cycle

Source: Trevino, S. A. and Anantatmula, V. S. (2008). Capitalizing from past projects: the value of lessons learned. Paper presented at PMI® Research Conference: Defining the Future of Project Management, Warsaw, Poland. Newtown Square, PA: Project Management Institute.

3.6 Culture and Social

For team members to generate ideas and share them for improvement or reporting matters related to lessons learned, they need to be motivated; therefore, the PM should ensure the motivation of the team members and build a trustworthy environment as well as be aware of individual's responsibility, expertise and influence, authority levels and their awareness regarding organization policies. Knowledge transfer and learning occur through social, situated learning, and de-coupling the lessons learned process significantly reduces their value [20]. Project teams' social capital is conducive to overcoming barriers to learning in project-based organizations [21].

As per [10], Face-to-face interaction is usually the most effective way to build the trusting relationships that are needed to manage knowledge, and there is an important role of political awareness as it helps the project manager to plan communications based on two aspects, which are project environment and political environment. Additionally, leadership should encourage project stakeholders to use the process, tools, and results [12]. The project manager is responsible for conducting lessons learned sessions for all projects with key internal and external stakeholders and the need to align environmental culture and project culture [22].

For example, the culture and method in the agile project lie in leveraging team knowledge and producing deliverables, which shows that some of the practice is inbuilt knowledge-sharing mechanisms. Also, this emphasizes that face-to-face interaction is more effective for knowledge sharing and building relationships of trust and respect among the team. This is supported by the finding of [5] as face-to-face interactions and agile approaches facilitate knowledge transfer and strong relationships between them.

Figure 4 shows the integral model for Enterprise Environmental Factors (EEF) that can influence managing project knowledge [23].

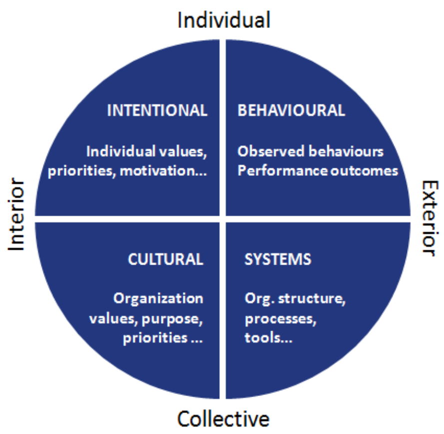


Figure 4: Model for enterprise environmental factors

Source: Rowley, J. (2018), 6th Edition PMBOK® Guide–Process 4.4 Manage Project Knowledge: Inputs. Available at: https://4squareviews.com/2018/02/03/6th-edition-pmbok-guide-process-4-4-manage-project-knowledge-inputs

Cultural - The need to share knowledge in a relationship of trust among all team members and stakeholders to allow managing knowledge

Systems - Show how the team's location affects the knowledge share and how its knowledge plan is to be accessed and shared, along with the level of confidentiality of project information and who to identify the people who specialize in knowledge management.

3.7 Project Management Information Systems (PMIS)

Lessons Learned Database Software is a project management solution, and cloud-based knowledge created by sector-solutions helps businesses capture experiential knowledge from key projects. It is a tool that discovers and manages valuable content and categorization and will be available anytime for staff. Project management information systems can be utilized as an effective tool for project planning and monitoring of phases. Furthermore, [24] argued that project management information systems are useful for project managers as they store and organize enormous amounts of data. Implementing such systems can facilitate the managing process of simple and complex projects; below (Figure 5) is an example of a project dashboard using Primavera PMIS, where we can fit project size based on its simplicity or complexity.

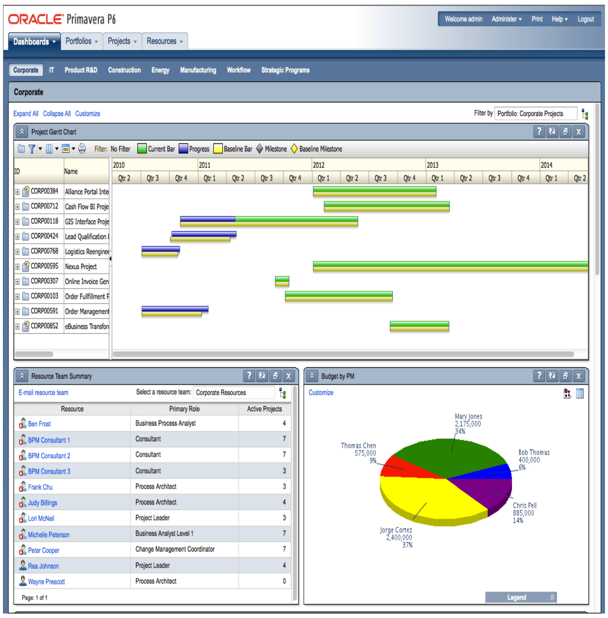


Figure 5: Primavera project management information systems

Source: Powerful Tools for Global Project Planning available at: https://www.oracle.com/ae/applications/primavera/products/project-portfolio-management

In addition, integrating the lessons learned through project management information systems can increase information availability, which is essential for short and long-term planning and decision-making. Furthermore, the collected data's size and accuracy affect the managerial decisions' efficiency. Additionally, [25] mentioned that self-hosting applications require a specialized person to effectively manage financial and human resources. The study covered a sample of 100 project managers selected from different countries in Europe, the Middle East, and Africa (EMEA). The results showed that the level of satisfaction with the current project management information systems was above 60% in Western Europe and the Middle East and 72% in Africa. The level of satisfaction per industry from the highest to the lowest: information system (79%), manufacturing (77%), construction and government (71%), energy (55%), and defense (45%). The participants mentioned some limitations correlated with the current systems, such as the lack of uniformity and integration with enterprise resource planning (ERP) and insufficient support to access open-source folders.

On the other hand, they mentioned some compelling features for lesson learning and data storing. For instance, they added that PMIS could be utilized to create risk registers, implement an earned value management that focuses mainly on evaluating the actual performance against the planned one, and provide a technical support during project planning.

Integrating all systems and tools for knowledge management to maximize the benefit is very important. Although some companies have systematic learning to provide competency, they fail to use it effectively [26].

3.8 Managing Project Knowledge Can Support Decision-Making and Risk Management

"Those that fail to learn from history are doomed to repeat it." Winston Churchill (1874-1965). Managing Project Knowledge will have updated Lessons Learned Register, which include type, description, risk level, occurrence, solution, proper process, etc. that can expedite decision-making and action to perform for similar concern. Regarding risk management, it supports overcoming issues by being proactive to eliminate issues from occurrence. For cases with multiple occurrences, it provides an indication to verify the root cause and ways to eliminate occurrence.

3.9 Project Knowledge Management Best Practices

Following is knowledge management as best practices that can be obtained from lessons learned, which are critical for projects and support decision-making and risk management [22].

1. Align with The Sponsoring Organization's Knowledge Management Practices - Avoid reinventing knowledge management and engage the sponsor company knowledge management practice
2. Use Existing Knowledge - Project teams must quickly familiarize themselves with the organization's knowledge as an essential aspect of the project. Failing to understand the organization's existing knowledge can lead to potentially explosive political situations. For instance, it's crucial to avoid allocating resources to a feasibility study that has already been completed or inadvertently contravening the organization's established strategy and best practices
3. Validate & Assess Knowledge - Validate and assess the existing knowledge prior to using it.
4. Be as open as possible with project knowledge - This is to ensure the reduction of any resistance to the project
5. Communicate & Socialize Knowledge Deliverables - Use the different methods for communication at the right time and in the right way
6. Establish Project Principles That Layout Your Knowledge Management Approach - This will be a guideline to help everyone follow
7. Identify the long-term owners of knowledge - Set a clear process to hand over the knowledge to the company as the project is transitional
8. Integrate Project Knowledge with Organizational Knowledge - Connect and combine both project knowledge & organization knowledge

4 Data Collection

The qualitative approach has been used to represent the participant's point of view since inductivism theory emerged from obtained data. Interviews with six subject matter experts in the project management field have been conducted. The candidate answered seven questions. These data, in addition to data obtained from PMBOK 6th edition [10], articles from the internet, and research papers, were used to support obtaining a more rounded and clearer picture regarding our research topic.

4.1 Source of Input for Lessons Learned and Expect to Know

Lessons learned captured during the course of the project from all stakeholders involved are essential to be managed perfectly by the project manager to have positive project progress as per [27] and this need to capture various incidents and stages during the project as it may have both positive and negative points, risks handled, etc. and can serve as a guide and reference to future project managers [28] [29].

As per the interviewees, there are several sources of input to lessons learned that can be helpful to add to the Lessons Learned Register such as;

- Failed projects, processes
- Mistakes made during projects
- Success during projects
- Alternative/solutions generated during projects
- Encountered problems and remedies implemented
- The previous Lessons Learned Register
- Staff meetings & different social source
- Detailed analysis of project variance or deviation by dedicated team
- Study of historical project data and processes.
- Subject matter expert advice

In [27], believes that Enterprise Environmental Factors (EEFs) must be considered as they could provide either opportunities or threats to any project. These details will be added to the Lessons Learned Register with some of the specifications and categories such as type, description, risk level, occurrence, etc.

4.2 How Do Lessons Learned Enhance the Managing Project Knowledge and the Benefits

In [28] and [29] advised that lessons learned serve as a reference and help set benchmarks for future projects as they document the events, risks, problems, solutions, etc.; [30] added that it enhances project managers' knowledge, experience and enhances risk management, which enhances project maturity. This is supported by implementing new processes and enriching the project knowledge as lessons learned build the knowledge base, [31, 32, 27] highlighted that lessons learned help avoid repeating the same errors committed in the previous project, analyze the scope of improvement, and reduce risk, cost & time. These incrementally add to project knowledge and support the continuous improvement of the project management processes. This results from the right storage and files to ease access to lessons learned and have data analysis for what was learned, according to [30].

[27] added an example for that: if an issue occurred while the project was implemented and the same was already recorded in the Lessons Learned Register, then fast response and proper action can be taken to put the project back on track while this may take more time in case of no previous knowledge was there.

4.3 Expected Outcomes from Lessons Learned

Based on multiple feedback from interviewees, the section below summaries the significant contributions of lessons learned;

- Lessons Learned Register that supports in reusing existing knowledge and creating new knowledge.
- Enriched Lessons Learned Register with new update
- Opportunities for improvement
- How to improve
- Enforces what has been done well
- Mistakes committed in the project
- Problems that could have been avoided
- How to archive data
- How to avoid future project risks
- New processes or frameworks of improvement
- New and improved ways to achieve/implement similar projects
- Achieving similar projects faster and more efficiency
- Savings in projects due to possible reduced timelines and wastage
- Create & update the company repository

4.4 Lessons Learned Process Used and How It Can Be Improved

The lessons learned started from day one of the project and even before that while lessons learned from previous projects have to be shared and communicated with team members in the kick-off meeting, as well as to be visible any time for them to obtain great results as advised by [27] [29]. It also includes other processes like change requests, reporting, documentation, etc., and classifying the lessons as all to be updated in the Lessons Learn Register. In [32] added that each project is implemented in accordance with expert opinions, decisions, and support from subject matter experts together with historical data or trends.

This will keep the register always updated including need modification to existing details; hence, this will be an input to Organization Process Assists (OPA) while taking into consideration that details are captured throughout the project life cycle as these lessons are observed and identified from staff who are handling project implementation or project progress report thus it is not only at the end of the project while each team member as the team may forget later what they need to update so this will give them clear guidance to do their works under consultation of project manager for any concerns and to ensure collected data is stored [27] [30].

In [27] added that after identifying the concern, it has to be recorded in the project Lessons Learned Register before starting to analyze them with the support of the subject matter expert, followed by appropriate corrective action to proceed further. After that, the register will be updated once the issue is cleared and finalized. It is important to use them correctly based on the right analysis, as per [30]. According to [29], projects should not be closed formally without documenting and reviewing lessons learned along with approval from relevant leadership as this creates governance; hence, this needs formal meetings to close any project along with minutes of meetings that are distributed and followed up.

Furthermore in [27] believes that the process in Etisalat varies among different sections due to the different nature of works and details, but generally, it's almost close to each other's. On the other hand, [30] advised that Project Management structures are independent in each division but supported by the back office for Technical Project Management (TPMO) responsible for unifying the processes of managing projects and having the right tools. The overall process summarized by [32], is identify, document, analyze, store, and retrieve while [31] added some steps as;

- Gather feedback from project team members by way of formal and informal interviews
- Use closed focus groups to develop lessons learned
- Keep an active register for lessons learned
- Share lessons learned with peers in other projects
- Continuous research into new methods for achieving similar projects
- Emails and meetings
- Collaborative tools such as portals, surveys, and data capture forms to gather lessons

Lessons learned process can be improved as per [27], by applying and implementing a Workforce Management system (WFM) as well as risk simulators; hence, there will be an indication of upcoming shortages of resources, time, or other factors and usage of big data and AI would improve the lessons learned process to the most as per [30]. Also, [31] added that ideal unification is required for shared knowledge within an organization; then on completion of every project, lessons learned are recorded, and team meetings are called to discuss various risks encountered, variance from the baseline of the project, and how we can bring about improvement, [32]. Lessons learned as per [29], should be incorporated into the KPIs to ensure that PMs actually look at and show sufficient evidence that they have tried to mitigate future risk based on the lessons learned.

4.5 Tool and Technique

According to [28], the process is done by recording the lessons in an Excel file with a template which includes a record of events, successes, failures, risks encountered, mitigation steps taken, and recommendations or improvements; thus data collection, data analysis and data storage is done based on this. [27] added that this is done by having a shared folder for each project as well as having a single database for project risk and stopper that is updated by the back office team, with respect to their observation, staff reporting from the field, or subject matter expert. He added such experience and correction action shared by different communication methods such as email, share folder, or verbal communication (call, meeting, etc.) with all concerned members. Furthermore, [31] added the use of collaborative tools such as portals, surveys, and data capture forms to gather lessons.

According to [27], there are different systems used but do not integrate all project parts and lack end-to-end solutions as each system has some portion of data, and the same is shared by [28]. Furthermore, [29] mentioned that ADCCI does not use Project Management Information Systems PMIS but they have developed a governance to provide data and analytics regarding project management to leadership as this is mainly in the form of manual reports that are based on MS Project schedules and weekly meetings.

Also [27] added that improvement can be obtained by having a PMIS system that can capture and provide an indication for abnormality, which also provides a guideline to overcome issues that were added from previous experience. In [30] is mentioned that Etisalat is in the process of procuring PMIS to support project managers to have smart and mobile access to multiple projects and program documents while [31] advised that this section already has PMIS which is fed by multiple projects' owners, allowing project managers to avoid similar mistakes made by others already.

4.6 How Managing Project Knowledge Can Support Decision-Making and Risk Management

According to [28] [27], managing project knowledge includes an effective recording of lessons learned and project experiences that will speed the decision-making, help deal with similar situations, anticipate possible risks, previous solutions, and effectively efficiently manage the project. Knowledge management provides a way to treat knowledge as an asset, transfer data, and benefit from lesson learning for future projects, as per [30]. He added this can support in reducing risks and continuous improvement in project performance. In [31] it was stated that the team will avoid issues and support to minimize risk by intelligence from data gathered from lessons learned.

In addition, [27] believes that it supports overcoming issues by being proactive to eliminate issues from occurrence. For cases with multiple occurrences, it provides an indication to verify the root cause and ways to eliminate occurrence. According to [29], a good repository of knowledge is vital to decision-making as proper decisions need to be supported with knowledge of past, present, and forecasting future events as this becomes a norm in an organization if there is Knowledge-Based decision-making and also applicable from a risk perspective as it provides the ability to estimate or foresee what may happen in the future by looking to previous situations, events, and experiences. He added that project knowledge is also important to estimate project budgets and contingency based on previous projects.

5 Analysis

5.1 Knowledge and How to Maximize Its Benefits by Project Manager

Every project creates new knowledge which can be either explicit or tacit, where both are important for lessons learned. It is worth mentioning that both are important as each of them complements each other and supports to cover the knowledge for contexts. This project management is "…is the application of knowledge, skills, tools, and techniques to project activities to meet the project requirements" [4] as it promotes an integrated approach to identifying, capturing, evaluating, retrieving, and sharing all information, which includes what was unrecorded such as expertise and experience. This is very important for the benefit of the project to avoid mistakes and help the project manager avoid reinventing the wheel every time they initiate a project as it leads to faster and more effective decision-making for the project manager, facilitating him to get needed approval. Subsequently, it makes the project team more competent with a better understanding of project requirements based on knowledge as this knowledge is a mixture of stored documents or data and interactions between people.

Thus, as a first step towards improving existing practices in the future, it is essential for the project manager to increase his understanding of how project management activities are supplemented by knowledge management activities. The knowledge and lesson learned should include both positive and negative experiences of a project as the project manager and his team should learn from project failures and project successes; thus lessons learned has to be captured as an ongoing effort throughout the life of the project and the project manager has an important role her to ensure that however to maximize the benefit, the project manager should look beyond database or record as cultural and processual aspects are very essential to consider to gain massive benefits and proper implementation. Putting attention to Enterprise Environmental Factors (EEFs) provide either opportunities or thread to any project as these details will be added to the Lessons Learned Register with some specification and categories such as type, description, risk level, occurrence, etc.

5.2 Source of Knowledge and Role of Project Manager

There was an alignment of data obtained from literature and interviewers that the project manager is responsible for verifying Manage Project knowledge MPK inputs like project management plan, project documents, Organizational Process Assets OPA, Enterprise Environmental Factors EEF, deliverables to make action to ensure project completed in respect to project constraints. All interviewers agreed that lessons learned can be utilized to minimize the percentage of errors in the future, so this can be obtained by better planning during the initial stage of new projects by eliminating risks and encouraging opportunities to occur. As elaborated by [28], it will help organizations enhance project execution by eliminating mistakes and problems that could occur. The same was echoed by [27], who mentioned by lessons learned we can increase the success rate for upcoming projects. Therefore, [29] considered having knowledge of previous projects to be key to future project success, so highly advise the importance for others to review available lessons learned at the beginning of any project. In [31] is advised the main area for a project manager to focus on while creating and adding inputs for the Lessons Learned Register:

- Mistakes made during projects
- Success factors during projects
- Processes forces towards failure or success of projects
- Alternative solutions generated during projects
- Encountered problems and remedies implemented
- How to avoid future project risks.

Before starting, the project manager can benefit from meeting with SMEs, and other project managers who have experience on the same project and from lessons learned and this could be before or during the kickoff meeting; however, he needs to ensure that managing knowledge is performed throughout the project life cycle help. The project manager should strongly encourage team members to participate and share their knowledge and experience to gain maximum benefit as this can be done via many mechanisms such as meetings, interviews, focus groups, etc. to support the transfer of project knowledge and share information across the team. There are several tools and techniques that the project manager or facilitator can depend on to gather data needed for knowledge management, which contain both explicit and tacit knowledge [10]. Furthermore [33] found that 64% of participants in an experiment reported having learned from the documents filed by their peers; however, only nine percent were willing to contribute to the documentation due to the time and level of effort required.

One of the project managers' main difficulties is managing lessons learned during the project. While 62.4% of organizations have formal procedures to document lessons learned, 89.3% of organizations are not doing it [15]. In [34] it was found that the main causes of the lack of documentation of lessons learned on projects are time, motivation, discipline, and skills. The project manager should ensure motivating the team members and build a trustworthy environment; hence, in order to create this environment, the project manager must know about team members, their positions and responsibilities, stakeholders, their level of expertise, and their ability to influence, about organizational policies around knowledge/information sharing, and his own authority to take necessary actions to create the environment in the team in addition to knowing if the sponsor has a dedicated team or department that help in knowledge management. These incrementally added to project knowledge and supported the continuous improvement of project management processes.

This is critical as project managers can't force people to share everything they know as even the best knowledge management tools will not help unless people feel motivated to share their knowledge. The ability to manage will reduce the burden on the project manager's role as the job becomes easier if he understands how to transfer project knowledge and share information across team members. He needs to focus more on face-to-face interaction as it is usually the most effective way to build the trusting relationships that are needed to manage knowledge as this supported practice in agile projects as the culture and method in agile projects lie in leveraging team knowledge and producing deliverables which show that some of the practice is inbuilt knowledge-sharing mechanisms.

5.3 Knowledge Process and Project Manager Role

The project manager has to focus on three levels for lessons learned which are;

- Level-1 is lessons learned process
- Level-2 is evaluation of lessons learned repository
- Level-3 is related to lessons learned metrics

In Level 1, there are five steps: identify, document, analyze, store, and retrieve as represented in figure 1 as this has to be a continuous cycle involved by all teams. The project manager should ensure governance of the participation of all concerned team members.

In Level 2, the project manager should ensure that there is an analysis of stored lessons learned and ensure to utilize them in order to provide action to raise project knowledge and implementation. Adding to this, the project manager needs to ensure the integration of existing knowledge with new knowledge and the removal of obsolete knowledge. Also, he needs to empower the person who is doing the analysis with a kind of power to enable and implement approved solutions or enhance the skill of the team through specific training.

In Level 3, the metric has to be designed and mapped to knowledge since the executive level is mainly interested in metric data for further approval and decision. Therefore, it's essential to convert obtained data from completed analysis to metrics. Overall, achieving an effective metric report depends mainly on the quality and type of data captured in lessons learned, which need to be consistent and maintained in a centralized repository.

The project manager plays a critical role in managing the project using project knowledge, which is very useful; hence if the project team suffers from a lack of efficient knowledge transfer, then this situation leads to wasted activity and poor project performance before and during project execute, the project manager needs to do something in parallel, collect, store and manage knowledge along with project deliverable to gain success. He must also ensure that capturing project lessons learned is part of procedures and an expected deliverable from your project management and product methodologies. They must ensure that project teams work together to document project best practices and areas of improvement for the next project. Then, identifying any concerns that have to be recorded in the project Lessons Learned Register prior to starting analyzing them with the support of a subject matter expert followed with an appropriate correction action to proceed further. After that, the register will be updated once issues are cleared and finalized. It is important to use them correctly based on the right analysis.

It is clearly observed from the interviews that there is no unified concept; hence, the Project Management Structure is independent in each division in Etisalat. For example, [30] elaborated on the need for back office support for Technical Project Management (TPMO), where such processes are core function for TPMO to unify the processes of managing projects and ensure having the right tools and resources.

5.4 Knowledge Tool and Project Manager Role

The interview report showed that different processes, tools, and techniques are being used for different sections in Etisalat and ADCCI. An agreement was observed to start lessons learned throughout the project life cycle and capture concerns related to threats and opportunities. Tools or techniques to capture and record the lessons learned included basic office document applications to specialized Project Management Information systems as this depends on project size and complexity and the company's maturity for project management practices. However, even using the basic method, the project manager must ensure a standard template for capturing lessons learned and predefined fields to record attributes to ensure ease of later retrieving and analyzing. Interviewees advised that all record is stored in shared folders to which the project team has access, which is very important to ensure knowledge transfer; however, this keeps data at risk of change without approval or lack of confidentiality of data. Also, it has been found that common major tools and techniques are:-

1. Gather feedback from project team members by way of formal and informal interviews
2. Use closed focus groups to develop lessons learned
3. Keep an active register for lessons learned
4. Share lessons learned with peers in other projects using Emails and meetings

The interview report showed that Etisalat and ADCCI are not using any PMIS system. On the other hand, they are using different systems but not integrating all project parts or providing end-to-end solutions. As per [35] [36] the use of technologies and applications provide better support to project manager in daily tasks and processes. Thus, if a company uses PMIS, then this would be a tool to manage valuable content and categorization. It will be available anytime for staff and benefit project managers as it is useful to store and organize enormous amounts of data. PMIS could also be utilized to create risk registers, implement an earned value management that focuses mainly on evaluating the actual performance against the planned one, and provide a technical support during project planning so it is recommended to use it by the project manager and organization. This is supported by the results obtained from the literature, which showed that the level of satisfaction with the current project management information systems was above 60% in Western Europe and the Middle East and 72% in Africa. The level of satisfaction per industry from the highest to the lowest: information system (79%), manufacturing (77%), construction and government (71%), energy (55%), and defense (45%).

In addition, it is very important to integrate all systems and tools for knowledge management to maximize the benefits [26]. This supports overcoming issues by being proactive to eliminate issues from occurrence and use the correlated data in the best form to support decisions. For cases with multiple occurrences, it provides an indication to verify the root cause and ways to eliminate occurrence. Also [29] provides insight into how big data and AI can support improvements. Big data and AI are improving the lessons learned process by utilizing as much data as possible and using them correctly based on the right analysis.

5.5 Benefits to Decision-Making and Risk Management

Past recorded experiences will support the PM in making fast decisions since they helps in dealing with similar situations, anticipating possible risks, and effectively and efficiently managing the project so that teams do not make the same mistakes. General feedback obtained as managing project knowledge includes the effective recording of lessons learned and project experiences that will speed up decision-making by better dealing with similar situations, anticipating possible risks & previous solutions, and effectively managing the project in an efficient way.

For example, if an issue occurred while the project was implemented and already included in the register, then fast response and proper action will take place to put the project back on track, which will take longer if it was not registered and correction actions were not identified. Only two interviewees linked lessons learned as an assist; [27] [30] elaborated more about how OPA can have better progress that will support PM to make a plan for eliminating risks or encouraging opportunities. Two candidates talked about how effective recording and managing project knowledge, including effective recording of lessons learned and project experiences will speed the decision-making, help in dealing with similar situations, anticipate possible risks and previous solutions, and effectively efficiently manage the project [27] [28].

6 Conclusions

Managing project knowledge is a vital cyclic process; hence, it helps use available data or create new knowledge. It stitches together project manager knowledge and experience acquired from the past, present, and future. This knowledge acts as a framework and is very powerful for decision-making and risk management. Knowledge can be obtained as explicit or tacit, but the project manager needs to ensure that tacit knowledge is converted as explicit to have it in codified matter for further utilization or analysis. Such knowledge provides a way to treat knowledge as an asset to transfer data, noting that a good knowledge repository is vital for better decision-making, reduced risks, and continuous improvement in project performance. So, we can say having knowledge of previous projects is key to future project success. Therefore, the project manager needs to review available lessons learned at the beginning of any project and transfer such knowledge to the project team either by eliminating threats or encouraging opportunities.

The project manager plays an important role in the lessons learned process since it starts from day one, so corrective actions must be taken immediately. These lessons are observed and identified by staff handling project implementation or any stakeholders. After identifying the concern, it will be documented under the project lessons learned template, and the project manager will start the analysis with the support of a subject matter expert. After that, it is very important that corrective actions take place to proceed further, and once the concern is cleared and finalized, the register will be updated for upcoming projects. Furthermore, the project manager should encourage the improvement culture and ensure the team members are motivated to add their values and input for further improvement. The project manager can also take management support to create such a culture and to move the organization towards improvement and a trustworthy culture.

It is important for a project manager to initiate and conduct lessons learned sessions during the project life cycle, not only at the end, to obtain the following outcome: New and improved ways to achieve/implement similar projects, new processes or frameworks of improvement, becoming as a reference for upcoming similar projects and set benchmarks for future projects. Additionally, continuous improvement is essential for better work practices. For lessons learned processes, tools, and techniques can be utilized using project management information systems (PMIS), which can be used effectively for project planning and monitoring of phases. Also, PMIS system can capture and provide an indication for abnormality, which also provides a guideline to overcome issues added from previous experiences.

Primavera is one of the good examples of PMIS, that can increase satisfaction; hence, results showed that the level of satisfaction for project management information systems was above 60% in Western Europe and the Middle East and 72% in Africa and the highest recorded industry satisfaction found within information system (79%). Finally, it must be emphasized that it is very important to integrate all systems and tools for knowledge management to maximize the benefit.

7 Research Limitations

One of the primary limitations of cross-sectional designs is that they cannot establish causality or infer the direction of relationships between variables. They only provide a snapshot of associations at one point in time. Additionally, they may not capture changes that occur over time. Furthermore, due to time constraints only sis subject matter experts were interviewed and this might limit the views, opinions, and suggestions offered. Moreover, another limitation lies in the geographical scope of the sample, which was exclusively drawn from the United Arab Emirates (U.A.E.). As a result, the findings may not be generalizable to broader or more diverse populations outside the U.A.E.

8 Future Research Recommendations

Researchers may want to consider expanding the scope of their study in future research efforts to encompass a more diverse range of geographical locations and cultural contexts for a more comprehensive understanding of the topic. It is also recommended that future research should capture a wider range of industries.

References

[1] I. Lurie, "Information is free. Knowledge is not," Available at: https://www.portent.com/blog/random/information-is-free-knowledge-is-not.htm, (2012).

[2] S. Surbhi, "Difference Between Information and Knowledge," Available at: https://keydifferences.com/difference-between-informa-tion-and-knowledge.html, (2018).

[3] I. Nonaka, "A Dynamic Theory of Organizational Knowledge Creation," *Organization Science*, 5(1):14-37, 1994.

[4] PMI, *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*, (5th ed.), Project Management Institute, Pennsylvania, USA, (2013a).

[5] Y. Karagoz, A. Korthaus, and N. Augar, "How Do ICT Project Managers Manage Project Knowledge in the Public Sector? An Empirical Enquiry from the Victorian Public Sector in Australia," *Australasian Journal of Information Systems*, 20:1-20, 2016.

[6] M. Colombi, "Top 5 Reasons Why Knowledge Management is Necessary" Available at; https://www.askspoke.com/blog/knowledge-management/knowledge-management-importance, 2016.

[7] Morten, "A.R.I.T.P., To Cite this Document: Asbjørn Rolstadås Iris Tommelein Per Morten Schiefloe Glenn Ballard," Understanding Project Success Through Analysis of Project Management Approach," *International Journal of Managing Projects in Business*, 7(4):638-660, 2014.

[8] D. Linman, "Project Knowledge Management – The Importance of Managing Project Knowledge," Available at; https://mymanagementguide.com/project-knowledge-management-the-importance-of-managing-project-knowledge/, 2018.

[9] J. R. Meredith and S. J. Mantel Jr, *Project Management: A Managerial Approach (7 utg.)*," John Wiley & Sons, Inc., Hoboken, New Jersey, 2010.

[10] Guide, P.M.B.O.K., "A Guide to the Project Management Body of Knowledge: PMBOK® Guide," *Int. J. Prod. Res*, 53:13-14, 2017.

[11] A. Rolstadås, I. Tommelein, P. Morten Schiefloe, and G. Ballard, "Understanding Project Success Through Analysis of Project Management Approach," *International Journal of Managing Projects in Business*, 7(4):638-660, 2014.

[12] S. F. Rowe and S. Sikes, "Lessons Learned: Taking it to the Next Level," Paper Presented at PMI® Global Congress 2006—North America, Seattle, WA., Newtown Square, PA: Project Management Institute, 2006.

[13] S. A. Trevino and V. S. Anantatmula, "Capitalizing from Past Projects: The Value of Lessons Learned," Paper Presented at PMI® Research Conference: Defining the Future of Project Management, Warsaw, Poland, Newtown Square, PA: Project Management Institute, 2008.

[14] S. F. Rowe, "Applying Lessons Learned," Paper Presented at PMI® Global Congress 2008—EMEA, St. Julian's, Malta, Newtown Square, PA: Project Management Institute, (2008).

[15] T. Williams, "How Do Organizations Learn Lessons from Projects—and Do They?" 55(2):248-266, 2008, [Online], Available at: https://www.researchgate.net/publication/3077089_How_Do_Organizations_Learn_Lessons_From_Projects-And_Do_They [Accessed 7

July 2019].

[16] P. Wyrozebski and R. Pawlak, "The Role and Meaning of Lessons Learned in Project Knowledge Management in Organizations in Poland," *Procedia Computer Science*, 192:2396-2405, 2021.

[17] F. Wijnhoven, "Operational Knowledge Management: Identification of Knowledge Objects, Operation Methods, and Goals and Means for the Support Function," *Journal of the Operational Research Society*, 54(2):194-203, 2003.

[18] A. Labib and M. Read, "Not Just Rearranging the Deckchairs on the Titanic: Learning from Failures Through Risk and Reliability Analysis," *Safety Science*, 51(1):397-413, 2013.

[19] V. Desai, "Constrained Growth: How Experience, Legitimacy, and Age Inuence Risk taking in Organizations," *Organ Sci*. 19:594-608, 2008.

[20] A. Hartmann and A. Dorée, "Learning Between Projects: More than Sending Messages in Bottles," *International Journal of Project Management*, 33(2):341-351, 2015.

[21] V. Bartsch, M. Ebers, and I. Maurer, "Learning in Project-based Organizations: The Role of Project Teams' Social Capital for Overcoming Barriers to Learning," *International Journal of Project Management*, 31(2):239-251, 2013.

[22] A. Mar, "9 Project Knowledge Management Best Practices," Available at: https://management.simplicable.com/management/new/9-project-knowledge-management-best-practices, (2013).

[23] J. Rowley, "6th Edition PMBOK® Guide–Process 4.4 Manage Project Knowledge: Inputs," Available at: https://4squareviews.com/2018/02/03/6th-edition-pmbok-guide-process-4-4-manage-project-knowledge-inputs, 2018.

[24] L. Raymond and F Bergeron, "Project Management Information Systems: An Empirical Study of Their Impact on Project Managers and Project Success," 26(2):213-220, 2008, Available at: https://www.researchgate.net/publication/222572029_Project_management_information_systems_An_empirical_study_of_their_impact_on_project_managers_and_project_success [Accessed 7 July 2019].

[25] M. A. B. Ilyas, M. K. Hassan, and M. U. Ilyas, "PMIS: Boon or Bane?" Paper Presented at PMI® Global Congress 2013—EMEA, Istanbul, Turkey, Newtown Square, PA: Project Management Institute, 2013.

[26] M. Chaves and G. Veronese, "A Proposal to Manage Lessons Learned in Projects: Web 2.0 Technologies to Promote Innovation," Available at: https://www.researchgate.net/publication/307850646_A_Proposal_to_Manage_Lessons_Learned_in_Projects_Web_20_Technologies_to_Promote_Innovation, 2014.

[27] Y. Samir, "Interview with Yusuf Samir from Etisalat," July-2019.

[28] B. Idicula, "Interview with Blesson Idicula from Etisalat," July-2019.

[29] L. Elsin, "Interview with Dr, Loay Elsin from Abu Dhabi Chamber of Commerce & Industry (ADCCI)", July-2019.

[30] M. Ramadan, "Interview with Mohammed Ramadan from Etisalat," July-2019.

[31] K. Collet, "Interview with Kudze Collet from Etisalat," July-2019.

[32] J. Jacob, "Interview with Jimin Jacob from Etisalat," July-2019.

[33] S. Petter and V. Vaishnavi, "Facilitating Experience Reuse Among Software Project Managers," *Information Sciences*, 178(7):1783-1802, 2008.

[34] M. Schindler and M. J. Eppler, "Harvesting Project Knowledge: A Review of Project Learning Methods and Success Factors," *International Journal of Project Management*, 21(3):219-228, 2003.

[35] M. Boulos, I. Maramba, and S. Wheeler, Wikis, "Blogs and Podcasts: A New Generation of Web-based Tools for Virtual Collaborative Clinical Practice and Education," BMC Medical Education, 6(41): **need page numbers** 2006.

[36] S. Cleveland, "Using Microblogging for Lessons Learned in Information Systems Projects," International Research Workshop on IT Project Management, pp. 122-128, 2012.

Ahmed Al-Nakeeb (photo not available) received his education in the United Kingdom and has a Ph.D. in Quality Management. He worked at various Universities in the U.K. and the UAE and taught various courses. He held several positions, such as Campus Coordinator (Deputy Dean), Head of Department, Department Chair, and Program Chair. He led the development of various undergraduate and postgraduate programs and developed curricula. Participated in various international conferences and published in well-known international journals.

Mounir El khatib (photo not available) has a long and varied experience in Digital Government and Digital Transformation in UK, Middle East, and UAE; worked as practitioner and in academics; in focused areas like Project and Program Management, Information Systems, Strategy and Operations, Business process and service management, providing teaching, training, consulting, advising, research and professional services.

Raed Abu Zitar (photo not available) (BS 88, MS 90, and PhD 93) is a senior research scientist in the context of Thales Endowed Chair of Excellence, Sorbonne Center of Artificial Intelligence (SCAI), Sorbonne University- Abu Dhabi (SUAD), UAE. Prof. Abu Zitar has joined SCAI-SUAD in Feb-2021. Prof. Abu Zitar is directing the research at Thales/SCAI in the fields of Intelligent detection and tracking of drones in cooperation with industrial experts from Thales and Aveillant (France and the UK). Prior to

that date, he was the founding coordinator for the Master of Artificial Intelligence Program (one of the first in UAE), Head of the Center of Artificial Intelligence, and Manager for the Teaching and Learning Center at Ajman University, UAE. He has more than 30 years of experience in teaching, research, and administration in the USA, Jordan, and the UAE. He has published more than 100 Journal and conference papers.

Abdulrahman Alhosani (photo not available) is an experienced Senior Manager @ Etisalat by e& with a demonstrated history of working in the telecommunications industry. Skilled in Operations Management, Access Network Deployment and Network Modernization. In addition to capabilities in Process Design, Partner Management, Digitalization Project & Data Analysis. Strong Engineering knowledge with a Bachelor focused in Applied Science in Engineering Management from Higher Colleges of Technology and a Masters in project management from Hamdan bin Mohamed Smart University.

Ibrahim Alhosani (photo not available) is a Senior Manager/Core Deployment @ Etisalat by e& with more than 20 years of experience in project implementation (Telecommunication industry). Handling various types of Core Networks Projects from its initiation till completion. Holds a Bachelor in Applied Science in Engineering Management from Higher Colleges of Technology and Masters in Project Management from Hamdan bin Mohamed Smart University.

Appendix

Lesson Learn Register Real Example

Source: Interview with MR. Yusuf Samir

Section: Transport and Backbone Projects

Lessons Learned Register			
Project Name	Data Center Modernaization	Project Manager Name	Yusuf Samir
Project Description	Data Center Modernaization project to meet 5G requirement of Buisness customers		
Lessons Learned	Resources	Severity Levels	Meduim
	Cost		High

ID	Situation			Identified On	Highlighted By	Severity	Impacts	Recommendations	Remarks
	Category	Project Phase	Description						
1	Resources	Implementation	Process of spare handover to maintenance team (part of project handover) intend to accept first by project in their store and then they need to handover to maintenance.	03/01/2019	Ali	Medium	1. Duplication of resources allocation, for 2 days 2. Duplication of works, like verification of Delivery note	1. Process to be changed to have direct delivery to maintenance with presence of project team. 2. To be agreed with all stake holders and then process to be modified.	Closed- 25/03/2019
2	Cost	Implementation	Existing spares process leading for un necessarily cost, by allocating staff for same task in 2 days instead of 1 as well as there could be a need for helpers to shift these spares from project to maintenance store	03/01/2019	Ali	High	1. Duplication of resources allocation, for 2 days 2. Un necessarily cost for helper charges and other cost encountered to shift the items	As mentioned above, such extra time can be eliminated by direct delivery since no cost will be encountered on project team.	Closed- 25/03/2019
3	Resources	Implementation	Missing installation materials	15/05/2019	Salem	High	Affecting project completion as per given time and scope.	BOQ to be verified during planning phase and to be compared with design doc.	Updated for upcoming projects

Managing Risks in the Adoption of Cybersecurity Technology in Manufacturing Enterprises: Identification and Assessment

Le Vinh Quang[*,†]
[1] Eastern International University, Binh Duong New City, VIETNAM
Industrial University Ho Chi Minh City, Ho Chi Minh City, VIETNAM

Tran Huu Đuc[‡], Narayan Chandra Debnath[§]
Eastern International University, Binh Duong New City, VIETNAM

Nguyen Ngoc Long[†]
Industrial University Ho Chi Minh City, Ho Chi Minh City, VIETNAM

Abstract

The adoption of cyber security technology has become an urgent challenge, particularly in the context where manufacturing enterprises increasingly rely on connectivity technologies and digitalization processes to enhance operational efficiency. However, the adoption of cybersecurity solutions poses numerous risks that require identification and assessment. This research aims to pinpoint the risks associated with the effective adoption of cybersecurity solutions in manufacturing business operations and employs the Analytic Hierarchy Process (AHP) for risk evaluation. The analysis results provide profound insights into investment priorities for cybersecurity technology and mitigation strategies within the manufacturing industry. The study focuses on analyzing risks related to strategy, organization, technology, finance, and human factors. This research contributes to existing knowledge by addressing risks in cybersecurity adoption within the manufacturing sector. These findings offer practical guidance for manufacturing organizations seeking to enhance the effectiveness of their cybersecurity deployment, allowing them to safeguard critical assets, ensure uninterrupted production processes, and protect sensitive information.

Key Words: Cybersecurity, risk management, manufacturing industry, average analytic hierarchy process, AHP.

1 Introduction

The manufacturing industry has progressively embraced digital technologies and interconnected systems, offering huge opportunities for growth and efficiency. But this shift to digitalization also brings threats and risks, particularly concerning cybersecurity (Volberda et al., [38]). Protecting vital assets, sensitive data, and guaranteeing uninterrupted production processes are crucial considerations as industrial businesses adopt Industry 4.0 technologies and digitize their operations (Ahmed et al., [1]). Given the potential ramifications of successful attacks on the industry's infrastructure and operations, the need to protect against cyber threats has never been more critical (Gunduz & Das, [18]).

As the business environment grows more complex concerning cybersecurity, risks in the manufacturing industry also increase (Yeboah-Ofori & Islam, [43]). Attackers continually develop new techniques to circumvent defenses as new technologies emerge and current ones are improved (Zlomislić et al., [45]). Due to the dynamic nature of cyber threats, cyber security actions must be continuously monitored, updated, and adjusted (Brass & Sowell, [10]), placing pressure on resources and making it important for organizations to be on full alert.

The manufacturing industry is increasingly aware of the importance of cybersecurity (Ani et al., [6]; Wells et al., [39]). However, previous studies have often provided generalized approaches and have sometimes overlooked the complexities of securing industrial control systems, supply chains, and sensitive data within this industry (Cheung et al., [12]; Knowles et al., [23]; Raimundo & Rosário, [27]). While the body of cybersecurity literature grows, there remains a pressing need for more comprehensive research to examine the complexities and risks associated with implementing strong cybersecurity technologies. Specifically, there is a critical need to further investigate the unique risks that the manufacturing sector faces.

This research study aims to fill this important gap by analyzing the adoption of cybersecurity technologies in the manufacturing industry. To assess and prioritize risks when adopting cybersecurity technologies, the Average Analytic Hierarchy Process (AHP) method has been employed. The study analyzes over 25 sub-risks affecting the adoption of cybersecurity solutions in manufacturing firms. The findings

* Becamex Business School.
† Business Administration Faculty .Email: quang.le@eiu.edu.vn,
, nguyenngoclong@iuh.edu.vn.
‡ Administration Assistant team. Email: duc.tran@eiu.edu.vn.
§ School of Computing and Information Technology . Email: narayan.debnath@eiu.edu.vn.

reveal that organizational risks are the most critical, followed by strategy, technology, finance, and human risks. The Average AHP methodology enables a quantitative examination of the importance and impact of each risk, providing useful information regarding priorities for cybersecurity investments and mitigation solutions.

This research makes significant contributions to the field of cyber security adoption within manufacturing enterprises. In an era where manufacturing increasingly relies on connectivity technologies and digitalization to boost operational efficiency, the pressing challenge of cybersecurity adoption looms large. To address this challenge, we aimed to pinpoint and assess the inherent risks in adopting cybersecurity solutions within the manufacturing sector, employing the Analytic Hierarchy Process (AHP) for precise risk evaluation.

The paper provides a literature review in Section 2 followed by research methodology in Section 3. A detailed explanation of the results is included in Section 4. Finally, the discussion and conclusion are presented in Section 5.

2 Literature Review

The adoption of cybersecurity technology is critical for organizations to protect their digital assets and sensitive data in an increasingly interconnected world. However, this process has challenges, as various risks can hinder its successful implementation. This literature review explores key research areas related to the identification and assessment of risks that organizations face during the adoption of cybersecurity technologies.

Several research studies have been conducted to investigate the risks that firms experience while adopting cybersecurity measures. One primary challenge is the technical complexities associated with implementing cybersecurity solutions. Research by Yan et al. [42] highlights that organizations often encounter difficulties in integrating these technologies seamlessly into their existing IT infrastructure, leading to compatibility issues and potential disruptions (Kimani et al., [22]). Financial considerations play a significant role in cybersecurity technology adoption. The financial burden includes initial implementation costs and ongoing maintenance (Argaw et al., [17]; Kabanda et al., [20]). Regulatory requirements and compliance standards can limit technology adoption. Brass & Sowell [10] discuss the challenges organizations encounter in aligning cybersecurity technologies with evolving regulatory frameworks, emphasizing the need for continuous monitoring and adaptation. A shortage of skilled cybersecurity professionals can hinder technology adoption efforts. Ghobakhloo et al. [17] highlight the importance of addressing the skills gap and ensuring organizations have the expertise required for effective technology deployment. The literature highlights a range of challenges and risks organizations encounter during the adoption of cybersecurity technologies. Understanding and addressing these risks is essential for successful technology implementation and improved cybersecurity posture.

Despite some existing research on risk assessment in the adoption of cybersecurity technology, there is a need for a comprehensive study that identifies and evaluates the full spectrum of key risks and sub-risks, transcending beyond the realms of finance, technology, or human factors. As a result, the purpose of this study is to identify and thoroughly analyze these risks using the research methodology presented by Kabra et al. [21], originally used to highlight challenges in the implementation of digital technologies.

The Analytical Hierarchy Process (AHP) is a well-known decision-making process that offers a systematic strategy for evaluating difficult situations with various criteria (Albayrak & Erensal, [4]). In the context of the risks associated with implementing cybersecurity technology in the manufacturing business, AHP provides a useful framework for quantitatively measuring the relative relevance of various risks. Discussing with experts and using AHP enables developing a prioritized list of risks, helping manufacturers allocate resources to the most critical issues. AHP is especially useful for evaluating problems in Information and Communication Technology (ICT) applications because it gives you a structured way to figure out how important these problems are based on several criteria (Kabra et al., [21]). However, a scarcity of research utilizing the Analytical Hierarchy Process (AHP) method for the analysis of cybersecurity implementation issues persists.

3 Research Methodology

This study utilizes the Average Analytic Hierarchy Process (AAHP) method to evaluate and prioritize the key risks influencing cybersecurity technology adoption in manufacturing enterprises.

3.1 Risk Identification

An extensive literature review of scholarly articles was first conducted to identify an initial set of cybersecurity adoption risk factors. These were mapped to the framework proposed by Kabra et al. [21] which categorizes risks into strategic, organizational, human, financial, and technological categories. Additional relevant risks were added based on the literature.

Input from a panel of 10 subject matter experts was then gathered through interviews to validate and refine the identified risks. The experts included CEOs of IT services companies, manufacturing company IT department heads, and university professors specializing in cybersecurity and information systems. Their insights helped modify the risk framework and ensure its applicability to the manufacturing sector.

3.2 AHP and Average AHP

Analytic Hierarchy Process (AHP) is a Multi-criteria decision-making (MCDM) technique proposed by (Wind & Saaty, [41]), which involves pairwise comparisons of multiple criteria to determine their relative importance. The AHP approach applied in this research consists of the following main steps:

- AHP Model Development: a hierarchical model was developed with the overall goal of evaluating cybersecurity adoption risks at the top level. The next level consisted of the 5 main risk categories. The lowest level contained the 25 sub-risk factors identified through the literature review and expert input.
- Pairwise Comparison Matrices: A questionnaire was designed to gather expert assessments of the relative importance of risks through AHP pairwise comparisons. Utilize a scale ranging from 1 (equal importance) to 9 (significantly more important) for this purpose and gather the resulting data.
- The eigenvector method was utilized to calculate the local weights of the sub-risks from the comparison matrices (Wind & Saaty, [41]). Global weights were obtained by multiplying local weights by the weights of their parent criteria.
- Consistency Verification: The consistency ratio (CR) was calculated to verify judgment consistency. CR ratios exceeding 0.10 indicated inconsistent comparisons. Any inconsistent matrices were discarded.

These steps will enable the calculation of weights for criteria, sub-criteria, and options based on the results of pairwise comparisons. In this research, we suggested the percentage scale (from 1 to 9) for the pair-wise comparison shown in Figure 1.

AHP comparison tables are generated for each expert to compute the relative weights of criteria, which are referred to as AHP values. If there are numerous experts (n experts), each expert's assessment (AHPEi) is performed, and the average AHP value of the experts (A(AHP)) is determined using the arithmetic mean, as shown in Equation (1).

$$A(AHP) = \frac{\sum_{i=1}^{n} AHPE_i}{n} \tag{1}$$

The AAHP risk weights were analyzed to determine priorities for cybersecurity adoption. A higher weight indicates a higher priority risk factor. This AAHP methodology enables a quantitative, systematic multi-expert evaluation of the key cybersecurity adoption risks for manufacturing enterprises.

4 Results

4.1 Identifying Challenges in Cyber Security Solutions Adoption

This study created a multi-hierarchical structure of cybersecurity adoption risks based on the technological adoption challenges stated by (Kabra et al., [21]). This was strengthened by incorporating recent literature insights and input from ten experts, including IT company CEOs, manufacturing IT managers, and ICT professors. The risks associated with cyber security technology advice will be classified into two tiers, as shown in Table 1.

4.2 Prioritization of the Cybersecurity Solutions Adoption with A(AHP)

The A(AHP) questionnaire was designed for evaluating and prioritizing the five key risk factors and their related sub-factors. As previously stated, data was gathered using the A(AHP) questionnaire. The survey included the participation of ten cybersecurity experts. The experts were instructed to use numerical scales ranging from 1 to 9 while making their choices to assess the priority of implementing cybersecurity . The questionnaire, which included a decision-making process and paired comparisons, took each participant between forty and fifty minutes to complete. Table 2 shows an example of paired criterion comparisons for a specific target defined by expert No.1. Meanwhile, Figure 1 depicts the attribution of importance on a scale of 1 to 9. As transversal values, the reciprocal values of these significance ratings were employed (aij = 1/aji). According to Expert No. 1, Strategy was three times more significant than Organization, providing a transversal value of one-third. Organization, on the other hand, was deemed three times less significant than Strategy, resulting in a reciprocal ratio of aij = 1/aji.

We meticulously developed a pairwise comparison matrix for the numerous risk variables in Table 3. This matrix was generated by dividing each element within it by the total of its column values. To better demonstrate this technique, consider an individual item in the matrix, such as 0.560. This number was calculated by dividing 1 (from Table 2) by the cumulative sum of column values for that specific entry, which is the sum of 1.00, 0.33, 0.11, 0.14, and 0.20, for a total of 1.79 (from Table 2).

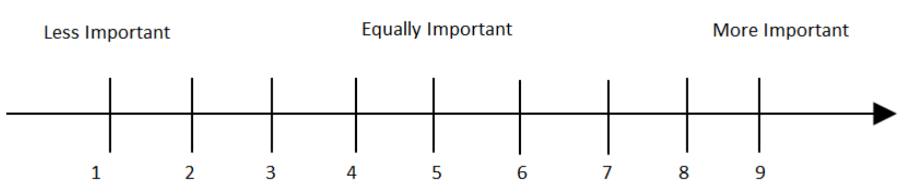


Figure 1: Shows the percentage scale (1–9) for pair-wise comparisons

Table 1: Identification of risk factors from the previous literature

Criteria	Sub-risks	Relevant studies
Main risks	Strategic risks	Gcaza & von Solms [16]
	Organizational risks	Kabanda et al. [20]
	Human risks	Bowen et al. [9]
	Financial risks	Kabanda et al. [20]
	Technological risks	De Bruijn & Janssen [13]
Strategy (S)	Lack of policies to adopt technology	Stewart & Jürjens [32]
	Inadequate policy awareness and support from the government	Kraemer et al. [24]
	Lack of management vision	Tsohou et al. [37]
	Lack of cross-organization development program	Zissis & Lekkas [44]
	Lack of supply chain understanding	Yeboah-Ofori & Islam [43]
Organization (O)	Conflicting short-term focus goal-oriented culture	Zissis & Lekkas [44]
	Not inviting end-user input	Zissis & Lekkas [44]
	Lack of cybersecurity personnel	Szczepaniuk et al. [33]
	Lack of pressure from other organizations	Singh & Alshammari [30]
	Lack of transparency in the utilization of funds	Sirisha et al. [31]
Human (H)	Lack of skills to use cybersecurity solutions	Tam et al. [34]
	Lack of education and training for the employees	Akter et al. [2]
	Lack of benchmarking about the knowledge of cybersecurity solutions	Holstein et al. [19]
	Workforce resistance to change	P. Kumar et al. [26]
	Lack of motivation to use cybersecurity solutions	Wessels et al. [40]
Finance (F)	Donor's support	Chang & Coppe [11]
	Lack of funds for investment in technology	Fielder et al. [15]
	High Cost	Alsuwian et al. [5]
	Competition for funding	Balon & Baggili [8]
	Fundraising expenses	Eusanio & Rosenbaum [14]
Technology (T)	Lack of awareness about exact technological solutions	Alahmari & Duncan [3]
	Lack of cybersecurity solutions enabling infrastructure	A. Kumar [25]
	Lack of customization	Alahmari & Duncan [3]
	Frequent updates of technology	Tawalbeh et al. [35]
	Incompatibility in cybersecurity facilities linked with different organizations	Ani et al. [6]

Table 2: Pairwise comparison matrix of expert No. 1's decision criteria (Risk factors) with respect to the goals

Criteria	S	O	H	F	T
S	1.00	3.00	9.00	7.00	5.00
O	0.33	1.00	7.00	5.00	3.00
H	0.11	0.14	1.00	0.33	0.20
F	0.14	0.20	3.00	1.00	1.00
T	0.20	0.33	5.00	1.00	1.00

Table 3: Priorities of main risk factors

	S	O	H	F	T	Weight	Rank
S	0.560	0.642	0.360	0.488	0.490	0.508	1
O	0.187	0.214	0.280	0.349	0.294	0.265	2
H	0.062	0.031	0.040	0.023	0.020	0.035	5
F	0.080	0.043	0.120	0.070	0.098	0.082	4
T	0.112	0.071	0.200	0.070	0.098	0.110	3

After completing the matrix computation, we calculated the row averages, which are critical in calculating the Eigenvectors or relative weights of the criterion, in this case, the risk factors listed in Table 3. To illustrate, consider the relative importance of a strategic problem. This weight was determined by adding the numbers in each row, namely 0.560, 0.642, 0.360, 0.488, and 0.490, and then dividing the total number of challenge elements or criteria, which in this case is 6. As a consequence, a strategic challenge's relative weight was judged to be 0.508. In this manner, we meticulously established a framework for assessing the significance and relative importance of various risk factors, providing a valuable tool for decision-makers and experts in the field to prioritize and address these factors effectively.

We employed Saaty [29] criteria to evaluate the consistency of the comparison matrix, and in doing so, we determined both the Consistency Index (C.I.) and the Consistency Ratio (C.R.). The C.I. was computed using the formula C.I. = (max - n) / (n - 1), where 'max' represents the largest eigenvalue of the pairwise comparison matrix, and 'n' is the number of criteria being compared. This was a crucial step in ensuring the validity of our assessment. Furthermore, the C.R. was determined by dividing the C.I. by the Random Consistency Index (R.I.), which is a pre-established value. For a five-by-five matrix, the appropriate R.I. value is set at 1.12, as indicated in Table 4 for reference. In our evaluation, it was imperative to consider that an assessment is deemed satisfactory if the C.R. does not exceed the threshold of 0.10 (10 percent), as stipulated by Saaty [29]. With these procedures in mind, we meticulously calculated the C.R. using the prescribed technique, and the resulting C.R. was found to be 0.072348987. This value did not surpass the 0.10 (10 percent) criterion. Consequently, we can conclude that the judgments provided by the experts were relatively consistent, instilling confidence in the appropriateness of these criteria for making informed decisions.

Ten experts utilized the A(AHP) approach to assess each of the adoption risks in Table 1 to itself. The average score was used to establish the priority of each adoption risk level, as indicated in Table 5. The arithmetic mean of the experts' A(AHP) values was used to get the AAHP value. To show, the average AHP of the strategic risk in the table was determined by adding the rows (0.508+ 0.508+ 0.529+ 0.264 + 0.035 + 0.070 + 0.508 + 0.260 + 0.278 + 0.236), providing a result of 0.320 and placing second. Table 6 highlights the findings of all risk factors' priority, including major and sub-risks.

5 Discussion

The primary objective of this study was to identify and assess the key risks influencing the adoption of cybersecurity technology. After conducting an extensive review of the literature on cybersecurity adoption risks, a total of 25 risk factors were identified. Subsequently, a survey was designed to gather input from professionals regarding the impact of these risks. An Analytic Hierarchy Process (AHP) model was constructed to analyze the data.

The AHP analysis revealed that among the main criteria, organizational risks emerged as the most significant challenges in the adoption of cybersecurity solutions. Within organizational risks, the absence of pressure from external organizations mandating the use of cybersecurity is a high-ranked risk factor that can delay this process. This outcome suggests that the lack of external pressure may lead to a reduced sense of urgency in adopting cybersecurity solutions. Additionally, insufficient motivation from peer organizations can undermine the prioritization and importance of cybersecurity implementation. These findings align with the research by Kabanda et al. [20], who highlight that external factors play a reinforcing role in the limited adoption of cybersecurity practices.

Table 4: Average random consistency index (R.I.)

n	1	2	3	4	5	6	7	8	9	10
R.I	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49
n	11	12	13	14	15					
R.I	1.51	1.48	1.56	1.57	1.58					

Table 5: Average of the AHP values of the experts for major risks

	Criteria weights											
	RES 1	RES 2	RES 3	RES 4	RES 5	RES 6	RES 7	RES 8	RES 9	RES 10	Average	Rank
SR	0.508	0.508	0.529	0.264	0.035	0.070	0.508	0.260	0.278	0.236	0.335	2
OR	0.265	0.264	0.185	0.505	0.503	0.502	0.264	0.503	0.410	0.448	0.374	1
HR	0.035	0.070	0.038	0.137	0.068	0.039	0.070	0.035	0.064	0.069	0.061	5
FR	0.082	0.038	0.077	0.045	0.134	0.223	0.038	0.068	0.098	0.101	0.088	4
TR	0.110	0.120	0.170	0.049	0.260	0.166	0.120	0.134	0.150	0.147	0.141	3

Table 6: Average of the AHP values of the experts for major risks

Criteria	Weights	Ranks
Main Factors		
Strategic Risk	0.320	2
Organizational Risk	0.385	1
Human Risk	0.062	5
Financial Risk	0.090	4
Technological Risk	0.143	3
Sub-Challenge Factors (Strategic)		
Lack of policies to adopt technology	0.162	4
Inadequate policy awareness and support from government	0.095	5
Lack of management vision	0.208	3
Lack of cross-organization development program	0.285	1
Lack of supply chain understanding	0.249	2
Sub-Challenge Factors (Organizational)		
Conflicting short-term focus goal-oriented culture	0.218	2
Not inviting end-user input	0.153	5
Lack of cybersecurity personnel	0.205	3
Lack of pressure from other organizations	0.224	1
Lack of transparency in the utilization of funds	0.200	4
Sub-Challenge Factors (Human)		
Lack of skills to use cybersecurity	0.257	2
Lack of education and training to the employees	0.169	4
Lack of benchmarking about the knowledge of cybersecurity	0.179	3
Workforce resistance to change	0.273	1
Lack of motivation to use cybersecurity	0.122	5
Sub-Challenge Factors (Financial)		
Donors support	0.205	2
Lack of funds for investment in technology	0.345	1
High Cost	0.115	5
Competition for funding	0.150	4
Fundraising expenses	0.185	3
Sub-Challenge Factors (Technological)		
Lack of awareness about exact technological solutions	0.226	2
Lack of cybersecurity enabling infrastructure	0.178	4
Lack of customization	0.145	5
Frequent updates of technology	0.261	1
Incompatibility in cybersecurity facilities linked with different organizations	0.190	3

In contrast to research focused on technology usage in manufacturing, one of the significant implementation barriers is related to financial or strategic risks (Kabra et al., [21]). Nonetheless, the current investigation indicates that financial challenges are ranked as only the fourth most significant. This suggests that the market provides a wide range of cybersecurity solutions, indicating that financial limitations may not serve as a major impediment for businesses when it comes to adopting these solutions. Most surprisingly, risks associated with human factors are considered the least prioritized which contradicts the findings of (Triplett, [36]), who highlighted humans as the weakest link in data security. Triplett identified specific careless and unintentional behaviors that were made worse by the lack of awareness among both leaders and employees.

In terms of technical risks, the findings of this study closely reflect previous studies undertaken in developing-country small and medium-sized firms (SMEs). Frequent updates of technology are ranked as the number one criterion in assessing technological risks. SMEs generally use simpler systems and infrequently update software and technology, which may make it difficult to implement rigorous protections (Rawindaran et al., [28]).

6 Conclusions, Limitations and Future Work

In conclusion, this study has shed light on the critical risks influencing the adoption of cybersecurity technology in organizations. Through an extensive literature review and a comprehensive survey, we identified 25 key risk factors and employed the AAHP to analyze their significance. Our findings highlight that organizational risks specifically the lack of external pressure mandating cybersecurity and the absence of motivation from peer organizations, pose significant challenges in the adoption of cybersecurity solutions. These results underscore the importance of external influences in shaping cybersecurity practices within organizations.

While this study provides useful insights into cybersecurity adoption risks, some limitations exist. Firstly, the sample of experts consulted was small at only 10 participants. A larger and more diverse expert panel could validate the results further. Secondly, the study focused solely on the manufacturing industry. Expanding the research across other sectors could reveal additional risks and challenges. Finally, the AHP technique has some shortcomings in terms of potential inconsistencies. Utilizing other multi-criteria decision-making methods like ANP could strengthen the analysis.

According to the findings, the pressure from external organizations is insufficient to persuade manufacturing enterprises to embrace cybersecurity solutions. Future investigations should further examine the role of external pressures from organizations and stakeholders in driving the urgency of implementing cybersecurity solutions. Furthermore, a shortage of funds for technological investment is viewed as a risk factor impacting this process. As a result, the recommendation for future study is to concentrate on solutions that may have a greater impact on focused companies and to give financing assistance to manufacturing organizations to improve cybersecurity in the present digital environment. Additionally, investigating differences in risk priorities across company sizes and developing vs. developed countries could provide more nuanced insights. Lastly, strategies for fostering collaboration among peer organizations to collectively elevate the importance of cybersecurity implementation and facilitate the sharing of best practices should be considered.

Acknowledgments

This research is financially supported by Eastern International University, Binh Duong Province, Vietnam.

References

[1] S. F. Ahmed, M. S. Bin Alam, M. Hoque, A. Lameesa, S. Afrin, T. Farah, M. Kabir, G. M Shafiullah,.and S. M. Muyeen, "Industrial Internet of Things Enabled Technologies, Challenges, and Future Directions, " *Computers and Electrical Engineering*, 110:16 pp., July 2023, https://doi.org/10.1016/j.compeleceng.2023.108847.

[2] S. Akter, M. R. Uddin, S. Sajib, W. J. T. Lee, K. Michael, and M. A. Hossain,. "Reconceptualizing Cybersecurity Awareness Capability in the Data-Driven Digital Economy," *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04844-8, 2022.

[3] A. Alahmari and B. Duncan, "Cybersecurity Risk Management in Small and Medium-Sized Enterprises: A Systematic Review of Recent Evidence," *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, Cyber SA 2020*, *March 2021*. https://doi.org/10.1109/CyberSA49311.2020.9139638.

[4] E. Albayrak and Y. C. Erensal, "Using Analytic Hierarchy Process (AHP) to Improve Human Performance: An Application of Multiple Criteria Decision Making Problem," *Journal of Intelligent Manufacturing*, 15(4):491-503, 2004, https://doi.org/10.1023/B:JIMS.0000034112.00652.4c.

[5] T. Alsuwian, A. Shahid Butt, and A. A. Amin, "Smart Grid Cyber Security Enhancement: Challenges and Solutions—A Review," *Sustainability (Switzerland)*, 14(21):1-21, 2022, https://doi.org/10.3390/su142114226.

[6] U. P. D. Ani, H. (Mary) He, and A. Tiwari, "Review of Cybersecurity Issues in Industrial Critical Infrastructure: Manufacturing in Perspective," *Journal of Cyber Security Technology*, 1(1):32-74, 2017, https://doi.org/10.1080/23742917.2016.1252211.

[7] S. T. Argaw, J. R. Troncoso-Pastoriza, D. Lacey, M. V. Florin, F. Calcavecchia, D. Anderson, W Burleson,. J. M. Vogel, C. O'Leary, B. Eshaya-Chauvin, and A. Flahault, "Cybersecurity of Hospitals: Discussing the Challenges and Working Towards Mitigating the Risks," *BMC Medical Informatics and Decision Making*, 20(1):1-10, 2020, https://doi.org/10.1186/s12911-020-01161-7.

[8] T. Balon and I. Baggili, "Cybercompetitions: A Survey of Competitions, Tools, and Systems to Support Cybersecurity Education," *Education and Information Technologies*, pp. 1-33, 2023.

[9] B. M. Bowen, R. Devarajan and S. Stolfo, "Measuring the Human Factor of Cyber Security," *2011 IEEE International Conference on Technologies for Homeland Security, HST 2011*, pp. 230-235, 2011. https://doi.org/10.1109/THS.2011.6107876.

[10] I. Brass and J. H. Sowell, "Adaptive governance for the Internet of Things: Coping with emerging security risks," *Regulation and Governance*, 15(4):1092-1110, 2021. https://doi.org/10.1111/rego.12343.

[11] L. Y. C. Chang and N. Coppel, "Building Cybersecurity Awareness in a Developing Country: Lessons from Myanmar," *Computers & Security*, 97:101959, 2020.

[12] K. F. Cheung, M. G. H. Bell and J. Bhattacharjya, "Cybersecurity in Logistics and Supply Chain Management: An Overview and Future Research Directions," *Transportation Research Part E: Logistics and Transportation Review*, *146*, July 2020, 102217. https://doi.org/10.1016/j.tre.2020.102217.

[13] H. de Bruijn and M. Janssen, "Building Cybersecurity Awareness: The Need for Evidence-Based Framing Strategies," *Government Information Quarterly*, 34(1):1-7, 2017, https://doi.org/10.1016/j.giq.2017.02.007.

[14] J. M. Eusanio and D. J. Rosenbaum, "Technology

Considerations for Not-for-Profit Organizations," *The CPA Journal*, 89(4):13-15, 2019.

[15] A. Fielder, E. Panaousis, P. Malacaria, C. Hankin, and F. Smeraldi, "Decision Support Approaches for Cyber Security Investment," *Decision Support Systems*, 86:13-23, 2016, https://doi.org/10.1016/j.dss.2016.02.012.

[16] N. Gcaza and R. von Solms, "A Strategy for a Cybersecurity Culture: A South African Perspective," *Electronic Journal of Information Systems in Developing Countries*, 80(1):1-17, 2017, https://doi.org/10.1002/j.1681-4835.2017.tb00590.x.

[17] M. Ghobakhloo, M. Iranmanesh, M. Vilkas, A. Grybauskas and A. Amran, "Drivers and Barriers of Industry 4.0 Technology Adoption Among Manufacturing SMEs: a Systematic Review and Transformation Roadmap," *Journal of Manufacturing Technology Management*, 33(6):1029-1058, 2022, https://doi.org/10.1108/JMTM-12-2021-0505.

[18] M. Z. Gunduz and R. Das, "Cyber-Security on Smart Grid: Threats and Potential Solutions," *Computer Networks*, 169, 2020, 107094. https://doi.org/10.1016/j.comnet.2019.107094.

[19] D. Holstein, T. W. Cease and M. G. Seewald, "Application and Management of Cybersecurity Measures for Protection and Control," *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 76-83, 2015.

[20] S. Kabanda, M. Tanner and C. Kent, "Exploring SME Cybersecurity Practices in Developing Countries," *Journal of Organizational Computing and Electronic Commerce*, *28*(3):269-282, 2018. https://doi.org/10.1080/10919392. 2018.1484598.

[21] G. Kabra, A. Ramesh, V. Jain, and P. Akhtar, "Barriers to Information and Digital Technology Adoption in Humanitarian Supply Chain Management: A Fuzzy AHP Approach," *Journal of Enterprise Information Management*, 36(2):505-527, 2023, https://doi.org/10.1108/JEIM-10-2021-0456.

[22] K. Kimani, V. Oduol, and K. Langat, "Cyber Security Challenges for IoT-based Smart Grid Networks," *International Journal of Critical Infrastructure Protection*, 25:36-49, 2019, https://doi.org/10.1016/j.ijcip.2019.01.001.

[23] W. Knowles, D. Prince, D. Hutchison, J. F. P. Disso, and K. Jones, "A Survey of Cyber Security Management in Industrial Control Systems," *International Journal of Critical Infrastructure Protection*, 9:52-80, 2015, https://doi.org/10.1016/j.ijcip.2015.02.002.

[24] S. Kraemer, P. Carayon, and J. Clem, "Human and Organizational Factors in computer and information security: Pathways to vulnerabilities," *Computers and Security*, *28*(7):509-520, 2009, https://doi.org/10.1016/j.cose.2009.04.006.

[25] A. Kumar,. "Cyber Physical Systems (CPSs) – Opportunities and Challenges for Improving Cyber Security," *International Journal of Computer Applications*, 137(14):19-27, 2016, https://doi.org/10.5120/ijca2016908877.

[26] P. Kumar, J. Bhamu, and K. S. Sangwan, "Analysis of Barriers to Industry 4.0 Adoption in Manufacturing Organizations: An ISM Approach," *Procedia CIRP*, *98*: 85–90, March 2021, https://doi.org/10.1016/j.procir.2021.01.010.

[27] R. J. Raimundo and A. T. Rosário, "Cybersecurity in the Internet of Things in Industrial Management," *Applied Sciences (Switzerland)*, 12(3):19 pp., 2022, https://doi.org/10.3390/app12031598.

[28] N. Rawindaran, A. Jayal, E Prakash, and C. Hewage, "Perspective of Small and Medium Enterprise (SME's) and their Relationship with Government in Overcoming Cybersecurity Challenges and Barriers in Wales," *International Journal of Information Management Data Insights*, *3*(2):100191, 2023, https://doi.org/10.1016/j.jjimei.2023.100191.

[29] T. L. Saaty, "Decision Making — the Analytic Hierarchy and Network Processes (AHP/ANP)," *Journal of Systems Science and Systems Engineering*, 13(1):1-35, 2004, https://doi.org/10.1007/s11518-006-0151-5.

[30] H. P. Singh and T. S. Alshammari, "An Institutional Theory Perspective on Developing a Cyber Security Legal Framework: A Case of Saudi Arabia," *Beijing Law Review*, 11(03):637-650, 2020, https://doi.org/10.4236/blr.2020.113039.

[31] N. S. Sirisha, T. Agarwal, R. Monde, R. Yadav, and R. Hande, "Proposed Solution for Trackable Donations using Blockchain," *2019 International Conference on Nascent Technologies in Engineering, ICNTE 2019 - Proceedings*, *Icnte*, pp. 1-5, 2019, https://doi.org/10.1109/ICNTE44896.2019.8946019.

[32] H. Stewart and J. Jürjens, "Information Security Management and the Human Aspect in Organizations," *Information and Computer Security*, 25(5):494-534, 2017, https://doi.org/10.1108/ICS-07-2016-0054.

[33] E. K. Szczepaniuk, H. Szczepaniuk, T. Rokicki, and B. Klepacki, "Information Security Assessment in Public Administration," *Computers and Security*, 90:11 pp. (2020). https://doi.org/10.1016/j.cose.2019.101709.

[34] T. Tam, A. Rao, and J. Hall, "The Good, the Bad and the Missing: A Narrative Review of Cyber-Security Implications for Australian Small Businesses," *Computers and Security*, 109:1-56, 2021, https://doi.org/10.1016/j.cose.2021.102385.

[35] L. Tawalbeh, F. Muheidat, M. Tawalbeh, and M. Quwaider, "Applied Sciences IoT Privacy and Security : Challenges and Solutions," *Mdpi*, pp. 1-17, 2020.

[36] W. J. Triplett, "Addressing Human Factors in Cybersecurity Leadership," *Journal of Cybersecurity and Privacy*, 2(3):573–586, 2022. https://doi.org/10.3390/jcp2030029.

[37] A. Tsohou, M. Karyda, S. Kokolakis, and E. Kiountouzis, "Managing the Introduction of Information Security Awareness Programmes in Organizations," *European*

Journal of Information Systems, 24(1):38-58, 2015, https://doi.org/10.1057/ejis.2013.27.

[38] H. W. Volberda, S. Khanagha, C. Baden-Fuller, O. R. Mihalache, and J. Birkinshaw, "Strategizing in a Digital World: Overcoming Cognitive Barriers, Reconfiguring Routines and Introducing New Organizational Forms," *Long Range Planning*, 54(5):102110, 2021, https://doi.org/10.1016/j.lrp.2021.102110.

[39] L. J. Wells, J. A. Camelio, C. B. Williams, and J. White, "Cyber-Physical Security Challenges in Manufacturing Systems," *Manufacturing Letters*, 2(2):74-77, 2014, https://doi.org/10.1016/j.mfglet.2014.01.005.

[40] M. Wessels, P. van den Brink, T. Verburgh, B. Cadet, and T. van Ruijven, "Understanding Incentives for Cybersecurity Investments: Development and Application of a Typology," *Digital Business*, 1(2):100014, 2021, https://doi.org/10.1016/j.digbus.2021.100014.

[41] Y. Wind and T. L. Saaty, "8002_Marketing_Applications_of_the_Analytic.pdf," *Management Science*, 26(7):641-658, 1980.

[42] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A Survey on Cyber Security for Smart Grid Communications," *IEEE Communications Surveys and Tutorials*, 14(4):998-1010, 2012, https://doi.org/10.1109/SURV.2012.010912.00035.

[43] A. Yeboah-Ofori and S. Islam, "Cyber Security Threat Modeling for Supply Chain Organizational Environments," *Future Internet*, 11(3):25 pp. 2019, https://doi.org/10.3390/fi11030063.

[44] D. Zissis,and D. Lekkas, "Addressing Cloud Computing Security Issues," *Future Generation Computer Systems*, *28*(3):583-592, 2012, https://doi.org/10.1016/j.future.2010.12.006.

[45] V. Zlomislić, K. Fertalj, and V. Sruk, "Denial of service attacks, defences and research challenges," *Cluster Computing*, 20(1):661–671, 2017, https://doi.org/10.1007/s10586-017-0730-x.

Le Vinh Quang (photo not available) is a senior lecturer at Becamex Business School, Eastern International University. He obtained his master degree in Business specializing in Technology Management, from Western Sydney University and is currently pursuing a Ph.D in Faculty of Business Administration at Industrial University of Ho Chi Minh City. His research interests encompass supply chain management, technology management, and enterprise risk management.

Tran Huu Duc (photo not available) is working as an Assistant to the University Council and Board of Presidents at Eastern International University in Vietnam. He graduated with a bachelor's degree in Supply Chain Management from Eastern International University. Currently, he focuses on research related to facilities management and technology adoption.

Narayan C. Debnath (photo not available) is currently the Founding Dean of the School of Computing and Information Technology at Eastern International University, Vietnam. He is also serving as the Head of the Department of Software Engineering at Eastern International University, Vietnam. Formerly, Dr. Debnath served as a Full Professor of Computer Science at Winona State University, Minnesota, USA for 28 years, and the elected Chairperson of the Computer Science Department at Winona State University for 7 years. Dr. Debnath has been the Director of the International Society for Computers and their Applications (ISCA), USA since 2014. Professor Debnath made significant contributions in teaching, research, and services across the academic and professional communities. He has made original research contributions in software engineering, artificial intelligence and applications, and information science; technology, and engineering. He is an author or co-author of over 500 research paper publications in numerous refereed journals and conference proceedings in Computer Science, Information Science, Information Technology, System Sciences, Mathematics, and Electrical Engineering. He is also an author of over 15 books published by well-known international publishers including Elsevier, CRC, Wiley, Bentham Science, River Publishing, and Springer. Dr. Debnath has made numerous teaching, research, and invited keynote presentations at various international conferences, industries, and teaching and research institutions in Africa, Asia, Australia, Europe, North America, and South America. He has been a visiting professor at universities in Argentina, China, India, Sudan, and Taiwan. He has been maintaining an active research and professional collaborations with many universities, faculty, scholars, professionals, and practitioners across the globe. Dr. Debnath is an active member of the IEEE, IEEE Computer Society, and a Senior Member of the International Society for Computers and their Applications (ISCA), USA.

Nguyen Ngoc Long (photo not available) is currently serving as the Vice Dean of the Faculty of Business Administration at the Industrial University of Ho Chi Minh City. He earned his doctorate degree from Hunan University in China in 2017. His research interests encompass behavior, leadership, and risk management.

Index

Authors

A

B-C

D-G

R-S

T-V

W-Z

Key Words

T

U-V

W-Z

Journal Submission

The International Journal of Computers and Their Applications is published four times a year with the purpose of providing a forum for state-of-the-art developments and research in the theory and design of computers, as well as current innovative activities in the applications of computers. In contrast to other journals, this journal focuses on emerging computer technologies with emphasis on the applicability to real world problems. Current areas of particular interest include, but are not limited to: architecture, networks, intelligent systems, parallel and distributed computing, software and information engineering, and computer applications (e.g., engineering, medicine, business, education, etc.). All papers are subject to peer review before selection.

A. Procedure for Submission of a Technical Paper for Consideration

1. Email your manuscript to the Editor-in-Chief, Dr. Ajay Bandi. Email: ajay@nwmissouri.edu.
2. Illustrations should be high quality (originals unnecessary).
3. Enclose a separate page (or include in the email message) the preferred author and address for correspondence. Also, please include email, telephone, and fax information should further contact be needed.
4. **Note**: Papers shorter than 10 pages long will be returned.

B. Manuscript Style:

1. **WORD DOCUMENT**: The text should be **double-spaced** (12 point or larger), **single column** and **single-sided** on 8.5 X 11 inch pages. Or it can be single spaced double column.

   LaTex DOCUMENT: The text is to be a double column (10 point font) in pdf format.
2. An informative abstract of 100-250 words should be provided.
3. At least 5 keywords following the abstract describing the paper topics.
4. References (alphabetized by first author) should appear at the end of the paper, as follows: author(s), first initials followed by last name, title in quotation marks, periodical, volume, inclusive page numbers, month, and year.
5. The figures are to be integrated in the text after referenced in the text.

C. Submission of Accepted Manuscripts

1. The final complete paper (with abstract, figures, tables, and keywords) satisfying Section B above in **MS Word format** should be submitted to the Editor-in-Chief. If one wished to use LaTex, please see the corresponding LaTex template.
2. The submission may be on a CD/DVD or as an email attachment(s). **The following electronic files should be included:**
   - Paper text (required).
   - Bios (required for each author).
   - Author Photos are to be integrated into the text.
   - Figures, Tables, and Illustrations. These should be integrated into the paper text file.
3. Reminder: The authors photos and short bios should be integrated into the text at the end of the paper. All figures, tables, and illustrations should be integrated into the text after being mentioned in the text.
4. The final paper should be submitted in (a) pdf AND (b) either Word or LaTex. For those authors using LaTex, please follow the guidelines and template.
5. Authors are asked to sign an ISCA copyright form (http://www.isca-hq.org/j-copyright.htm), indicating that they are transferring the copyright to ISCA or declaring the work to be government-sponsored work in the public domain. Also, letters of permission for inclusion of non-original materials are required.

Publication Charges

After a manuscript has been accepted for publication, the contact author will be invoiced a publication charge of **$500.00 USD** to cover part of the cost of publication. For ISCA members, publication charges are **$400.00 USD** publication charges are required.

Revised 2020