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 31, No. 2, June 2024** **ISSN 1076-5204**

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

## EDITOR-IN-CHIEF

**Ajay Bandi**
Associate Professor
School of Computer Science and Information Systems
Northwest Missouri State University
800 University Drive, Maryville, MO, USA 64468
Email: ajay@nwmissouri.edu

## EDITORIAL BOARD

# Editorial

It is my distinct honor, pleasure, and privilege to serve as the Editor-in-Chief of the International Journal of Computers and Their Applications (IJCA) since 2022. I have a special passion for the International Society for Computers and their Applications. I have been a member of our society since 2014 and have served in various capacities. These have ranged from being on program committees of our conferences to being Program Chair of CATA since 2021 and currently serving as one of the Ex-Officio Board Members. I am very grateful to the ISCA Board of Directors for giving me this opportunity to serve society and the journal in this role.

I would also like to thank all the editorial board, editorial staff, and authors for their valuable contributions to the journal. Without everyone's help, the success of the journal would be impossible. I look forward to working with everyone in the coming years to maintain and further improve the journal's quality. I want to invite you to submit your quality work to the journal for consideration for publication. I also welcome proposals for special issues of the journal. If you have any suggestions to improve the journal, please feel free to contact me.

Dr. Ajay Bandi
School of Computer Science and Information Systems
Northwest Missouri State University
Maryville, MO 64468
Email: AJAY@nwmissouri.edu

In 2024, we are having four issues planned (March, June, September, and December). The next latest issue is taking shape with a collection of submitted papers.

I would also like to announce that I will begin searching for a few reviewers to add to our team. We want to strengthen our board in a few areas. If you would like to be considered, don't hesitate to get in touch with me via email with a cover letter and a copy of your CV.

Ajay Bandi, Editor-in-Chief
Email: AJAY@nwmissouri.edu

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**:** 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:

TESNIM KHELIFI, NOURHENE BEN RABAH, and BENEDICATE LE GRAND from Universite Paris 1 Pantheon Sorbonne, France present their work "Explainable Learning Analytics Dashboard: Enhancing Understanding of Insight derived from Educational Data". This article introduces the Explainable Learning Analytics Dashboard (EX-LAD), designed to present learning analytics data on student performance, engagement, and perseverance in an accessible way for both teachers and students. The goal is to make this data comprehensible, even for those without data analysis expertise, enabling students to self-assess and address weaknesses promptly while helping teachers identify students' specific needs. By emphasizing explicability, the EX-LAD aims to boost user confidence and engagement. A case study using real data from ESIEE-IT, an engineering school in France, during the 2021-2022 academic year, demonstrates the dashboard's effectiveness.

PRATIK SHRESTA, CHIRAG PARIKH, and CHRISTIAN TREFFTZ from Grand Valley State University, USA, "Accelerating the execution of the Partition Problem: A Comparative Study of Various Techniques for Efficient Computation". This paper examines the inefficiency of exponential- time algorithms for solving intractable NP-complete problems, such as the partition problem, compared to polynomial-time algorithms for tractable problems. It highlights the importance of understanding these limitations for software developers. Exact algorithms can only solve small instances of NP-hard problems efficiently, making large instances practically intractable. To address this, the paper explores using parallel processing with NVIDIA T4 GPU and PYNQ FPGA board to speed up the evaluation of partition problem solutions. Four different FPGA overlays are created, and their performance is compared with the GPU using Python and the numba/cuda library.

ALI IBRAHIM AHMED, BDUL SATTAR M. KHIDHIR, SHATHA A. BAKER, OMAR I and ALSAIF IBRAHIM AHMED SALEH from Al-Noor University, Iraq present their work "Enhancing Cyber Security by Relying on a Botnet Attack Tracking Model Using Harris Hawks Optimization". This paper addresses the threat of botnet attacks and proposes using the Harris Hawks Optimization (HHO) algorithm to enhance detection and mitigation efforts. The HHO algorithm is used as a feature selector to analyze anomalous network traffic and improve botnet IP positioning. The study includes sections on attack path analysis, system testing, and experimental results. After configuring the network topology and determining attack paths with HHO, the algorithm's performance in correcting attack paths and preventing IP spoofing is verified, demonstrating effective results.

HICHAM AMELLAL from, IBNOU ZOHR University, Morocco present their work “Improving Communication Security Against Quantum Algorithms”. This paper analyzes the threat of Shor’s algorithm to classical cryptographic protocols, particularly HTTPS. It proposes a Quantum Intrusion Prevention System (QIPS) to protect against quantum threats, using components like beam splitters and detectors. The QIPS enhances the resilience of classical cryptographic protocols, bolstering network security against quantum algorithms.

BAGHDADI AMMAR AWNI Abbas, MOHAMMED AL-MUKHTAR and MAHA ABDILAMEER from the University of Baghdad, Iraq, and NAJEEB ABBAS Al-SAMMRAIE from Al MADINAH International University Kuala Lumpur, presented their work “Unsupervised Interactive lecture evaluation using the Kano Model”. The paper discusses the importance of interactivity in effective e-learning and presents an interactive lecture course for undergraduate students on computer networks, created using MATLAB AppDesigner. This system was implemented for second-year students at the College of Mass Communication, Baghdad University. A questionnaire based on the Kano model was used to evaluate student satisfaction and future expectations regarding the interactive lectures as a supplement to traditional lectures. The analysis revealed that interactive lectures significantly satisfy students' learning needs and could serve as a valuable backup to conventional lectures, as well as a training and testing method. The approach is applicable to various subjects both scientific and social.

SAPNA ARORA from, IILM University, RUCHI KAWATRA from, SRM University, and NARAYANA C.DEBNATH India present their work “White and Black Box Techniques towards Deploying a Prediction Model In Educational Data Mining”. This study explores the use of predictive systems in Educational Data Mining to enhance learning and teaching methodologies. By reviewing white and black box models, the research evaluates their advantages and limitations. Using surveys and techniques like ID3, CART, XGBoost, and MLP, an adaptive self-assessment case study reveals that black box models, especially XGBoost and MLP, provide superior accuracy. The study offers practical recommendations for building predictive systems and presenting data.

ANAL KUMAR and ABM SHAWKAT ALI from, Fiji National University, Fiji present their work “Big Data Visualization In Digital Marketplaces – A Systematic Review and Future Directions”. The paper explores the significance of big data visualization in digital marketplaces, emphasizing its role in transforming complex data into actionable insights for better decision-making. It discusses how visualization techniques help businesses optimize operations, identify patterns, and understand customer behavior. The study highlights the collaborative aspect of visualization, fostering innovation and performance improvements. By reviewing literature from 2010 to 2022, the paper identifies gaps and challenges in big data visualization. The findings indicate an increase in research on the topic, providing a comprehensive overview for future investigations.

As guest editors, 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 Editors:

Ajay Bandi, Northwest Missouri State University, USA

**June 2024**

# Explainable Learning Analytics Dashboard: Enhancing Understanding of Insights derived from Educational Data

Tesnim Khelifi, Nourh'ene Ben Rabah *
Universite Paris 1 Pantheon Sorbonne 90 rue de Tolbiac, Paris, France
and B´en´edicte Le Grand †
Universit´e Paris 1 Panth´eon Sorbonne 90 rue de Tolbiac, Paris, France

June 23, 2024

**Abstract**

The integration of Learning Analytics into educa- tional environments can improve the learning process. However, to be used effectively, these tools need to be both explainable and comprehensible. This article introduces a novel dashboard known as the Explainable Learning Analytics Dashboard (EX-LAD), designed to present learning analytics data relating to student performance, engagement, and perseverance in a clear and accessible way. The main aim of this study is to make this information easily understandable for both teachers and students, even for those without in- depth knowledge of data analysis. The EX-LAD primarily empowers students to self- assess by tracking their progress. This enables them to better target their weaknesses and try to remedy them quickly and effectively, thus avoiding any risk of failure. Teachers, meanwhile, can identify students' specific needs, and detect any learning difficulties. By emphasizing explicability, we aim to boost user confidence in the analyses generated by the system and encourage their engagement in the process of continuous improvement of the educational experience. To showcase the effectiveness of our dashboard, we conducted a case study using real data collected from ESIEE-IT, an engineering school in France, during the 2021-2022 academic year .

**Keywords:**Explainable Learning Analytics, Dash- board, Higher Education

## 1 Introduction

During the COVID-19 pandemic, distance learning systems emerged as a crucial means of ensuring teaching continuity in a virtual environment provided by the World Wide Web. Despite initial reservations, teachers and students have widely adopted these e-learning solutions. Today, while the situation has improved and allowed a return to the classroom, many higher education institutions still wish to maintain certain aspects of distance learning [1], particularly by leveraging Learning Management Systems(LMS). LMSs are commonly used in institutional academic environments to deliver educational content and enhance the learning experience of teachers and students. However, it is important to note that there are many LMSs available on the market, such as Moodle, widely used in universities, and Black Board Learn [2], which is of interest in our study. Although these platforms provide learning analytics dashboards to showcase valuable information, they often face two significant challenges. Firstly, they tend to prioritize student performance which measures the students's level of achievement, their ability to assimilate knowledge and demonstrate academic skills, as well as their positioning with regard to their peers in terms of academic results [3], [4] such as grades obtained in various activities, exams, projects, presentations etc. Regrettably, this narrow perspective often neglects other vital indicators like engagement encompassing cognitive, behavioral, social, and emotional aspects. includes the conclusion.

- Behavioral engagement refers to students' consistent presence and dedication to diverse learning activities. It is expressed through assiduous participation in class, where students ask questions, interact with learning materials, and contribute constructively to discussions.
- Cognitive engagement showcases the students' active mental involvement, going beyond simple physical presence. It embraces creativity, critical analysis of information and problem-solving, reflecting a deep investment in the learning process.
- Social engagement manifests through students' social interactions within their educational environment, as well as through their participation in collaborative activities. It goes beyond the boundaries of academic learning, fostering a sense of connection and belonging among learners.
- Emotional engagement refers to the students' emotional desire, motivation and satisfaction during the course (enthusiasm, feeling of being valued). It can be perceived by the student's interest in the course and his/her relationships with classmates and teachers. As a result, there is a pressing need for a more comprehensive approach that takes into consideration the multiple dimensions of students' learning and provides a holistic view of their educational experience. Another challenge that arises when using analytical dashboards is that users, including teachers and students, may not necessarily have

*Researcher Phd Candidate Email:Tesnim.Khelifi, Nourhene.Ben-Rabah@univ-paris1.fr.
†Professeur en Informatique Universite Paris 1 Pantheon Sorbonne 90 rue de Tolbiac. Email: Benedicte.Le-Grand@univ-paris1.fr.

In-depth knowledge of data analysis. Dashboards with complex hard-to-understand graph can result in either limited future usage of these tools or incorrect interpretation of the data. This can lead to erroneous conclusions or unfortunate interventions. Visualization techniques in general, and Learning Analytics Dashboards (LADs) in particular, have proved effective in visually communicating the data. Visualization techniques are used to graphically represent data that appears complex to simplify it and make it more comprehensible to users. They also enable results to be communicated clearly and effectively to a varied audience, relationships and trends to be identified, and decision-making to be supported. There are several types of visualization, the most common of which are as follows: bar charts and histograms, often used for comparisons between categories; pie charts, used to represent proportions or parts of a whole; and scatter plots, used to present relationships between several variables etc. However, they are often considered difficult to understand and interpret [5]. To address this thinking, a new field called "Explainable Learning Analytics " [6,7],has been introduced. Therefore, our research questions are the following:

- RQ1: What indicators are essential for supporting both students and teachers in utilizing LMS effectively?
- RQ2: How can we create a Learning Analytics dashboard that is understandable and interpretable for individuals without expertise in data analysis? To address these research questions, we developed an EXplainable Learning Analytics Dashboard (EX- LAD) that presents learning analytics data on student performance, engagement, and perseverance in a clear and easily understandable manner.The objective of EX-LAD is to make this information accessible not only to teachers but also to students, who may not have extensive knowledge in data analysis.

This dashboard empowers teachers to gain valuable insights into their students' progress, identify at-risk learners, and provide targeted support. Similarly, students can utilize this dashboard to track their learning journey, identify strengths and weaknesses, and make informed decisions to enhance their academic performance. To demonstrate the effectiveness of our dashboard, we conducted a case study using real data collected from ESIEE-IT, an engineering school in France, throughout the academic year 2021-2022. This case study serves as concrete evidence of the impact and value our dashboard brings to the educational context.The paper is organized as follows: Section 2 presents a review of some recent learning analytics dashboards in higher education. Section 3 describes the proposed EX-LAD. Section 4 illustrates our approach by providing answers to the research questions, section 5 discusses the results obtained in our study and finally, section 6 concludes our work and presents our future works.

## 2 Related Work

In this paper, we focus on the usefulness of learning analytics dashboards for monitoring students and detecting the risk of failure or drop-out. In this context, we considered Search towards the global minimum by emulating the predatory behavior of Harris hawks [7].

Various research works for our literature review, including those from the Learning Analytics (LA) and Educational Data Mining (EDM) communities. We conducted keywords-based queries such as 'learning analytics', 'dashboard', 'learner', 'Indicators', 'online learning environment', and 'data visualization' while specifying the research area, higher education. We discarded articles published before 2019 as we wanted to focus on recent works. These keyword-based queries returned over 670 research articles. We read their abstracts and selected those that presented empirical research on Learning Analytics in higher education all over the world. We excluded review articles and theoretical articles that focus on the Learning Analytics Dashboards aspects. Following this methodology, we finally selected nine papers that we analyzed in depth. In [8], the authors introduce the 'TELA system,' a Learning Analytics dashboard designed to enhance the performance and engagement of students enrolled in distance learning courses.Its primary goal is to simulate students' motivation to continue their studies by providing them with the opportunity to monitor their progress and grade evolution while comparing their performance with that of their peers. To achieve this objective, the system offers a diverse range of learning indicators, including measures of engagement such as cognitive engagement, assessed by the number of activities completed and resources accessed by the student; behavioral engagement, determined by the frequency of interactions; and social engagement, calculated based on the volume of messages exchanged in discussion forums. Additionally, the system incorporates performance metrics derived from students' grades. In [9], the authors not only provide a descriptive overview of the results but also expand their perspective to include predictive and prescriptive elements.The objective is to enhance student engagement by offering detailed explanations of predictions for each learner. This dashboard specifically focuses on a critical aspect of engagement: cognitive engagement, inferred from students' resource usage, along with academic performance, assessed through each student's GPA (Grade Point Average). By incorporating these predictive and prescriptive features, the dashboard aims to give students a proactive outlook on their learning, encouraging them to optimize their academic success. 'Tabat' [10], is a Learning Analytics dashboard designed for both educators and students. It offers an in-depth analysis of learning data, aiming to simplify monitoring and control of the learning process. Their main objective is to use this tool to enhance the engagement and success rates of online learners. The 'PLD prescriptive dashboard' [11] guides students in improving their academic performance. It aims at presenting students with a variety of learning indicators such as behavioral engagement, cognitive and social engagement as well as a performance indicator calculated from the students' grades. These indicators are grouped by type and each page is

dedicated to a specific type of indicators. This dashboard offers personalized recommendations for each student depending on the difficulties he/she faces and clusters students who share the same learning behavior into different profiles. The dashboard introduced in [12] diverges from typical daily dashboards by adopting a personalized learning support approach. It focuses on face-to-face interactions, with particular emphasis on collaborative argumentation between students.This platform enables teachers to identify groups of students facing similar argumentation difficulties, by providing exclusively social engagement indicators. The dashboard presented in [13]is specifically dedicated to teachers entered around behavioral engagement and performance indicators. Its primary goal is to offer behavioral process-oriented feedback in online courses. The visualizations are brought together in an interface, offering a global view of the indicators. The authors in [14] developed a Learning Analytics dashboard that allows students to evaluate their cognitive engagement as well as their performance and influences their motivation in distance learning environments. This dashboard offers a global view of these indicators by grouping visualizations in one inter- face which facilitates interpretations. It also generates personalized messages for each student according to their weekly report. The authors in [15] dedicate their dashboard only to students. It offers a single type of visualizations, i.e. a progress bar showing the student's grade for each notion of the course and using only one learning indicator which is performance calculated using grades, number of correct answers and question response time. This dashboard also recommends resources for each chapter of the course that can help the student having a problem in this chapter. Finally, the dashboard developed in [16] proposes a learning analytics approach, known as 'Student Inspection Facilitator (SIF)'. It assists instructors in identifying students requiring special attention based on their numerical data. SIF could be integrated into institutional systems to effectively interpret student behavior and classify them for intervention while leaving the choice of whether to intervene to the instructor. We established a set of criteria for comparing various existing works in the field. This methodology allows us to conduct a thorough analysis and discern the strengths and weaknesses inherent in each approach. Table 1 provides a summary of the chosen studies based on five primary criteria: (a) target users, (b) data protection, (c)learning indicators, (d) visualization, and (e) actionable insights:

a) **Target users** (TU) represent the final users of the dashboard who can be students (S) and/or teachers(T). This is an important criterion since it guarantees the dashboard's relevance and usefulness to those who need it.By defining the dashboard's endusers, we can customize and design it to meet their specific needs and identify the indicators that are most relevant to them.

b) **Data protection (DP)** indicates whether the researchers have guaranteed the ethical use of data by teachers and the educational team as the collected data raises legitimate concerns about confidentiality and privacy. We therefore proposed rigorous measures to ensure data integrity and security in line with the General Data Protection Regulation (GDPR), highlighting four fundamental requirements which are (R1) data confidentiality, (R2) Informed Consent, (R3) data Anonymization, (R4) transparency, and (R5) diversity.

• Data confidentiality requirement aims to ensure the protection of the information of users participating in the study, in compliance with the rules established by the (GDPR). It aims to minimize any potential risk of disclosure of sensitive data.

• Informed consent requirement ensures that participants are provided with transparent information on the final use of their data, thereby guaranteeing their consent and agreement to the use of their data and fostering the establishment of a relationship of trust.

• Data anonymization requirement aims to remove all personal information that could identify individuals and reveal their identity, giving absolute priority to the protection of privacy.

• Transparency requirement emphasizes the trans- parency of the experimental results obtained, as well as the explicability and comprehensibility of the approach used by the participants to foster mutual trust.

• Diversity requirement ensures the inclusion of diverse data representing a variety of demographic, social and cultural groups.

c) **Learning Indicators** represent the specific type of indicators used in the dashboard that may include performance indicators (P), cognitive engagement indicators (CE), behavioral engagement indicators (BE), social engagement indicators (SE), and more.We proposed this comparison criterion based on our first research question.

d) **Visualization** is described based on three main criteria which are: (i) Number of visualizations and chosen techniques, (ii) explainability and (iii) objective of visualization referring to our second research question.

This criterion is proposed considering the importance of visualization techniques in a dashboard, as previously highlighted, as well as their ability to simplify the presentation of information for different users. • Number of visualizations and type:

This criterion focuses on the variety of the visualizations proposed in the dashboard (for example scat- ter plots, bar charts, pie charts, etc.).

• Explainability:This criterion assesses whether the provided visualizations are understandable and easy to interpret by non-experts in data analysis.It can be achieved either by offering an explanatory text, meaningful color coding such as traffic code colors, or through the number of proposed interfaces.

• Objective of visualization:This criterion presents the idea that each visualization aims to convey to the user. It could include showing change over time (temporary evolution), comparing group values (comparison), establishing relationships between variables, or displaying value distributions.

e) **Insightful Actions** represent the types of actions delivered to the users of the dashboard following the visualizations

such as personalized recommendations or notifications. These recommendations are designed to support learners on their learning path by providing personalized support and advice tailored to their individual needs. For instance, they may include pedagogical suggestions like proposing specific activities or resources to students, as well as personalized learning path recommendations that adjust to individual student needs. Notifications within learning analytics dashboards play a crucial role as well, delivering pertinent information to diverse users and enhancing their interaction with the tool. They empower learners to actively engage with their studies, offering updates on course progress and various activities to help them monitor their advancement. Notifications also stimulate social engagement by alerting users to new group discussions, fostering active participation, and encourage behavioral engagement by reminding students of impending activity deadlines to ensure timely assignment submissions. Additionally, teachers can benefit from notifications that highlight any issues with a student, aiding in the identification of those at risk.

Based on the works we studied, we made some observations. First, we observe that all of the studies uses the performance indicator, which is derived from student grades (see [17]) except [12]. We also note a diversity in the proposed engagement indicators. For example, works [8], [9], [10] and [14] focus on cognitive engagement, while learning analytics dashboards in [8], [10], [11], [13] and [16] deal with behavioral engagement, and [8], [10], [11] and [12] address social engagement. Most of these works are limited to two indicators, namely performance and an engagement indicator, except for [8] and [10], which combines all four indicators. However, most studies opted for a straightforward presentation of data in the form of visualizations, without developing the formulas for calculating indicators or clearly identifying engagement and performance. One exception is [10], which developes several scores to facilitate the understanding of each indicator. Among these scores, we may find the participation score, which is calculated according to the duration of interaction with the platform, thus reflecting the student's behavioral engagement. Another score, called the section progress score, indicates each student's level of progress in each section of the course. They also offer the Course Progress Score, which reflects overall progress in the course. We also find the social interaction score, calculated from messages exchanged between students. Finally, there's the Successful Progress Score, which provides an estimate of the learner's level of success. Nevertheless, although several different learning indicators were proposed, visualization options remain limited. Most studies rely mainly on bar charts, curves, or even tables and lists.

These visualizations are not suitable for handling complex information, as they may not capture all nuances and complexities effectively. In addition, they are not suitable for the comparison of multiple variables, which can restrict the depth of analysis and lead to misinterpretation. A few exceptions, however, introduce scatter and radar plots, as

| Ref | TU | DP | Learning Indicators | | | | Visualizations | | | Actions |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | P | BE | SE | CE | Number& Type | Explainability | Objective | |
| [8] | S | √ | √ | √ | √ | √ | 5 Bar charts, 1 Linear Graph, 5 Line charts, 1 Gauges, 1 Tree graph | × | Comparison, Evolution | × |
| [9] | S | √ | √ | × | × | √ | 5 Line charts, 1 histogram | Text | Evolution | × |
| [10] | S/T | √ | √ | √ | √ | √ | 6 Tables,3 Line charts, 1 Bar chart, 1 Pie chart | × | Comparison, Evolution | Notifications |
| [11] | S | √ | √ | √ | √ | × | 2 Bar charts,1 Gauge, 1 chart, 2 Lin charts e Colum n | × | Comparison | Recommendations |
| [12] | S/T | √ | × | × | √ | × | 1 Radar chart, 1 Network Graph, 1 Bar chart | Text, Color Coding | Data distribution | × |
| [13] | T | √ | √ | √ | × | × | 2 Bar charts,3 Tables | Color Coding | Data distribution | × |
| [14] | S | √ | √ | × | × | √ | 1 Pie chart, 1 List, 1 Table | × | Data distribution | × |
| [15] | S | √ | √ | × | × | × | 1 Radar chart, 1 List, 1 Scatter Plot | × | Data distribution, Evolution | Recommendations |
| [16] | T | √ | √ | √ | × | × | Radar charts, 2 box plots | Text Color Coding | Data distribution | Recommendations |

Figure 1: Comparative table between existing learning analytics dashboards
Target Users (TU), Data Protection (DP), Students(S), Teachers (T), ()Yes, (×) No

referenced in articles [15], [12] and [16]. It is observed that the works presented do not pay particular attention to the comprehensibility or explicability of their visualizations. Given the limited choice of available visualizations, there is a risk that users will find it difficult to understand the presented results. However, we note a few exceptions, notably in works [9], [12], and [16], where text descriptions are provided, and sometimes significant color choices were used, such as traffic light colors in works [12], [13] and [16]. Finally, it is important to note that only three studies provide their users with insightful actions. [15], [11]and [16] deliver personalized recommendations to the students using their dashboards and [10]'s dashboard as well as offered notifications to the students for each indicator allowing them to identify their strengths and weaknesses and make informed decisions to improve their academic performance. To guarantee these objectives, we place great emphasis on clarity, providing visualizations that are easily understood by all users, accompanied by explanatory text for the indicators presented. Our solution also respects privacy and ensures the protection of the personal data used. To propose adequate support actions, we suggest different profiles of students based on the learning indicators that will be defined later. Another observation is that the presented learning ana- lytics dashboards share an important common feature: the protection of the data used in their visualizations. The authors ensured the data used is anonymized to respect ethical requirements and preserve the privacy of the concerned individuals but there is no indication

that the other requirements were respected. In the next section, we describe our proposed Explain- able Learning Analytics Dashboard EX-LAD.

## 3 The proposed EX-LAD

In this section, we introduce the participants in our study, describe our case study in detail to demonstrate the effectiveness of our approach, and finally present the steps of our proposed Learning analytics dashboard.

### 3.1 Study Context

We conducted a case study with real data collected from the LMS used by an IT school called ESIEE- IT [18]. ESIEE-IT is based in France. It offers several computer science programs of different specialties such as artificial intelligence, cybersecurity, and information systems dedicated to different student profiles such as bachelor, engineer, and master. The participants in this study were 128 students who took a programming course with Python. Among these students, 22 were enrolled in Master Green, 48 took an engineering course, 29 BTS, and 29 followed a Master in Big Data. There were 117 male and 11 female students participating in this study. The dataset was collected during the 2021-2022 academic year. While collecting these data, we proceeded to data anonymization to ensure that it could be used in accordance with ethical principles.

The Python programming course is taught in a hybrid way, i.e., 80% of the course time is online and 20% of the course time is face-to-face. In practice, during online lessons, the student must follow the course through the LMS of the school which is Blackboard Learn [2]. During the face-to-face session, the student must be present at school to interact with teachers and ask questions related to the course. The course on Blackboard is composed of a set of sequences. Each sequence can contain four types of resources which are the following:

- (a) the course in a video format,
- (b) the notes allowing the student to constitute exploitable resources in different formats such as text, video or audio that can be used in addition to the course,
- (c) the documents containing instructions for the exercises along with corrections either as an attachment or directly in the document,
- (d) the quizzes composed of 5 to 10 questions delivered as assessment activities and a final test made of 20 questions. Student interactions with BlackBoard Learn [2] were recorded in the Snowflake data warehouse. These interactions include data such as number of clicks, time spent on the platform, number of accesses to the platform, and other information that will be detailed later. In the following section, we present the different steps of our dashboard.

### 3.2 Steps of the proposed EX-LAD

In this section, we present the four steps of our solution for our dashboard which are: data collection, data pre-processing, data analysis, and data visualization as shown in Figure 2.

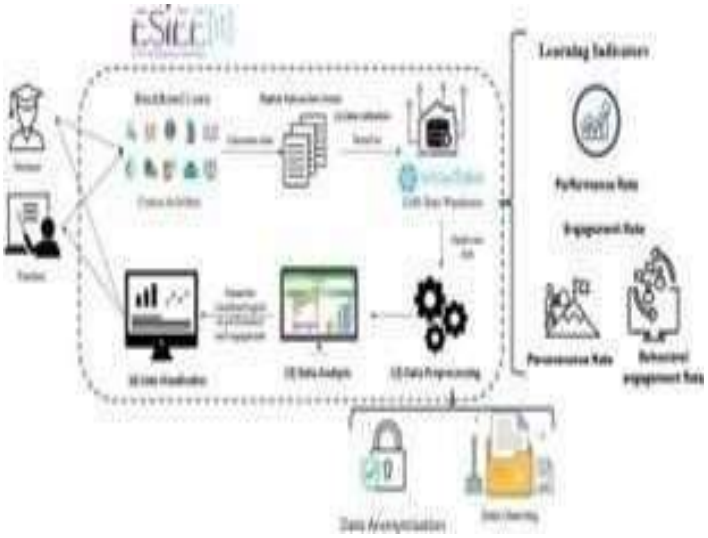


Figure 2: The dashboard development steps.

#### 3.2.1 Step 1. Data collection

In the first step, we collected digital learning traces resulting from the learner's interactions and stored in the SnowFlake data warehouse. Our dataset contains 128 instances and 106 features of the student .Table 2 describes these different features. It is made up of 26 features organized into five groups describing the various features of our dataset.

The first group includes the student's personal data (SF) which is name (1), e-mail address (2) , public (3) and course of study (4). The second part (AF) from feature number 5 to 8 is related to the student's access to the platform, such as 'Course Access Connection' and 'Course Access Minutes'. The following part from 9 to 18 (PF) concerns academic performance, including grades, ranks and average score. Engagement indicators (EF) are described in the next section (from 19 to 25):

'Interaction Oriented Investment (IOI)' , 'Course Access Connection Oriented Investment (CA- COoI)' and 'Course Access Count Oriented Investment (CACoI)'. Finally, the last feature 'Difficulty' (DF) contains four values representing the different profiles of students according to the problems they encounter which are as follows: 'E+P+', 'E+P- ', 'E-P+' and 'E-P-'. All these indicators are described according to equations presented below: The **Performance** indicator is calculated through the grades of the student in the executable activities using his/her grades in the quizzes (Q) (50%) and the final exam (50%) using the following score:

$$\text{performance} = 0.5 \times \text{Average}(Q_1, Q_2, \ldots, Q_n) + 0.5 \times \text{Final Score}$$

A student is considered successful if his or her average

exceeds 50 and failing if it does not. we must mention that there are two types of activities in Blackboard:

| Cat | FN | Feature Name | Type | Feature Meaning | Value Example |
|---|---|---|---|---|---|
| SF | 1 | Student | O | The student's name and last name | TOTO TATA |
| | 2 | Email | O | The student's academic email address | TOTO.TATA @edu.esiee-it.fr |
| | 3 | Public | O | Level and branch of studies | M2I, IA |
| | 4 | Course Name | O | The name of the course | Python |
| AF | 5 | Course_Access Connection | I | The number of accesses to the course | 10 |
| | 6 | Course Access Minutes | I | The access time to the course in minutes | 662 |
| | 7 | First_Course_Access | T | First access to the course | 2021_10_18 05:38:56 |
| | 8 | Last_Course_Access | T | Last access to the course | 2021_02_09 2:23:25 |
| PF | 9 | Rating SiQ1 | F | Score of quiz n° 1 in the sequence number 1 | 80 |
| | 10 | Rank_SiQ1 | I | Rank of the student in the quiz n°1 in the sequence number 1 | 6 |
| | 11 | Diff_Rating_SjQ1 | F | The difference of score between the actual quiz in the actual sequence and the last one | 20 |
| | 12 | Diff_Ranking_SjQ1 | I | The difference of rank of the student between the actual executable activity in the actual sequence number j and the last one | -5 |
| | 13 | Rating Final Exam | F | Score of the final exam | 75 |
| | 14 | Rank Final Exam | I | Rank of the student in the final exam | 3 |
| | 15 | Diff_Rating_Final Exam | F | The difference of score between the final exam and the last executable activity | -20 |
| | 16 | Diff_Ranking_Final Exam | I | The difference of rank of the student between the final exam and the last executable activity | 23 |
| | 17 | Avg Rating | F | The average score in all executable activities | 38,75 |
| | 18 | Rank | I | The rank of the student in the class | 20 |
| EF | 19 | SiQ1 Exe Submission Count | I | Number of attempts in the executable activity Quiz number 1 of the sequence number I | 2 |
| | 20 | FE Exe Submission Count | I | Number of attempts in the final exam | 1 |
| | 21 | T Exe Submission Count | I | Total number of attempts in quizzes | 10 |
| | 22 | Interaction Oriented Investment (IoI) | F | A score that measures the interaction-oriented investment of the student in all the executable and non-executable activities | 37,5 |
| | 23 | Course Access Connection Oriented Investment (CACOoI) | F | A score that measures the investment of the student related to the access count to the course | 32,95 |
| | 24 | Course Access Count Oriented Investment (CACoI) | F | A score that measures the investment of the student related to the time spent in the course | 23,66 |
| | 25 | Engagement | F | The average of the four investment scores to measure the engagement of the student | 66,84 |
| DF | 26 | Difficulty | O | Type of difficulties of each student depending on the calculated scores. | E+P+, E+P-, E-P+ E-P- |

Figure 3: Dataset Features.

nonexecutable activities which are the resources offered to students (pdf, video, etc.) and executable activities (quizzes, exams, etc.).

Engagement is defined as 'the active involvement of learners in a learning activity and any interaction with teachers, other learners or learning content through the use of digital technology' [19]. To calculate it, we compute four different scores.

• Interaction oriented investment (IOI): This learning indicator aims to evaluate student engagement by considering the number of interactions with the LMS, compared to the most active student. It should be noted that on the BlackBoard platform, an interaction refers to the number of clicks done by the student throughout the course executable activities (quizzes, exams) and nonexecutable activities (consultation of documents or videos). It is calculated as follows:

$$IOI = \frac{\text{Total number of interactions for each student}}{\text{maximum number of interactions for a student in the class}}$$

• Course Access Connection Oriented Investment (CACOoI): This indicator assesses students' behavioral engagement, focusing particularly on the amount of time they spend on the platform, compared with the most active student on the platform. This measure provides a better understanding of students' level of involvement and interaction with the resources and activities offered online. It is calculated as follows:

$$CACOoI = \frac{\text{Time spent on the platform by the student}}{\text{Max time spent on the platform by a student in the class}}$$

It should be noted that on Blackboard, students had the option of retaking their quiz before submitting it in order to improve their results. However, this indicator is primarily designed to assess learners with mediocre results, to find out whether they really made an effort to improve their scores, or whether they were satisfied with a single attempt, which could reflect their level of motivation and commitment. On the other hand, a low perseverance value for a student who succeeded brilliantly on his first attempt should not be interpreted as a sign of disengagement. Instead, it could be a sign of course understanding and self-confidence.

of course understanding and self-confidence.

In our case study, the only data available regarding the three scores defined above is the overall number of clicks of connections and connection time over the whole course; we do not have the values over time and this is one of the limitations of our dashboard as the indicators and visualizations provided by the dashboard can only be based on the raw data collected from the LMS. On the other hand, we could collect the number of attempts a student made for each quiz during the course. We refer to this indicator as the perseverance score and may analyze its evolution during the course.

### 3.2.2 Step 2. Data preprocessing

In this step, we prepare the raw data for the following steps which are analysis and visualization. As our data was collected from different tables and stored in a single dataset, we have proceeded to cleaning incorrect and mislabeled data. We removed incomplete and duplicate data from our dataset to avoid false results that lead Finally, we have ensured that our data is anonymized in compliance with the requirements of the General Data Protection Regulation (GDPR). We eliminated all the information that could help identify the participant such as his/her email address or his/her name.

### 3.2.3 Step 3. Data visualization

We proposed in our dashboard a set of visualizations that meet certain criteria and offer a set of features as shown in table

3. This table explains how we presented the indicators that we calculated. We used various forms of presentation, including raw data (scores, ranks, etc.) and indicators grouped together in graphs to provide an overview. We used various types of graphs, such as bar charts and line graphs, to show the tem- poral evolution of data and make comparisons between different indicators such as engagement indicators in grouped bar charts as shown in table 3. We also used scatter diagrams to show relationships between variables like the scatter plots that show evolution of student's profiles through the quizzes. The choice of chart types was made with the target audience and clarity of presentation in mind. We also ensured that our graphs were explainable, i.e., easy to interpret by a normal dashboard user and does not require any knowledge in the field of data science. We provided text descriptions for some charts like the radar charts (see table 3) and used color coding to express the level of severity of situations. In short, we developed a dashboard that is practical, user-friendly, and easy to understand by all stakeholders.

In the following section, we present the actions to be taken from this dashboard

#### 3.2.4 Step 4. Insightful actions

The main goal of Learning Analytics dashboards is to offer different stakeholders actionable insights. Our dashboard provides clear information to students and teachers so that they can take suitable actions. The student can compare his individual level to the level of the whole class in real time and catch up. The dashboard also allows teachers to identify the students who share the same learning behavior and face the same difficulties to provide them with adequate assistance according to their specific needs. We grouped the students into four profiles based on the perseverance score noted E for engagement and performance rate that we defined previously:

- Profile 1 (E+P+): The student has a high engagement score (above the median value of the class) with a positive performance, which means that this student succeeds through hard work. He/she seems to be invested in these studies and makes a remarkable effort to get good grades. The teacher can detect potential problems by providing special follow-up to students belonging to this category.
- Profile 2 (E-P+): The student has a positive performance score and a low engagement score. This student easily succeeds the quizzes as he/she can have a good mark even from the first attempt. This means that this student does not require special help as there is no risk of failure currently. However, it is important to monitor whether this student remains sufficiently stimulated in his/her studies to avoid boredom or disinterest.
- Profile 3 (E+P-): The student belonging to this category, has a low performance score despite his high engagement. This student is really dedicated to his studies, but he/she fails despite his/her efforts, therefore needs academic support in the topics in which he has difficulties.
- Profile 4 (E-P-): The student belonging to this profile has.

serious problems related to both performance and engagement. This leads us to conclude that the student may be disinterested because of problems related to the course itself which affects his results or because of external factors which may be psychological problems, family, or a bad choice of academic program. A quick intervention is then needed to avoid the risk of dropping out.

In the following section, we present the different dashboard interfaces.

## 4 Experimental Results

In this section, we present the experimental results, which we have organized according to the research questions they answer.

- **RQ1: What are the necessary indicators to support students and teachers when using LMS?**

To answer our first research question, we present how we displayed the learning indicators in our dashboard for both students and teachers. To assess student performance, we choose grouped bar charts. These diagrams illustrate the evolution of the student's grades throughout the course, from the quizzes to the final exam. They enable the student to compare his or her grades with the best and lowest marks obtained. In this way, students can see where they stand in relation to their classmates. The grouped bar charts presented in Figure 2 show the evolution of Student 7's grades through the course. We notice that this student managed to get consistently good scores for the first 4 quizzes but then suddenly he/ she had zeros for the following five quizzes (quizzes 5,6,7,8,9) which means that he/ she is no longer performant and that he/she has serious problems knowing that this student has a global performance score equal to 37,87.

We also provide students with an evolutionary view of their engagement score for each quiz during the course. The bar chart presented in Figure 3 shows the engagement score of student number 7. By comparing this figure with the previous one, we understand the reason why this student got the lowest score of 0 for the quizzes from 6 to 10. In fact, he didn't even try to answer these quizzes which proves the relevance of the indicators we have proposed. Student 7 has an overall engagement score equal to 16.35. We can conclude from these scores that he/she does not log on regularly to the LMS, does not spend enough time there and does not interact sufficiently with the different activities. These results further explain the grades he/she obtained in the various quizzes which illustrates the relationship between our different indicators for analyzing the student's behavior and deducing the main reasons for the difficulties he is facing.The teacher also has a detailed view of his students performances, as shown in the bar charts in figure 4. These charts enable him/her to analyze in detail the evolution of students' grades throughout the course and to compare the obtained results. This visualization provides the teacher with valuable information for assessing student performance. The

| LA indicators / Data | Visual Charts | Comparative data | Objective | Explainability |
|---|---|---|---|---|
| Student's ranks | Bar chart | Individual and class scores | Comparison, Temporal evolution | None |
| Student's grades | Bar chart | | | Color coding |
| Ranks and grades | Bar chart | | | None |
| Perseverance score | Bar chart combined with line chart | Individua values and class l median | Comparison Temporal evolution | None |
| | Bar chart | Compared with grades | Comparison Temporal evolution | Text |
| Engagement indicators (IOI, CACOoI, CACoI) | Grouped bar charts | Individual and class scores. | Comparison | Text |
| Performance | Bar chart | Individual and class scores | Comparison Temporal evolution | Text |
| Engagement and performance | Radar chart | Individua and average class l scores | Comparison | Text |
| Students' profiles | Scatter plot | None | Relationship between variables | Text |
| | Pie chart | Average class scores | Comparison | Text, color coding |

Figure 4: Visualizations of EX-LAD and their distinctive characteristics.

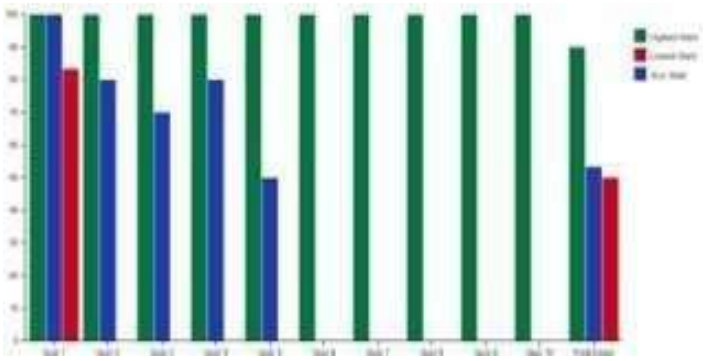


Figure 5: Evolution of the student 7's grades through the course

Bar chart presented in Figure 4 shows a comparison of students' scores and ranks in quiz number 5 which is an intermediate quiz.To view students' grades in a specific quiz, the teacher can select the desired quiz from the adjacent drop-down list (see figure 5).

This feature allows the teacher to monitor student's progress and analyze the evolution of their results through the course as he/ she can detect the dropor the progress in the student's performance from one quiz to another. Then using a drill-down

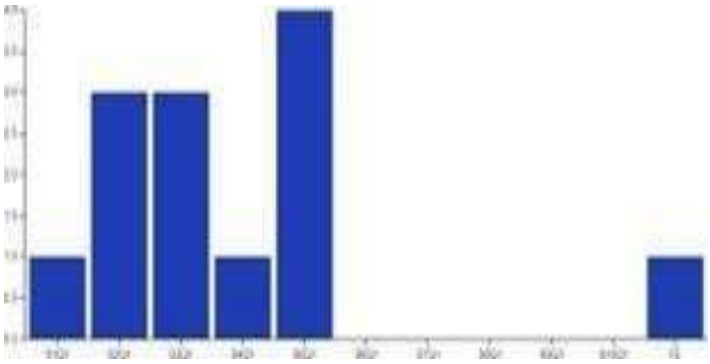


Figure 6: Evolution of the student 7's perseverancescore through the course

operation, the teacher is allowed to navigate from the whole class to visualize each student and compare his/ her values to the others as shown in Figure 5. Figure 5 shows three stacked histograms where each bar represents an engagement indicator score: IOI, CACOI and CACOoI

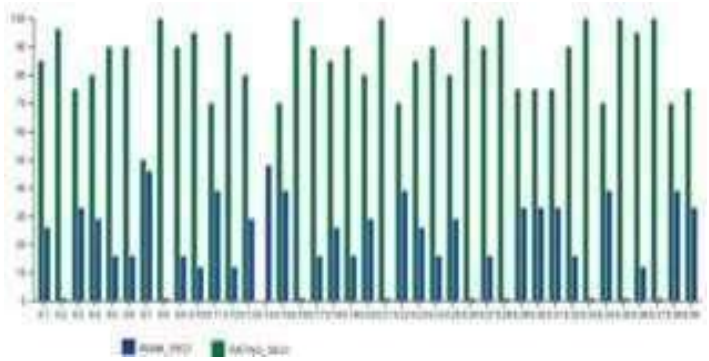


Figure 7: Comparison of students' grades and ranks for the Quiz n°5

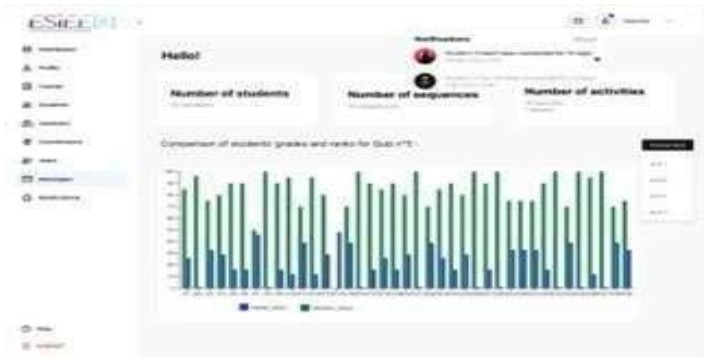



Figure 8: Teacher's interface

respectively. To ensure the readability and clarity of the visualization, we chose to present only 15 students. The teacher may wish to have an overall view of the engagement of each student over the time spent on the platform, the number of

connections and the number of clicks made online which reflects whether the student has done activities or consulted resources over the course. We have chosen to represent these three engagement indicators combined in a single figure to provide a comprehensive overview of student behavior on the e- learning platform. By visualizing these three indicators simultaneously, we can identify correlations between the number of interactions, the time spent on the platform and the frequency of connections. For example, an increase in time spent on the platform may be associated with an increase in the number of interactions, as demonstrated in the case of student 11. On the other hand, opposite scenarios can also occur, as observed with students 6 and 14. Similarly, an increase in the number of connections does not necessarily guarantee that the student spends more time on the platform, as shown by the cases of students 4 and 13. Without the combination of these three indicators, we could have falsely concluded that these two students were among the most engaged, when this was not the case. In summary, this combined representation offers a deeper and more accurate understanding of students' behavior on the platform, enabling a better assessment of their actual engagement. In this figure, we have intentionally chosen not to include the perseverance indicator, be- cause as we have already discussed above, this indicator is more relevant to students with mediocre results and is closely associated with academic performance. By focusing on the three engagement indicators combined, we aim to provide a more holistic perspective on student behavior on the e-learning platform. We can see from

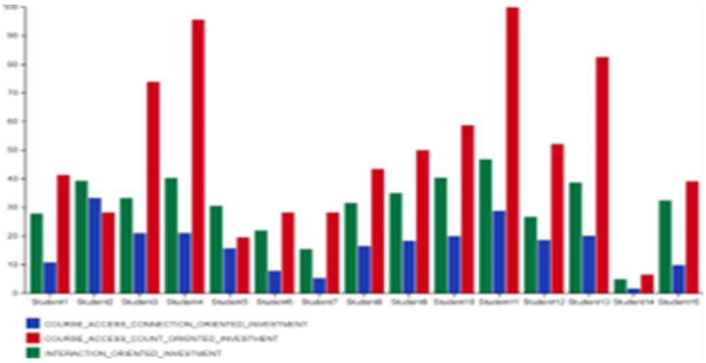


Figure 9: Overview of engagement indicators for the class

Figure 9 that Student 4 used the platform extensively as did Student 13. Both had a similar perseverance score since they made 2 attempts on quiz 5. We can then conclude that these indicators are complementary to properly characterize student engagement. In this section, we presented the various visualizations that allow us to display the indicators to our dashboard users. We demonstrated the effectiveness of these indicators and their relevance in allowing the teacher to clearly identify students with difficulties and easily conclude the type of difficulty they are experiencing, enabling him/her to intervene at the right moment and to adapt this intervention to the student's specific needs. Students can also understand their own difficulties through these detailed indicators making it easier for them to overcome these problems. However, the ability of users to understand and interpret these graphs directly may vary. This leads us to our second research question in the next section.

• **RQ2: How can we create a dashboard that is understandable and interpretable by nonspecialists in data analysis?**

To address this research question, our study focuses on the explainability of learning analytics through different graphs that are easy to understand and interpret by the different dashboard users .We demonstrated the importance of our proposed learning indicators in the previous section. This section is dedicated to the remaining criteria. First, we ensured our dashboard offered comparative views for both teachers (see figure 9) and students as shown in figure 10. Figure 10 offers a

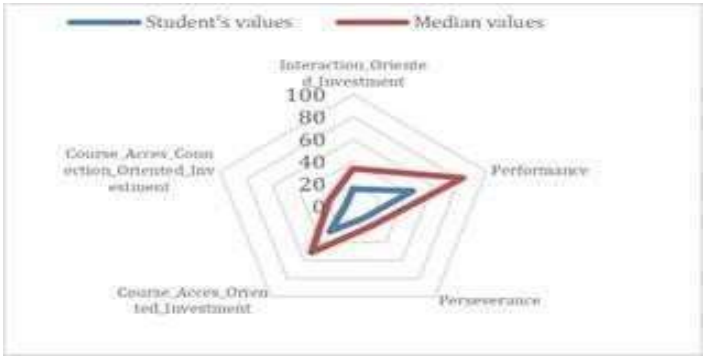



Figure 10: Global view of the student 7's engagement and performance indicators.

using the proposed formulas, through a radar graph. This radar graph highlights performance indicators, perseverance and engagement scores, comparing them with median scores. This individualized view helps students situate themselves in relation to their peers and analyze efficiently their own academic problems. They can therefore understand their results which enables them to adopt the right measures to improve their academic performance. The bar chart in figure 8 demonstrates the relation- ship between the engagement and performance globalindicators for the whole class .This enables the teacher to confirm the results we have seen in the detailed views and thus take the right decision since he can understand that not only academic performance shouldbe used to evaluate the student as engagement may also influence these results. Another important criterion for

achieving EX-LAD is to transform recommendations and predictions into actionable steps. In other words, it is not enough just to provide information, but also to facilitate decision-making and action based on this information. In fact, we also considered the feasibility of actions in our solution .We proposed different student profiles calculated according to their performance and engagement indicators. Instead of applying similar interventions to all students, we focused on tailoring actions to these profiles. These profiles may be detected with

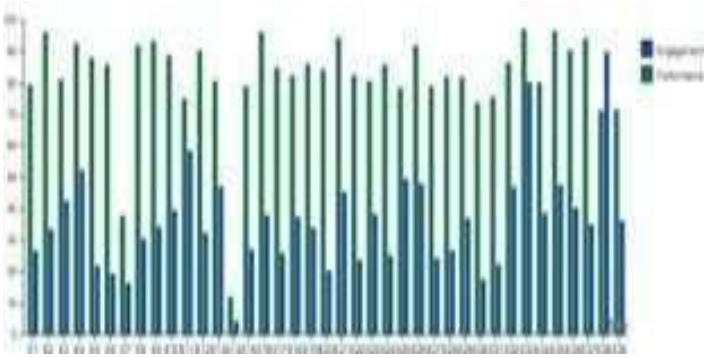


Figure 11: Overview of engagement and performance global indicators for the class.

the scatterplots shown in figure 12. Figure 12 shows students' profiles' evolution through the course quizzes highlighting the relationship between performance and perseverance. This allows the teacher to identify specific students of a given profile and follow his/her individual evolution. Our goal is to help teachers to identify the students who share the same learning behavior and face the same difficulties to provide them with adequate assistance according to their specific needs. In addition,

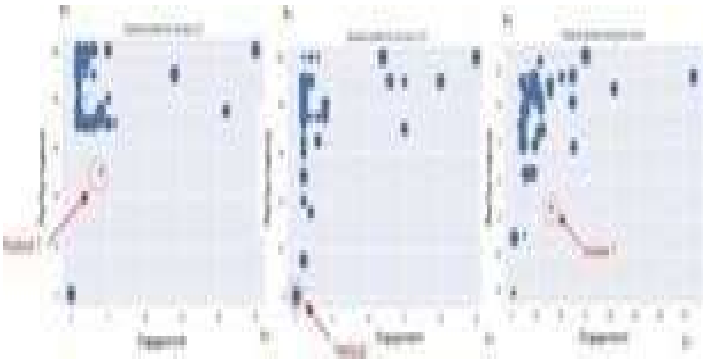


Figure 12: Overview of engagement and performance global indicators for the class.

we also adopted the use of significant color coding in certain figures to emphasize the seriousness of the situation. This allows users to quickly grasp key information and identify important aspects of the data presented. The Bar charts in Figure 10 presents the evolution of this student's grades and perseverance score as well as his grades and his rank in each quiz. Student 7 had good grades for the first four quizzes however his results decreased for the following tests despite his efforts shown by his numerous attempts to respond correctly. We proposed a specific color code to highlight the significance of the presented values. Red was used to express seriousness of the situation and that an immediate intervention should be done after these dissatisfactory results. Green was used to express positive results. The choice of traffic lights' colors allows users to easily identify the indicators that need particular attention which facilitates the interpretation and decision-making. Our dashboard offers a variety of visualizations, each aimed at a specific objective, making it easier to interpret the displayed results. We have opted for bar charts or radars to provide a comparative view, scatter plots to demonstrate relationships between variables as well as pie charts.

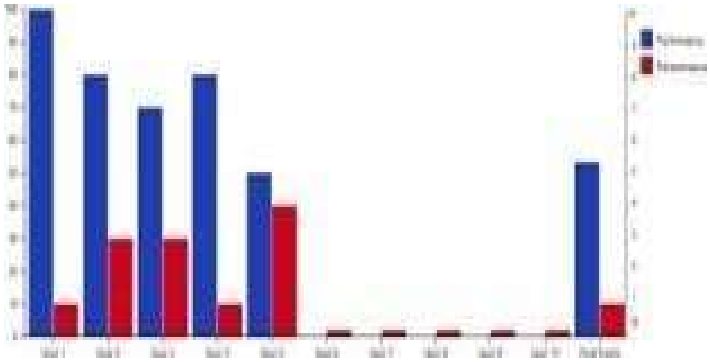


Figure 13: Overview of engagement and performance global indicators for the class.

Our dashboard offers a personalized approach that facilitates the identification of problems that are common for each group of students and allows the teachers to provide them with specific interventions tailored to their needs. This enables the students to improve their academic results and boosts their engagement and motivation.

## 5 Discussion

We have successfully developed a student-centered dashboard aimed at empowering students to self-assess and enhance their learning journey, while equipping teachers with the necessary tools to monitor progress and identify those at risk of academic setbacks, enabling timely intervention. Ensuring the dashboard's accessibility to all stakeholders was a key priority to maximize its effectiveness. However, we encountered several challenges along the way. Understanding the database structure of the Learning Management System (LMS), particularly Blackboard Learn, proved to be a significant hurdle. Efficiently accessing and extracting data and metadata necessitated an in-depth examination of the system and data management practices. Moreover, the dashboard has certain limitations associated with the available raw data. For instance, some indicators cannot be recalculated over time, hindering the representation of longitudinal trends. For example, data on clicks and LMS accesses were only available for the entire course duration, rather than at different time intervals. Furthermore, we faced challenges related to compliance with GDPR regulations.Securing consent from all students can be challenging, resulting in a limited dataset and potentially compromising result quality. In cases where certain student profiles are under or over-represented in the data, biases may be introduced. It is crucial to consider these challenges when

implementing data-driven techniques.

## 6 Conclusion and Future works

A crucial aspect of our proposed dashboard is to ensure that the proposed visualizations are comprehensible to all users, as part of the Explainable Learning Analytics (EX-LA). This means that the presented information are clear and easy to interpret, enabling every user, whether student or teacher, to quickly draw relevant conclusions from data analysis. By integrating explicit and intuitive visualizations, we strive to ensure that our dashboard is truly informative and useful for all players involved in the learning process. We attach great importance to trust and transparency in the use of data. Therefore, our dashboard offers a textual explanation of the indicators calculated and used in the visualizations. User friendliness of the dashboard is an essential consideration. Ethics is a fundamental aspect of our solution. Although we provide students with comparative visualizations to encourage them to situate themselves in relation to their peers, we took care not to mention the name of any student when displaying best and worst grades. In this way, we respect the confidentiality and protection of students' personal data. We integrated as well, a chat section enabling students to decide whether they wish to communicate directly with their teachers and receive personalized interventions. Our solution aims to maximizing the success of all students, not just those experiencing difficulties. This is demonstrated by the assistance offered to students with the E+P+ profile who have no difficulties. We value equal opportunities and promote success for all. In this article, we analyze the evolution of student performance over time. However, due to the insufficient temporal granularity of the raw data, we are unable to conduct an in-depth study of the evolution of student engagement. In our ongoing research, we aim to utilize richer data with a finer temporal granularity to align with the objectives of studying indicator evolution. Our objective is to enhance our ability to detect student difficulties early. While this article presents representations of indicators based on measured data, our future direction involves leveraging these data to predict the evolution of student difficulties using machine learning techniques. We are committed to maintaining transparency in these predictions, ensuring that the criteria used for predictions are clearly communicated to end-users, whether they are students or teachers. This approach fosters understanding and confidence in the predictive processes. The outlined requirements and concerns underscore the importance of having a large dataset with a substantial number of observations, allowing for the calculation of numerous indicators over time. Given the complexity of this task, exploring alternative solutions such as leveraging existing datasets is under consideration. However, comprehensive comparisons of available datasets, including their characteristics and ethical assurances, are lacking. Therefore, a detailed assessment of available datasets remains a major focus of our ongoing work.

## 7 Acknowledgements

The authors would like to thank ESIEE-IT for al- lowing us to collect the educational raw data resulting from students' interactions with the Blackboard-Learn platform.

**References**

[1] Samantha L Schneider and Martha Laurin Coun- cil. Distance learning in the era of covid-19. Archives of dermatological research, 313(5):389– 390, 2021.

[2] Blackboard learn — responsive advanced lms — blackboard, https://www.blackboard.com/.

[3] Shiful Islam Shohag and Masum Bakaul. A machine learning approach to detect student dropout at university. International Journal, 10(6), 2021.

[4] Monica Ciolacu, Ali Fallah Tehrani, Leon Binder, and Paul Mugur Svasta. Education 4.0- artificial intelligence assisted higher education: early recognition system with machine learning to support students' success. In 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging(SIITME), pages 23–30. IEEE, 2018.

[5] Muhammad Adnan, Asad Habib, Jawad Ashraf, Shafaq Mussadiq, Arsalan Ali Raza, Muhammad Abid, Maryam Bashir, and Sana Ullah Khan. Predicting at-risk students at different percentages of course length for early intervention using machine learning models. Ieee Access, 9:7519– 7539, 2021.

[6] Kelly Linden, Neil van der Ploeg, and Noelia Roman. Explainable learning analytics to identify disengaged students early in semester: an intervention supporting widening participation. Journal of Higher Education Policy and Management, pages 1–15, 2023.

[7] Tinne De Laet, Martijn Millecamp, Tom Broos, Robin De Croon, Katrien Verbert, and Raphael Duorado. Explainable learning analytics: challenges and opportunities. In Companion Proceedings of the 10th International Conference on Learning Analytics Knowledge LAK20 Society for Learning Analytics Research (SoLAR), pages 500–510, 2020.

[8] Rangana Jayashanka, E Hettiarachchi, and KP Hewagamage. Technology enhanced learning analytics dashboard in higher education. Electronic Journal of e-Learning, 20(2):151–170, 2022.

[9] Teo Susnjak, Gomathy Suganya Ramaswami, and Anuradha Mathrani. Learning analytics dashboard: a tool for providing actionable insights to learners. International Journal of Educational Technology in Higher Education, 19(1):12, 2022.

[10] Yassine Safsouf, Khalifa Mansouri, and Franck Poirier. Tabat: design and experimentation of a learning analysis dashboard for teachers and learners. Journal of Information Technology Education, 20:331–350, 2021.

[11] Mehmet Koko¸c and Arif Altun. Effects of learner interaction with learning dashboards on academic performance in an e-learning environment. Be- haviour & Information Technology, 40(2):161–175, 2021.

[12] Jeongyun Han, Kwan Hoon Kim, Wonjong Rhee, and Young Hoan Cho. Learning analytics dashboards for adaptive support in face-to- face collaborative argumentation. Computers Education, 163:104041, 2021.

[13] Raphael A Dourado, Rodrigo Lins Rodrigues, Nivan Ferreira, Rafael Ferreira Mello, Alex Sandro Gomes, and Katrien Verbert. A teacher-facing learning analytics dashboard for process-oriented feedback in online learning. In LAK21: 11th International Learning Analytics and Knowledge Conference, pages 482–489, 2021.

[14] Fatma Gizem Karaoglan Yilmaz and Ramazan Yilmaz. Learning analytics as a metacognitive tool to influence learner transactional distance and motivation in online learning environments. Innovations in Education and Teaching International, 58(5):575–585, 2021.

[15] Rebecca L Sansom, Robert Bodily, Caroline O Bates, and Heather Leary. Increasing student use of a learner dashboard. Journal of Science Education and Technology, 29(3):386–398, 2020.

[16] Shiva Shabaninejad, Hassan Khosravi, Solmaz Abdi, Marta Indulska, and Shazia Sadiq. Incor- porating explainable learning analytics to assist educators with identifying students in need of attention. In Proceedings of the Ninth ACM Conference on Learning@ Scale, pages 384–388, 2022.

[17] Tesnim Khelifi, Nourh'ene Ben Rabah, Ibtissem Daoudi, B´en´edicte Le Grand, and Farah Barika Ktata. Intelligent prediction- intervention approach to support students' success in web-based learning environments: A case study in higher education. In 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI- IAT), pages 256–262. IEEE, 2022.

[18] L´´ecole de l'expertise num´erique, https://www.esiee-it.fr/fr.

[19] Lisa R Halverson and Charles R Graham. Learner engagement in blended learning environments: A conceptual framework. Online Learning, 23(2):145–178, 2019.

# The Execution of the Partition Problem: A Comparative Study of Various Techniques for Efficient Computation

Pratik Shrestha *
Grand Valley State University, Grand Rapids, Michigan, USA.
Chirag Parikh †
Grand Valley State University,Grand Rapids, Michigan, USA.
Christian Trefftz ‡
Grand Valley State University, Grand Rapids, Michigan, USA.

June 20, 2024

**Abstract**

Exponential-time algorithms for solving intractable problems are considered inefficient when compared to polynomial-time algorithms for solving tractable problems. The reason being that the execution time for former grows rapidly as problem size increases. A problem is considered NP- complete when a problem is non-deterministic polynomial (NP) and all other NP-problems are polynomial-time reducible to it. The partition problem is one of the simplest NP- complete problems. Many real-life applications can be modeled as NP-complete problems, and it is important for software developers to understand the limitations of existing algorithms that can solve those problems. Solving the partition problem is a time-consuming endeavor. Exact algorithms can find solutions, in a reasonable amount of time, only for small instances of these problems. Large instances of NP-hard problems will take so long to solve with exact algorithms, that for practical purposes those large instances should be considered intractable. The execution time required to find a solution to instances of the partition problem is greatly reduced by using parallel processing counterparts such as Graphics Processing Unit (GPU) and Field Programmable Gate Array (FPGA). In this paper, we talk about the use of the NVIDIA T4 GPU and PYNQ FPGA board in conjunction with an overlay to accelerate the execution of a function that evaluates if a partition is a solution to an instance of the partition problem. To assist with the evaluation, four different overlays are createdfor FPGA and performance comparison among them with GPU using python and the numba/cuda library is then presentedin the paper.

Parallel processing is a topic of growing importance in the computing world. The exponential growth of processing and network speeds means that parallel architecture is not just a good idea but now a necessity. Many problems require an enormous amount of time to be solved. For e.g., exponential-time algorithms take longer for solving intractable problems in comparison to their polynomial- time algorithm counterpart for large problem sizes. In addition, parallel systems have proven to be the only alternative to obtain solutions in a reasonable amount of time. Hence, there is lot of recommendations for curricula of computer science undergraduate degrees to emphasize on topic of parallel processing. Introductory courses in parallel processing include surveys of different computer architectures: Shared memory machines with microprocessors comprising of several cores, Graphics Processing Units (GPUs) and clusters of computers, among others. Field Programmable Gate Arrays (FPGAs) on the other hand have proven to be efficient accelerators for the execution of many different applications [1]. Hence, it is of benefit to have the topic of FPGAs be included in a course in parallel processing. The challenge faced by an instructor who wants to cover FPGAs in a parallel processing course is that programming FPGAs requires a very strong background and skills in hardware design that most computer science students lack. To assist with this, Xilinx has created a board called PYNQ [2] for pedagogical purposes that can be easily programmed using Python without the need to be proficient in hardware design. Figure 1 shows the PYNQ board.

*Reasearch Assistant Email: shrestpr@mail.gvsu.edu.
†Professor and Chair of Computer Engineering program at Grand Valley tate University. Email: parikhc@gvsu.edu.
‡Professor at the College of Computing at Grand Valley State University . Email: trefftzc@gvsu.edu.


**Figure 1: PYNQ Development kit from Xilinx**

PYNQ board contains an FPGA device with a built-in Arm microprocessor that has two cores and a programmable fabric. PYNQ board runs a custom version of Linux and are therefore considered as a stand-alone computer. A PYNQ board can connect to a traditional computer through an Ethernet cable and a USB cable. Xilinx has chosen Jupyter notebooks to provide a very convenient way of interacting with a PYNQ board. The PYNQ board can run a web server that interacts with a python interpreter. The user can start a browser on his/her computer and access web pages on the server running on the PYNQ board. Those web pages may contain python code that will be executed on the PYNQ board. The Python interpreter on the PYNQ board can interact with overlays, which are configurations of the programmable fabric of the FPGA that can execute specific functions. On the other hand, a GPU is a specialized electronic circuit designed to accelerate graphics rendering, primarily used in rendering images and videos for computer displays [7]. Over time, GPUs have evolved beyond their original graphics-centric purpose, finding significant applications in parallel processing tasks. Due to their parallel architecture and capability of high-performance computing,GPUs excel in handling large amounts of data simultaneously, making them highly efficient for parallelizable tasks such as scientific simulations and artificial intelligence (AI) computations. The environment used to implement the version of the program that uses a GPU was Google COLAB. COLAB executes on a dedicated machine with a Xeon microprocessor.

and a NVIDIA T4 GPU. [8]. A T4 GPU has 2560 cores. It has 16 gigabytes of main memory, and it is connected to the host computer using a PCI Express bus (version 3.0 x16). In this work, the GPU is utilized to solve the NP-Complete problem through parallelization. A problem is considered NP-complete when a problem is non-deterministic polynomial (NP) and all other NP-problems are polynomial-time reducible to it. In this paper, we describe the process of creating an overlay to accelerate the execution of a python program that finds a solution to the partition problem, a problem that belongs to the "NP- complete" category of problems. Algorithms to find exact solutions to problems in this category are very time consuming The rest of this paper is structured as follows: The partition problem is described in section 2 followed by a "brute force" approach to solve the Partition problem outlined in section 3. process of creating the overlay is described in section 5 followed by experimental results and conclusions in sections 6 and 7 respectively.

## 2 The Partition Problem

In the world of computer science, partition problem or sometimes called as number partitioning [3] is the task of deciding when given a multi-set of positive integers S, if it can be partitioned into two sub multi-sets S1 and S2 such that the sum of the elements is S1 is equal to the sum of the elements in S2? Consider the following example: Let S be the multi-set 4,5,9. In this particular case it is evident that the answer to the problem is yes: We partition the multi-set into two sub multi-sets S1: 4,5 and S2: 9. The partition problem is one of the simplest NP-complete problems. NP-complete problems are very interesting for several reasons. Many real-life applications can be modeled as NP-complete problems, and it is important for software developers to understand the limitations of existing algorithms that can solve those problems. Exact algorithms can find solutions, in a reasonable amount of time, only for small instances of these problems. Large instances of NP-hard problems will take so long to solve with exact algorithms, that for practical purposes those large instances should be considered intractable. Other alternatives are available (heuristics, approximation algorithms) but the solutions produced by these alternatives are likely to be sub-optimal. NP-complete problems are yes/no questions. The letters NP stand for Non-deterministically Polynomial. These problems have the characteristic that it is possible to write an algorithm with Polynomial execution time that will be able to determine if a candidate solution is indeed a solution to the problem or not. The challenge for this family of problems

Implementation of the partition problem using NVIDIA T4 GPU is described in section 4. is to generate the appropriate candidate solution. To this day, the exact time complexity of algorithms that solve NP-complete problems is not known. So far, the only algorithms that generatethe proper candidates for a problem have exponential time complexity. The consensus among most practitioners is that it

is not feasible to find algorithms with better time complexity to generate the proper candidates. NP-complete problems have another interesting property: Algorithms to transform the input of one NP-complete problem to other NP-complete problems exist. Those algorithms are called “reductions” and they have polynomial complexity. When a researcher comes across a new problem and wants to show that this is an NP-complete problem, the proof is the description of a “reduction” to an existing NP-complete problem. Thus, if anybody were to write an exact algorithm with polynomial time complexity to solve an NP-complete problem, all the problems in the class could be solved in polynomial time as well, thanks to the existing reduction algorithms. It is important for software developers to be aware of the existence of NP-complete problems. Some problems from real life are in this category. It is important to be able to tell the users that only small instances of problems in this class can be solved exactly in reasonable amounts of time. Large instances of these problems are intractable, and it becomes necessary to use other alternatives, to use approximation algorithms or to use procedures that may not produce optimal solutions. The next section talks about the exact algorithm to solve the partition problem.

## 3 An Exact algorithm to solve the Partition problem.

Woeginger [4] has observed that there is a subset of NP-complete problems that can be solved by brute-force by enumerating exhaustively all the possible subsets (the power set) of a particular set of elements. For each of those possible subsets, one uses a function that evaluates if that subset is a solution to the problem of interest. One then proceeds to choose, among the subsets that are possible solutions, the one that works best. Other NP-hard algorithms that can be solved using the same brute-force approach include the maximum-clique problem, the maximum independent set problem and the minimum dominating set problem. If we wanted to explore the power set of the multi-set S, we could do it by observing that the binary representation of the integers between 1 and 2n-1 - 1 encode the possible subsets of interest. Notice that the other values between 2n-1 and 2n - 2 are symmetrical to the values considered. Table 1 illustrates the values for the example in the previous section: S = 4,5,9. The indices for the different encodings of the subsets are listed on the first column, Index, on Table 1. The binary encoding is listed on the second column. The rightmost digit encodes to the subset to which element 1 belongs, the middle digit encodes the subset to which element 2 belongs and the leftmost digit encodes the subset where node 3 belongs. Take the entry that corresponds to 3: 011. This is interpreted as subset 1 (encoded by 0) containing element 3 and subset 2 (encoded by 1) containing elements 1 and 2. The table contains all the integers between 0 and 7(23 - 1), but it is not necessary to consider the value 0, nor the value 7. Observe that the values between 0 and 3 are symmetrical to the values between 4 and 7; the values are each other’s complements, 1 (001) is the complement of 6 (110), 2 (010) is the complement of 5 (101), and 3 (011) is the complement of 4 (100).

| Index | Binary encoding | Solution |
|---|---|---|
| 0 | 000 | No |
| 1 | 001 | No |
| 2 | 010 | No |
| 3 | 011 | Yes |
| 4 | 100 | Yes |
| 5 | 101 | No |
| 6 | 110 | No |
| 7 | 111 | No |

**Table 1.** Indices, subsets, and solutions for an instance of the partition problem

Notice that the set of possible subsets of interest is encoded by the set of integers in the range between 1 and 2n-1 - 1. As soonas an algorithm finds a possible partition of the multiset, the algorithm can stop and the answer for this particular instanceof the problem yes. If all possible partitions are considered and no possible satisfying partition is found, the answer for thisparticular instance of the problem is No. The outline of the main algorithm is shown in Figure 2. As can be observed, the complexity of the algorithm is O(2n), exponential.

This algorithm can be easily parallelized using environments like OpenMP, for shared memory machines, Thrust, for GPUs, or MPIfor clusters or computers. The evaluation of each possible partition can be carried out independently from the evaluation of the other possible partitions. On computing platforms with several processors, every processor can evaluate a possible partition in parallel with other processors evaluating other possible partitions [5]. Most of the execution time of theprogram is spent in the function that evaluates if a particular partition is a solution for the problem. In the next section, we use NVIDIA T4 GPU to speed-up the execution of the function.

**Algorithm 1** Algorithm to solve instances of the Partition problem

```
input: n size of the problem, array: values in the multiset
output: true or false
indexOfPossiblePartition = 1
while indexOfPossiblePartition < 2^n − 1 do
    if evaluatePossiblePartition(indexOfPossiblePartition, n, array) then
        return true
    else
        indexOf PossiblePartition + +
    end if
end while
return false
```

**Algorithm 2** Algorithm to evaluate if partition is the solution

```
Input: n, array that contains the values, indexOf PossiblePartition
Output: true or false
sumOfValuesInPartition0 = 0
sumOfValuesInPartition1 = 0
index = 0
while index < n do
    if bit index in the binary representation of indexOfPossiblePartition is 0 then
        sumOfValuesInPartition0+ = array[index]
    else
        sumOfValuesInPartition1+ = array[index]
    end if
    index + +
end while
if sumOfValuesInPartition0 = sumOfValuesInPartition1 then
    return true
else
    return false
end if
```

## 4 Implementation using NVIDIA T4 GPU

The NVIDIA T4 Tensor Core GPU is a powerful accelerator designed for various cloud workloads, including deep learning, machine learning, data analytics, and video transcoding. Even though the NVIDIA T4 Tensor Core GPU is not specifically designed to solve NP-Complete problems directly, however, it can significantly accelerate certain aspects of solving such problems due to its parallel processing capabilities and high throughput. In terms of performance, the T4 delivers up to 40 times higher performance than CPUs [9]. This implementation was performed in Google Colab using Nvidia T4 GPU as shown in Figure 4 below.

Python lacks native support for GPU programming, but developers can leverage the NUMBA/CUDA library to write code tailored for NVIDIA GPUs. This library harnesses the

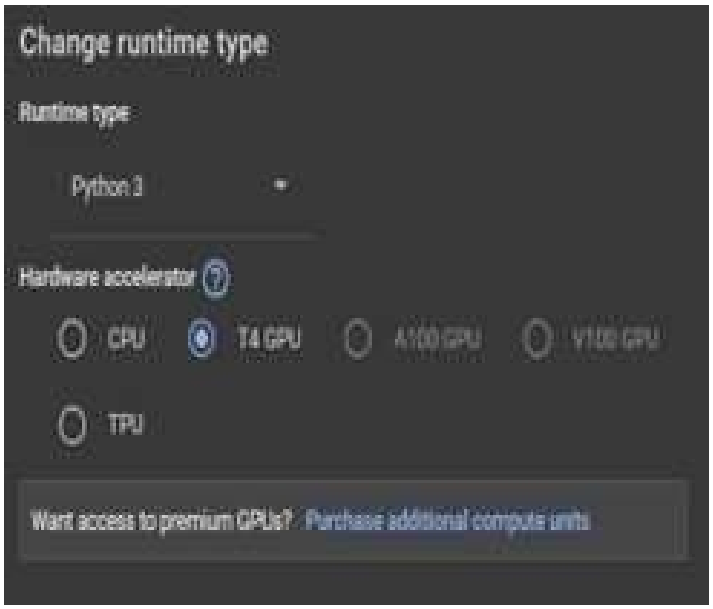



Figure 1: Selection of Runtime type

LLVM compiler infrastructure to generate binary executable code. Through the NUMBA library, specific functions within a Python program can be compiled into native binaries. Decorators are applied to signal which functions should undergo compilation, guiding the library towards optimization. Upon successful compilation, function calls execute as binary code, resulting in accelerated performance compared to the interpretation process of standard Python code. The NUMBA/CUDA library, a subset of NUMBA, specializes in producing GPU- executable code. Decorators are utilized just before functions intended for GPU execution. The evaluate partition has been decorated as shown below:

```
@cuda.jit
def evaluatePartition(array:DeviceNDArray,result:DeviceNDArray,n:np.dtype=np.int64):
```

Figure 2: Decorator used in the implementation.

In addition to utilizing decorators, developers must employ additional functions when working with GPUs. GPUs function as distinct computing units with their own dedicated memory. Hence, it is essential to transfer variables, typically arrays, from the host memory to the GPU's memory. Once the relevant variables have been transferred to the GPU's

memory, the code designated for GPU execution is invoked. This requires calling a function that has been compiled with the necessary decorators. Developers must specify the size of the arrays that the functions will operate on. As previously mentioned, the code executing on the GPU corresponds to

a function marked with the apropriate decorator. Upon completion of the GPU code, the program retrieves the desired results and transfers them back to the host's memory. figure below highlights the most important actions in the code to interact with the GPU:

- Moving arrays to the GPU memory
- Executing the code on the GPU
- Finally copying the results from the GPU memory to the host memory

```
# Copy variables to the GPU memory
arrayGPU = cuda.to_device(array)
resultGPU = cuda.to_device(result)
# Execute the function in the GPU
evaluatePartition.forall(nPartitions)( arrayGPU,resultGPU, n)
# Copy the result array back to the CPU
resultGPU.copy_to_host(result)
```

Figure 3: Important steps while interacting with GPU.

As discussed in the previous section, the essence of the algorithm is to execute a function that evaluates if a particular integer value is an encoding of a partition that solves the instance of the problem. All the integers that encode subsets (elements) from the power set can be evaluated in parallel. If the number of elements to be evaluated is larger than the number of cores available in the GPU, the GPU operating system takes care of executing the code the necessary number of times so that the function is executed on all the elements. On the COLAB notebook mentioned before, an instance of size 25 took 1.024 seconds to execute. In the next section, we use the programmable fabric of the FPGA to accelerate the execution of that function

## 5 Implementation on an FPGA using an Overlay

Overlays, also known as Hardware libraries, are programmable/configurable FPGA designs that

extend the user application from the Processing System into the programmable logic [6]. They are extremely useful to accelerate a piece of software using a hardware platform for a particular application. The software programmer can use an overlay in a similar way to a software library to run some of the applications on an FPGA as overlays can be loaded into the FPGA dynamically. This allows software programmers to take advantage of FPGA capabilities without having detailed knowledge about the low-level hardware design. All they have to worry about is the top-level program. Creating an Intellectual Property (IP) core using High Level Synthesis (HLS) is the very first step required to create a custom overlay. For the HLS portion of this design, Xilinx's Vivado HLS was used. Different pragmas were inserted in a C program to boost the efficiency. After the successful creation of the IP core, the IP component is imported into the Vivado Suite. In the block diagram shown in Figure 4, the Zynq processor is connected to the custom IP. For this work, the High- performance AXI bus is chosen explicitly to boost up the execution. After successful synthesis of the overlay, the bitstream is then generated. This step produces .BIT and .HWH files which are then stored in the working directory inside the PYNQ board.

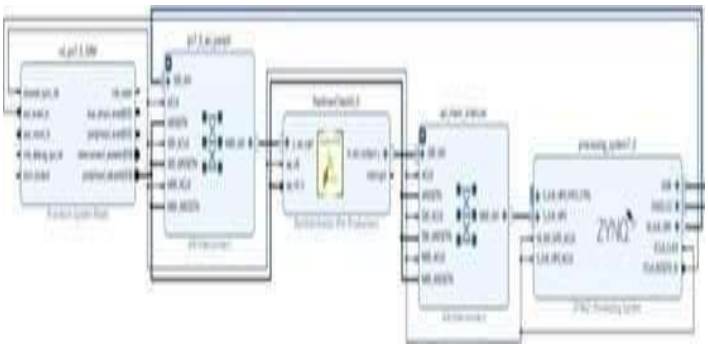


Figure 4: Block design of the overlay

To interact with the IP, first the overlay must be loaded into the Jupyter notebook which contains the IP. The PYNQ board must be physically connected to the PC for this step as all the rest of the process will be done in PYNQ board. This step has been depicted in Figure 8 below using the Python code. Here, the overlay "PartitionCheckII" has been imported. Then the next line indicates that the overlay consists of an IP PartitionCheckII0 which is the IP of interest here

```
In [1]: from pynq import Overlay
        from pynq import Xlnk
        import numpy as np

        ol=Overlay('PartitionCheckII.bit')
        sqrt_ip=ol.PartitionCheckII_0
```

Figure 5: Import Overlay

This overlay can be thought as a block, as shown in Figure 6, which takes an array as the input and produces single output, 1 or 0, indicating if the given numbers can be partitioned or not. The very first element of the array indicates the total numbers present in the array.

As can be seen clearly from the overlay block in Figure 6, there are two ports in total. Each of them has their own physical memory address used as Memory Mapped Input


Figure 6: Overlay block

Output (MMIO) for I/O operation. In the code snippet shown in Figure 10 below, the address 0x18 is used as the input address for the array and 0x10 is used as output address. The bit value 1 in the address 0x00 indicates beginning of the process.

```
In [43]: import time


         numbers=[25,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,24]
         length=len(numbers)
         inpt=Xlnk().cma_array(shape=(length,),dtype=np.int32)
         np.copyto(inpt,numbers)
         start = time.time()
         sqrt_ip.write(0x18,inpt.physical_address)

         ##sqrt_ip.write(0x18,length)
         sqrt_ip.write(0x00,1)

         #wait untill the result is ready in the memory. Sets the ap_idle to 1 when ready
         while sqrt_ip.read(0x00)!= 0x04:
             pass

         end = time.time()
         print("Done!! total:",end-start)
         sqrt_ip.read(0x10)

         Done!! total: 15.033666849136353

Out[43]: 1
```

Figure 7: Implementation of the overlay

Execution time is another important aspect of this work as the main goal is to accelerate the partition problem using FPGA. To measure the execution time, the "time" module is imported and used. In Figure 10 above, the array "numbers" include 26 elements where the first element, 25, denotes that there are a total 25 numbers which are to be partitioned. The output "Done" indicates the execution has been completed with the time consumption of 15.03 seconds. Along with the above-mentioned overlay, three additional overlays were created for this work bringing the total to four overlays. The next section discusses the remaining overlays in brief along with their implementation results.

## 6 Experiments, Results and Analysis

Altogether four different versions of overlays were created for this work, every

overlay had slight modifications. Thus, the experiments were conducted with four different methods. For reference purposes, a partition program was created in native python to compare with the obtained result. The followings are the details about methods used in the project:

1. Method 1: In this method, an overlay was created with inputs 'n' and 'array '. The S_AXILITE Bus was used instead of the High- Performance AXI bus. A loop was used to assign every array element to every memory location which made this overlay significantly slower. This method performed well with up to 'n=20' but with 'n=25' the execution time took so long that the execution had to be stopped forcefully
2. Method 2: In this method, the overlay was created with 'n=25' defined (hardcoded) inside the overlay. Thus, the only input was 'array []'. Instead of assigning every array element one by one into the memory addresses in the FPGA, it uses the AXI Burst method (High performance AXI Bus) which improves the execution time drastically. As this method explicitly uses 'n=25' inside the overlay, this overlay performs efficiently only for 'n=25'. The numbers in array can be changed. This method produces result efficiently for instances of the problem of this specific size, but the user does not want a software implementation with this restriction. 3. Method 3: in this method, the overlay was created with inputs 'n' and 'array []'. This overlay is similar to the one created in method 1 but this time the bus used is High Performance AXI bus instead of S_AXILITE bus. Also, this overlay uses AXI burst method to transfer array numbers into the memory addresses instead of using a loop and transferring data one by one. This method is also significantly faster than method 1 but not as much as method 2 for 'n=25'. The good thing with this method is that it provides the user flexibility to change the value of 'n' unlike method 2 which only works efficiently with 'n=25'. 4. Method 4: In this method, the overlay uses 'array []' as the only input. The very first element of the array denotes the value of 'n' in this method. Then rest of the elements denotes the numbers that needs to be partitioned. If thefirst element in the array is 25, which means 'n=25' and there are 25 numbers after the first element in the array. This method uses High Performance Bus for data transfer and uses AXI Burst method instead of transferring one data at a time. This method produced the same results as method 3. Table 2 below summarizes the results obtained with the methods discussed above. Similarly, Table 3 illustrates the execution time achieved using various methods for different values of 'n'.

It can be deduced from Table 4 shown below that as the value of 'n' increases the speed factor increases. Hence, computing this problem in FPGA and NVIDIA GPU is much more efficient if the instance of the partition problem is larger. Particularly for n=25, NVIDIA T4 GPU turned out to be the fasted among all, which is 979 times faster than pure python executed in CPU. If

| Methods | Data Transfer method | Inputs | Execution Time |
|---|---|---|---|
| Method 1 | Loop | n, Array () / | - |
| Method 2 | AXI Burst | Array () with n=25 fixed | 8.05 sec |
| Method 3 | AXI Burst | n, Array () | 15.03 sec |
| Method 4 | AXI Burst | Array () with first element as 'n' | 15.03 sec |

**Table 2.** Results obtained for n=25

| | Python (CPU) | Method 1 (FPGA) | Method 2 (FPGA) | Method 3 (FPGA) | Method 4 (FPGA) | NVIDIA T4 (GPU) |
|---|---|---|---|---|---|---|
| n = 10 | 0.01 | 0.218 | 0.038 | 0.0059 | 0.0047 | 0.338 |
| n = 15 | 0.62 | 9.58 | 0.019 | 0.022 | 0.02 | 0.355 |
| n = 20 | 25.83 | 389.21 | 0.25 | 0.39 | 0.38 | 0.39 |
| n = 25 | 1002.59 | n.a. | 8.05 | 15.03 | 15.03 | 1.024 |

**Table 3.** Execution time for various values of n in seconds

the size of the instance problem is smaller, then it might not be significantly faster. From Table 4, one can see that method 2 is124 times faster than pure python code when n= 25. Table 4. Execution time speed factor versus the pure python code

| | n = 10 | n = 15 | n = 20 | n = 25 |
|---|---|---|---|---|
| Method 1 | x 0.045 | x 0.06 | x 0.066 | Too long |
| Method 2 | x 2.63 | x 32.63 | x 103.32 | x 124.47 |
| Method 3 | x 1.69 | x 28.18 | x 66.23 | x 66.66 |
| Method 4 | x 2.12 | x 31 | x 67.97 | x 66.66 |
| NVIDIA T4 | x 0.029 | x 1.74 | x 66.23 | x 979.09 |

**Table 4.** Execution time speed factor versus the pure python code

Figure 8 below shows the graphical representation of theresults achieved using various methods. From the above result,

method 1 is the slowest among all the methods as it uses loop technique to transfer the array values into the memory address into the overlay. In method 1, the execution for 'n=25' took too long so eventually the process had to be stopped. Thus, there is no data for that particular size. Also, it can be concluded from the Table 4 and Figure 8 that the methods 2, 3 and 4, which use HLS as well as AXI Burst technique for data transfer, are faster in execution than compared to pure python code without HLS. Additionally, particularly for n=25, NVIDIA T4 GPU is the fastest among which has the execution time of 1.024 sec.

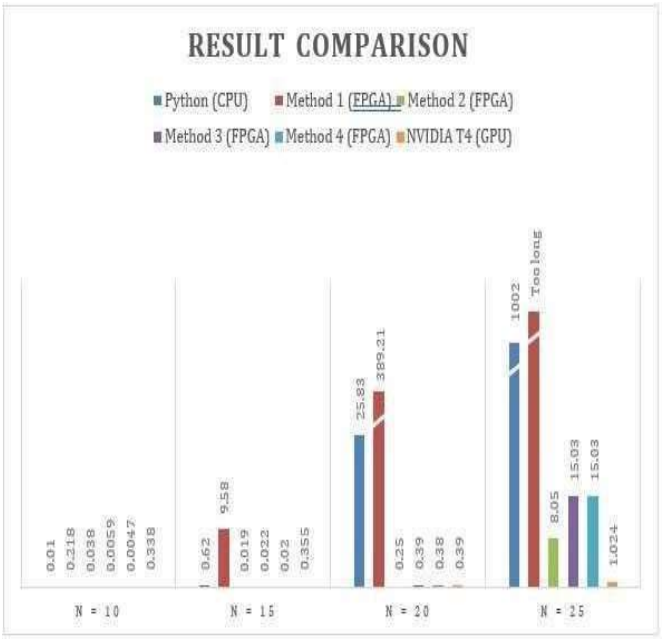



**Figure 8: Graphical illustration of Table 3**

From Figure 8, it can be seen that FPGA outperforms GPU for all cases except n=25. We think this behavior was observed due to the following factors:

• Architecture and Design o FPGAs and GPUs have fundamentally different architectures. FPGAs are highly customizable hardware that can be tailored to specific tasks, but their design and optimization can be complex. GPUs, on the other hand, are designed for parallel processing and may have better-suited architectures for certain types of algorithms. It is possible that the algorithm we used for this NP- hard problem is better suitable for GPU than FPGA, specifically the FPGA we used. • Optimization and Compiler Efficiency o GPUs have more mature and well-optimized compilers, and their ecosystems, such as CUDA, have extensive community support. This contributes to better compiler optimization and overall efficiency, potentially resulting in improved performance. Whereas in FPGA, we used pragmas manually for optimization. Thus, it is possible that the built-in

optimizer of T4 GPU compiler provided by Google outperformed our manual optimization.

## 7 Conclusion

In this paper, a comprehensive comparative analysis was conducted on various techniques aimed at accelerating the execution of the Partition Problem, a well- known NP-complete problem. The potential of parallel processing architectures, specifically GPUs and FPGAs, was explored to enhance computational efficiency. The effectiveness of GPU and FPGA implementations in accelerating the execution of the Partition Problem was demonstrated through experimental evaluation. Significant reductions in execution time were observed, particularly for larger problem instances, such as n=25, where the GPU implementation on the NVIDIA T4 GPU outperformed FPGA implementations and traditional CPU-based approaches. Insights into design considerations and optimization strategies pertinent to both GPU and FPGA implementations were provided. While GPUs offer mature compilers and extensive community support, FPGAs boast highly customizable hardware tailored to specific tasks. Thus, the choice of which technology to use depends on various aspects of the problem. Overall, this study contributes to advancing the understanding of efficient computation techniques for NP-hard problems. It serves as a valuable resource for researchers and practitioners interested in leveraging parallel processing architectures for computational acceleration. A GitHub repository has been set up to assist interested audiences with all the resources necessary to use the software: https://github.com/pratikstha/PartitionProble mUsingFPGA article graphics verbatim.

## References


1. M. Gokhale and P. Graham, ”Reconfigurable computing: Accelerating computation with field programmable gate arrays.” 2006. [Online]. Available: Springer Science and Business Media.
2. ”XUP PYNQ,” [Online]. Available: https://www.xilinx.com/support/university/boardsportfolio/xupboards/XUPPYNQ.html.”Partition problem” Wikipedia [Online]. Available: https://en.wikipedia.org/wiki/Partition
3. G. Woeginger, ”Exact algorithms for np-hard problems: A survey,” in Combinatorial optimization-eureka, you shrink!, Springer, 2003, pp. 185-207.
4. C. Trefftz, J. Scripps and Z. Kurmas, ”An introduction to elements of parallel programming with java streams and/or thrust in a data structures and algoriths course,” Journal of Computing Sciences in Colleges, 2017, 33(1):11-23.
5. ”Introduction to Overlays - Pyhton Productivity For Zynq (Pynq) V1.0,” Pynq.readthedocs.io 2020 [Online]. Available:https://pynq.readthedocs.io/en/v1.4/6 overlays.html.
6. ”NVIDIA T4,” [Online].Available: https://www.nvidia.com/en-us/data-center/tesla-t4.”NVIDIA
7. Turing GPU Architecture,” [Online Available: https://images.nvidia.com/aem-dam/en- zz/Solutions/design-visualization/technologies Whitepaper.pdf
” What is a Graphics Processing Unit (GPU)? Definition and Examples,” [Online]. Available: https://www.investopedia.com/terms/g/ graphics-processing-unit-gpu.asp.


## Biography / Biographies

Dr. Chirag Parikh earned his Master’s and Doctoral degreesfrom the University of Texas at SanAntonio in 2003 and 2007 respectively. Currently, he is a Professor and Chair of Computer Engineering program at GrandValley State University. His research interests are embedded system design, cryptography, and FPGA-based system design.

Dr. Christian Trefftz earned a master’s degree in Computer Science from Western Michigan University in 1989 and a Ph.D. Degree in Computer Science from Michigan State University in 1994. He is a Professor at the College of Computing at Grand Valley State University. His research interest is in the area of parallel processing.

Parthik Shrestha holds a Master’s degree in Computer Science (2020) and a Master’s degree in Electrical and Computer Engineering (2019) from Grand Valley State University Currently, he is a software Engineer at CTDI. His areas of expertise and interest include machine learning systems and software development.

# Enhancing Cybersecurity by relying on a Botnet Attack Tracking Model using Harris Hawks Optimization

Ali Ibrahim Ahmed *
Al-Noor University,Mosul, Iraq.
AbdulSattar M. Khidhir †
Northern Technical University, Mosul, Iraq.
Shatha A. Baker ‡
Northern Technical University, Mosul, Iraq.
Omar I. Alsaif §
Northern Technical University, Mosul, Iraq.
Ibrahim Ahmed Saleh ¶
Mosul University, Mosul, Iraq

**Abstract**

A botnet attack is a major cybersecurity threat that involves coordinated control of a network of infected computers, enabling large-scale distributed denial of service (DDoS) attacks, malware spreading, and other cybercrime activities. Proactive security measures and advanced threat intelligence systems are essential to detect and mitigate these assaults. This paper proposes the Harris Hawks Optimization (HHO) algorithm, which employs exploration and exploitation techniques to find optimal solutions for analyzing botnet attack pathways. The proposed approach involves HHO as a feature selector for extracting features from anomalous network traffic. The algorithm's impact on botnet IP positioning performance is analyzed, considering different optimization modes and control center accuracy. The paper is organized into sections covering attack path establishment and analysis, system testing and verification, and a central leadership entity controls it [1]. Botnets are created based on the use of malicious software packages to infect important and sensitive devices in the network, thus making servers, computers, and Internet of Things devices vulnerable [2]. To detect these attacks and limit their impact requires many proactive security measures such as strong network security settings, regular software upgrades, etc. [3]. HHO is a powerful method that has the potential to solve many functional optimization problems and provides a suitable environment for engineering applications, as it mimics the exploration and exploitation phases during the foraging process of Harris Hawks [4]. A model based on HHO algorithm. is proposed in this paper that has the ability to track and analyze bot attack paths by extracting a set of features during abnormal network traffic. The results were analyzed and their impact on the performance of robot networks was discussed, based on the use of different

experimental results. After configuring the network topology and determining the attack path based on HHO, the performance of the algorithm and its effectiveness in preventing IP addresses from being spoofed are verified. The results showed convergence in being able to correct attack paths and effective performance in repelling the interference of fake IP addresses.

**Keywords:** Botnet attack, Harris Hawks Optimization, zombie virus

## 1 Introduction

Botnet attacks pose a major threat to cybersecurity because they have the potential to allow criminals to launch large-scale DDoS attacks, launch spam campaigns, and engage in various forms of cybercrime. It is coordinated based on a network of infected devices called "zombies" or "bots", and

optimization modes, determining the IP position, and the extent of the influence of the control center's accuracy. The rest of the paper is organized as follows: Section 2 provides an overview of the HHO algorithm's concept and formulation, while Section 3 provides a discussion of the proposed materials and methods. The experiments and results analysis are presented in Section 4. Section 5 finally includes the conclusion.

## 2 Harris Hawks Optimization

This is an algorithm proposed in 2019 by Heidari et al and is a descriptive heuristic [5]. It draws inspiration from the

*Reasearch Assistant
†Department of Electronic Technologies Northern Technical University.
‡Department of Electronic Technologies Northern Technical University.
§Department of Electronic Technologies Northern Technical University.
¶3Department of Software, College of Computer and Mathematics.

predation behavior of Harris hawks, specifically their hunting technique of capturing prey, such as hares. In a variety of optimization tasks and problem domains, HHO outperforms other well-known methods [6].In this optimization algorithm, the preysymbolizes the search for the optimal solution, and the candidate solutions are represented by the Harris hawks. Figure 1 illustrates the two primary stages of the algorithm, the exploration phase, the algorithm focuses on discovering new potential solutions by exploring unknown regions of the search space. While, the exploitation phase aims to refine and improve the solutions that have shown promising results during the exploration phase. In order to arrive at an optimal solution, convergence requires striking an equilibrium between exploration and exploitation. Through this iterative process, HHO gradually guides the search towards the global minimum by emulating the predatory behavior of Harris hawks [7].

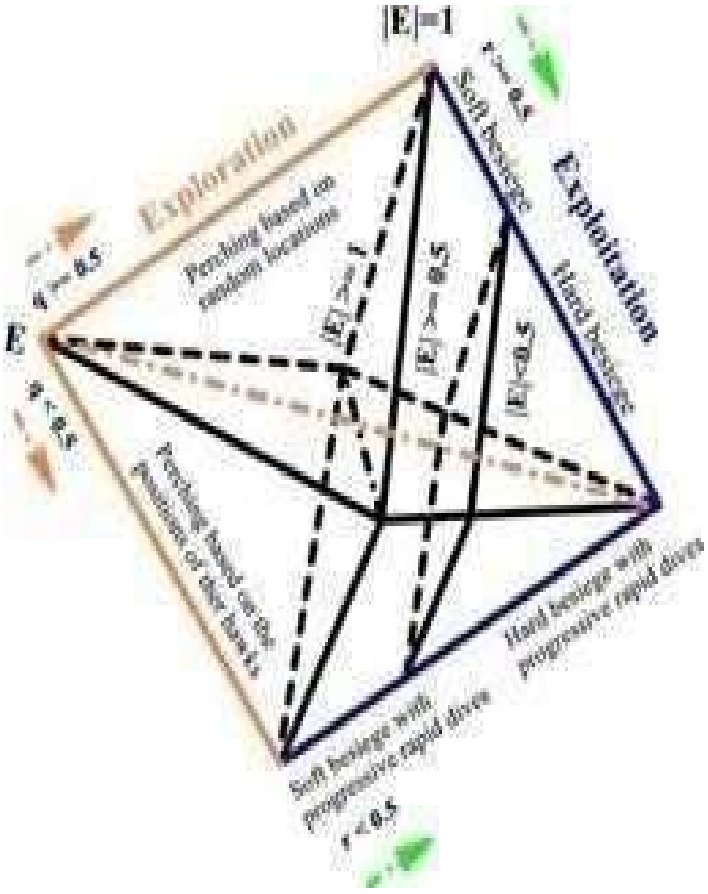



Figure 1: HHO Phases

### 2.1 Exploration Phase

Every Harris hawk in the population is considered a potential fix. During each iteration, every possible solution's fitness value is assessed in relation to the intended prey. In this case, exploitation refers to local research conducted within the area identified through exploration stage. Harris hawks initially wait, observing and evaluating the search space know by the upper bound (ub) and lower bound (lb) of the problem domain. They then engage in random searches for prey using two different strategies. The position update during the iteration is influenced by a probability parameter (q), determining the likelihood of a particular movement. The mathematical expression for this update process is utilized to guide the hawks in their exploration and exploitation efforts [8].

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 X(t)| & q \geq 0.5 \\ (X_{rabbit}(t) - X_m(t)) - r_3 (LB + r_4 (UB - LB)) & q < 0.5 \end{cases} \quad (1)$$

LB and UB upper and lower bounds of variables; Xm represents the average position of the current hawk population; Xrand(t) represents a randomly selected hawk from the population; Xrabbit (t) rabbit position; X(t) current position of hawks; and r1, r2, r3, r4, and q are random numbers between (0,1), which are updated in each iteration. Eq. 4 is used to get the hawks' average position:

$$X_m(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t) \quad (2)$$

### 2.2 Exploration to Exploitation Transition

In swarm optimization algorithms like the HHO, maintaining a balance between exploration and exploitation is crucial for effective problem-solving. The Harris hawk is renowned for its flexible hunting tactics, since it can transition between various forms of predation based on the energy levels of its victims. If the prey is in a state of escape, its energy, which is symbolized by the symbol E, will gradually diminish. The HHO included the concept Escape energy to make the transition easy between the exploitation and exploration stages. The escape energy equation was relied upon in the algorithm to regulate this conversion process [9].

$$E = 2E_0(1 - t) \quad (3)$$

The symbol E stands for the prey's escape energy, the symbol T for the maximum number of repeats, and the symbol E0 for the prey's initial energy condition [10].

### 2.3 Exploitation Phase

When Harris' hawks execute a surprise pounce, they target prey that was identified during the preceding phase. Because prey frequently flee from hazardous circumstances, numerous pursuit techniques have developed. In the HHO, four potential tactics are put forth to simulate the offensive stage [11-12].

• Soft Besiege Because of the energy it has, the prey in this scenario can flee when $|e| \geq 0.5$ and $r \geq 0.5$. The Harris hawk then relies on a soft siege strategy for the purpose of gradually depleting the energy of the prey. The primary objective of this strategy is to select the optimal position from

which to launch raids and dives, effectively capturing the prey. The following equation controls the location update during the soft siege strategy [13]:

$$X(t+1) = \Delta X(t) - E\,|JX_{rabbit}(t) - X(t)| \quad (4)$$

$$\Delta X(t) = X_{rabbit}(\mathrm{t}) - X(t) \quad (5)$$

$$X(t+1) = \Delta X(t) - E\,|JX_{rabbit}(t) - X(t)| \quad (4)$$

$$\Delta X(t) = X_{rabbit}(t) - X(t) \quad (5)$$

where $\Delta X\,(t)$ is the variation between the rabbit's position vector and its current location in the iteration, and $J = 2(1 - r_5)$ indicates the rabbit's random leap strength throughout the escape phase. During each repetition, the $J$ value fluctuates at random to mimic the characteristics of rabbit motions [**?**].

- **Hard Besiege**

The prey's energy is greatly reduced when —E— ¡ 0.5 and r ¿ 0.5, at which point the Harris hawks execute a surprise pounce attack. At this stage, the Harris hawks no longer engage in extensive encircling maneuvers but instead make a sudden represented by Z in this equation, while the Levy function is represented by LF(d). A random vector of size 1 * D is called S. Y denotes the position ascertained by the gentle siege approach. The following formula is used to calculate the Levy function:

$$LF(x) = 0.01 \times \frac{u \times \sigma}{|v|^{1/\beta}} \quad (8)$$

Here, $\sigma$ is a calculated value, $\beta$ is a constant with a value of 1.5, and $v$, $u$ are random numbers that range from zero to one. The HHO method simulates the prey's escape behavior by adding a stochastic element to the location updates through the use of the Levy function. This enables the Harris Hawks to modify their positions in response. This dynamic location update approach improves convergence towards optimal solutions and the algorithm's ability to catch elusive prey [**?**].

- **Hard besiege with progressive rapid dives**

The prey still has a chance to escape when —E— ¡ 0.5 and r ¡ 0.5, but its escape energy E is insufficient. The Harris hawk uses a hard besiege tactic in this instance, which is typified by increasingly quick dives. This approach involves initiating a hard besiege prior to launching an attack, gradually decreasing the distance between

and decisive attack [15]: The position is updated using the following equation

$$X(t+1) = X_{rabbit}(t) - E\,|\Delta X(t)| \quad (6)$$

- **Soft besiege with progressive rapid dives**

The HHO method models prey escape patterns and leapfrog movements statistically by utilizing the levy flight (LF) idea. LF imitates the irregular, sudden, and fast dives of hawks around the fleeing prey, as well as the true zigzag misleading maneuvers of rabbits during the escaping phase. The hawks circle the rabbit quickly as a team, making multiple attempts to adjust their position and trajectory. Real data from various competitive scenarios in nature provide evidence for this mechanism [17]. Hawks use the following rule to choose their next step when executing a gentle besiege:

$$Z = Y + S + LF(d) \quad (7)$$

The Harris Hawks' updated position vector is the hawk and the prey[19]. The position update equation governing this phase is as follows:

$$X(t+1) = \begin{cases} Y : X_{rabbit,t} - E\ \ JX_{rabbit,t} - X_{m,t} & \text{if } F(Y) < F(X(t)) \\ Z : Y + S \cdot LF(D) & \text{if } F(Z) < F(X(t)) \end{cases} \quad (1)$$

The Harris hawk executes a hard besiege with progressive rapid dives, continuously adjusting its position towards the prey to increase the chances of capturing it successfully. By employing different attack mechanisms based on the prey's escape energy and the factor r, the HHO algorithm effectively solves optimization problems [20-23].

## 3 Materials and Methods

### 3.1 Network Topology

Network topology refers to the arrangement and connectivity of nodes in a network. Traditionally, network topology has been created using either completely random or completely regular methods of layout and connection paths. However, in order to simulate real network environments, this research utilizes a random graph generator. The topology design involves placing $v$ nodes within a square of size $M \times M$ and then randomly connecting them with a certain probability to establish the network topology. The probability of connecting two nodes, $i$ and $j$, is determined using a formula that takes into account their Euclidean distance and a maximum possible distance between nodes. By adjusting the control variables $\eta$ and $\gamma$ within specific intervals, the average number of nodes connected and the average distance between node connections can be influenced. Unlike previous network topologies used for IP traceability, where the attacker and victim were positioned on the periphery, this study randomly selects the locations of the two ends to closely resemble real network scenarios. In the context of this paper, the establishment of the attack path analysis model involves generating a network topology between the attacking node and the victim node by randomly placing nodes within a square region and connecting them based on the probability equation (10).

$$P_{ij} = \frac{1}{1 + e^{-(d_{ij} - \eta)/\gamma}} \quad (2)$$

The selection of the attacking and victim nodes within the topology is also randomized to reflect real-world scenarios.

### 3.2 Reconstruction and statistical characteristics of attack path

The reconstruction of the attack path involves generating 30 sets of random topologies with different numbers of nodes. Monte Carlo Simulation is employed to simulate hackers performing one-to-one attacks on the victim, generating attack paths. The path, nodes, and number of packets are recorded as judgment criteria for backtracking the attack path. The reconstruction component of the attack path utilizes the HHO algorithm. Equation (11) serves as the state transition rule for path exploration, while formulas (6) and (7) are used to update the path. To effectively guide the Harris search along the correct attack path, certain research parameters are set.

$$p_{ij} = \sum_{k \in \text{neighbor}(i\ j)} \frac{[r_{i\,j}(t)]^{\alpha}\,[\eta_{i\,j}(t)]^{\beta}}{ij \cdot \eta\,(t)}\,[r \cdot \eta]^{\beta}_{ij} \tag{11}$$

### 3.3 Detection of fake IP

The detection of fake IP addresses involves identifying their distinct behavior compared to normal IP addresses. To modify the HHO algorithm to prevent counterfeiting IP, the following steps are taken:

• **Non-existent fake IP:** Internet Control Message Protocol (ICMP) command tracers are used to detect the attacker and victim by sending requests to all nodes along the path. Alternatively, network security and intrusion detection policies are implemented using the Challenge Handshake Protocol (CHAP) authentication process can be employed. If there is no response or the authentication fails, it is determined to be a fake IP [24]. • **Existence of fake IP:** The HHO algorithm is used for path backtracking, and the ”Amount of attack information” at each node is considered for determining the presence of counterfeit IP. The detection of counterfeit IP can be categorized into two main categories: a. Inconsistent path: If there is no direct link between a node in the path and the destination (predicted attacker), or if the connection path leads to an unreachable node, it can be avoided depended on HHO characteristics. b. Abnormal amount of node attack information: This is primarily determined based on a sharp drop in the number of attacks. While the attack volume is set during the attack, in a real network environment, it is challenging to determine an upper-bound threshold for attack volume. Therefore, a fixed threshold should not be used to judge abnormal attack volume. Instead, the ratio of node attack amounts is examined. If the attack amount at a particular Harries-passing node is lower than that of the previous path, a penalty function is applied by multiplying the amount of node attack by a threshold and adding it to the cost of the path node. For example, if the threshold is set to 0.4 and the previous path node has an attack volume of 1000, but the selected path node has an attack volume of 350, it is considered lower than $1000\times0.4 = 400$. Therefore, if the algorithm determines that this point is an attacker, it is classified as a fake IP or an incorrect attacker. Otherwise, a penalty function is added to the exploration at this point.

$$r_{ij}(t+1) = (1-\rho)\cdot r_{ij}(t) + \mu_{ij} \tag{12}$$

$$\mu_{ij} = \begin{cases} \rho\Delta r & \text{if } \Delta r > 0 \\ 0 & \text{if } \Delta r \le 0 \\ \text{if AttPackets}_i - \text{AttPackets}_j & \text{otherwise} \\ -\rho\Delta r & \text{otherwise} \end{cases}$$

Where $\mu_{i\,j}$ is the penalty function and $\text{AttPackets}_i$ is the number of attack packets transmitted by node $i$. Its value depends on the number of attack packets collected. This adjustment in the algorithm helps in identifying nodes where the attack activity significantly decreases, indicating a wrong path or a fake node. The control threshold can be determined using a cubic spline function, where the cubic-spline interpolation formula predicts a reasonable value for $\text{AttPackets}_m$, and $\pm2$ standard deviations represent a 95% confidence interval. The upper and lower confidence limits $[\lambda, \omega]$ can be calculated using Equation (13):

$$[\lambda, \omega] = \text{AttPackets}_m \mp 2\sigma_{\text{AttPackets}}$$

If the amount of attack information falls below the lower limit $\omega$, it can be determined that the attacker is a fake IP.

## 4 Testing and Verification of Proposed Approach

The simulation network topology was generated using a random graph, where $\eta$ and $\gamma$ are weight variables representing two types of control topologies. The value of $\gamma$ plays a crucial role in controlling the average distance of the path. Increasing $\gamma$ results in stronger connections between nodes, thereby reducing the average distance. The concept of average distance is based on network models. To ensure consistency with real network environments and avoid excessively large path distances, a parameter setting similar to other studies [10] fixes $\gamma = 0.1$. Conversely, the average number of connections for every node is determined by the value of $\eta$. Higher values of $\eta$ increase the average connections, while too low values lead to insufficient paths, potentially causing excessive bottlenecks and reducing the number of feasible paths. Tables 1 to 3 provide insights into the influence of these parameter settings on the network topology. Figure 2 showcases a series of topology diagrams generated by setting the number of nodes to 100 and configuring $\gamma = 0.1$ and $\eta = 1.5$. In this figure, point A represents the randomly generated victim end, while point B represents the attack end.

The attack path reconstruction part is mainly built by the HHO algorithm. This algorithm performs a backtracking of the attacker based on the generated topology and monitors the

**Table 1:** Topology generates average data (number of nodes=50)

| Topology setting | Max. No. of connections | Min. No. of connections | Average No. of connections | Maximum Distance(m) | Average Distance |
|---|---|---|---|---|---|
| η=0.5 | 6.30 | 0.00 | 1.2 | 46.64 | 6.12 |
| η=1.0 | 5.85 | 0.05 | 2.39 | 59.77 | 8.61 |
| η=1.5 | 7.75 | 0.15 | 3,47 | 59.94 | 8.71 |
| η=2.0 | 10.15 | 0.55 | 4.47 | 59.11 | 8.72 |
| η=2.5 | 12.05 | 0.70 | 5.52 | 61.98 | 10.01 |

**Table 2:** Topology generates average data (number of nodes = 100)

| Topology setting | Max. No. of connections | Min. No. of connections | Average No. of connections | Maximum Distance(m) | Average Distance |
|---|---|---|---|---|---|
| η=0.5 | 6.65 | 0.00 | 2.37 | 61.88 | 8.77 |
| η=1.0 | 11.35 | 0.35 | 5.10 | 64.17 | 8.60 |
| η=1.5 | 13.85 | 1.30 | 6.95 | 68.39 | 8.61 |
| η=2.0 | 18.05 | 1.75 | 9.15 | 70.11 | 8.75 |
| η=2.5 | 21.05 | 2.70 | 11.05 | 76.10 | 9.43 |

**Table 3**: Topology generates average data (number of nodes=200)

| Topology setting | Max. No. of connections | Min. No. of connections | Average No. of connections | Maximum Distance(m) | Average Distance |
|---|---|---|---|---|---|
| η=0.5 | 8.35 | 0.05 | 3.44 | 67.12 | 8.77 |
| η=1.0 | 14.95 | 0.85 | 7.10 | 75.24 | 8.63 |
| η=1.5 | 19.95 | 2.25 | 10.52 | 76.01 | 8.71 |
| η=2.0 | 24.50 | 3.00 | 13.82 | 76.88 | 8.81 |
| η=2.5 | 28.65 | 4.70 | 16.44 | 80.85 | 9.49 |

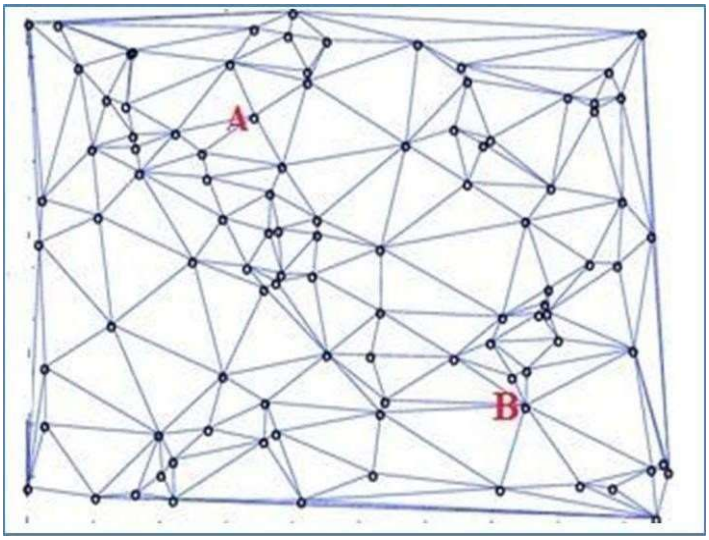



Figure 2: Simulation topology (100 nodes)

parameters. $\alpha$ and changes on the attack path convergence. The Harris size of 30, a decay rate of 0.5, $\alpha$ set to 0.7, $\beta$ set to 1.3, and performs 27 generations of execution. The HHO algorithm uses the number of transmissions of the assault packet as its search criterion. Additionally, the inference criterion for finding the victim is utilized to judge the selection of the attack path regarding the statistical characteristics of the attack.

The judgment of counterfeit IP can be classified into two categories: path inconsistency and abnormal node attack information. Among them, if there is no direct path between node in path and predicted attacker or node is unreachable, it can be calculated by HHO calculation. The Hurries of the law cannot reach and avoid being Fake IP spoofing; and the abnormal amount of node attack information is part of this study. It is verified by equation (12) and equation (13), and the threshold is set as 0.5. For attack detection part of counterfeit IP, this research is based on. There are 20 sets oftopologies with different numbers of nodes to execute 5 times each, select Set any node in the topology on the non-attack path as the node of the fake IP Point to test whether HHO algorithm will be counterfeited when searching for a path, the attacking end interferes, and correct attacking end cannot be found. Figures 3 illustrate performance of HHO algorithm modes simulation with node size of 100 points that which used in analysis with execution is perform for 30 times, as seen the criterion of execution performance is based on the probability of the average number of attack packets on the path searched by the algorithm. The performance is when increase number of executed that increase probabilities fake IP. The different topology sizes are used for analysis, Figure 3 and Figure 4 illustrate execution algebra and average of topological size 100 and 200 nodes respectively. Searching the relationship graph of the error rate, it can be observed that as the execution algebra increases .Therefore, it can be observed from Figure 4 and Figure 5 that regardless of the topology, it can converge to a search error rate, and the topology is smaller Because of feasible solution is less, which will increase the probability of resetting. The test of counterfeit IP through this step shows that at the algorithm initial stage of the execution, the algorithm not be able to find correct attacker due to influence of the counterfeit IP. However, as the execution algebra increases, the algorithm will still refer to the correct attack path. Attack the number of packets, and pull the search path back to the correct attack path. It is evident that the HHO algorithm with adjustment of formulas (11) and (12) is effective in preventing counterfeit IP It has a very high implementation effect.

## 5 Conclusion

This paper analyzes the botnet attack path traceability model based on the HHO. Aiming at the shortcomings of the HHO algorithm that is not easy to escape after converging to the best solution in the area. To analyze the attack path of the botnet, and explore the computing resources required by the botnet control center in reverse traceability, a completely

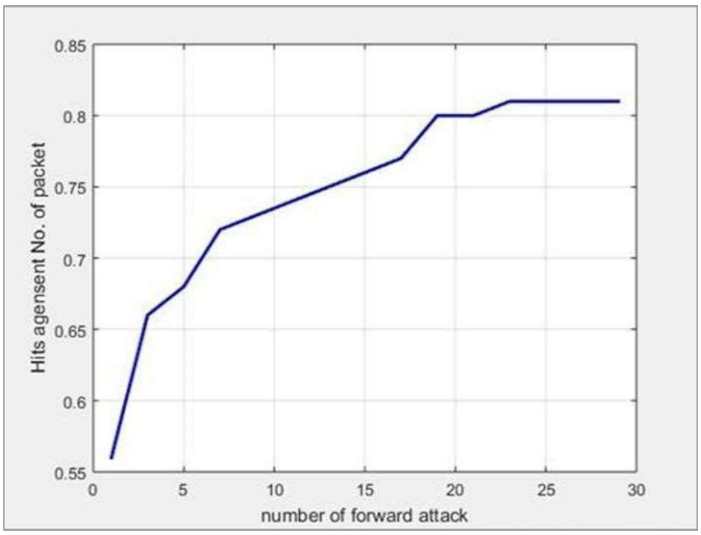



Figure 3: HHO probabilities for explored IP fake

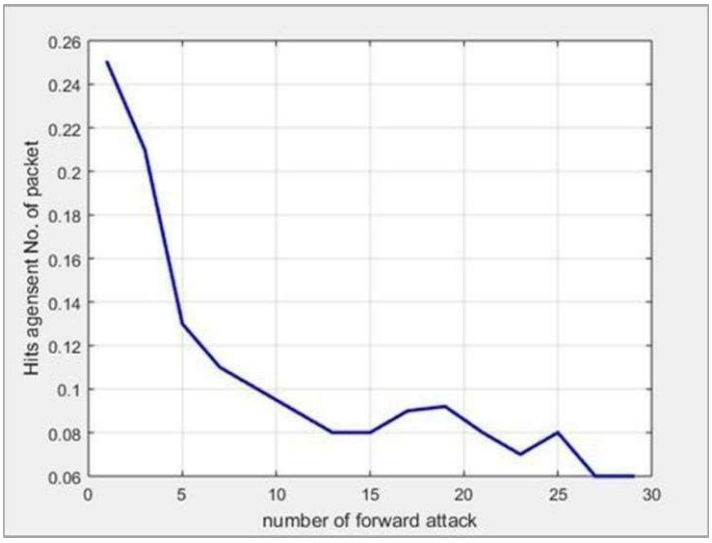



Figure 4: Algorithm search error rate of fake IP (100 nodes)

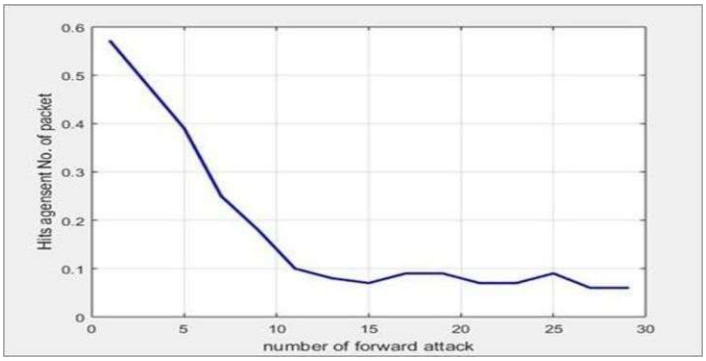



Figure 5: Algorithm search error rate of fake IP (200 nodes)

random communication method is used. The proposed approach has proven effective in preventing fake IP interference in the exploration and exploitation phases, and it has the ability to adapt and respond to the dynamic nature of bot attacks, strengthening cybersecurity measures against advanced cyber threats, as proven by experimental results. The paper also showed the importance of making modifications to the algorithm based on a set of information about the attack package to be able to identify nodes with significantly low activity, which results in improving the accuracy of the attack path.

## References

[1] Sokkalingam, Sumathi, and Rajesh Ramakrishnan. ”An intelligent intrusion detection system for distributed denial of service attacks: A support vector machine with hybrid optimization algorithm based approach.” Concurrency and Computation: Practice and Experience 34.27 (2022): e7334.

[2] Baker, S. A., Nori, A. S. “Internet of things security: a survey”. Advances in Cyber Security: Second International Conference, ACeS 2020, Penang, Malaysia, December 8- 9, 2020, Revised Selected Papers 2. Springer Singapore, 2021.

[3] Baker, S. A., Nori, A. S. “A secure proof of work to enhance scalability and transaction speed in blockchain technology for IoT”. In AIP Conference Proceedings (Vol. 2830, No. 1). AIP Publishing, 2023.

[4] Alabool HM, Alarabiat D, Abualigah L, Heidari AA. Harris hawks optimization: a comprehensive review of recent variants and applications.Neural Comput Appl. 2021;33(15):8939-8980.

[5] Heidari, Ali Asghar, et al. ”Harris hawks’ optimization: Algorithm and applications.” Future generation computer systems 97 (2019): 849-872.

[6] H. Moayedi, A. Osouli, H. Nguyen and A. Rashid, “A novel Harris hawks’ optimization and k-fold cross-validation predicting slope stability”, Engineering with Computers, 2019.

[7] H. Moayedi, M. Abdullahi, H. Nguyen and A. Rashid, “Comparison of dragonfly algorithm and Harris hawk’s optimization evolutionary data mining techniques for the assessment of bearing capacity of footings over two-layer foundation soils”, Engineering with Computers, 2019.

[8] Alabool, Hamzeh Mohammad, et al. ”Harris hawks’ optimization: a comprehensive review

of recent variants and applications.” Neural Computing and Applications 33 (2021): 8939-8980.

[9] Shehab, Mohammad, et al. ”Harris hawks optimization algorithm: variants and applications.” Archives of Computational Methods in Engineering 29.7 (2022): 5579-5603.

[10] Li, ChenYang, et al. ”Enhanced Harris hawks optimization with multi-strategy for global optimization tasks.” Expert Systems with Applications 185 (2021): 115499.

[11] Al-Betar, Mohammed Azmi, et al. ”A hybrid Harris Hawks optimizer for economic load dispatch problems.” Alexandria Engineering Journal 64 (2023): 365-389.

[12] Dhawale, Dinesh, Vikram Kumar Kamboj, and Priyanka Anand. ”An improved Chaotic Harris Hawks Optimizer for solving numerical and engineering optimization problems.” Engineering with Computers 39.2 (2023): 1183-1228.

[13] Fan, Qian, Zhenjian Chen, and Zhanghua Xia. ”A novel quasi-reflected Harris hawks optimization algorithm for global optimization problems.” Soft Computing 24 (2020): 14825-14843.

[14] C¸ etınbaS¸, ˙Ipek, Bu¨nyamın Tamyu¨rek, and Mehmet Demırtas¸. ”The hybrid Harris hawks optimizer-arithmetic optimization algorithm: A new hybrid algorithm for sizing optimization and design of microgrids.” IEEE Access 10 (2022): 19254-19283.

[15] Li, Wenyu, Ronghua Shi, and Jian Dong. ”Harris hawks optimizer based on the novice protection tournament for numerical and engineering optimization problems.” Applied Intelligence 53.6 (2023): 6133-6158.

[16] Yu¨zgec¸, Ugur, and Meryem Kusoglu. ”Multi-objective harris hawks optimizer for multiobjective optimization problems.” BSEU Journal of Engineering Research and Technology 1.1 (2020): 31-41.

[17] Fan, Qian, Zhenjian Chen, and Zhanghua Xia. ”A novel quasi-reflected Harris hawks optimization algorithm for global optimization problems.” Soft Computing 24 (2020): 14825-14843.

[18] Gupta, Shubham, et al. ”Opposition-based learning Harris hawks optimization with advanced transition rules: Principles and analysis.” Expert Systems with Applications 158 (2020): 113510.

[19] Zhang, Yang, Xizhao Zhou, and Po-Chou Shih. ”Modified Harris Hawks optimization algorithm for global optimization problems.”Arabian Journal for Science and Engineering 45 (2020): 10949-10974.

[20] Maray, A. H., Alsaif, O. I., & Tanoon, K. H. (2022). Design and implementation of low- cost medical auditory system of distortion otoacoustic using microcontroller. J. Eng. Sci. Technol, 17(2), 1068-1077.

[21] Mohammed, N. L., Aziz, M. S., & AlSaif, O. I. (2020). Design and implementation of robot control system for multistory buildings. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(5), 2682-2689.

[22] Thanoon, K. H., Q Hasan, S.,& I Alsaif, O. (2020). Biometric information based on distribution of arabic letters according to their outlet. International Journal of Computing and Digital Systems, 9(5), 981-991.

[23] Yuan, Yongliang, et al. ”An adaptive instinctive reaction strategy based on Harris hawks optimization algorithm for numerical optimization problems.” AIP Advances 11.2 (2021).

[24] Baker, S. A., Mohammed, H. H., & Alsaif, O. I. (2024). Docker Container Security Analysis Based on Virtualization Technologies. International Journal for Computers & Their Applications, 31(1).

## BIOGRAPHY / BIOGRAPHIES

**Ali Ibrahim Ahmad** earned his bachelor’s and master’s degrees from the University of Mosul, Faculty of Computer Science and Mathematics, Department of Software Engineering in 2019 and 2022 respectively. Currently, he works as a Lecturer at the Al-Noor University College, Mosul, Iraq. His research areas encompass artificial intelligence applications, security, and image processing. Email: ali.ibrahim@alnoor.edu.iq

**AbdulSattar M. Khidhir** (Born 9th Jan. 1959) is an assistant professor at Electronics Technology Department - Mosul Technical Institute - Northern Technical University in Iraq. He obtained his B.Sc. (1981) and M.Sc. (1989) both in Electronics and Communications Engineering from University of Mosul. His Ph.D. (2000) was obtained in Communications Engineering from University of Mosul too. He supervised many Ph.D. and M.Sc. theses in different scientific and engineering areas. He was a member of scientific committees for many Ph.D. and M.Sc. students. He published many researches in various fields of science and engineering (see google scholar). He reviewed many scientific papers for journals and conferences. Email : abdulsattarmk@ ntu.edu.iq, abdulsattarmk@gmail.com

**Shatha A. Bakr** has a Bachelor’s degree in Computer Science from the University of Mosul, which she obtained in 1997. In 2013, she earned a master’s degree in Computer Science from the same university. Later, in 2022, she completed her Ph.D. from the University of Mosul. Dr. Bakr worked as a Lecturer at the Northern Technical University in Mosul, Iraq. Her research interests encompass mobile phone programming, information security, multimedia communications, and artificial intelligence.

**Omar I. Alsaif** is currently a lecturer in the Mosul Technical Institute/ Northern Technical University in Mosul, Iraq. He received his B.Sc. in electrical engineering from the University of Mosul in 1992. In 2005 and 2018, he obtained his M.Sc. and Ph.D. degrees in Electronics and Microelectronic Engineering from Mosul University, respectively. His research interests encompass microelectronic and solid-state systems, renewable energy, and nanotechnology devices. Email: omar.alsaif@ntu.edu.iq.

**Ibrahim Ahmed Saleh** was born in Mosul - Iraq in 1963. He received his MSc. degree (in signal and image processing) from the University of Mosul, Iraq in 2003 and in 2013 he received his PhD inartificial techniques and computer networking from Mosul University. He became professor in 2021. From 1997 to 2005, he worked at computer center in Mosul University/Iraq. Currently he is lecturer at the Dept. of Software Engineering, College of Computer Sciences and Math, and University of Mosul, Iraq. He can be contacted at email: i.hadedi@uomosul.edu.iq.

# Improving communication security Against Quantum Algorithms Impact

Hicham Amellal *
LabSIV, Department of Computer Science
Faculty of Sciences Agadir, Ibnou Zohr University
Agadir, Morocco.

**Abstract**

In this paper, we explore the impact of quantum algorithms on classical network security. Our analysis focuses on Shor's algorithm, which excels in factorizing large prime numbers, presenting a significant threat to classical cryptographic protocols. We conduct an in-depth analysis of Shor's algorithm's potential effects on HTTPS to highlight its disruptive capabilities. Moreover, to fortify classical cryptographic protocols against quantum threats, we introduce a novel Quantum Intrusion Prevention System (QIPS) scheme. Leveraging basic components like beam splitters and detectors, this solution serves as a dedicated hardware interface between the classical network and external quantum networks. Our proposed QIPS scheme offers enhanced resilience to classical cryptographic protocols, mitigating the vulnerabilities posed by quantum algorithms and reinforcing network security in the face of evolving threats.

**Key Words**:Quantum algorithms, Network security, Quantum IPS, HTTPS, RSA, Shor's algorithm.

## 1 Introduction

Quantum information, which relies on certain phenomena of quantum mechanics, is considered one of the most powerful solutions proposed for information processing in recent years, at least theoretically. Unlike classical computing, which utilizes bits to represent information as either 0 or 1, quantum computing operates on quantum bits or qubits, which have the ability to exist in superposition states. This means that a qubit can represent both 0 and 1 simultaneously, enabling quantum computers to perform certain computations exponentially faster than classical computers [1, 2]. Quantum computing has the potential to revolutionize a wide range of industries, including finance, logistics, drug discovery, and materials science. This is due to the different algorithms that can reduce the time required to solve complex mathematical problems.

One of the most important fields in information security is classical cryptography, which is a type of encryption that is based on the complexity of mathematical calculations. In classical cryptography, plaintext is transformed into ciphertext using a cryptographic algorithm and a secret key. The goal of encryption is to make it difficult for unauthorized parties to read the plaintext without the secret key. Classical cryptography algorithms include symmetric key algorithms, such as the Data Encryption Standard (DES) and Advanced Encryption Standard (AES), and asymmetric key algorithms, such as the Rivest-Shamir-Adleman (RSA) algorithm. These algorithms rely on the computational complexity of mathematical problems, such as factoring large numbers or solving the discrete logarithm problem, to provide security.

However, with advances in computing power and new cryptographic attacks, many classical cryptography algorithms are no longer considered secure. The emergence of this threads has spurred the creation of innovative cryptographic algorithms and protocols that are designed to be resistant to attacks by quantum attacks strategies, which are expected to be able to break many classical cryptographic algorithms.

The paper is structured as follows: Section 2 provides an introduction to Shor's algorithm. Section 3 discus some classical protocols based on cryptography. Section 4 introduces the proposal quantum intrusion prevention system schema (QIPS). Section 5 analyzes the performance of the proposed "QIPS", followed by the conclusion in the final section of the paper.

## 2 Exploring Shor's Algorithm

The algorithm was developed by mathematician Peter Shor in 1994. He demonstrated that a quantum computer could efficiently factor large integers exponentially faster than any known classical algorithm [3]. The algorithm works by exploiting the properties of quantum mechanics, such as superposition and entanglement, to perform the factorization of an integer into its prime factors. Specifically, the algorithm utilizes a quantum Fourier transform in conjunction with a subroutine designed for efficient identification of the periodicity, allowing for the swift determination of the factors of an integer.

The algorithm's performance is measured by its asymptotic running time, which is polynomial in the size of the input,

*LabSIV, Department of Computer Science, Faculty of Sciences Agadir. Email: hi.amellal@uiz.ac.ma

whereas the best known classical algorithms for factoring are exponential in the size of the input. This means that for sufficiently large integers, Shor's algorithm can factor them in a reasonable amount of time on a quantum computer, while classical algorithms become infeasible. Shor's algorithm has important implications for cryptography, as many modern cryptographic protocols rely on the assumption that factoring large integers is computationally infeasible for classical computers. However, the development of a large-scale, error-corrected quantum computer capable of running Shor's algorithm remains a significant challenge.

We can summarize the key steps used by Shor's algorithm as follows:

- Quantum Fourier transform: The first step of Shor's algorithm is to apply a quantum Fourier transform to a superposition of possible solutions to the factoring problem. This transform effectively measures the frequency of the period of the integer being factored.
- Period-finding subroutine: The next step is to use a period-finding subroutine to determine the period of the function that maps a value to its modular exponentiation with the number to be factored. This step is critical for the success of the algorithm, as it allows us to find the factors of the number being factored.
- Continued fractions: Once the period of the function has been found, it can be used to construct a continued fraction approximation of the ratio of the two factors of the number being factored.
- Finally, the greatest common divisor of the original number and the factors obtained from the continued fraction approximation is computed to obtain the prime factors of the number.

Therfore, the algorithm consists of 2 parts:

- The classical segment of this algorithm is employed to transform the task of integer factorization into the quest for determining the period of a specific function. This period can be efficiently computed using a classical computer.
  In the first stage of Shor's algorithm, a number $a$ is randomly selected from the interval between 1 and $N$ — 1, ensuring that it is relatively prime to $N$. It then computes the period r of the function $f(x) = a^x \bmod N$. The period r can be found efficiently using the quantum part of Shor's algorithm. Once the period r is known, Shor's algorithm uses classical methods to compute the factors of N. Specifically, if r is even and $a^{(\frac{r}{2})}$ is not equal to — 1 mod $N$, Subsequently, the factors of $N$ can be derived as: $gcd(a^{\frac{r}{2}} + 1, N)$ and $gcd(a^{\frac{r}{2}} - 1, N)$. If $r$ is odd or $a^{\frac{r}{2}}$ is equal to — 1 mod $N$, then a new random number a must be selected and the process repeated.
- Quantum part used to find the period using the Quantum Fourier Transform (QFT).
  The quantum part of Shor's algorithm is used to efficiently find the period $r$ of the function $f(x) = a^x \bmod N$, where a is a randomly chosen integer between 1 and $N$— 1, and $N$ is the number to be factored. The quantum part of the algorithm utilizes a quantum computer and the utilization of the "QFT" enables the algorithm to effectively ascertain the period denoted as "r" of the function f(x) with remarkable efficiency. The "QFT" is a quantum analogue of the classical Fourier Transform, which is a mathematical tool used to analyze signals and identify their frequencies. To use the QFT in Shor's algorithm, we first initialize two quantum registers: one to store the input values of the function f(x), and the other to store the output values of the QFT. The input register is prepared in a uniform superposition of all possible input values, and the output register is initialized to a state of all zeros. Afterwards, a sequence of quantum gates, including the modular exponentiation gate, is applied to accomplish the intended task., to the input register to create a superposition of all possible values of f(x). Next, we apply the QFT to the input register to transform this superposition into a superposition of all possible periods $r$. Finally, we measure the output register to obtain a period r with high probability. If the measured period is even and $a^{\frac{r}{2}}$ is not equal to —1 mod $N$, then we can use the classical part of the algorithm to obtain the factors of $N$.

Overall, the classical part of Shor's algorithm is essential in obtaining the final factorization of the composite number N, but the quantum part is crucial in finding the period $r$ efficiently. In addition, the quantum part of Shor's algorithm is crucial in efficiently finding the period r using the QFT, which is exponentially faster than classical methods.

## Utilizing Shor's Algorithm in Practical Applications

Suppose we want to factor the number $N = 35$, which is the product of two prime numbers 7 and 5.

- In the first step we choose a arbitrary number $a$ among 1 and $N - 1$. Let's choose $a = 3$.
- In the second step we use the quantum part of Shor's algorithm to find the period $r$ of the function $f(x) = a^x \bmod N$. This is done by applying the Quantum Fourier Transform (QFT) to a superposition of states $|\alpha\rangle$, where $x$ ranges from 0 to $N$— 1, and measuring the result. The probability of measuring a state corresponding to a period $r$ is given by:
  $QFT \ \frac{1}{N}\sum x = 0^{N-1} |a^x \bmod N\rangle \ ^2$.
  In this case, we get the result $r = 4$ with high probability (around 50%for $N = 35$).
- Check if r is even and if $a^{\frac{r}{2}} + 1$ and $a^{\frac{r}{2}} - 1$ are not multiples of $N$. If they are not multiples of $N$, we can find the prime factors of $N$ as $gcd(a^{\frac{r}{2}} + 1, N)$ and $gcd(a^{\frac{r}{2}} - 1, N)$.
  In this case, we have $a^{\frac{r}{2}} + 1 = 3^2 + 1 = 10$ and $a^{\frac{r}{2}} = 3^2 - 1 = 8$, which have common factors with $N = 35$. Therefore, we need to try again with a different value of a until we get

a period $r$ such that $a^{\frac{r}{2}}+1$ and $a^{\frac{r}{2}}+1$ are not multiples of $N$.

- Repeat steps 1-3 until we find the prime factors of $N$. In practice, this can take many iterations and may require a large number of qubits and quantum gates.

In practice, using Shor's algorithm to factor large numbers on a quantum computer requires a large number of qubits and quantum gates, which are not yet available on current quantum computers. Therefore, factoring a number like $p = 15592110483128760 63$ using Shor's algorithm is not yet possible with current technology. Note that Shor's algorithm is only efficient for factoring large numbers on a quantum computer. For small numbers, classical algorithms are faster and more efficient.

## 3 Cryptographic Algorithm-Based Security Protocols

There are several web protocols related for secure the communication on the internet, and it is used by millions of websites worldwide to protect their users data. The most well-known and widely used ones is SSL (Secure Sockets Layer), TLS (Transport Layer Security), HTTPS (Hypertext Transfer Protocol Secure), , DNSSEC (Domain Name System Security Extensions) and SSH (Secure Shell). hese protocols are essential for protecting user's data and ensuring the security of web applications and services.

### 3.1 The Working Principle of Secure Sockets Layer (SSL)

SSL (Secure Sockets Layer) is a protocol that provides secure communication between two parties over the internet. It is used to establish an encrypted connection between a web server and a client (such as a web browser) to ensure that any data transmitted over the connection is protected and cannot be read by anyone who intercepts it [4]. The principle of work of SSL involves a series of steps that occur during the establishment of the encrypted connection:

- The client sends a request to the server to initiate an SSL connection.
- The server responds by sending a digital certificate to the client, which contains the server's public key and other information
- The client checks the certificate to ensure that it is valid and issued by a trusted authority.
- If the certificate is valid, the client generates a random symmetric key and encrypts it with the server's public key. This key is used to encrypt and decrypt data during the SSL session.
- The client sends the encrypted symmetric key to the server.
- The server decrypts the symmetric key using its private key.
- The server sends a message to the client, indicating that the SSL session has been established and encrypted communication can begin.
- The client and server can now exchange encrypted data over the SSL connection.

During the SSL session, all data transmitted between the client and server is encrypted using the symmetric key that was exchanged during the initial SSL handshake. This ensures that any data intercepted by an attacker is unreadable without the symmetric key.

Overall, the principle of work of SSL involves the exchange of digital certificates and symmetric keys to establish an encrypted connection between a client and server, ensuring secure communication over the internet (see Figure.1).

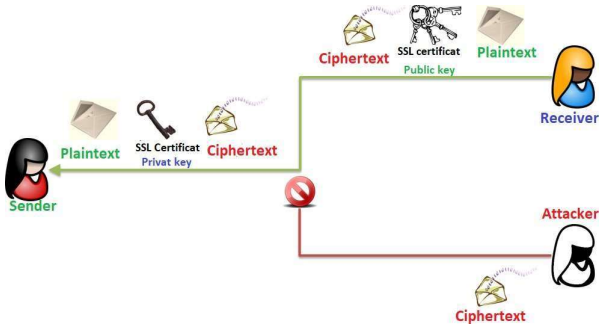



Figure 1: Principle of work of SSL

### 3.2 The Mechanism Behind the HTTPS Protocol

HTTPS is the secure version of HTTP, the protocol used for transferring data between a web browser and a website. It uses encryption to protect the data being transmitted, making it more difficult for attackers to intercept and steal sensitive information such as passwords, credit card numbers, and personal data. HTTPS uses SSL (Secure Sockets Layer) or TLS (Transport Layer Security) to establish an encrypted connection between the web server and the client (the web browser). The encryption ensures that any data transmitted over the connection is protected and cannot be read by anyone who intercepts it. In summary, HTTPS is a vital web protocol for secure communication on the internet, and it is used by millions of websites to protect their user's data [5].

#### 3.2.1 Cryptographic Algorithms Employed by HTTPS

HTTPS is a protocol that uses a combination of different cryptographic algorithms to provide secure communication over the internet. The main cryptographic algorithms used in HTTPS are:

- Symmetric-key encryption: HTTPS uses symmetric-key encryption to encrypt the data being transmitted between the client and server. The most commonly used symmetric-key encryption algorithms in HTTPS are AES (Advanced Encryption Standard) and 3DES (Triple Data Encryption Standard)

- Public-key encryption: HTTPS uses public-key encryption to establish a secure connection between the client and server. This is done through the use of digital certificates, which contain a public key that is used to encrypt data and a private key that is used to decrypt data. The most commonly used public-key encryption algorithm in HTTPS is RSA (Rivest-Shamir-Adleman).
- Hash functions: HTTPS uses hash functions to ensure data integrity and authenticity. Hash functions generate a unique digital fingerprint, or hash, of the data being transmitted, which is used to ensure that the data has not been tampered with or altered during transmission. The most commonly used hash functions in HTTPS are SHA (Secure Hash Algorithm) and MD5 (Message Digest 5).

Overall, HTTPS uses a combination of symmetric-key encryption, public-key encryption, and hash functions to provide secure communication over the internet, ensuring that data transmitted between the client and server is protected and cannot be read or altered by anyone who intercepts it. In this paper, we focus on the public-key encryption step, which is based on the RSA algorithm as mentioned below.

### 3.2.2 The Operational Principle of RSA Encryption

RSA (Rivest-Shamir-Adleman) stands as a renowned public-key encryption algorithm that plays a vital role in ensuring secure data transmission across the internet. Its inception dates back to 1977 when it was jointly developed by Ron Rivest, Adi Shamir, and Leonard Adleman. Even today, RSA remains one of the most widely utilized encryption algorithms. This cryptographic scheme relies on the principles of modular arithmetic and the challenge of factoring large composite numbers. By leveraging two large prime numbers, RSA generates a public key and a private key. The public key serves the purpose of encrypting data, while the private key is employed for decrypting the data.[6].
The process of generating a public key and a private key in RSA is as follows:

- Choose two large prime numbers, $p$ and $q$.
- Calculate $n = p \times q$
- Calculate $\varphi(n) = (p-1) \times (q-1)$.
- Choose an integer e such that $1 < e < \varphi(n)$ and $e$ is coprime with $\varphi(n)$. $e$ is the public key exponent.
- Calculate d such that $d \times e \equiv 1 (\bmod \varphi(n))$. $d$ is the private key exponent.

The public key is then $(n, e)$, and the private key is $(n, d)$.

To encrypt a message using RSA, the sender uses the recipient's public key to encrypt the message. The encryption process involves converting the message into a numerical value, raising it to the power of the recipient's public key, and then taking the result mod $n$. The resulting number is the encrypted message.

To decrypt the encrypted message, the recipient uses their private key to perform the reverse calculation. They raise the encrypted message to the power of their private key and then take the result mod$n$. The resulting number is the original message.

The security of RSA is based on the fact that it is computationally infeasible to factor large composite numbers into their prime factors. The public key in RSA consists of two large prime numbers, and it is difficult to determine these prime numbers from the public key alone. This makes it difficult for an attacker to decrypt the encrypted message without the private key. Overall, RSA provides a secure way to encrypt and decrypt data, making it an important tool for secure data transmission over the internet(see Figure.2).

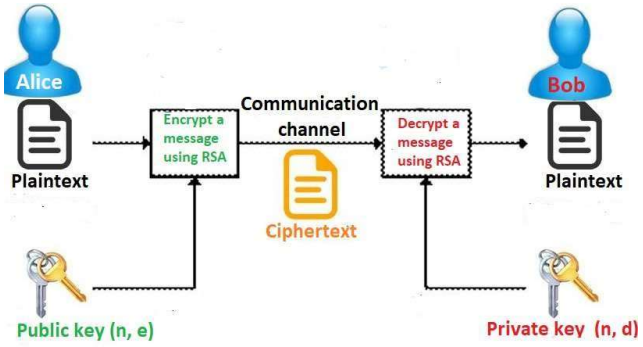



Figure 2: Principle of work of RSA algorithms

## 3.3 Security Analysis of HTTPS in Light of Shor's Algorithm

In section 3.2, we discussed that the security of HTTPS based on different cryptographic algorithms. Also, we mentioned that RSA is a widely used algorithm in HTTPS for public key generation, and any vulnerabilities or attacks that can compromise RSA could potentially weaken the security of HTTPS.

RSA is founded on the computational complexity of factoring large composite numbers into their prime factors. Shor's algorithm is a quantum algorithm that can efficiently factor large numbers, which could potentially break the security of RSA.

In the context of security analysis of RSA against Shor's algorithm, there are two main aspects to consider: the vulnerability of RSA to Shor's algorithm and the impact of this vulnerability on the security of systems that use RSA.

On the first aspect, it has been shown that Shor's algorithm can efficiently factor large numbers using a quantum computer, which means that RSA could be vulnerable to quantum attacks[7, 8, 9, 10]. However, it is important to note that building a large-scale quantum computer capable of running Shor's algorithm is still a challenging task, and there are still many technical and practical limitations that need to be overcome. On the second aspect, the impact of RSA's vulnerability to Shor's algorithm on the security of systems that use RSA depends on

various factors, such as the size of the key used in RSA, the sensitivity of the information being protected, and the available resources of the attacker. Theoretically, the RSA algorithm is vulnerable to attacks using Shor's algorithm. This means that if a large enough quantum computer were built, it could be used to break RSA encryption.

To secure RSA against attacks using Shor's algorithm, several post-quantum cryptographic schemes have been proposed. These schemes use mathematical problems that are believed to be hard for quantum computers to solve, such as the learning with errors (LWE) problem and the code-based McEliece cryptosystem. There are also efforts underway to develop quantum-resistant versions of RSA itself, which would involve modifying the RSA algorithm to make it resistant to attacks using quantum computers. However, this is still an area of active research, and it remains to be seen whether quantum-resistant versions of RSA can be developed that are as efficient and practical as the current version.

In the context of enhancing the security level of RSA against quantum attacks, we propose in this paper a quantum IPS that may help to secure RSA.

## 4 Description of the proposed quantum intrusion prevention system (QIPS)

In order to perform a quantum attack on a cryptographic system like RSA, a quantum computer alone is not enough. The attacker also needs to establish a secure communication channel with the target system using a quantum communication protocol, such as the BB84 protocol, in order to exchange information securely.

In a quantum communication system, information is encoded in quantum states, such as the polarization of photons [11]. These quantum states are then transmitted over a physical channel, such as an optical fiber, and measured at the receiving end. By using the principles of quantum mechanics, it's possible to detect any attempts to intercept or eavesdrop on the communication, as any observation of a quantum state changes its state.

Therefore, using a secure quantum communication protocol like BB84 can help to ensure the security of the communication channel, which is essential for performing a quantum attack on a cryptographic system like RSA.

### 4.1 Reviewing the Quantum Key Distribution Protocol: BB84

The BB84 protocol, formulated by Charles Bennett and Gilles Brassard in 1984, is considered as the most known quantum key distribution (QKD) protocol[12, 13]. The protocol is designed to allow two parties, traditionally named Alice and Bob, to establish a secure shared secret key over an insecure communication channel.

The protocol uses the properties of quantum mechanics to ensure the security of the key distribution process[14, 15]. The sender and the receiver both have access to a source of single photons that can be in one of four possible states, represented by two non-orthogonal bases. The sender randomly chooses one of the two bases to encode each photon, and sends the resulting sequence of photons to the receiver over the insecure communication channel. The receiver also randomly chooses one of the two bases to measure each photon upon reception.

Due to the non-orthogonality of the bases, the receiver's measurements are not always guaranteed to be correct. However, if the sender and the receiver choose their bases independently and at random for each photon, they can identify the presence of a malicious interceptor, traditionally named Eve, by comparing a subset of their measurement results. If the attacker has intercepted any of the photons to measure them, her presence will cause errors in the receiver's measurements, which the sender and the receiver can detect by comparing a subset of their results. They can then discard the corresponding key bits and establish a shorter, secure shared key from the remaining bits.

The BB84 protocol provides information-theoretic security, meaning that it is secure against any amount of computational power that the attacker may have. The protocol has been implemented experimentally and is widely considered to be a significant milestone in the field of quantum cryptography

- The sender transmits a sequence of random bits to the receiver by choosing either the "Horizontal/Vertical" or "Diagonal/Antidiagonal" bases to encode each bit
- The receiver randomly selects either the "Horizontal/Vertical" or "Diagonal/Antidiagonal"' basis to measure the states received from the sender,
- After transmitting the quantum states, the sender and the receiver communicate classically to exchange the bases they used for encoding and measuring the states. They then discard any bits in their shared key for which they used different bases during transmission.
- To improve the security of the shared key, the sender and the receiver publicly communicate a subset of the remaining bits.

The BB84 protocol utilizes single qubits to transmit key bits from the sender to the receiver. Each qubit is encoded in one of two orthonormal bases, which are conjugate to each other. When the sender uses the H/V bases, the signal states take on the following form:

$$\begin{aligned} |Horizontal\rangle &= \frac{1}{\sqrt{2}}(|0_Z\rangle + |1_Z\rangle) \\ |Vertical\rangle &= \frac{1}{\sqrt{2}}(|0_Z\rangle - |1_Z\rangle). \end{aligned} \tag{1}$$

When the sender utilizes the "Diagonal/Antidiagonal" bases, The signal states exhibit a varied form:

$$\begin{aligned} |Diagonal\rangle &= \frac{1}{\sqrt{2}}(|0_Z\rangle + i|1_Z\rangle) \\ |Antidiagonal\rangle &= \frac{1}{\sqrt{2}}(|0_Z\rangle - i|1_Z\rangle). \end{aligned} \tag{2}$$

### 4.2 Description of Unambiguous state discrimination (USD)

The subject of "USD" (Unambiguous State Discrimination) exhibits a strong connection to both QKD protocols and entanglement swapping protocols. It proves particularly useful in the realm of quantum communication, specifically when two signal states, which have yet to be implemented in solid-state systems, become non-orthogonal after traversing a channel. In the domain of quantum state discrimination, the primary goal is to design a measurement technique that effectively distinguishes a specified set of states. While the minimum-error measurement, known as the Helstrom measurement, is employed to differentiate between two equiprobable non-orthogonal states through a projective measurement, the optimal USD (Unambiguous State Discrimination) measurement is achieved through a generalized measurement known as the Ivanovic-Dieks-Peres (IDP) measurement. [16].

Suppose we have the simple example of comparing two coherent states that are different from each other: $|\alpha\rangle$ and $|\varphi\rangle$. A coherent state is a state for which $\hat{x}|\varphi\rangle = \varphi\,|varphi\rangle$, where $\hat{x}$ is the annihilation operator. In this case, we have no knowledge of the phase or amplitude of $|\varphi\rangle$ and $|\lambda\rangle$, only that the states are coherent. To compare the two states, we can use a 50%/50% beam splitter, as shown in Figure.3.

$$\hat{x}_{result} = \frac{1}{\sqrt{2}}(\hat{x}_{initial} + \hat{y}_{initial}) \tag{3}$$

$$\hat{y}_{result} = \frac{1}{\sqrt{2}}(\hat{y}_{initial} - \hat{y}_{initial}) \tag{4}$$

After performing some calculations, we found that $|\alpha\rangle$ and $|\beta\rangle$ transforms in the following way:

$$|\varphi\rangle_{x\ initial} \otimes |\lambda\rangle_{y\ initial} \Rightarrow \frac{\varphi+\lambda}{\sqrt{2}}_{x,result} \otimes \frac{\varphi-\lambda}{\sqrt{2}}_{y,result} \tag{5}$$

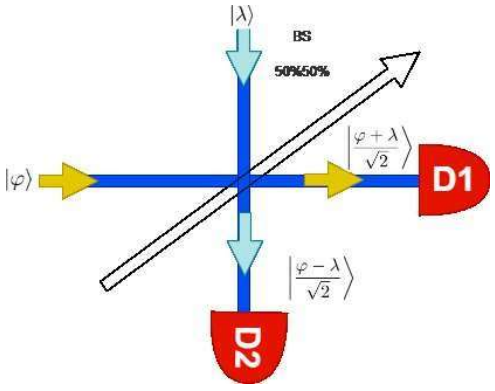



Figure 3: The beam splitter mixes the two input fields

- In the absence of dark counts in the detectors and when the values of $\varphi$ and $\lambda$ are equal, the result mode $y$ will solely consist of vacuum. Accordingly, the presence of any signal detected in result $y$ confirms the non-identity of the phase and amplitude of $\varphi$ and $\lambda$.
- In the case when $P_{dark-counts} /= 0$ in the detectors, we can no longer definitively confirm that $\varphi /= \lambda$ in this scenario.
- The detector's efficiency against dark counts is not flawless, resulting in a decrease in its effectiveness and causing a probability of detecting a distinction between $\varphi$ and $\lambda$ that is less reliable.

The success probability to detect a variation between $\varphi$ and $\lambda$ is equal to the probability of detecting at one oEr more qubit in result mode $y$, where the coherent state $\frac{\varphi-\lambda}{\sqrt{2}}$ is present. Since the probability of detecting no qubits in this mode is given by: $p(0) = \exp\left(-\frac{1}{2}|\varphi-\lambda|^2\right)$, the success probability can be expressed as follows:

$$Probability_{-success} = P(0) = 1 - e^{-(1/2)|\varphi-\lambda|^2} \tag{6}$$

### 4.3 Designing the Quantum Intrusion Prevention System (QIPS)

In theory, the sender use one of the four possible states at random to transmit a key bit. The receiver measures the received qubit in both the "Horizontal/Vertical" and "Diagonal/Antidiagonal" bases, which are selected with equal probability. Subsequently, They establish key reconciliation by utilizing a classical channel and selectively preserving the key bits where matching bases were employed. The resultant key is then amplified. However, in practical implementation, the sender utilizes Dim-Laser pulses for transmitting the states via an optical fiber. In this investigation, we analyze the sender's origin states to prevent unauthorized signal propagation across the quantum channel. Furthermore, we express one of the four states by employing a photon mode pair, denoted by the annihilation operators $a_R$ and $a_S$.

$$\begin{aligned}
|Horizontal\rangle &= e^{-|\varphi|^2} e^{\varphi(a^\dagger + a^\dagger)} |Vac\rangle_R \otimes |Vac\rangle_S = |\varphi\rangle_R \otimes |\varphi\rangle_S \\
|Vertical\rangle &= e^{-|\varphi|^2} e^{\varphi(a_R^\dagger - a^\dagger)} |Vac\rangle_S \otimes |Vac\rangle_S = |\varphi\rangle_R \otimes |-\varphi\rangle_S \\
|Diagonal\rangle &= e^{-|\varphi|^2} e^{\varphi(a^\dagger + ia^\dagger)} |Vac\rangle_R \otimes |Vac\rangle_S = |\varphi\rangle_R \otimes |i\varphi\rangle_S \\
|Antidiagonal\rangle &= e^{-|\varphi|^2} e^{\varphi(a^\dagger_R - ia^\dagger)} |Vac\rangle_S \otimes |Vac\rangle_S = |\varphi\rangle_R \otimes |-i\varphi()7\rangle_S
\end{aligned}$$

In this study, we focus on mode S, which is considered the "signal" pulse and used to encode the sender's information. We will examine a transmission scenario where a signal state $\chi|\ \rangle$ passes through our proposed device. When the incoming signal is received, certain detectors will click while others won't, depending on the state of the signal. In the following paragraph, we will describe various transmission scenarios and clarify how the quantum IPS either permits or blocks the input signal based on the detectors behavio (see Figure.4).

In the processing stage of the "QIPS" system, the original signal denoted as $|\chi\rangle$ is divided into 2 parts. The first part, $|\frac{\sqrt{\chi}}{2}\rangle$, undergoes analysis through the "QIPS", while the second part $|\frac{\sqrt{\chi}}{2}\rangle$, depending on the filtering rules, is either transmitted to detectors of received or rejected [17, 18]. During the processing stage, the initial portion is divided into four sub-fractions using beam splitters. The initial modes $|\frac{\sqrt{\chi}}{2}\frac{}{2}\rangle$ are combined with

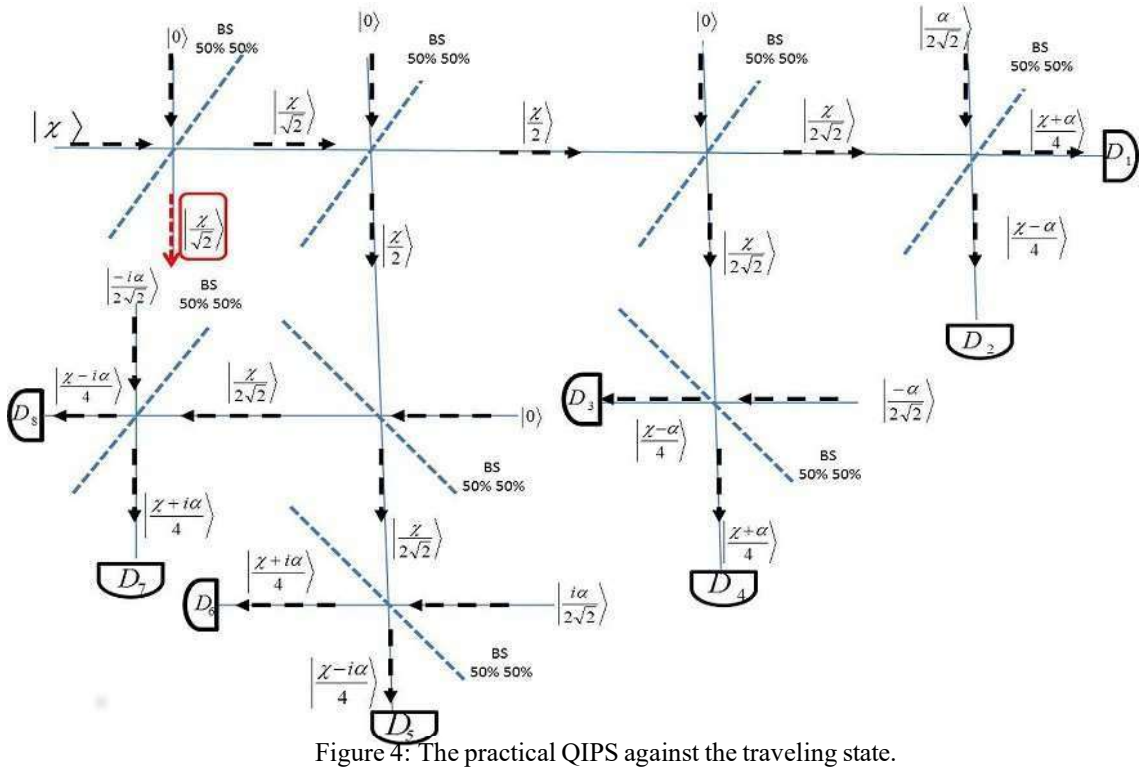



Figure 4: The practical QIPS against the traveling state.

special pulses such as $|\frac{\sqrt{\alpha}}{2}\rangle$, $|-\frac{\alpha}{2\sqrt{2}}\rangle$, $|i\frac{\sqrt{\alpha}}{2}\rangle$, and $|-i\frac{\alpha}{2\sqrt{2}}\rangle$ as the second beam-splitter input modes. The resulting output modes are $|\frac{\chi+\alpha}{4}\rangle$, $|\frac{\chi-\alpha}{4}\rangle$, $|\frac{\chi+i\alpha}{4}\rangle$ and $|\frac{\chi-i\alpha}{4}\rangle$ which are then directed to detectors $D_i$ (where $i \in 1, 8$) to measure the incoming signal amplitude.

### 4.4 Description of the 'QIPS' device

In the processing stage of the "QIPS" system, the original signal denoted as $|\chi\rangle$ is divided into 2 parts. The first part, $|\frac{\sqrt{\chi}}{2}\rangle$, undergoes analysis through the "QIPS", while the second part $|\frac{\sqrt{\chi}}{2}\rangle$, depending on the filtering rules, is either transmitted to detectors of received or rejected [17, 18]. During the processing stage, the initial portion is divided into four sub-fractions using beam splitters. The initial modes $|\frac{\sqrt{\chi}}{2}\rangle$ are combined with special pulses such as $|\frac{\alpha}{\sqrt{2}}\rangle$, $|-\frac{\alpha}{2\sqrt{2}}\rangle$, $|i\frac{\alpha}{\sqrt{2}}\rangle$, and $|-i\frac{\alpha}{\sqrt{2}}\rangle$ as the second beam-splitter input modes. The resulting output modes are $|\frac{\chi+\alpha}{4}\rangle$, $|\frac{\chi-\alpha}{4}\rangle$, $|\frac{\chi+i\alpha}{4}\rangle$ and $|\frac{\chi-i\alpha}{4}\rangle$ which are then directed to detectors $D_i$ (where $i \in 1, 8$) to measure the incoming signal amplitude. The detectors will either produce a click or not, depending on the output states. In this case, there are two possibilities:

- If certain detectors fail to click, it can be inferred that the input state originates from a legitimate sender, and as a result, the second part can be permitted to proceed through the receiver's measuring devices.
- In the event that all detectors click, it can be inferred that the input signal may have been intercepted by a spy. As a result, the second part of the split signal will be rejected. The receiver can quickly detect the presence of an eavesdropper using this method.

Suppose that the incoming signal to be analyzed using the proposed "QIPS" is $|\alpha\rangle$. In this case, only detectors 1, 4, 5, 6, 7, and 8 will click, while detectors 2 and 3 will remain silent, as shown in Figure.5. This clearly indicates the presence of an eavesdropper trying to intercept the transmission, enabling the receiver to detect their presence easily. We can summarize these simple rules in the following table:

## 5 Security analysis of the 'QIPS' system

In this section, we examine the impact of the proposed"QIPS" on various incoming signals. To simulate a real quantum communication scenario, we assume that the receiver cannot determine the source of the incoming signal (see Figure.5). This situation can be represented by three different scenarios:

- The first scenario involves the sender's signal source, which represents a typical communication between two legitimate correspondents. In this case, the sender transmits information using the four states mentioned earlier: $|\alpha\rangle$, $|-\alpha\rangle$, $|i\alpha\rangle$, and $|-i\alpha\rangle$.
- The second scenario involves the attacker's signal source in the context of a quantum attack. In this situation, the attacker creates pulses based on her measurements to send

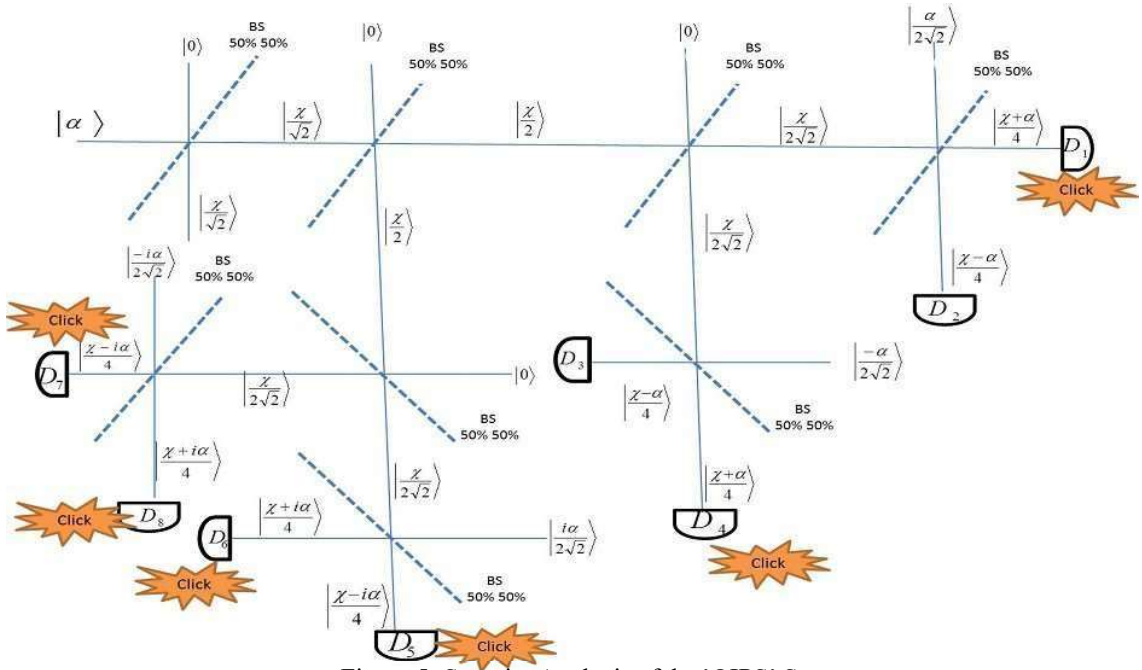



Figure 5: Security Analysis of the 'QIPS' System.

Table 1: Detector Responses Under Quantum IPS Operation.

| Incoming signal | Detectors that will click | Quiet detectors | The source of the signal |
|---|---|---|---|
| $\|\alpha\rangle$ | $D_1$, $D_3$, $D_5$, $D_6$, $D_7$, and $D_8$ | $D_2$ and $D_4$ | Legitimate sender |
| $\|-\alpha\rangle$ | $D_2$, $D_4$, $D_5$, $D_6$, $D_7$, and $D_8$ | $D_1$ and $D_3$ | Legitimate sender |
| $\|i\alpha\rangle$ | $D_1$, $D_2$, $D_3$, $D_4$, $D_6$, and $D_8$ | $D_5$ and $D_7$ | Legitimate sender |
| $\|-i\alpha\rangle$ | $D_1$, $D_2$, $D_3$, $D_4$, $D_5$, and $D_7$ | $D_6$ and $D_8$ | Legitimate sender |
| $\|\beta\rangle / = \{\|\pm\alpha\rangle, \pm i\alpha\rangle\}$ | all detectors | — | Illegitimate sender |

to detectors of the receiver. As a result, the attacker can potentially select the same signal amplitude as sender's.

- The third scenario involves the attacker's source signal for blinding the receiver's detectors. To carry out the quantum attack, the attacker must blind the receiver's detectors by using a special signal. This involves shining continuous light into receiver's detectors and manipulating the pulse strength or amplitude to control when they click. The attacker can also use this technique to prevent the receiver's detectors from detecting the legitimate input signals.

Based on the aforementioned scenarios, we will now analyze the behavior of the proposed 'QIPS' incoming pulses. We will consider each case individually based on the source of signal, and observe:

- In the first scenario, the sender's will use one of the following states to create a confidential key with the receiver based on the BB84 protocol $|\pm\sqrt{2}\alpha\rangle, |\pm i\sqrt{2}\alpha\rangle$ (see Figure.6).

Once the sender has prepared the random key, he sends the bit value corresponding to her chosen bases. As the signal travels,

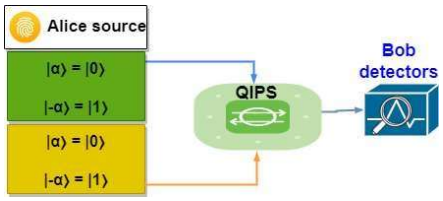



Figure 6: The normal use of the QIPS in a BB84 communication scheme.

it will be transformed by the "QIPS" process analysis (see Figure.7). From this analysis, we can conclude that the quantum state prepared by the sender passes the proposed 'QIPS' test and can then be measured in the receiver's devices according to the BB84 protocol.

In the first case, we consider the scenario in which the attacker attempts a quantum attack to obtain the secret key that the sender intends to share with the receiver. To carry out this attack, the attacker must first clone the architecture of both the sender and the receiver, as shown in the Figure.8:

In the second scenario, the attacker attempts to perform a

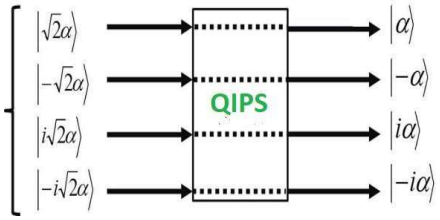



Figure 7: The evolution of the traveling qubits under the 'QIPS'.

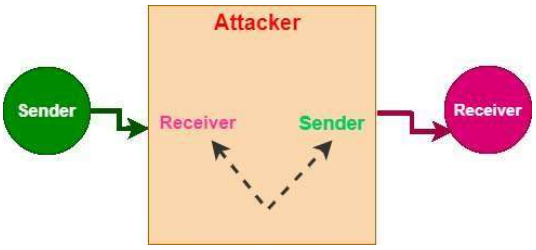



Figure 8: A simple scheme of quantum Attack.

quantum attack to obtain the secret key that the sender wants to share with the receiver. To achieve this, he clones the sender's and receiver's architecture and measures the intercepted states to prepare special pulses that will be sent to the receiver. The attacker may also clone the sender's states with the same amplitude, $|\pm\sqrt{2}\alpha\rangle, |\pm i\sqrt{2}\alpha\rangle$, in the first stage of the attack. In the next stage, she tries to blind the receiver's detectors using another light pulse, $|\beta\rangle$, that renders the blinded devices working in a linear mode. However, this state cannot bypass the proposed "QIPS" device and reach the target, as described earlier. Therefore, the receiver's attempt to blind the receiver's detectors will fail, and this attack will be ineffective in the presence of the proposed quantum device.

In the final scenario, we assume that the attacker prepares his states differently from sender's states. Analogously to the second case, it can be concluded that the attacker's states will also be blocked in the first attack stage before the blinding step, demonstrating the effectiveness of the proposed quantum IPS against such attacks.

From the results provided above, it is clear that the proposed Quantum IPS can distinguish between legitimate and illegitimate quantum signals. Additionally, we demonstrated that in order to use quantum computing to attack classical networks, a quantum network must be utilized first to initiate the attack. Therefore, if we can secure the quantum network against attack strategies based on quantum computing, we will also secure the classical network. As a simple implementation of the proposed "QIPS," we will deploy it between the classical and quantum networks, as depicted in the following figure (see Figure.9).

In the description of Shor's algorithm, we demonstrated that it consists of both classical and quantum steps to factorize a prime number. However, our analysis focuses on the quantum steps. The quantum part of Shor's algorithm begins by preparing

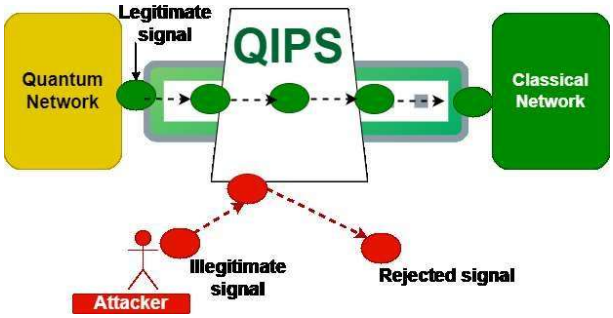



Figure 9: A Simple Scheme for Implementing Quantum Intrusion Prevention System (QIPS).

a quantum superposition of states using qubits. This is achieved by applying Hadamard gates to create a uniform superposition. The key quantum step in Shor's algorithm is the modular exponentiation, where repeated modular multiplications are performed using a controlled unitary operation to compute the values of $a^x$ mod $N$. This step utilizes quantum gates. Following the modular exponentiation, a quantum Fourier transform is applied to the output qubits to measure the periodicity encoded in the quantum state. The final step involves measuring the quantum state, resulting in a superposition of possible period values.

It is evident that Shor's algorithm, based on quantum computing, is capable of factorizing prime numbers and compromising the security of classical algorithms like RSA. To address this, the use of the Quantum Intrusion Prevention System (QIPS) can effectively detect the utilization of Shor's algorithm and reject the corresponding signals, thereby enhancing the security of classical networks.

## 6 Conclusion

In this study, we introduced a Quantum Intrusion Prevention System (QIPS) employing a superposition of elementary devices such as detectors and beam-splitters. To perform a thorough security analysis, we present diverse scenarios involving different sender sources. The objective is to assess the capabilities and limitations of the proposed 'QIPS' under challenging conditions and in hostile environments.

We delved into the security aspects of our proposed system against quantum attacks. We showed how the eavesdropping pulses are rejected during transmission and are detected by the receiver. Furthermore, our proposed communication method using the "QIPS" outperforms the standard protocol, providing an ideal balance between secure transmission and the simplicity of the physical setup. In conclusion, we have proposed a practical quantum IPS scheme that can contribute to preserving confidentiality and reducing the risk of eavesdropping by quantum algorithms.

## References


[1] P.A.M.Dirac, The Principles of Quantum Mechanics, 3rd ed.Oxford: Clarendon Press (1947).

[2] Kollmitzer.c, Pivk.M, Applied quantum cryptogtraphy, Lect.Not.Phys 797,ISBN 978-3-642-04829-6, Springer (2010).

[3] Peter W. Shor , SIAM J.Sci.Statist.Comput. 26,1480 (1997).

[4] Dierks, Tim and Eric Rescorla (2002). The TLS Protocol Version 1.1; IETF Internet-Draft.

[5] Mohamed G. Gouda, Elements of Network Protocol Design 1st Edition, Kindle Edition, 2008.

[6] Bleichenbacher, Daniel (1998). "Chosen ciphertext attacks against protocols based on RSA encryption standard PKCS#1." Advances in Cryptology—CRYPTO'98, Lecture Notes in Computer Science, vol. 1462, ed. H. Krawczyk. Springer-Verlag, Berlin.

[7] B.Qi, Fung.C.-H.F, Lo.H.-K and Ma.X, Time-shift attack in practical quantum cryptosystems. Quant.Inf.Comp. 7, 73,2 (2007).

[8] Lu¨N.tkenhaus, Security against individual attacks for realistic quantum key distribution. Phys.Rev.A 61, 052304 (2000).

[9] V.Scarani, A.Acin, G.Ribordy and N.Gisin, Quantum cryptography protocols robust against photon number splitting attacks for weak laser pulse implementations. Phys.Rev.Lett. 92, 057901 (2004).

[10] Y.Zhao, C.-H.FFung, B.Qi, C.Chen and H.-K.Lo, Quantum hacking: experimental demonstration of time-shift attack against practical quantum-key- distribution systems. Phys. Rev. A 78, 042333 (2008).

[11] W.Y.Hwang, Quantum key distribution with high loss:Toward global secure communication. Phys.Rev.Lett. 91, 057901 (2003).

[12] C.H.Bennett, G.Brassard, In International Conference of Computers in Systems and Signal Processing, Bangalore, India (IEEE, New York 1984) 175 (1984).

[13] C.H.Bennett, G.Brassard and N.D.Mermin Phys. Rev. Lett. 68 (1992).

[14] A.Meslouhi, H.Amellal, Y.Hassouni, and A.El Allati, A Secure Quantum Communication via Deformed Tripartite Coherent States Journal of Russian Laser Research,35, pages369–382 (2014).

[15] P.W.Shor, and J.Preskill, simple proof of security of the BB84 quantum key distribution protocol. Phys.Rev.Lett 85, 441, 44 (2000).

[16] H.Amellal, A.Meslouhi, Y.Hassouni et M. El Baz. A quantum optical firewall based on simple quantum devices. Quantum Inf Process 14, 2617–2633 (2015).

[17] V.Makarov, Controlling passively quenched single photon detectors by bright light. New J. Phys. 11, 065003 (2009).

[18] V.Makarov, Anisimov.A and Skaar, J, Effects of detector efficiency mismatch on security of quantum cryptosystems. Phys. Rev. A 74, 022313 (2006).

# Unsupervised Interactive lecture evaluation using the Kano Model

Baghdadi Ammar Awni Abbas1*
University of Baghdad, College of Mass Media Baghdad, Iraq.
Najeeb Abbas Al-Sammarraie †
Al Madinah International University Kuala Lumpur, M Malaysia
Mohammed Al-Mukhtar ‡
Computer Center, University of Baghdad, Baghdad, Iraq.
Maha Abdulameer §
University of Baghdad, College of Mass Media Baghdad, Iraq.

June 21, 2024

**Abstract**

Interactivity is the most critical factor in effective e- learning in web-based learning environments. This paper presents an interactive lecture course for undergraduate students built using MATLAB AppDesigner from scratch. The subject of the lectures was the computer network. The system was presented to a second-year College of Mass Communication/Baghdad University. Aquestioniere was deployed for the students, according to the Kano model, to measure their students' satisfaction with the interactive lectures and their future expectations for such a type of lecture to be used systematically as a supplement to the ordinary lecture. Kano Model Analysis was used to measure students' satisfaction with the interactive lectures. The results showed that interactive lectures have great potential for satisfying students' learning needs. The suggested lectures may serve as a backup for the ordinary lectures, or as a training and testing method for students. All types of materials, scientific or social, can be taught in this manner. Keywords: Interactive lecture, e-learning, Kano model, Matlab AppDesigner, Learning Management System.

## 1 Introduction

Interactivity is the ability to use some functions or activities for those functions or operations that are accessible to the users, which allows them to utilize the content provided in a graphical user interface of the software and obtain feedback. Interactive lectures were divided into the following categories: 1- Supervised Interactive lecture: The lecturer has direct contact with students via an interactive device such as a tablet, telephone, computer, or voting device 2- Unsupervised Interactive lecture: The lectures are designed and programmed previously with a fixed material and tests

Both of the previous types are considered interactive; the former induces brainstorming and discussion, while the latter is more time-flexible and generates self-learning. The main techniques used in interactive learning models are brainstorming, discussions, debates, multimedia (audio and video) web interference, projects, and games. Interactive lectures help students learn, develop their critical thinking and analytical skills, set logical connections, and make decisions with the necessary arguments. Interactive lectures induce self-learning, communication skills(supervised), and creativity, thereby improving the quality of the material. However, it requires high personal adaptation skills and high Teacher Qualifications and skills, which may create psychological discomfort. The Kano model is an effective scheme for any product maker who wants an efficient method to compute and prioritize product features. This is a useful tool for enhancing any product or service. The Kano model was presented by Noriaki Kano in a paper published in 1984 at Tokyo University of Science. Enhancing popular features and complaint processing are ways to improve and maintain customer loyalty. However, Professor Kano wanted to find out if there were other ways to improve that matter. According to his hypothesis, customer loyalty depends on five types of emotional responses to product features. By experimenting with a sample of selected participants, he was able to create a reaction graph(Table 1) to anticipate the emotional responses, on which he proved that customer satisfaction relies on the complexity of an available function, which causes more emotional responses. The Kano model can be used to measure customer satisfaction and increase it, determine if the current features cause high customer satisfaction, and enhance the current feature to an optimal level. Kano analysis is performed when resources and time are limited, which can save money and identify priority areas in a product that need attention because it is underperforming.

*University of Baghdad, College of Mass Media Baghdad
†Al Madinah International University Kuala Lumpur, M Malaysia
‡Computer Center, University of Baghdad, Baghdad, Iraq.
§University of Baghdad, College of Mass Media Baghdad

## 2 RelatedWork:

Interactive lectures have been the subject of many papers for two decades, and with the widespread use of personal handheld devices such as mobile phones and tablets, the subject has acquired great momentum. [1] Give essential building blocks for programming questions in MATLAB. [2] Propose an interactive lecture method was proposed using the CDEARA model. [ 3] presented a blended learning model that can accurately forecast how blended learning will be employed in the event of a pandemic. [4] Introduce A framework for collaborative mobile learning that can be used to create mobile learning environments that incorporate free-to-use social networking, software, and communication tools. The [5] study investigates the variables that influence students' intent and preparedness to use mobile learning in Jordanian higher education. Researchers have examined the needs and interests of students regarding the design and implementation of m-learning. The literature on E-learning systems has been reviewed intensively and a comprehensive model is made to evaluate the learning system degrees of achievement concerning several different success factors[6]. [7]Compare didactic lectures with interactive lectures for learning enhancement in third- year BDS students at Nishtar Medical University, Mu ltan, and conclude that"interactive lectures are more popular and beneficial than didactic lectures".[8] Uses Matlab to grade multiple-choice questions in paper- based exams.[9] Presents a mixture of quantitative and qualitative surveys to find gaps and patterns in the literature relating to blended mobile learning in education.[10] Discovers that the dynamic trajectories shown in Kano's model are useful for mid-term customer preference prediction and are partially confirmed. [11] explains a project in progress to create computer science courseware modules that are accessible via the World Wide Web and incorporate interactive elements within the curriculum, to increase student interactivity. In [12] the unified theory of acceptance and use in technology (UTAUT) model was used to perform a descriptive and regression analysis on the feasibility of mobile recognition of the need to address inadequate infrastructure and limited access to high-quality education. The interactivity of course-management systems (CMS) has been explored, focusing on Taiwanese students' perceptions, uses, and evaluations[13]. In a study[14] they used the Kano Model and a "relationship marketing perspective" to suggest an approach for creating an online "non-academic" course that can lead to "student satisfaction".[15] used a voting system to increase interactivity between lecturers and students.

The Kano model with Quality Function Deployment was used to identify customer needs [16]. This study examines the effects of using an interactive board, Bluetooth broadcasting system, voting system, notepad, free Internet access, computer-based exams, and interactive classroom technology in modern classroom technology [17]. This chapter examines the emergence of mobile learning initiatives and their contributions [18].[19] Describes the use of interactive learning in developing countries. (UTAUT) model to investigate students' behavioral purpose in implementing and utilizing mobile learning in postsecondary education in East Africa[20]. In [21], the authors reviewed the entire spectrum of interactive lecture formats and highlighted the severe dearth of research that can guide considerations of what is a very popular teaching method in higher education. By asking students about their opinions on the use of interactive video lectures in online classrooms, the effectiveness of the engagement opportunities provided by the virtual classroom was evaluated. [22]. In[23] a wireless interactive learning device was used to provide desktop devices with greater transformative potential. There was good acceptance of interactive teaching among students when compared to traditional didactic lectures [24]. In [25] a precise description of dynamic interactive systems is given. In [26] researchers introduced a new teaching paradigm based on Ubiquitous Computing to increase the interactivity between students and teachers, and the outcomes revealed a notable rise in interactivity due to the use of this method. A course at Eotvos Lorand University, Budapest, was presented. The main aim of this course is to raise the understanding of participating lecturers about how to make their lectures more memorable[27].[28] Create Clicker questions as a form of interactivity, the student can see an explanation of why the answer is correct or wrong. In[29] a comparison was made between the utility of modified versions of the unified theory of acceptance and use of technology (UTAUT) for mobile learning adoption in a developing nation's education (Guyana). [30] Examine medical education technology applications with an emphasis on interactive learning. In [31], a voting device was used, and a two-year study concluded that interactive lectures are the most promising approach for teaching. The Kano model is used to measure the degree of satisfaction of students using a virtual 3D medical device [32], and the results showed that these types of devices can induce positive satisfaction by learners. The Kano model is used to evaluate student satisfaction with an interactive medical device called a virtual 3D electroencephalogram [33]. A thorough model and tool to gauge student satisfaction with asynchronous e- learning systems were created in this study. The methods for designing the questionnaire, creating the items, gathering information, and verifying the multiple-item scale have been explained [34].

Interactive lectures were built using MATLAB AppDesigner. The lecture had a very simple introductory interface, as shown in Figure (1). Students can choose any lecture to attend or go directly to the testing pages.


Figure 1: Graphical user interface for the lectures

The user can go from the main window to the lectures in Figure(2), where he can choose the part that he can study, or he can take the exam directly as shown in Figure (3).

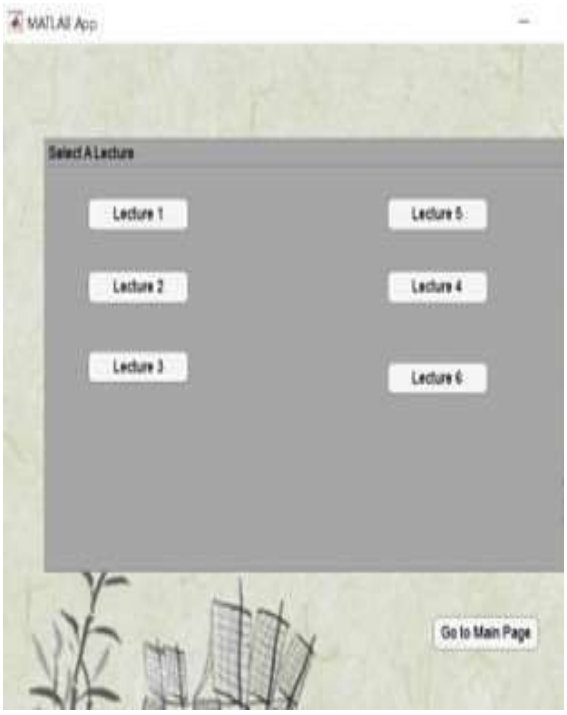



Figure 2: Lecture interface

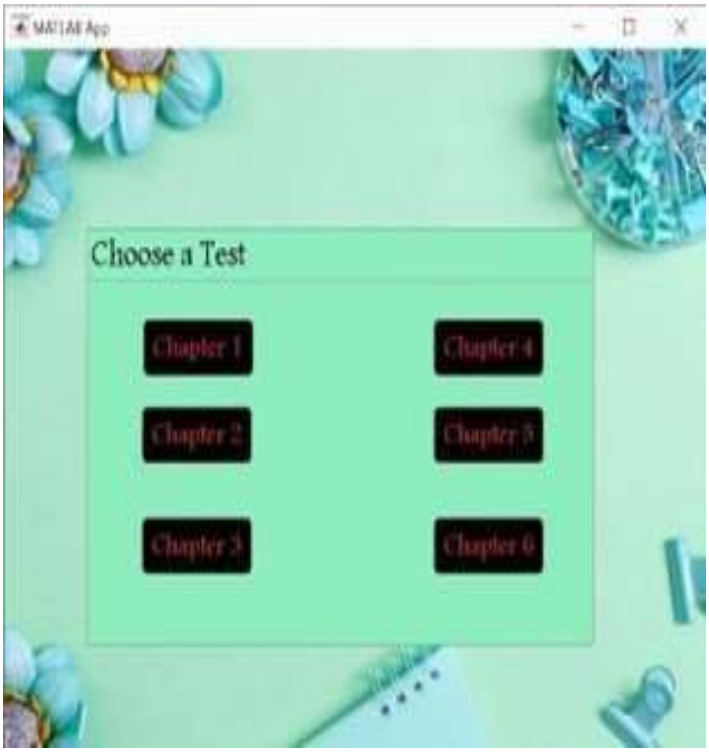



Figure 3: Take Exam.

When a student selects a lecture the first slide of the lecture appears. The slides range from simple text- only slides (Figure (4)) to complicated slides containing hyperlinks to a video or a website Figure(5(a)); pressing the pushbutton in that figure will lead to displaying the video shown in Figure(5(b)).


Figure 4: simple text-only slides.


Figure 5: Slides containing hyperlinks

All the slides have common features that include a push button, which leads to the previous slide; 2-Next push button, which leads to the next slide; 3- Take a Quiz push button, which leads to the end of the chapter exam directly without passing through the rest of the slides; and 4-Optional push buttons to display a video, audio or open a website. To create interactivity in the slides in each lecture (which normally consists of 15 - 35 slides), have a small quiz every 1-3 slides. The quiz types are Multiple Choice Questions, True or False, and Fill the Blank as shown in Figure 6 (a,b, and c). The most

popular forms of questions were programmed using MATLAB AppDesigner[34],[35].

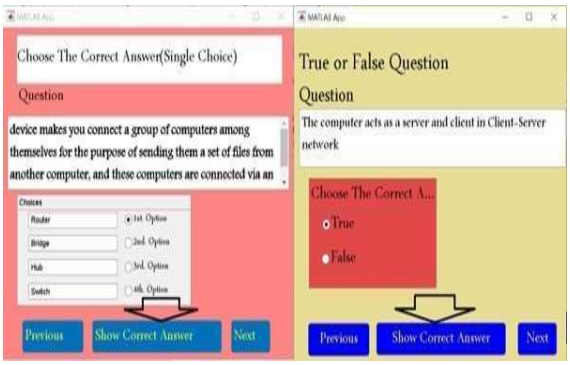



Figure 6

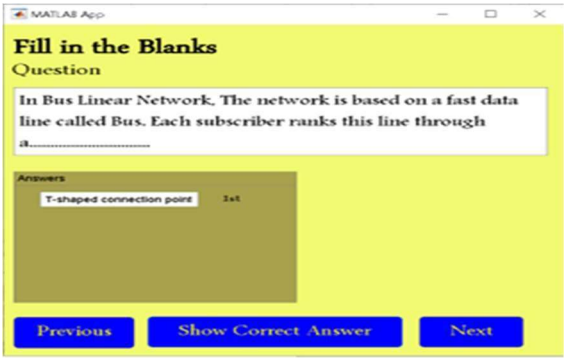



Figure 7: Pressing the Show Correct

However, the end of the chapter test contained Essay questions, where the answer appeared in a message box, as shown in Figure (8).

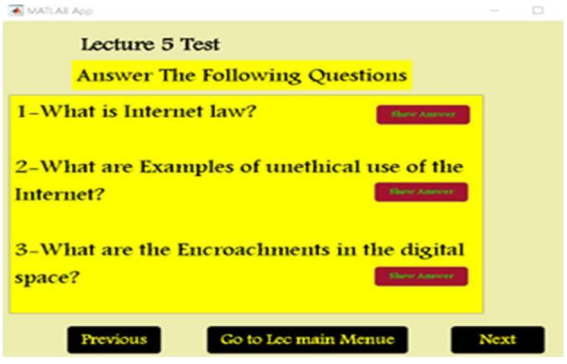



Figure 8: Essay Questions.

The Lectures were built using MATLAB AppDesignerand, and divided into six chapters. The lectures covered the material for the first semester of computer science for second-year students in the Department of Radio Television /the College of Mass Communication. An executable file was shared with students in Google Classroom. Upon completion of the course, the students completed a questionnaire. Each aspect of the interactive lecture properties was stated as a functional/dysfunctional pair. The Kano model analysis was depicted on the system to measure the satisfaction of the students (or customers according to the Kano model) with the product( the interactive lectures).

**Kano Analysis Model:** Any analysis process must undergo three stages, First: Selecting features, themes, and users for analysis; second: Obtaining the best information from customers and Third: Analyzing the results. Many researchers have exploited the flexible nature of the Kano analysis model to measure the degree of satisfaction of their customers. Software products, such as any product, have all the aspects that make them a good candidate to be subjected to the Kano model as a satisfaction-measuring tool. The students were considered the final customers as in[4]. The quality dimensions and items (themes) of interactive lectures can be categorized into (System Features, Learner interface, Personalization, and Future Enhancement)[21], as follows:

**System Features(SF):**

1- Stable system.
2- Work offline.
3- Unupervised, no direct interaction with the teacher, promotes self-learning.
4- Bilingual GUI, choose from two languages(English and Arabic).

**Learner interface(LI):**

1- There are many forms of questions at the end of the lecture, such as Essay questions multiple choice questions (single and multiple selections), and true or false questions.
2- Multimedia(video ,audio and web hyperlinks) to support the slides.
3- Scrambled answers each time you open the lectures

**Learner Community(LC):**

1- The learning system makes it easy to share what you learn with other students.
2- The learning system makes it easy to share what you learn with your teacher

**Personalization(P):**

1 - Easy to choose what you want to learn.
2-Choosing the time I want to learn.
3- Easy to choose how much you want to learn.

**Future Enhancement(FE):**

1-New Forms of questions like matching hot spot.
2-Certificate for the final test.
3-Slides have audio recordings for the original text.

4- System work on multiple platforms (PC, Tablets, and smartphones).

A Google Form questionnaire containing functional and dysfunctional features was created. The form was distributed to second-year undergraduate students from the Department of Radio and Television /College of Mass Media/ Unversity of Baghdad(who tested the software) and they were asked to answer them. These students were introduced to the interactive lectures as part of the materials of the first semester in computer techniques class( along with the other classes). The Multiple Choice Questions options measure student satisfaction. The responses ranged from Delighted (having full satisfaction) to Frustrated. Between these two are (Satisfied, Neutral, and Dissatisfied) values to measure the exact rate of satisfaction. From the previous theme, a questionnaire containing functional and dysfunctional features was created as follows:

**System Features(SF):**

1- Stable system. Functional: How would you feel about having a stable system? Dysfunctional: How would you feel about having an unstable system with pauses and glitches?

2- Work offline. Functional:

How would you feel about having the system working offline? Dysfunctional: How would you feel about having the system working online only?

3- Unupervised, no direct interaction with the teacher, promotes self-learning. Functional: How would you feel about having the system be unsupervised, having no direct interaction with the teacher, and promoting self-learning? Dysfunctional: How would you feel about having the system supervised, with direct interaction with the teacher?

4- Bilingual GUI, choose from two languages(English and Arabic). Functional: How would you feel about having a system that works with English and Arabic language choices? Dysfunctional: How would you feel about having the system with a single-language GUI?

**Learner interface(LI):**

1- There are many forms of questions at the end of the lecture such as Essay questions multiple choice questions (single and multiple selections) and true or false questions. Functional: How would you feel if you had many forms of questions at the end of the lecture like Essay questions multiple choice questions (single and multiple selections), plus true or false questions?

Dysfunctional: How would you feel if you had only one form of a question at the end of the lecture(only multiple-choice questions for example)?

2- Multimedia(video ,audio and web hyperlinks) to support the slides. Functional: How would you feel about using Multimedia(video, audio, and web hyperlinks) to support the slides? Dysfunctional: How would you feel about not using Multimedia(video, audio, and web hyperlinks) to support the slides, (static text slides only)?

3- For MCQs the answers are scrambled each time you open the lectures. Functional: How would you feel about scrambling the answers to the multiple-choice questions each time you open a lecture? Dysfunctional: 'How would you feel if the answers for the multiple- choice questions were static each time you open a lecture?

**Learner Community(LC):**

1- The learning system makes it easy to share what you learn with other students. Functional: Does the learning system make it easy to share what you learn with other students? Dysfunctional: This learning system does not make it easy to share what one learns with other students.

2- The learning system makes it easy to share what you learn with your teacher. Functional: Does the learning system make it easier to share what you learn with your teacher? Dysfunctional: The learning system does not make it easy for you to share what you learn with your teacher.

**Personalization(P):**

1- Easy to choose what you want to learn. Functional: Does the system let you easily choose what you want to learn? Dysfunctional: The system does not let you easily choose what you want to learn.

2- Choosing the time I want to learn. Functional: Does the system let you choose the time you want to learn?

Dysfunctional: The system does not let you choose the time you want to learn.

3- Easy to choose how much you want to learn. Functional: Does the system let you choose how much you want to learn? Dysfunctional: The system does not let you choose how much you want to learn.

**Future Enhancement(FE):**

1- New Forms of questions like matching hot spot. Functional: Do you want to add new forms of questions like matching hot spots? Dysfunctional: How would you feel if you had only the standard forms of questions MCQ T or F and Essay?

2- Certificate for the final test. Functional: Do you want to get a Certificate for each test? Dysfunctional: How would you feel if you didn't have a Certificate for each test?

3- Slides have audio recordings plus the original text. Functional: How do you feel if the slides have audio recordings that the teacher reads and explains the text in the slides? Dysfunctional: How would you feel if the slides had text only, without audio recordings from the teacher?

4- System work on multiple platforms (PC, Tablets and smartphones) Functional: How would you feel about having the system work on multiple platforms (PCs, Tablets, and smartphones)? Dysfunctional: How would you feel about having the system work on a single platform (PC alone)?

**Results:**

The collected questionnaire answers were subjected to the Kano evaluation table as shown in Table(1). The table joins

together the functional/dysfunctional answers in its rows and columns to obtain one of the Kano categories (Attractive(A), One Dimensional(O), Must be(M), Indifferent(I), Reverse(R), Questionable(Q)). Each answer pair leads to one of these categories.

Table (1): Kano Evaluation Table.

| | | Dysfunctional(Feature absent) | | | | |
|---|---|---|---|---|---|---|
| Dysfunctional (Feature absent) | | Dysfunctional(Feature absent) | Dysfunctional(Feature absent) | Dysfunctional(Feature absent) | Dysfunctional(Feature absent) | Dysfunctional(Feature absent) |
| | Like it | Q | A | A | A | P |
| | Expect it | R | I | I | I | M |
| | Don't care | R | I | I | I | M |
| | Live with | R | I | I | I | M |
| | Dislike | R | R | R | R | Q |

The questionnaires were asked to give a value for how important categories are to them on a scale from one to nine, where nine denotes the category is extremely important and 1 is not at all important. The results of this part of the questionnaire are shown in the figure(9).

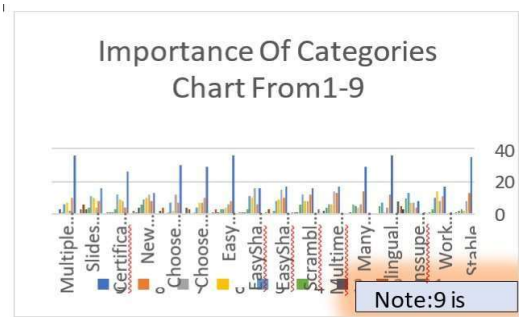



Figure 9: category importance

| Main Categories | No. | SYMBOL | SUM | M | O | A | R | I | Q | FINAL RESULT | BETTER | WORSE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SYSTEM FUNCTIONS | 1 | STABLE SYSTEM | 65 | 6 | 17 | 0 | 18 | 23 | 1 | A | 64.0625 | -37.5 |
| | 2 | WORK OFFLINE | 65 | 0 | 20 | 4 | 7 | 34 | 0 | A | 67.21311475 | -11.47540984 |
| | 3 | UNSUPERVISED | 65 | 0 | 47 | 13 | 0 | 2 | 3 | I | 4.091632653 | 0 |
| LEARNING INTERFACE | 4 | BILINGUAL LANGUAGE | 65 | 1 | 20 | 0 | 14 | 28 | 2 | A | 66.66666667 | -23.80952381 |
| | 5 | MANY FORMS OF Q | 65 | 3 | 22 | 5 | 10 | 24 | 1 | A | 57.62711864 | -22.03389831 |
| | 6 | MULTIMEDIA SUPPORT | 65 | 1 | 34 | 1 | 5 | 24 | 0 | I | 45.3125 | -9.375 |
| LEARNER CHARACTERISTICS | 7 | SCRAMBLED ANSWER | 65 | 1 | 40 | 6 | 6 | 11 | 1 | I | 29.31034483 | -12.06896552 |
| | 8 | EASY SHARE W STUDENTS | 65 | 2 | 36 | 2 | 8 | 16 | 1 | I | 38.70967742 | -16.12903226 |
| | 9 | EASYSHARE W TEACHER | 65 | 1 | 40 | 1 | 4 | 18 | 1 | I | 34.92063492 | -7.936507937 |
| REMEMBRANCE | 10 | EASY CHOOSE W LEARN | 65 | 5 | 21 | 0 | 18 | 21 | 0 | A | 60 | -35.38461538 |
| | 11 | EASY CHOOSE TIME TO LEA | 65 | 4 | 19 | 1 | 18 | 22 | 1 | A | 63.49206349 | -34.92063492 |
| | 12 | EASY CHOOSE HOWMUCH LE | 65 | 1 | 29 | 1 | 20 | 13 | 1 | I | 52.38095238 | -33.33333333 |
| | 13 | NEW FORMS OF Os | 65 | 0 | 39 | 6 | 4 | 16 | 0 | I | 33.89830508 | -6.779661017 |
| FUTURE EMPLOYMENT | 14 | CERTIFICATE AT END | 65 | 3 | 22 | 1 | 18 | 21 | 0 | I | 60.9375 | -32.8125 |
| | 15 | SLIDES WITH AUDIO | 65 | 1 | 38 | 6 | 5 | 14 | 1 | I | 32.75862069 | -10.34482759 |
| | 16 | MULTIPLE PLATFORMS | 65 | 2 | 20 | 1 | 26 | 14 | 2 | O | 64.51612903 | -45.16129032 |

**Table (2): Questionnaire Results**

The last two columns (Better/Worse); Where the Better represents the degree of customer satisfaction if the feature is present, and the Worse presents the degree of dissatisfaction if the feature is absent.

## 3 Discussion

The features selected in this paper were chosen by the author after a thorough discussion with the students and many members of the faculty of colleges at the University of Baghdad. Three categories were of top priority to students (Easy to choose how much you learn, the multiple platforms for the lectures, and Multi-lingual GUI). These features had a value of nine(Top Priority) for more than half of the students. Not surprisingly choosing how much you want to learn represents the essence of the unsupervised part of the unsupervised interactive lectures where the students can learn any part of the lecture at any time. On the other hand, having to choose the device for learning from three options (PC, Tablet, and Mobile phone) resulted in the ability to change the user from Arabic to English.

According to the Kano evaluation table, the ability to choose what to learn with bilingual GUI categories is considered attractive(Delighters or Exciters) by the students, because of the positive reactions felt by them. The multiple platform is considered a performance or one-dimensional category (where the more you have from these categories the more the customer is satisfied). Most of the other categories are considered indifferent to the students, such as scrambling the answers or the ability to share what they learn, mainly because the students are new to this type of education. The table also marks the absence of the must-be category, which occurred due to the formation of the questionnaire, where the questions were designed to be as compact as possible so as not to distract the students and get their full attention without boring them. Most of the answers are categorized as indifferent, which is mainly due to the infrequent use of this type of lecture by the students.

## 4 Reference:

1- Abbas, B.A.A., Al-Mukhtar, M.. Building Computer-Based Test (CBT) using MATLAB: Programming the Essential Types of Questions. International Journal of Computers and their Applications.Vol. 30, No. 3, PP. 311 – 324. 2023.

2- Abd Rahman N., Masuwai A A. Transforming the Standard Lecture into an Interactive Lecture: The CDEARA Model, International Journal for Innovation Education and Research. 2(10), 158-168. 2014.

3- Abdulhussien R., Najeeb H., Improving Measurement of effectiveness of blended learning In Iraqi education Using SVM. Iraqi Journal of Science 63(9),4057-4066. 2022.

4- Akhigbe, J.. Leveraging Collaborative Mobile Learning Instructional Pedagogy in The Era of COVID-19. Academia Letters, Article 2320. 2021. https://doi.org/10.20935/AL2320.

5- Al-Adwan1A.S., Al-Madadha A., Zvirzdinaite Z.. Modeling Students' Readiness to Adopt Mobile Learning in

Higher Education: An Empirical Study, International Review of Research in Open and Distributed Learning Vol 19, No.1,222-241. 2018.

6- Al-Fraihat, D., Joy, M., Masa'deh, R., Sinclair, J. . Evaluating E-learning systems success: an empirical study. Computers in Human Behavior, 2019(102), 67- 86. doi:10.1016/j.chb.2019.08.004 ISSN 0747-5632. 2020.

7- Ashraf S, Khan SA, Ahmad M, Rind SM, Fatima A, Safdar S.(2023). Comparison of didactic lecture with interactive lecture for learning enhancement in third- year BDS students at Nishtar Medical University, Multan. Professional Med Journal. 30(1), 129-135.

8- Baghdadi, A..An Automatic System to Grade Multiple Choice Questions Paper-Based Test (2009) J. of Al-Anbar University for Pure Science, 3 (1), pp. 174-181.

9- Baran, E.,. A Review of Research on Mobile Learning in Teacher Education. Educational Technology Society, 17 (4), 17–32. 2014.

10- Borgianni, Y.. Verifying dynamic Kano's model to support new product/service development, Journal of Industrial Engineering and Management (JIEM), 11(3), 569-587. 2018.

11- Carlson D., Gttzdial M., Kehoe C., Shah V., Stasko J.. www interactive learning environments for computer science education, ACM SIGCSE Bulletin28(1), 290–294. https://doi.org/10.1145/236462.236558. 2008.

12- Chaka J.G., Govender I.. Students' perceptions and readiness towards mobile learning in colleges of education: a Nigerian perspective, South African Journal of Education, 37( 1), 1-12. 2017.

13- Chou C., Peng H., Chang C.. The technical framework of interactive functions for course- management systems: Students' perceptions, uses, and evaluations, Computers Education, 55(1), 1004– 1017. 2010.

14- Dominici, G. Palumbo, F.. How to build an e- learning product: Factors for student/customer satisfaction. Business Horizons, 56(1), 87-96. 2013.

15- Draper, S.W. , Brown, M. I.. Increasing interactivity in lectures using an electronic voting system, Journal of Computer Assisted Learning 20, 81–94. 2004.

16- Gupta P., Srivastava R.K.. Customer Satisfaction for Designing Attractive Qualities of Healthcare Service in India using Kano model and Quality Function Deployment, MIT International Journal of Mechanical Engineering. 1(2), 101-107. 2011.

17- Hashim A., Kareem N. K. Effect Of Technology Learning As A Supplement To Traditional Technology On Student's Achievement, Journal of Engineering 18(12), 1439-1445. 2012.

18- Kukulska H., Sharples A., Milrad M., Arnedillo-S´anchez M., Inmaculada , Giasemi V.. The genesis and development of mobile learning in Europe. In: Parsons, David ed. Combining E-Learning and M- Learning: New Applications of Blended Educational Resources. Hershey, PA: Information Science Reference (an imprint of IGI Global), pp. 151–177. 2011.

19- Lamptey HK., Boateng R.. Mobile Learning in Developing Countries: A Synthesis of the Past to Define the Future. World Academy of Science, Engineering, and Technology International Journal of Educational and Pedagogical Sciences,11(2), 448-455. 2017.

20- Mtebe J.S., Raisamo S. Investigating students' behavioral intention to adopt and use mobile learning in higher education in East Africa, International Journal of Education and Development using Information and Communication Technology (IJEDICT), 10(3), 4-20. 2014.

21- Murphy R., Sharma N. What Don't We Know About Interactive Lectures?, International journal of media, technology and lifelong learning 6(1), 111-120. 2010.

22- Ottusch, T.Jordan A.C.. Students' perspectives on the use of interactive video lectures in online classes, Family Science Review. 26(3), 1-20. 2022.

23- Pea, R. D., Maldonado,H.. WILD for learning: Interacting through new computing devices anytime, anywhere. The Cambridge Handbook of the Learning Sciences, New York: Cambridge University Press, 852-886, hal-00190630. 2006.

24- Preethi BP, Sunitha M., Abhay K.C. Large Group Interactive learning among first-year medical students, Medica Innovatica. 11(2), 125-129. 2022.

25- Sabry, K., Barker, J.. Dynamic interactive learning systems. Innovations in Education and Teaching International, 46(2), 185-197.2009. https://doi.org/10.1080/14703290902843836.

26- Scheele N. et al. The Interactive lectures: A new teaching paradigm based on Ubiquitous computing, conference proceedings University of Mannheim, Germany, 1-162. 2005

27- Szabo E.. Interactive presentations and lectures in Higher Education, IV. International Conference From theory to practice language for specific purposes, PP.323-335. 2019.

28- Teese R.B. et al.Interactive online Lectures for asynchronous delivery, Journal of Physics.: Conf. Ser. 2297 012004. 2022.

29- Thomas T.D., Singh L. and Gaffar K.. The utility of the UTAUT model in explaining mobile learning adoption in higher education in Guyana. 2013.

30- Tuma F. The use of educational technology for interactive teaching in lectures, Annals of Medicine and Surgery. 62(1), 231-235. 2021.

31- Utomo P., Utama M M. Enhancing students' engagement using interactive lectures. The Indonesian Journal of Medical Education. 11(3) 326-331. 2022.

32- Vezzetti E., Violante, M M. Virtual interactive e- learning application: an evaluation of a student satisfaction, computer application in engineering science, ISSN 1061-3773. -, 72-91. 2013.

33- Violante M.G.; Vezzetti E.. Virtual Interactive E- Learning Application: An Evaluation of the Student Satisfaction. Computer Applications in Engineering Education, 23 (1), 72-91. ISSN 1061-3773. 2013.

34- Wang, Y.,(2003). Assessment of learner satisfaction with asynchronous electronic learning systems, Information Management, 41(1), 75-86.

# White and Black Box Techniques towards Deploying a Prediction Model In Educational Data Mining

Sapna Arora *
IILM University, Gurugram
Ruchi Kawatra†
SRM University, Haryana.
Narayan C. Debnath ‡
Eastern International University Thu Dau Mot, Vietnam

**Abstract**

An effective predictive system is essential within Educational Data Mining, integral to shaping learning and teaching methodologies. Its true utility lies in its ability to decipher progress, identify trends and behavioral patterns, and pinpoint the origins of educational challenges. By capturing and disseminating critical success factors to students and faculty, educational institutions ensure their achievements are duly acknowledged. This study delves into the deployment of both white and black box models in education through a comprehensive literature review, evaluating their respective advantages and limitations. Through the utilization of questionnaire-based surveys and various techniques including ID3, CART, XGBoost, and MLP, the researchers conducted an adaptive self-assessment case study. The findings indicate that black box models, particularly XGBoost and MLP, exhibit superior accuracy compared to their white-box counterparts. Drawing from recent research, the study offers practical recommendations for the construction of predictive systems and the presentation of output data. This research provides valuable insights for educators and scholars, shedding light on the efficacy of black box models in educational prediction.

## 1 Introduction

With the contemporary age, education is vital because it shapes both the individual and the community. It offers the resources required to comprehend and value history, culture, and other significant facets of human existence. Gaining this knowledge makes people more capable of making significant contributions to society, which promotes advancement and creativity. Education teaches people how to lead with empathy, integrity, and true values—qualities that are essential for effective leadership—in addition to academic subjects. In the modern world of rapid change, education is also a key factor in building sustainability. It serves as a primary means of communication and serves as the cornerstone for cultivating a ”sustainability mindset.” The concept encompasses a systemic approach to comprehending the interconnectedness of various components that contribute to a healthy ecosystem and a thriving society; it goes beyond simply knowing technical information. According to Inga[15], the sustainability mindset incorporates complex systems thinking and encompasses a holistic understanding that extends beyond the fundamentals. Education develops the critical thinking, creativity, and problem-solving abilities that are necessary to meet the challenges of the contemporary world. It promotes inquiry, discovery, and creativity among people, which propels social and technological progress. By fostering inclusive behaviours and teaching respect for diversity, education also helps to foster social cohesion. This contributes to the creation of a just and equitable society where everyone can prosper. By equipping people with the knowledge and abilities needed to successfully negotiate the challenges of modern life, education empowers people. It raises prospects for the economy, raises quality of life, and opens doors to opportunities. Societies can invest in their future by putting money into education, which makes sure that the next generation is equipped to face today’s problems with compassion and intelligence.

Effective teaching and learning practices are increasingly shaped by the relationship between education and educational data. Educational data is essential for shaping educational strategies and policies. It includes a wide range of metrics, from student performance and attendance to engagement levels and socio-emotional indicators. Through methodical data collection and analysis, educators and administrators can obtain thorough insights into avariety of facets related to the teaching and learning process. By using data to identify patterns, areas for growth, and strengths, this data-driven approach helps educators create more effective, individualized lessons that are catered to the individual needs of every student[33]. One of the primary advantages of utilizing educational data is the capacity to implement evidence-

*School of Computer Science Engineering Reasearch Assistant Email: sapnaarora023@gmail.com
†Department of Computer Science Engineering, SRM University, Haryana
‡School of Computing and Information Technology,

based practices. Through data analysis, educators can move beyond anecdotal evidence, making decisions rooted in concrete information. For example, teachers can use data to identify the most effective instructional strategies for specific student groups, thereby improving learning outcomes. Additionally, data can highlight disparities in achievement among different student demographics, leading to targeted interventions aimed at closing achievement gaps and promoting equity within the educational system. The use of educational data also fosters a culture of accountability and continuous improvement within educational institutions. By consistently monitoring and evaluating educational outcomes, schools and districts can ensure they are meeting their goals and progressing towards enhancing student achievement. This ongoing feedback loop not only aids in refining instructional methods but also in optimizing resource allocation, ensuring that resources and efforts are directed towards the most impact areas[9]. Essential to the support of differentiated instruction is educational data. Teachers are able to create and execute lessons that accommodate different learning styles and capacities when they have access to comprehensive data on student performance and learning preferences[7][10]. By addressing the various needs of students, this individualized approach contributes to the effectiveness and inclusive of education. Additionally, information on the socio emotional health and engagement of students can direct the creation of supportive interventions, resulting in a more comprehensive learning environment that fosters both academic and personal development.When it comes to crafting and executing policies that are responsive to the realities of the educational landscape, educational data gives decision-makers the tools they need. Data driven policies, since they are founded on a comprehensive understanding, can result in more effective and efficient educational systems.

The growing usage of data mining in the education industry has necessitated the development of high-quality data mining courses for students and faculty. Data mining is a technique for detecting patterns and correlations in data to enhance decision-making. It is a multidisciplinary field that combines statistics, evolutionary computation, artificial intelligence, databases[22], machine learning, neural networks[23], pattern recognition, information visualization, and knowledge acquisition approaches. Many approaches for data analysis are available through data mining.[5][8]. Without the use of automated analytical techniques, the vast volume of data already in educational databases exceeds the individual ability to determine and extract the most valuable data. The process of extracting implicit, unknown, and possibly relevant information from a big database is known as knowledge discovery (KDD). The analysis of personnel data to improve work performance by providing different parameters, as well as to promote job sustainability and job satisfaction by understanding their behavior and other associated qualities, is a perfect use of data mining in the educational sector[16][24]. Yet, several researchers use white or black-box machine learning models[12] to predict work performance, but none of them have provided a comparative evaluation of the same in the educational sector. The model was created to cover a variety of pedagogical facets, and it was expanded to include perceived understanding and its impact on how white and black box algorithms are adopted. A survey experiment was carried out with several educators to examine their behavior. From the 1598 responses, the black-box and white-box algorithms as well as the educators' cognitive styles were examined. the outcomes of how cognitive types influence perceived behavioral knowledge. The research was carried out using Jupyter notebook, a web-based data-mining platform that supports both white-box and black-box approaches.

An efficient prediction system is a powerful tool in Educational Data Mining systems, as it is associated with learning and teaching methodologies. To be truly useful as a teaching resource, a prediction model must include tools for properly interpreting progress, detecting trends behavioral patterns, and determining the source of

teaching and learning issues. Educational Data Mining (EDM) delivers a wealth of useful data and a more complete view of individuals and their learning activities. It analyses educational data[29]and solves educational challenges[13][28] using data mining methods through two types of models. One of them is a white box and the other is a black box. These two Data Mining models collect information from educational data that is interesting, comprehensible, helpful, and unique. On the one hand, EDM is devoted to the development of strategies for incorporating unique data into educational systems. These techniques(White Box and Black Box), on the other hand, are used to increase comprehension of educational phenomena, entities associated with them, and the contexts in which they learn and behave.

Machine learning models are frequently classified as white-box or black-box, depending on how transparent the learned models are. White box models show richer algorithm characteristics than black box models within that context, and they also allow users to create algorithms by connecting algorithm-constructing elements. White-box models can be easily explained in terms of how they work, how they make predictions, and what the important determinants are. The most simple models to comprehend, these linear and monotonic models are better suited for applications like health, education, accounting, and other areas where forecasts must be unambiguous[31]. A white-box model is one whose underlying logic, functioning, and scripting stages are straightforward, making its decision-making process understandable[4].

Black box models are the kind of models that can only predict the inputs and outputs that are expected from them. Deep Neural networks and boosting algorithms are the most common examples of ML Black-Box models. Patrons of black box models can enter parameters to obtain models that assist users in finding patterns in data by applying predefined algorithms. This makes it easier to use because the algorithm's complexities are concealed from the user. Like white- and

black-box techniques [25]have produced appropriate results for many logistical considerations, but while one is appropriate for a task and produces high accuracy, the other typically produces negative outcomes.

By utilising a thorough review that includes both white- and black-box models for forecasting educator work performance in the educational domain, all of these

problems are addressed throughout this research. As a result, the following are the key contributions of this paper:

• A realistic comparison of white and black models with two different viewpoints, including their pros and cons. • From a practical standpoint, an evaluation of the most remarkable models using the white and black-box approach. • Using both techniques to conduct a questionnaire-based survey of instructor/educator data and analyze the effects.

The paper is organized as follows. Section II provides a high-level review of the Work performance challenge, its relevance, and the historical work done in terms of applying machine learning white box and black box models to tackle it. Section III shows the working methodology. Section IV describes the experimental design in terms of dataset features, pre-processing, and the measures used to compare accuracy. The study's findings and the accompanying discussion are presented in Section V. The study closes with a recommendation for the best machine learning model in the educational area in Section VI.

## 2. Literature Review

Predicting the performance of entities within educational settings, particularly in college and university contexts, stands as a significant challenge in educational data mining. Both offline and online learning environments can greatly benefit from accurate predictions in this regard. White and black machine learning models offer diverse methods to address this challenge, yet the literature highlights research gaps, particularly regarding black-box approaches. There are various methods under white and black machine learning models for achieving this goal, and hence the literature has examined the research gaps as well. Black-box approaches[24]- [26]have received less attention from researchers throughout the last few years of study on regression test efficiency[11], sentimental categorization[2], and performance prediction. Black-box approaches have recently made strides in encouraging diversity in test cases[27], with results being published for test case generation. In rigorous experimental research, these approaches haven't been compared to each other or to more established white-box methods. As a result, it is undetermined that how well the BB models perform when compared to one another and to the more established WB alternatives. The proposed study examines various machine learning models that higher education institutions can use. Through an analysis of both white and black box model effectiveness in predicting faculty performance, the study seeks to provide insights into their comparative efficacy. Table 1 summarizes findings from existing research, offering a comprehensive overview of the performance of different machine learning models in.

Table 1. A summarization of various machine learning models based on analysing the performance and behaviour of faculty.

| OBJECTIVE | PARAMETERS | ALGORITHM AND ACCURACY | RESEARCH GAPS |
|---|---|---|---|
| To offer instructors feedback to aid and improve their performance. [6] | Time, subject, Laboratory interaction, teaching methods, Class Control , and helping attitude, | Naïve Bayes- 95.9% Logistic Regression 97% Decision Tree - 97% SVM- 95.9% NNET- 95.9% | For a better assessment, parameters such as course completion and student involvement should be included. Only the student's perspective and the result are used to evaluate faculty performance. |
| Assess teacher performance quality of teaching[19] | Topic dictionary as input and attention mechanism | Precision rate - 80% Recall rate – 79% F1 – 79% | Analyzing quantitative does not help get better results |
| Analysis of how peer assessment improves academic performance[14] | Providing task-relevant feedback, traditional outcomes, and practical skills. | g = 0.31, SE = .06, 95% CI = .18 to .44, p < .001 | Make a distinction between being an assessee vs the outcome of an assessment. |
| Academic qualifications and experience have an impact on performance. The disparity between faculty members who are NET qualified and those who are not[30] | Faculty profile, Student feedback. | F1-score=0.94 Precision = 0.93 Recall = 0.94 | More factors need to be analyzed before giving faculty feedback. |
| Flipped classroom teaching[1] | Quiz, videos, and out-of-the-class learning activities. | - | More concrete ways are needed to decide the efficacy of flipped classrooms |
| Effective teaching via classroom management and organization[21] | Teacher's education, care, management, planning, implementation of instructions | Residual mean score, M=133.49 | Student time spent on self-study should also be considered |
| Recognize that study application strategy, black box or white box, is preferable[20]. | The model provides more accuracy by combining decision trees and CNN with white and black box techniques. | - | For better results, a good way to analyze which model to apply is required. Fusion should be done only if necessary. |

The challenge of predicting performance in higher education settings underscores the need for effective educational data mining techniques. Both offline and online learning environments can significantly benefit from accurate performance predictions. While white-box and black-box machine learning models offer various methods to addressthis challenge, there are notable research gaps, especially regarding black-box approaches. Studies have shown that black-box models, despite their potential, have received less attention in areas such as regression test efficiency, sentiment categorization, and performance prediction [25][11] [2].

The proposed study aims to address these gaps by examining the effectiveness of both white-box and black- box machine

learning models in predicting faculty performance in higher education institutions. as shown in Table1. Through a detailed analysis of these models, the study seeks to provide valuable insights into their relative efficacy. Table 1 in the research summary offers a comprehensive overview of existing findings, presenting a comparative analysis of different machine learning models in educational contexts. The study's thorough evaluation of both types of models aims to fill this research gap, ultimately helping higher education institutions select the most effective machine learning tools for performance prediction. The insights gained from this analysis could lead to better decision-making and enhanced educational outcomes, benefiting both educators and students. The Successive section emphasize on the related methodology.

## 3 Methodology

Throughout the research, we established a comprehensive set of questions to aid in the comprehension of meaningful information from various education providers. These education providers are evaluated on the basis of professional, societal, cognitive and tempramental qualities[3]. The dataset contains self- reports on many elements of their performance, such as professional experience and designation, publications for research information, social interaction with students and colleagues, and real-life experiences for emotional aspects. All of this personal information is obtained directly from educators via feedback.

Without the use of technology, analyzing and clarifying a huge quantity of data is a difficult task for individuals, especially in the case of organizations. Different machine learning models may be employed to define the educator's behavior based on their involvement and to discover the elements that impact their success. EDM is a field of study that relies on mining educational data to uncover intriguing patterns and knowledge in educational institutions. Given the vast amount of data collected, manual analysis and interpretation would be daunting, particularly for organizations. The research proposes the use of machine learning (ML) models to analyze educator behavior and identify factors influencing their success. Educational Data Mining (EDM) serves as the overarching framework, leveraging data mining techniques to uncover patterns and insights within educational settings. The methodology outlined in Figure 1 illustrates how ML models can effectively analyze the collected data. This involves preprocessing the data to ensure quality and consistency, selecting appropriate ML algorithms, and training these models using the gathered dataset. The trained models are then used to predict and understand educator performance based on various input variables. By employing ML techniques within the EDM framework, the research aims to provide actionable insights into educator performance. This approach allows for a systematic and data-driven evaluation of education providers, considering not only their professional expertise but also their societal contributions, cognitive abilities, and emotional intelligence. The research seeks to enhance our understanding of factors contributing to educator success and inform strategies for improving educational outcomes.

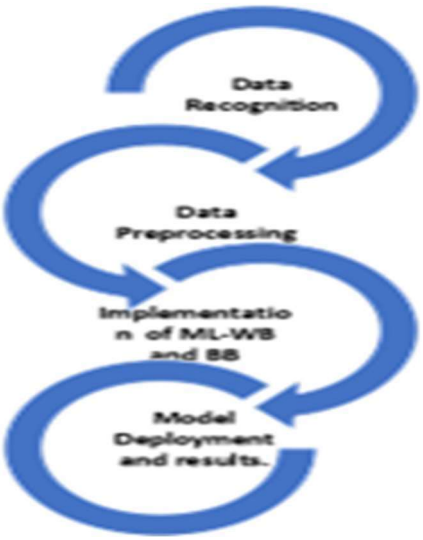



Figure 1: Working Methodology of WB and BB model in ML

### 3.1 Data Recognition

The section emphasise the importance it is to thoroughly assess teachers' performance along a range of dimensions in order to develop a comprehensive picture of their efficacy. The assessment framework takes into account factors such as professional expertise, cognitive abilities, societal contributions, and temperamental behaviour in recognition of the complex nature of educator roles. This method recognises that social interactions, research projects, emotional intelligence, and subject matter expertise are all important components of effective teaching. In order to support this thorough assessment, a four-tier analysis methodology is suggested, incorporating knowledge from both seasoned researchers and previously published studies. Every phase of the analysis explores distinct facets of the behaviour of educators, such as their academic accomplishments,social engagement, professional competency, and emotional intelligence. The importance of educators' capacity to comprehend and regulate their own feelings and relationships is underscored by the inclusion of emotional intelligence as a fundamental element. This capacity is essential for cultivating a positive learning environment. This study supports earlier research that highlights the value of emotional intelligence in educational settings, such as that done by Jais [17]. This study attempts to provide a nuanced understanding of educators' performance by using a multidimensional approach to educator assessment. This will enable targeted interventions and professional development strategies.

### 3.2 Data Preprocessing

1700 educators from different institutions and ranks participated in the research, and their contributions resulted

in a dataset of 1598 documents that was used for analysis. Following examination of the data, which was provided in an Excel spreadsheet format, 447 submissions were rated as ”Excellent,” 709 as ”Good,” and 422 as ”Need improvement” in the performance status categories. Statistical analysis and visual aids like charts and schematics were used to obtain a complete understanding of the dataset. Prior to diving into more complex data mining models and algorithms, this phase of data exploration is essential for helping writers and researchers understand the subtleties of the dataset. Finding patterns and trends in the data can be rendered easier by data analysis, and communicating insights in an understandable way is made easier by visualisation techniques. Researchers can create a strong basis for further analyses and model development by thoroughly analysing the dataset using statistical analysis and visualisation. This preliminary exploration stage is an essential prelude to more complex data mining activities, guaranteeing that further analyses are founded on a comprehensive comprehension of the features and intricacies of the dataset.1700 educators from different institutions and ranks participated in the study, and their contributions resulted in a dataset of 1598 documents that was used for analysis. After careful examination, 447 submissions were rated as ”Excellent,” 709 as ”Good,” and 422 as ”Need improvement” in performance. The data was provided in an Excel spreadsheet format.Statistical analysis and visualisation tools like charts and schematics were used to obtain a complete understanding of the dataset. Prior to diving into more complex data mining models and algorithms, this phase of data exploration is essential for helping writers and researchers understand the subtleties of the dataset. Finding patterns and trends in the data is made easier by statistical analysis, and communicating insights in an understandable way is made easier by visualisation techniques. Researchers can create a strong basis for further analyses and model development by thoroughly analysing the dataset using statistical analysis and visualisation. This initial investigation stage is an essential prelude to more complex data mining activities, guaranteeing that further analyses are founded on a comprehensive comprehension of the features and intricacies of the dataset.

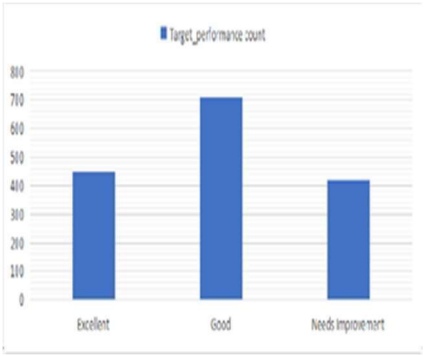



Figure 2: Count measurements of different target variables.

A model’s reliability needs to be confirmed before it is put into practice, and this is usually done by using the Cronbach alpha method. This approach works especially well with questionnaires that use Likert scales. After the dataset was examined for reliability in this context, the Cronbach alpha value was found to be 0.74. This number shows that the dataset has a respectable degree of reliability, indicating that it can be successfully used for additional research and model implementation. The internal consistency of a collection of items or variables is assessed using the Cronbach alpha technique, which yields a measure of the group’s degree of relatedness. The dataset’s items appear to be sufficiently consistent, as indicated by a value of 0.74, which offers a trustworthy foundation for further analysis and decision-making.

## 4 Implementation of ML models

On the dataset, the white box algorithms ID3, CART, and the black box algorithms XGB and MLP were employed.

**ID3:** Information entropy is the basis of the ID3 algorithm, a classification method. Classifying samples into discrete groups according to varying values within a set of condition attributes is its fundamental idea. Finding the best classification attribute among these condition attribute sets is the main goal of the algorithm. Until all samples within a branch are assigned to the same category, this process iterates across each branch, creating new nodes and branches. ID3 efficiently ascertains the characteristic that most effectively divides the data into homogeneous groups by utilising information entropy, thereby enabling precise classification. The algorithm is able to recursively partition the dataset thanks to this iterative approach, and in the end, it creates a decision tree that illustrates the hierarchical relationships between attributes and the classes that correspond with them. Through its systematic classification process, ID3 provides a powerful tool for analyzing and understanding complex datasets, aiding in various fields such as machine learning, data mining, and pattern recognition. Branching can be done based on the attribute values. The splitting attributes are chosen using the concepts of entropy and information gain[18].

**CART:** Prediction instability in regression models can occasionally be brought on by variables such as modestly variable predictors or insignificant variables. Machine learning (ML) techniques, like ensembles of trees, work well in these situations. To overcome these challenges, the study used the Classification and Regression Trees (CART) method, which was implemented using a Jupyter notebook. By utilising ensemble techniques, CART’s algorithm is especially skilled at managing prediction instability, improving the accuracy and robustness of regression models in situations where more conventional approaches might not hold up.

**XGB:** Extreme Gradient Boosting (XGBoost) belongs to the supervised branch of Machine Learning and operates as a tree-based technique. While it can effectively handle both classification and regression tasks, its strength lies primarily in classification. XGBoost addresses the challenge of creating highly accurate prediction rules by leveraging boosting, a concept introduced by Schapire in 1997[32]. Boosting involves amalgamating rough and somewhat imprecise rules of thumb to generate a remarkably precise prediction rule. XGBoost employs gradient boosting, iteratively refining the model's performance by minimizing errors. What distinguishes XGBoost is its utilization of a more regularized model formalization, which effectively mitigates overfitting—a common issue in machine learning models. This regularization ensures that the model generalizes well to unseen data, thereby enhancing its performance and robustness. Consequently, XGBoost consistently outperforms other gradient-boosted machines, making it the preferred choice for various classification tasks. Its capacity to strike a balance between accuracy and generalization renders XGBoost a formidable asset in the toolkit of machine learning practitioners, contributing significantly to advancements in predictive modeling and data analysis.

**MLP:** : The multi-layer perceptron (MLP) consists of an input layer, hidden layer(s), and an output layer. Input signals are received and processed in the input layer, while prediction and categorization tasks are managed by the output layer. Hidden layers, positioned between the input and output layers, utilize neurons as their computational units. Training of neurons within the MLP is facilitated by the backpropagation learning algorithm. MLPs are engineered to approximate continuous functions and address non-linear problems effectively. In the research context, the primary focus was on evaluating the accuracy and other pertinent metrics of the MLP. By leveraging its capability to approximate complex functions and tackle non-linear challenges, the MLP emerges as a versatile tool for various prediction and classification tasks, providing researchers with a robust framework to analyze diverse datasets.

**Model Deployment (the implementation phase) and Results:** Model deployment, a critical phase in Educational Data Mining (EDM), involves applying previously acquired knowledge to new input. Throughout the construction process, development and deployment models are refined iteratively to meet the desired objectives. Performance evaluation typically categorizes performance into three levels: exceptional, good, and requiring improvement, with exceptional performance denoted as "Excellent." The model considers four independent parameters (P1, P2, P3, and P4) and assigns three values to the goal class: 0 for exceptional, 1 for good, and 2 for poor (indicating a need for improvement). Using self-reported data from educators acquired through feedback, the study sought to thoroughly evaluate education providers based on professional, sociocultural, cognitive, and temperamental traits. After their performance was assessed using a dataset of 1598 documents from 1700 participating educators, 447 entries were scored as "Excellent," 709 as "Good," and 422 as "Need improvement." The idea is to use machine learning (ML) models within the framework of Educational Data Mining (EDM) to find patterns and insights because it would be difficult to manually analyse such a large dataset.

Prior to examining intricate data mining models and algorithms, it was essential to conduct data exploration using statistical analysis and visualisation equipment. This stage enables it easier to comprehend the nuances of the dataset, spot trends, and enable additional analysis. Figure 1 illustrates the methodology, involving preprocessing the data to guarantee its quality and consistency, choosing suitable machine learning algorithms, and training these models on the dataset.A key component of analysing educator behaviour and determining the variables affecting their success is the use of machine learning models.

Through the application of machine learning algorithms, insights can be obtained from a variety of aspects of the performance of educators, including publications, professional experience, social contacts, and real-world experiences. The last objective is to make sense of and comprehend relevant data from the dataset.Charts and schematics are instances of visual aids that are necessary for effectively presenting insights. By thoroughly analysing and visualising datasets, researchers may create a solid basis for further investigations and model building. This initial investigation guarantees that subsequent analyses are based on a thorough comprehension of the features and nuances of the dataset. Educators' self- reported data is gathered, their performance is assessed, and machine learning models are applied within the EDM framework to find patterns and insights. To fully comprehend the dataset, statistical analysis and visualisation techniques are used, which makes it easier to conduct further investigations and construct models. To assess the efficacy of white box versus black box models, various metrics are utilized. These metrics include the confusion matrix, which offers insights into classification performance, as well as measures like accuracy, root mean square error (RMSE), Kappa score, precision, recall, and F1-score. Accuracy gauges overall correctness, while RMSE quantifies average prediction error. The Kappa score evaluates agreement between predicted and actual classifications, considering random agreement. Precision assesses the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives correctly identified. Lastly, the F1-score provides a balanced evaluation of precision and recall. These metrics collectively inform the comparison between white box and black box models, guiding decision-making in model selection and deployment within the EDM framework

## 5 Results and Discussion

The suggested model was built using a Jupyter notebook and four different machine-learning algorithms: ID3, CART,

XGB, and MLP. Several performance measures were used in order to compare the models' performances. Of these metrics, accuracy is particularly important since it shows the percentage of correct forecasts compared to all possible projections. The measure of accuracy offers valuable information about the model's overall predictive accuracy, facilitating an evaluation of its efficacy in identifying the underlying patterns present in the data. Researchers can obtain important insights into the relative advantages and disadvantages of each algorithm by using accuracy as a performance metric. This allows for well-informed decision-making during the model selection and deployment processes.The following performance measures were used to conduct a comparative examination of the models' performance.

**Accuracy:** Accuracy is a pivotal metric in evaluating theefficacy of machine learning algorithms, measuring the model's ability to make correct predictions among all possible outcomes. It represents the proportion of accurate predictions generated by the model relative to the total predictions made. In Educational Data Mining (EDM) contexts, accuracy plays a crucial role in assessing the effectiveness of algorithms in predicting student outcomes or classifying educational data. A higher accuracy score indicates greater reliability in the model's predictions, while a lower score suggests potential inaccuracies. However, accuracy alone may not provide a comprehensive understanding of a model's performance, particularly in instances of imbalanced class distributions or when certain prediction errors hold greater significance than others. Therefore, supplementary metrics like precision, recall,F1-score, and AUC-ROC are often employed to offer a morenuanced evaluation. Ultimately, accuracy serves as a valuable guiding factor for educators and researchers in selecting and deploying machine learning models within educational settings, aiding in the improvement of teaching methodologies and student learning outcomes

$$\text{Accuracy} = \frac{\text{Total Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

The **kappa statistic:** The kappa statistic holds significance within machine learning algorithms as it provides a metric for assessing agreement between predicted and observed classifications, while accounting for chance agreement. It is particularly valuable in contexts involving multiple raters or classifiers, offering insights into inter-rater reliability. By quantifying agreement beyond chance levels, kappa illuminates the robustness of classification models. Employed to evaluate inter-rater reliability, the kappa statistic reflects the alignment between collected data and measured variables. In machine learning, a higher kappa value signifies stronger agreement between predicted and actual classifications, indicating enhanced reliability in predictions. In Educational Data Mining (EDM), kappa statistics play a pivotal role in assessing the consistency between predicted student outcomes and observed performance metrics. Educators and researchers utilize kappa to evaluate the reliability of predictive models in categorizing students into performance groups. A high kappa value in EDM suggests close alignment between model predictions and actual student performance, bolstering confidence in the model's utility for educational decision-making and interventions.

$$\text{Accuracy} = \frac{\text{Total Correct Predictions}}{\text{Total Number of Predictions}} \quad (2)$$

where $k$ stands for Cohen's Kappa.

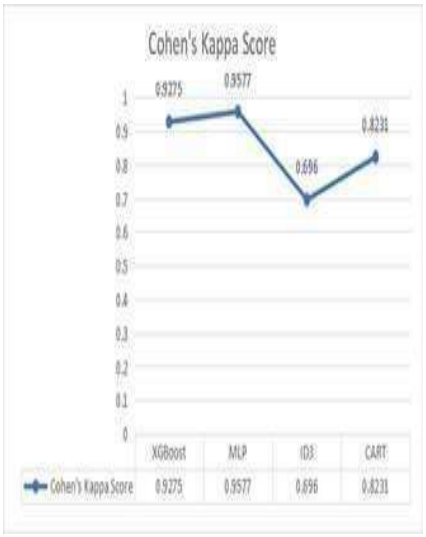



Figure 3: Cohen's Kappa score variation w.r.t. White and Black Box models

Precision and recall serve as fundamental metrics for evaluating classification model performance. Precision reflects the accuracy of positive predictions by measuring the proportion of true positives (TP) among all instances predicted as positive. It is calculated as TP divided by the sum of TP and false positives (FP). Conversely, recall, also known as sensitivity, gauges the model's ability to correctly identify all positive instancesin the dataset. It is computed as TP divided by the sum of TP and false negatives (FN). Precision and recall offer complementary insights into a model's effectiveness, with high precision indicating few false positive errors and high recall indicating comprehensive capture of positive instances. Balancing precision and recall is crucial as improving one metric may compromise the other. Achieving this balance is essential for developing robust classification models capable of accurately identifying positive instances while minimizing false positives and false negatives.

Precision is computed as TP divided by the sum of True Psotivie and False Positives. Recall is computed as TP divided by the sum ofTP and false negatives (FN).

$$\text{Precision} = \frac{\text{Total Positives}}{\text{Total Positives} + \text{False Positives}} \quad (3)$$

$$\text{Recall} = \frac{\text{Total Positives}}{\text{Total Positives} + \text{False Negatives}} \quad (4)$$

Table 2 depicts that-

• Performance metrics accuracy indicates a higher value for XGB and MLP as compared to ID3 and cart, indicating that black-box models are more effective to work with to employ than white-box models.

• White-box models provide prediction findings along with influencing factors, making prediction fully explicable, in contrast to black-box models.

• In the case of black box models, precision and recall provide more precise values.

• The mean of the all each (Precision and Recall) values is used to get the weighted-average based F1 score, which is highest for MLP with values of 0.97 and 0.97.

**Table2.** Performance measures used in WB and BB models

| Algorithm | Accuracy | Cohen's kappa score | Confusion Matrix | Measurement Metrics |
|---|---|---|---|---|
| **XGBoost (Gradient BoostedDecision Tree)** | 95.36% | 0.9275 | [[128 2 0]<br>[ 3 210 10]<br>[ 1 6 114]] | precision recall<br>0 0.97 0.98<br>1 0.96 0.94<br>2 0.92 0.94<br>accuracy<br>macro avg 0.95 0.96<br>weighted avg 0.95 0.95 |
| **MLP** | 97% | 0.9577 | [[130 0 1]<br>[ 0 204 5]<br>[ 1 6 127]] | precision recall<br>0 0.99 0.99<br>1 0.97 0.98<br>2 0.95 0.95<br>accuracy<br>macro avg 0.97 0.97<br>weighted avg 0.97 0.97 |
| **ID3** | 80.379% | 0.6960 | [[117 14 3]<br>[ 5 182 3]<br>[ 1 67 82]] | precision recall<br>0 0.95 0.87<br>1 0.69 0.96<br>2 0.93 0.55<br>accuracy<br>macro avg 0.86 0.79<br>weighted avg 0.84 0.80 |
| **CART** | 88.396% | 0.8231 | [[117 14 3]<br>[ 5 171 14]<br>[ 1 18 131]] | precision recall<br>0 0.95 0.87<br>1 0.84 0.90<br>2 0.89 0.87<br>accuracy<br>macro avg 0.89 0.88<br>weighted avg 0.89 0.88 |

## 6 Conclusion

The study conducted a thorough examination of white-box (WB) and black-box (BB) models within the realm of educational data mining, focusing particularly on managing educator performance. By utilizing environments closely resembling educator datasets, the study aimed to evaluate these models' efficacy in providing feedback on educator performance. Using four key variables from the PSRE study as predictors, the analysis produced promising findings. Notably, black- box models like XGBoost and MLP demonstrated superior performance, achieving accuracy scores of 95.36 percentage and 97 Percantage respectively. This highlights the superiority of black box classification algorithms over white-box models in educational data mining contexts. Additionally, MLP exhibited the highest inter- reliability with a rating of 0.9577, emphasizing its robustness in capturing the nuances of educator performance. The results also underscored the significance of various educator-related factors, such as professional, research, and emotional attributes, in driving improved performance.

**Limitations**

The significant dataset offers a strong basis for the investigation. This vast amount of data provides a foundation for finding important trends and insights that can advance the discipline. The comprehensive nature of the dataset ensures that a wide range of educational contexts are still represented, even though it might not fully capture the global diversity of educational practices and institutional types, that is one of the limitation of the research. This allows for useful generalisations and may even inspire future research with more diverse samples. The utilization of sophisticated models such as XGBoost and MLP, which are renowned for their exceptional accuracy, exemplifies the innovative methodology employed in this study. Although the fact that these models are frequently regarded as "black boxes", their outstanding success shows how powerful algorithms can be when used to produce amazing outcomes.

Furthermore, the effectiveness of these models may stimulate the creation of fresh methods and resources to improve interpretability, which could increase the educational community's comprehension and confidence in predictive modelling. The use of educators' self-reported data makes it possible for learners to obtain in-depth, first-hand information that would be challenging to get otherwise. By ensuring that the viewpoints and experiences of individuals who are actively engaged in education are recorded, this method produces insightful qualitative data that enhances quantitative results. The comprehensive feedback from educators can result in a deeper understanding of educational practices and issues, ultimately driving more effective and focused interventions, even though self- reported data can add biases.

Future Scope Looking ahead, there are opportunities for further extension in this research. By leveraging a diverse arrayof classification methods and data mining approaches, such as genetic algorithms, ensemble methods, and the Naive Bayes classifier, alongside data from additional universities, the study can deepen its exploration of personal characteristics associated with educator performance. Investigating the correlation between performance metrics and personality types holds promise for informing future applications. Such endeavors could lead to meaningful advancements in educator performance management, facilitating targeted interventions aimed at enhancing teaching efficacy and student outcomes. Overall, this study lays a solid groundwork for future research endeavors aimed at harnessing machine learning and data mining techniques to optimize educational practices and elevate educator performance. Future research should focus on expanding the dataset to include a more diverse range of educational institutions globally, ensuring a comprehensive understanding across different contexts. Additionally, efforts should be made to enhance the interoperability of black

box models without compromising their accuracy, making them more accessible and trustworthy for educators. This involves developing new techniques which gives more precise results, Finally, exploring alternative data collection methods to mitigate biases inherent in self- reported data, such as utilizing observational or automated data gathering techniques, could further strengthen the research outcomes and provide a more accurate representation of educational environments.

**References**

[1] Adawi, C. S. (2018). Flipped classroom research: From 'black box' to 'white box'evaluation. Educ. Sci, 8(1), 2-5. doi:10.3390/educsci8010022

[2] Alok Kumar, R. J. (2018). Faculty Evaluation System. Procedia Computer Science, 125, 533-541. doi:https://doi.org/10.1016/j.procs.2017.12.06 9

[3] Arora, S. A. (2022). PSRE Self-assessment Approach for Predicting the Educators' Performance Using Classification Techniques. Communications in Computer and InformationScience, 1546. doi:doi.org/10.1007/978-3-030-95711-7$_3$4

[4] B. Delibas˘ic´, M. V. (2013). White-box or black-box decision tree algorithms: Which to use in education? IEEE Trans. Educ, 56(3), 287–291. doi:doi: 10.1109/TE.2012.2217342.

[5] Bezerra, L. N. (2020). Educational Data MiningApplied to a Massive Course . International Journal of Distance Education Technologies, 18(4), 17-30. doi:10.4018/ijdet.2020100102

[6] Bhatnagar, S. (2018). Analysis of Faculty Performance Evaluation Using Classification.Int. J. Adv. Res. Comput. Sci., 9(1), 115-121. doi:10.26483/ijarcs.v9i1.5260

[7] Brown C, W. R. (2021). Teachers as educational change agents: what do we currently know? findings from a systematic review. Emerald Open Research. doi:https://doi.org/10.35241/emeraldopenres.1 4385.1

[8] Campagni, R. M. (2015). Data mining modelsfor student careers. Expert Systems with Applications, 42(13), 5508–5521. doi:doi:10.1016/j.eswa.2015.02.052

[9] Datnow, A. . (2016). Teacher capacity for andbeliefs about data-driven decision making:

A literature review of international research. Journal of Educational Change, 17(1), 7-28. [10] Datnow, A. (2020). The role of teachers in educational reform: A 20-year perspective. J Educ Change, 21, 431-441. doi:https://doi.org/10.1007/s10833-020-09372-5

[11] E. Engstr¨om, P. R. (2010). A, systematic review on regression test selection,techniques. Info. Softw. Tech, 52(1), 14–30.

[12] Emmanuel Pintelas, I. E. (n.d.). A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability. Algorithms 2020, 13(1). doi:https://doi.org/10.3390/a13010017

[13] Greller, W. E. (2014). Learning Analytics: From Theory to Practice – Data Support for Learning and Teaching. Communications in Computer and Information Science. 439. Springer. doi:https://doi.org/10.1007/978-3- 319-08657-6$_8$

[14] Hopfenbeck, J. A. (2020). The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. Educ Psychol Rev , 32, 481–509. doi: https://doi.org/10.1007/s10648-019-09510-3

[15] Inga Z˘ale˙niene˙, P. P. (2021). Higher Education For Sustainability: A Global Perspective,Geography and Sustainability. 2(2), 99-106. doi:https://doi.org/10.1016/j.geosus.2021.05.00 1.

[16] Jae Young Chung, S. L. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346-353. doi:10.1016/j.childyouth.2018.11.030

[17] Jais, M. M. (2016). Emotional Intelligence and Job Performance: A Study among Malaysian Teachers. Procedia Econ. Financ, 35, 674–682. doi:10.1016/S2212-5671(16)00083-6

[18] K. Adhatrao, A. G. (2013). Predicting Students' Performance Using ID3 And C4.5 Classification Algorithms. Int. J. Data Min. Knowl. Manag. Process, 3. doi:10.5121/ijdkp.2013.3504

[19] L. Yang, P. R. (2023). Automatic scoring of student feedback for teaching evaluation basedon aspect-level sentiment analysis. Educ Inf Technol , 28, 797–814. doi:https://doi.org/10.1007/s10639-022- 11151-z

[20] Loyola-Gonza´lez, O. (2019). Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access, 154096–154113. doi:doi: 10.1109/ACCESS.2019.2949286

[21] M. A. Mun˜oz, J. R. (n.d.). Investigating the 'black box' of effective teaching: The relationship between teachers' perception andstudent achievement in a large urban district. Educ. Assessment, Eval., 25(3), 205-230. doi:10.1007/s11092-013-9167-9

[22] Mr. B. Narendra, M. K. (2016). Sentiment Analysis on Movie Reviews: A ComparativeStudy of Machine Learning Algorithms and Open Source Technologies. I.J. Intelligent Systems and Applications, 8, 66-70. doi:10.5815/ijisa.2016.08.08

[23] Oludare Isaac Abiodun, A. J. (2018). State- of- the-art in artificial neural network applications:A survey. Heliyon, 4(11). doi:10.1016/j.heliyon.2018.e00938

24] Onan, A. (2019). Mining opinions from instructor evaluation reviews: A deep learning approach. Computer Applications in Engineering Education. doi:10.1002/cae.22179

[25] Oyebade K. Oyedotun, S. N. (2015). DataMining of Students' Performance: TurkishStudents as a Case Study. I.J. Intelligent Systems and Applications, 9, 20-27.

[26] R. Muhamedyev, K. Y. (2020). The use of machine learning "black boxes" explanation systems to improve the quality of school education. Cogent Engineering. doi:10.1080/23311916.2020.1769349

[27] Raheela Asif, g. M. (2015). Predicting Student Academic Performance at Degree Level: A Case

Study. I.J. Intelligent Systems and Applications, 49-61. doi:10.5815/ijisa.2015.01.05

[28] S. ARORA, D. R. (2020). PSE ASSESSMENT-BASED E-LEARNING: NOVEL APPROACH TOWARDS ENHANCING EDUCATIONIST PERFORMANCE. In D. M. Cosmena, New Paradig. eLearning Technol. Aris. Due To Covid-19 Cris (pp. 11-26). EPFRA.

[29] S. Arora, M. A. (2021). Comparative Analysis of Educational Job Performance Parameters for Organizational Success: A Review. Algorithms for Intelligent Systems (pp. 105-121). Springer, Singapore. doi: https://doi.org/10.1007/978- 981-15-7533-4$_9$

[30] S. Arora, R. K. (2021). An Empirical Study - The Cardinal Factors towards Recruitment of Faculty in Higher Educational Institutions using Machine Learning. IEEE (pp. 491-497). Noida, India: IEEE. doi:10.1109/SPIN52536.2021.9566057

[31] Sapna Arora, M. A. (2018). Empowerment through Big Data : Issues Challenges. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(5).

[32] Schapire, Y. F. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci., 55(1), 119–139. doi:10.1006/JCSS.1997.1504

[33] Williamson, B. (2017). Williamson Big Data Education Chapter1 2017.

## Authors

Dr. Sapna Arora currently serves as an assistant professor at the School of Computer Science within IILM University's Gurugram Campus. She obtained her PhD in Computer Science Engineering from Banasthali Vidyapith in Rajasthan,India. Dr. Arora has actively engaged in numerous Faculty Development Programs and workshops, showcasing her commitment to professional development. Her scholarly contributions include the publication of over ten research papers across various journals and conferences, along with chapters in Scopus indexed books. Her research interests span Big Data, machine learning, and data mining. With over a decade of teaching experience, Dr. Arora prioritizes student learning and understanding of their aspirations. She is recognized for her optimistic approach to learning and patient work ethic, underscoring her dedication to academic excellence and nurturing the potential of her students.

Dr. Ruchi Kawatra is a computer science and information technology practitioner with 16 years of expertise. She earned a Ph.D. in information technology from Banasthali Vidyapith, Rajasthan, India. She has published numerous papers in peer- reviewed journals and at both national and international conferences. She has also written numerous book chapters on the Internet of Things, data analytics, and machine learning. She has studied in prestigious colleges and universities in addition to her current position as Associate Professor at SRM University. Additionally, she has conducted training sessions on advanced Excel and data/statistical analytics using Python and R.

Professor Dr. Narayan C. Debnath 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 received numerous Honors and Awards. He served as the elected Chairperson of the Computer Science Department at Winona State University.

# BIG DATA VISUALIZATION IN DIGITAL MARKETPLACES – A SYSTEMATIC REVIEW AND FUTURE DIRECTIONS

Anal Kumar *
Fiji National University, Nadi, Fiji
ABM Shawkat Ali †
Fiji National University, Nadi, Fiji

June 27, 2024

**Abstract**

ABSTRACT As the digital landscape continues to evolve, digital marketplaces have become critical platforms for businesses to connect with customers and thrive in the highly competitive market. Amidst this growing complexity and influx of data, the role of big data visualization has emerged as a powerful tool for extracting meaningful insights and could also help with predictive analysis in digital marketplaces. Digital marketplaces have revolutionized the way businesses operate, creating vast streams of data generated by various transactions, customer interactions, and market dynamics. Navigating this data deluge presents a challenge, as businesses strive to uncover valuable insights that can inform strategic decision-making. Big data visualization has emerged as a powerful approach to transforming complex data into visually appealing representations that enable better understanding, analysis, and utilization of information in digital marketplaces. This paper explores the significance of big data visualization in the context of digital marketplaces. It highlights the growing importance of visualization techniques to unlock the hidden potential of massive datasets and facilitate data-driven decision-making. By employing innovative visualization tools and technologies, businesses can gain a comprehensive view of their marketplace, identify patterns, and extract actionable insights to optimize their operations. Additionally, the paper highlights the benefits of big data visualization for stakeholders involved in digital marketplaces. It emphasizes how visualization empowers decision-makers to identify emerging trends, understand customer behavior, and make data-informed strategic choices. Moreover, it addresses the collaborative aspect of visualization, enabling teams to share insights, foster innovation, and drive performance improvements across the marketplace ecosystem. This paper offers a multidisciplinary overview of the research problems and developments in big data and the tools and strategies used for its display. The primary goal is to give creative solutions for problems relating to the present state of big data visualization and highlight obstacles in visualization approaches for existing big data. Complex data visualization design projects frequently require collaboration between individuals with various visualization-related talents. For instance, many teams combine designers who produce fresh visualization concepts with engineers who put the resultant visualization software into practice. The authors pinpoint gaps that present difficulties for designer-developer teams trying to produce new data visualizations. Data for this study came from papers published between 2010 and 2022 and obtain using a comprehensive literature procedure (12 years). For this study, several publications from a variety of sources are utilized using the specified inclusion, exclusion, and quality criteria. The focus is primarily on the research regarding big data visualization in the context of digital marketplaces and the methods used for data visualization. The current study compiles and arranges the published literature on big data visualization in digital marketplaces that is currently available. The findings of this study indicate that there has been a rise in the number of papers published annually and that there are several studies on big data in digital marketplaces. The study will aid academics in understanding the research that is now accessible on big data in digital marketplaces and will ultimately be utilized as support in other investigations

**Keywords:** Big Data, Visualization, Data Visualization Tools, Digital Marketplace, and Systematic Literature Review

## 1 Introduction

In today's digitally connected world, online marketplaces have experienced a remarkable surge in popularity. These platforms bring together buyers and sellers from across the globe, facilitating transactions, and offering a wide array of products and services. Behind the scenes of these bustling digital marketplaces lies a hidden treasure trove of data, known as Big Data. The utilization of Big Data has revolutionized the way these platforms operate, enabling businesses to make data-driven decisions, enhance customer experiences, and unlock valuable insights that drive growth and success.

Big Data refers to vast volumes of structured and unstructured data that are generated at an unprecedented pace. These data sets are characterized by their variety, velocity, and

*Department of Computing Sciences Information System Reasearch Assistant Email: anal.kumar@fnu.ac.fj
†Department of Computing Sciences Information System, Email:shawkata@unifiji.ac.fj

volume, making them challenging to process and analyze using traditional methods. However, digital marketplaces have harnessed the power of Big Data to gain a competitive edge and meet the ever-evolving demands of the modern consumer. Data is now an essential component of social interactions, history, politics, science, economics, and corporate organizations. Social media platforms like Facebook, Twitter, and Instagram, where users regularly create a massive flood of diverse data (music, photographs, text, etc.), are blatant examples of this tendency [64].

In the realm of digital marketplaces, Big Data plays a pivotal role in numerous aspects of operations, ranging from inventory management to pricing strategies, personalized recommendations, fraud detection, and customer engagement. By leveraging the vast amounts of data generated by users' interactions, transactions, and behaviors, these platforms can gain deep insights into consumer preferences, market trends, and supply chain dynamics. Massive amounts of data are produced daily by businesses and social media platforms, and these data are typically represented in formats that are consistent with illogical databases: weblogs, text files, or machine code, such as geospatial data that may be gathered in different stores even outside of a business or organization [65,66,73-81]. One of the primary advantages of utilizing Big Data in digital marketplaces is the ability to enhance customer experiences. Through advanced analytics and machine learning algorithms, platforms can analyze vast amounts of customer data to create personalized recommendations, tailored marketing campaigns, and targeted promotions. This level of personalization not only increases customer satisfaction but also boosts sales and customer loyalty.

Moreover, Big Data enables digital marketplaces to optimize their pricing strategies. By analyzing historical sales data, competitor prices, and market trends, platforms can dynamically adjust prices to maximize revenue and maintain a competitive edge. These data-driven pricing strategies can lead to improved profitability, increased market share, and improved customer satisfaction. Furthermore, the application of Big Data in digital marketplaces enables effective fraud detection and prevention. By analyzing patterns and anomalies in transactional data, platforms can identify suspicious activities and potential fraud attempts in real-time[68,70]. This proactive approach helps protect both buyers and sellers, ensuring a secure and trustworthy environment for conducting business. In conclusion, Big Data has revolutionized the way digital marketplaces operate, allowing businesses to leverage vast amounts of data to make informed decisions, enhance customer experiences, and drive growth. The utilization of Big Data enables platforms to optimize pricing strategies, provide personalized recommendations, and detect fraudulent activities. As the digital marketplace landscape continues to evolve, the importance of Big Data analytics will only grow, empowering businesses to stay competitive and meet the ever-increasing expectations of the modern consumer.

The following are the main contributions of this research: To highlight the research work done from January 2010 till January 2022 in the field of visualization of big data in digital marketplaces To present a summary of the techniques used for the visualization of data in digital marketplaces To highlight the benefits of visualizations in the field of digital marketplaces with an indication of the limit of power

The paper's organization is as follows; Section 2 shows the detailed process of the research used to conduct the systematic literature review. Results, discussions, and answers to the research questions are presented in Section 3. The limitations and conclusion of the present research work are given in Section 4.

## 2 Research Method

The methodology for the research topic involves a structured approach to gather, analyze, and synthesize relevant literature and empirical evidence. Initially, a comprehensive search was conducted across electronic databases including, IEEE Xplore, Science Direct, Scopus, and Google Scholar, utilizing keywords such as "big data visualization," "digital marketplaces," "data analytics," "Digital AND "Visualizing marketplace big data" "OR" data visualization "OR" big data visualization "OR" Digital marketplaces data visualization". Selection criteria were established to include studies based on relevance, publication date, language, research methodology, and scope. Following the selection process, a systematic framework for data extraction was implemented to gather pertinent information such as research objectives, methodologies, key findings, visualization techniques, and proposed future directions from the selected literature. Quality assessment tools were employed to evaluate the rigor and credibility of the included studies, ensuring the validity and reliability of the systematic review. Thematic analysis techniques was then applied to identify common themes, patterns, and trends across the synthesized literature, while also pinpointing gaps and limitations that warrant further investigation.

Furthermore, the documentation and reporting phase adhered to established guidelines such as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) to ensure transparency and reproducibility. By following these methodological steps, the systematic review aims to provide a comprehensive overview of the current state of research on big data visualization in digital marketplaces, offering insights into future research directions and practical implications for stakeholders in the field.

## 3 Research Questions

The following are the key research questions identified for conducting the proposed study:

RQ 1. What research has been done from January 2010 till January 2022 in the field of visualization of big data in Digital Marketplaces?

RQ 2. What techniques are used for the visualization of data in Digital Marketplaces?

RQ 3. What are the benefits of visualizations in Digital Marketplaces?

## 4 Search Strategy

A well-formulated search process devises it promising to thoroughly execute the resources available to identify all the associated existing studies that meet the defined search criteria. To maintain the standard of systematic literature review and conduct this study a proper search process has been done to identify the related materials which are published in the given well-reputed libraries. The proposed study uses keywords related to digital marketplaces' big data visualization based on the research questions. The following are the libraries that were used for the search process of the defined keywords for the related studies to the proposed research;

a. ScienceDirect
b. Taylor and Francis Online
c. IEEE Explore

## 5 Search String

Initially, we decided to choose the libraries and appropriate keywords related to the present research. The scope of the searched terms was defined to be in the range of the current research. The keywords defined include ("Digital) AND ("Visualizing marketplace big data" "OR" data visualization "OR" big data visualization "OR" Digital marketplaces data visualization".

The formulation and confirmation of the key search phrases then took place through the use of the information and detail gleaned from the sources based on the keywords. Then, these terms were adjusted because different sources have different search syntax. Figure 2 displays the details of the phrase that was searched for as well as the results. Journal articles, book chapters, books, conference proceedings, and other online resources are among the information gathered from many sources. Table 3 displays the complete list of the articles received. While Figure 2 shows the original, filtered by title, filtered by abstract, and filtered by content. The phases of the search are depicted in Fig 3.

## 6 Publication Inclusion and Exclusion Criteria

Numerous journal articles, books, conferences, seminars, and other published resources were discovered throughout the search process. The pre-defined keywords were manually searched in each of the aforementioned libraries. The Endnote reference management program was used to manage the necessary references and bibliographic data [71]. The bibliographic data in the Endnote library consists of the author's name, the title of the article, the name of the conference or journal, the year the piece was published, and the page numbers of that particular article. Figure 2 depicts the specifics of the general search procedure carried out by the specified keywords in the available libraries. The initial search, inclusion and exclusion, and filters by title, abstract, and full text are all included in this.

The authors decided to include the paper with the following inclusion criteria shown in Table 1. The authors decided to exclude the papers with the following exclusion criteria shown in Table 2. Figure 4 shows the initial results obtained from the search process of the proposed research. The study selection process in the proposed research was performed in different stages. Initially, the authors reviewed the articles' titles based on the defined criteria of inclusion and exclusion. The exclusion criteria were used to exclude the papers that weren't pertinent. Following that, the articles were screened by reading the abstracts, which led to the exclusion of several publications that were irrelevant to the stated research topics. The list of papers that were chosen based on the inclusion criteria is shown in Table 3. Only the papers that met the specified inclusion and exclusion criteria were chosen during the procedure [72]. Table 3 lists the chosen papers, titles, and citations. According to the trend in Table 4, there is a year-by-year increase in research and articles, indicating the field's growing importance and applicability. The quantity of papers in the chosen year range is shown in Figure 5.

*Table 1. Overview of the Inclusion Criteria.*

| Table 1. Inclusion Criteria |
| --- |
| The papers were published between Jan 2010 – Jan 2022 |
| The full content of the article is available |
| The papers were in English |
| The paper gives details about the use and application of Big Data Visualisation in Digital marketplaces. |
| The article exists in the databases defined in the search Strategy |
| The paper provides the background which is used to answer the research questions. |

*Table 2. Overview of the Exclusion Criteria.*

| Table 2. Exclusion Criteria |
| --- |
| The papers were not in the range of Jan 2010 – Jan 2022 |
| The full content of the article is not available |
| Several same versions of the paper |
| Not in English |
| Not existing in the databases defined in the search Strategy |
| Not associated with the defined research questions. |

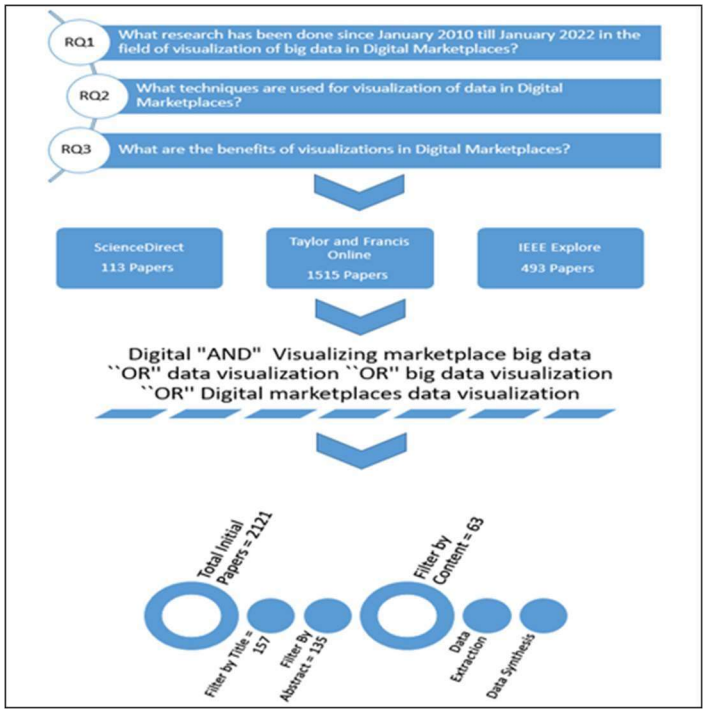



Figure 1: Protocol phases designed to carry out SLR

## 7 Systematic Literature Review

Critically analyzing the graphical presentation of big datasets that are generated from accumulative business transactions, digital marketplace reports, and publications in the digital marketplace [ 57]. Cota et al [62] highlighted electing perfect graphic visual static representation of visualized crucial big datasets that allows users interactive tool to control from the frontend of interaction, to keep track and relate the logic of extracted information [62]. As stated by scholars, visualization of dataset representation should be such that would produce an instinctive inner sense of understanding of specific datasets' behavior [63]. In 2016, Leung et al. [63] outlined on PyramidViz visualization accompanied by association rule mining that offers a hierarchical layout of informative datasets. Big data visualization tools evaluate the final report of taken datasets to boost business, government, and educational performances, and transactions, eliminate mistakes, and assist in a bright future [7, 51, 57]. Big data Visualization tools accordingly present descriptive information, digital materials, administrative information, and statistical data, where users can easily assume evaluated reports on the vigorous in-flows of generated datasets [8, 30, 40]. The most recent approach to visualizing big datasets formulated from mathematical models [20, 57]; namely Euclidean distances, correlation coefficient, using conditional probabilities, Gaussian functions, and applying eigenvalue decomposition to find distance represented datasets along with K-nearest neighbor (KNN) algorithms [20]. Mohammed et al. [31] examined big data visualization tools like Tableau, Qlikview, Sisense, Domo, Microsoft Power BI, Klipfolio, Plotly, Chartio, Geckoboard, Datawrapper, Infogram, Chart Blocks, D3.js, Google Charts, Fusion Charts, Chart.js, Grafana, Chartist.js, Sigmajs, Polymaps in various domains concerning the respective task.

Data is a vital component of any industry from educational to service industries which supports making wise choices within the continuous processes of industries in the digital platform [5, 6, 7, 40, 51, 14].

The rapid integration of datasets within digital transformation requires respective appropriate big data visualization tools to learn the linking patterns in taken datasets from various fields and domains [30, 51, 57, 14]. Islam and Jin [43] highlighted categories of visualizations like chartsBlock, Google Charts, Infogram, and Datawrapper to visualize hidden patterns and movements and chances for later scrutiny [43]. Visualizing big datasets enables data to be utilized most efficiently, swiftly outlines reports, and enables users to absorb information seamlessly [44]. In 2021, Ishika and Mittal [44] reviewed approaches to information visualization of massive datasets. To represent digestible reports, Ishika and Mittal highlighted tree maps, circle packing, parallel coordinates, and stream graph methods [44]. According to [1,15,40], projected multiples of applications and tools to collect information for visualization of data in the learning process within an open digital platform. Namely of multiple information sources for big data visualization purposes is enterprise content management system, online analytical processing tools like IBM Cognos, Oracle OLAP, and Oracle Essbase, enterprise architecture tool, decision tree techniques, neural network techniques, Na¨ıve Bayes techniques, K- Nearest neighbor techniques within educational data mining tool [1, 15] for efficiently providing

Table 3. Evaluation of common visualization tools.

| Name | Usage | Software category | Visualization structure | O.S. | License | Scalability | Extensibility |
|---|---|---|---|---|---|---|---|
| **Tableau** | Presentation | Desktop App., cloud hosted | Various Charts, graphs and maps | Windows 7 or later, OSX 10.10 or later | Commercial and Academic license | Hadoop and cloud | DBs Drivers, API for Matlab, R, Python and Javascript |
| **Infogram** | Presentation | Desktop App., cloud hosted | Charts, maps, images and even videos | Windows 7 or later, OSX 10.10 or later | Commercial and educational license | Cloud | API for Matlab, R, Python and Javascript |
| **QlikView** | Presentation | Desktop App., cloud hosted | Various Charts, graphs and maps | Windows 7 or later, OSX | Commercial | Hadoop and cloud | API for Matlab, R, Python and Javascript |
| **Plottly** | Presentation + Developers | Web tool, JavaScript and Python library | Charts, plot and maps | Web Based | Commercial and Community | Cloud | API for Matlab, R, Python and Javascript |
| **Power BI** | Developers | Desktop App., cloud hosted | Various Charts, graphs and maps | Windows 7 or later, OSX 10.10 or later | Free, Pro, and Premium Per User | Cloud and Hadoop | DBs Drivers, API for Matlab, R, Python and Javascript |
| **Ember-charts** | Developers | JavaScript library | Charts | Web Based | Open-source | Cloud | - |
| **Google charts** | Developers | JavaScript library | Charts, tree map, timeline and gauge | Web Based | Open-source | - | e Chart Tools Datasource protocol |
| **Fusion Charts** | Developers | JavaScript library | Charts | Web Based | Commercial | - | - |
| **Chart.js** | Developers | JavaScript library | Charts | Web Based | Open-source | - | - |
| **Leaflet** | Developers | JavaScript library | Map | Web Based | Open-source | - | Extensive plugin repository |

with the valuable data for further interpretation by users along with various digital applications and devices. In the digital marketplace, big data visualization intuitionally [1] brings awareness of updated organization tasks and involved or taken transactions and that further improves in reasoning abilities of users. Big data visualization represents explicit knowledge for the viewers on a cloud-based platform to make inferences critically using tact knowledge [1] along with learning algorithms and analytical algorithms. Hybrid information infrastructure enables an understanding of the concepts for the learner and academic coordinator at a faster pace [1,15] within real-time data analytics. Big educational data visualization provides a new direction for learners, businesses, academics, and professors to learn and understand specific fields' transactions [7]. The daily production of educational data within the Internet of Things, social media platforms, learning management systems, massive open online courses, open courseware, and open educational resources [7, 15] promotes an easy and flexible way for learners to continue with education [7]. Ang et al. [7] revised techniques of visualizing educational data that is through distributed architecture, five-layered architectures, cloud-based architecture, big data architecture, and logging architecture for education that evaluates the learner's pace of learning in the digital marketplace within predictive analytics and learning analytics [7]. Meanwhile, Dai et al. examined smart big educational data which is being generated from the Internet of Things by applying the visualization software CiteSpace [13] to configure which application is best to enhance e-learning in the digital marketplace with reliability and that provides useable, useful usability. Moscoso-Zea et al. also examined descriptive analytics that focuses on historical and present datasets for visualization and reviewed the decision support system and conceptual blueprint of the institute [15] to enrich judgments, evaluations, and smooth control of systems in the e-learning platform. Consequently, educational institutes produce students data and educational-related resources big datasets from educational management information systems in the digital platform [16]. Feng et.al [16] indicated analytical discipline for big educational data visualization from educational management information systems [16] to control, monitor, and boost e-learning for academicians and learners. The pace and patterns of acquiring skills in the digital marketplace are studied using process techniques within clustering algorithms and support vector machine algorithms [16] to improve the usability of the system for both learners and administrators. The clinical complex report is publically accessible from a cloud-based platform to easily understand the course of symptoms through the Multiple Imputation Visualization-Aided Validation Index techniques in the digital marketplace [2]. It searches similar patterns of clinical data sets on the web and groups them accordingly using unsupervised clustering learning algorithms [2]. The web-generated clinical data set visualization [2], enables the public to justify and take precautionary measures of incurring similar

symptoms early on. It is the holistic treatments given to patients from different generated treatment effects data sets [2] for patients to become responsible for health issues. The patient applies tact knowledge from visualized reports to make assumptions without having background knowledge on different health-related issues [2]. According to the scholar, designing the various big data visualization tools concerning its task is the one of most challenging tasks to map with intuition reasoning [1, 2]. The web-generated clinical report data visualization aims to improve statistical formulation and reliability of data within Multiple Imputation Fuzzy grouping and authentication methods [2]. Globally, every second the academician publishes quality scientific or art articles, literature, conference papers, magazines, books, and business reports with digital databases and libraries [3]. Vigorously, researchers

Table 4: Techniques of visualizing data by researchers.

| | |
|---|---|
| **Five-layered Architectures** | Cloud-based architecture, big data architecture, and logging architecture |
| **Internet of Things** | Visualization software CiteSpace |
| **Descriptive Analytics** | Decision support system and conceptual blueprint of institute |
| **Analytical Discipline** | Big educational data visualization |

need to look into the massive generated scholarly data [3] to acquire knowledge in specified fields based on officially acceptable findings. The Big data on the writer, main terms, reference, and summary is considered for visualization using rule-based metadata extraction [3]. Researchers are vigilantly well-informed of specialist details from structured semantic profile visualization tools [3]. Systematically, enables a researcher to simulate real-time interconnection of concepts using programming visualization tools with Java scripts; D3.js, Chart.js, FusionChart, FlotChart, ZingChart and Gephi, Nodebox3, Ggplot2, Processing, JpGraph also applying non-programming visualization tool like Tableau, ICharts, Infogram, Raw Graphs, Visualize Free to visualize scholar datasets [3]. The motive is to extract the proficiency of content aligned with real-life task solutions across wide disciplines [3]. The big dataset on body scan reports along with advice and monitoring provided electronically to long distances patients through mobile devices where the body scan reports are generated from computed tomography scans and magnetic resonance imaging machines [4]. For body scan reports visualization, [4] indicated a lightweight progressive transmission algorithm that supports a full-scale report with reliability and further promotes mHealth in the digital marketplace [4]. The projected big data visualization tool [4] upholds a healthy lifestyle that supports health-related issues electronically and maintains generated body scan reports datasets confidentially over the on-demand availability of the internet [4]. Nazir et al. [6] applied a holistic approach to analyze Digital Marketplaces' big data retrieval under digital transformation that further assist academicians and practitioners in diagnosing any recent heart disorders or cardiovascular system along with data mining algorithms that determine meaningful patterns of reports. [6]. Thus counsel the heart patient accordingly through the health information system (HIS) to minimize mistakes, leftovers, and care costs [6]. Meanwhile, structured, semi-structured, and unstructured large volume of data from various sources uses advanced analytical techniques for quickly reaching conclusion and forecasting upcoming consequences along with predictive analytics algorithms and artificial neural networks algorithms [5]. In this digital age, various Businesses, Government agencies, and Educational institutes' daily transactions and sensitive information are placed on cloud-based platforms. For instance, a loss of data occurs while transferring to the cloud platform, leading to misconduct intentionally or unintentionally. While big data analytics examines rising crime datasets in the digital marketplace which enables investigators to study the pattern of linked cases, determine the primary cause of misconduct, and place control on possible upcoming crimes in a specific country or location [5]. According to scholars, placing data security measures in the digital marketplace is a major concern of study [5, 7]. Henceforth, lengthy datasets generate from Internet of Things (IoT) applications in the digital marketplace [8]. The authors [8] reviewed various data analytics like real-time analytics, off-line analytics, memory-level analytics, business intelligence analytics, and massive analytics concerning the Internet of Things applications that further empower new strategies in the digital marketplace. Marjani et al. [8] highlighted that visualization of the increasing dataset from the Internet of Things applications is quite difficult to achieve accurate outcomes. Consequently, the Internet of Things big datasets visualization tools generates poor results in terms of clarity [8] along with real-time interactive charts and pie charts, scatterplots, line graph, and bar charts that are very critical for processing and submitting a graphical presentation for assumptions. Since the integration of data sciences, [9] restructured courses for colleges and institutes to digital platforms within data analysis and visualization tools to improve the pace of learning amongst learners and facilitators.

As stated by [10], enormous smart meter data, digital image, and video data are generated from smart grid applications within Lambda architecture in the digital marketplace to maintain the continuous and quality distribution of power within smart grid applications in the digital marketplace. The real-time visualization of smart grid application usage of power data using the Hadoop tool for spark connection to Tableau and Matlab software along with data mining clustering K-means algorithms [10]. Sqed et al. [11] examined big-time usage of the two-way flow of energy and data produced from real-time automatic payment systems, real-time consumer appliances along with Internet of Things devices. The tracking and tracing of smart grid data based on occurrence, state, customer operational, business, and signal task analytics [11]. Willingness of reaching their destination on time, Bangkok residents opt for a mathematical model of transportation system which assists in real-time taxi pick-up that is generating trip time, distance, and speed data [12]. The real-time trace of movement services examined by PANICHPAPIBOON et al. [12] using the satellite-based radio navigation system. The on-demand trip, time, distance, and speed data are visualized using a visual curve fitting to forecast upcoming wanted trips, the most well-known place that the visitor visit in Bangkok [12]. The best-fit approach model improves taxi facilities and infrastructure of

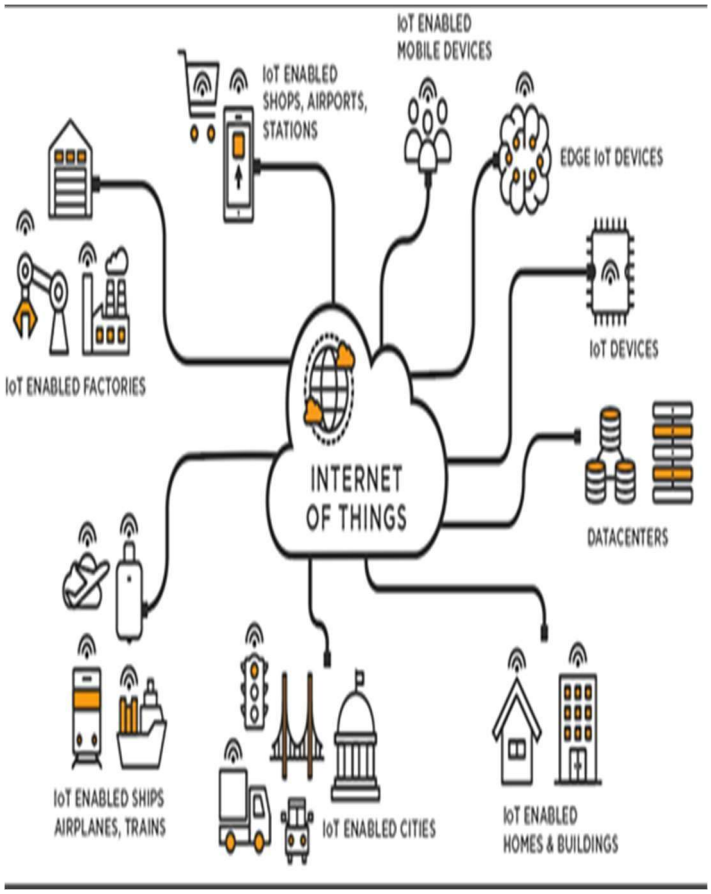



Figure 2: The illustration by Business Tech Michigan Ross shows various points of connection of the IOT.

places [12] to bring growth in economic development and gross domestic product in Bangkok. In 2018 Liao et al. examined open source code and software projects code big dataset that is generated in GitHub for software development on demand over digital platform [17]. Liao et al. reviewed the version control and Git;s big dataset visualization using the GiLA GitHub label analyzer to track source code matters in terms of recognizing the most used labels, determining maximum vigorous and expert users around each label, the time frame in which source code matters arises along with TL-IV Analysis Model [17], as a result to master and design clean software development source codes. Henceforth the authors [17] proposed C-SFS visualization, stacked Flow Visualization, and applied Sunburst Visualization techniques to trace source code matters in GitHub. The upcoming study [17] focuses on timeframe verification in organizing, preserving, and implementing the available source code dataset in web and cloud-based platforms [17]. The availability of source code in cloud-based platforms needs to be critically utilized in terms of constructing, handling, testing, and assimilating within version control systems [19]. In 2014 Liu et al. [19] tested large collections of source code datasets by applying Team Watch as the big data visualization tool for control version in software development [19] from GitHub, Git, GitLab, Apache Subversion, CVS, Mercurial, and Montone in the digital marketplace. Liu et. al [19] intend to create history attentiveness on the control version system of source code by using Seesoft, Evospace, COOP/Orm, BSCW, Xia, Augur, and Rationalizer [19] to increase competence levels and to create smooth learning for software developers in designing specific application in the digital marketplace. In everyday use of applications for prediction or evaluation in digital platforms for specific tasks, predictive analytics generates unlabeled datasets from given label datasets within Artificial intelligence [18]. In 2020 Hartono, [18] proposed new hierarchical neural network techniques for visualizing big unsupervised and supervised datasets for better presentation of predictive reports. Hartono outlined the future directions for using the novel approach of applying a soft-supervised topological autoencoder to improve the learning abilities of the learner in an e-learning platform along with a user-friendly presentation [18]. Cyber-related datasets generate every second in the Cyber marketplace between two connected nodes [21]. Jiang et al. in 2022 [21] reviewed cyber data visualization on network security analysis and malware analysis visualization to create cyber scenarios alertness and to protect business computer information systems, business computer networks, and infrastructure from offensive cyber threats or cyber risks in the digital marketplace. Lex et al. proposed on Upset techniques for general inter-linked dataset visualization which supports detecting the linkage of represented datasets within the Euler and Venn Diagram for multiple domains [22]. As stated by Kuhail et al. [23] that software developers in the digital marketplace visualize advanced programming datasets with statistical graphs by using the Uvis visualization tool that offers user interactions [23]. The graphical presentation of programming datasets efficiently improves developers' programming skills [23]. The continuous existence of events taken as well as endless three-dimensional events' relative location and path is being examined by He et al. [24] in the study of sciences, organisms, environment, living organisms, forecasting the weather, climate change, biodiversity, curing of diseases, the pace of movements. He et al. [24] recommended analytical reasoning for a multidisciplinary field that offers interactive user interfaces by using Spatiotemporal trajectory visualization techniques [24] for the duration, location, and direction of events' datasets. The motive of interactive visualization techniques is to understand an actual event's conditional behavior based on its changing value which is depending on occurrences and circumstances to make fair logistics on multidisciplinary' s event occurrences [24]. In online forum discussions, detecting unethical common interest interaction with transactions, the application's dataset is dissecting task in analytical reasoning to understand the complex user interaction. [25]. Shi et al. [25] proposed visualization techniques for a daily user to derive transaction datasets by using tracing and observing methods, investigation, and controlling movements, applying tacit knowledge, scrutinizing human-computer interaction, and refinement and identification methods to visualize in terms

of graphs, charts, manuscript, physical feathers of areas of event and symbols [25]. Visualizing the usage pattern of unethical user interaction, assist in placing controls on online network access and services from cyberattacks and defamation in internet-based social media platform [25]. Windhager et al. [26] 2019 proposed InforVis visualization techniques for the heritage of tangible and intangible heritage datasets of specific cultures in software-based online infrastructure [26]. The visualizing techniques used to preserve one's culture in digital transformation for today's generation that is to maintain integrity, manifestations of human social and professional behavior, find the linkage of one's tangible and intangible cultural heritage dataset to the other, and proficiently transform valuable skills from one generation to other [26]. Wiktorski et al. [27] studied dataset behavior on genetic utility, however, recommended physical activity monitoring datasets visualization approach within the UCI machine learning repository in the digital marketplaces field for effective results in the diagnosis process [27]. In the digital marketplaces field of visualization, Ma et al. [28] 2021 proposed K-means clustering segmentation visualization techniques to improve the decision-making process in diagnosing digital marketplaces report dataset patterns [28] in three-dimensional vision. The segmentation rule in visualizing digital marketplaces dataset reports is in terms of observing and extracting tissue's color, image, texture, and edges [28]. Zhiyuan et al. [29], 2017 reviewed the pace of movement data within Shanghai, China by using Echart.js and D3.js as the visualization tool to maintain travel feasibility, travel time, and travel convenience [29]. The travel datasets visualize in terms of connections of two nodes, which direction, particular areas, and final destination [29]. The automated system generates a continuous flow of traffic event datasets, mainly on the traveling details of travelers in Shanghai, China [29]. Liang et al. [31] proposed high-dimensional data visualization using k-means clustering algorithms, multi-source diverse visualization, time series visualization, predictive analytical system, and extensibility system to uncover the correlated patterns in specific datasets in various disciplines [31]. AN et al. [32] applied D3.js visualization techniques on film big datasets for film ranking and evaluation purposes. The film's big datasets are generated on search engines along with Python language [32]. For evolving film industries, Hu et al. [34] proposed a film expert evaluation system for visualizing film big datasets [34]. The film expert evaluation system enables producers and fans to refine ideas dynamically for later film production [34]. Ahmed et al. [33] proposed a web-based visualization platform for personal datasets generated on web-based and mobile-based applications. The web-based visualization platform provides easy access to users for task analysis of data in the form of graphical presentation [33] along with JQuery and JavaScript. Wei [35] proposed 3D electric power datasets visualization on real-time power consumption within the statistical parameters [35]. The 3D electric power visualization empowers healthy growth within digital transformation that evaluates actual time spent within the human-computer interaction process [35]. Atta et al. [36] applied real-time analysis and outlined the structure of "Spatial-Crowd" datasets that are generated on events taking place based on social media platforms [36]. Visualizing the bulk of transactions on social media platforms with real-time analysis improves seamless interaction with applications and prevents cyber deception in the digital marketplace [36]. According to Min et al. [37], big datasets proliferates within innovative technologies. The onward modes of visualizing big data in China turn to apply 3D visualization, information visualization, and research visualization in practice with echart, python, gi, Hadoop, OpenGL, and Matlab [37]. Henceforth develop big data visualization methods along with Neural Networks, Cluster Analysis, Complex Networks, and Regression Analysis algorithms [37]. Sergeevich et al. [38] proposed programs that have access to the internet linked through hypertext transfer protocol for visualizing real-time big datasets in the digital marketplace along with Advanced Data Extraction Infrastructure (ADEI) [38]. Exponentially growth in technology and data, Erraissi et al. [39] outlined big data meta-model visualizations layers to hierarchically visualize the compositional and structural datasets pattern for assumptions [39]. Raghav et al. [40] highlighted aspects of big data visualization in terms of enabling users to understand the specific course of action in different layouts with respective interactive details [40]. Hirve et al. [41] highlighted conventional visualization techniques which give the ability to understand and make an assumption based on datasets patterns generated from real-world and computer-generated content also based on stimulated-generated datasets [41]. Galletta et al. [42] examined health-related services datasets over the internet within MonogoDB by applying Geo-JSON for clear visualization of long distances in patients' digital marketplace reports [42] along with a decision support system. Big data visualization for telemedicine offers configurable services with available resources, and previous and current real-time data and maintains engagements with collaborative tools [42]. Internet content delivers ubiquitous access to information to mobile devices [45]. In 2019, Grujic´ et al. [45] examined massive telecommunication data visualization using high-level, general-purpose programming language and Quantum Geographic information system along with Application programming interfaces [45]. The stand-alone Python Quantum Geographic Information system maintains data integrity and trust components in the visualization of mobile data [45]. Fahad and Yahya highlighted visualizing structured and unstructured datasets within scientific methods, processes, and algorithms to extract and extrapolate knowledge and make insight assumptions [46]. The practice of visualizing meaningful noisy, structured, and unstructured datasets by using structured, object-oriented, and functional programming language and free software environment for statistical computing and graphics [46]. The graphical representation of extracted datasets in the form of a Bar, line chart, Box plot, Heat map, Histogram, map visualization,

Mosaic plots, and Scatter plot to diversely analyze clean visualization and assist in monitoring and forecasting structured and unstructured datasets [46, 82 - 93]. In 2022, Wang reviewed airlines' big data visualization by using high-level, general-purpose programming language including structured, object-oriented, and functional programming language to detect and trace patterns for inter-related components in flight datasets that lead to postponement of flights in the United States [47]. Leung et al. [48] highlighted visualizing Covid-19 datasets by applying sophisticated tools to scrutinize Covid-19 datasets along with visual representations of the most recent updates on infected Covid-19 countries, number of life losses, and recoveries. The visual analytics tools provide a deep understanding of global pandemic scenarios globally along with charts, graphs, and maps and which countries need more digital marketplace care attention urgently [48]. Nazir et al. [49] reviewed disorders of the heart and the cardiovascular system dataset visualization within the healthcare information system and electronic digital marketplaces record which assists practitioners in diagnosing heart-related issues. In the future looking forward to applying advanced practices to evaluate on digital marketplace big datasets [49]. Soklakova et al. [50], in 2016 highlighted on visualizing education big data by using data-driven document JavaScript along with hypertext markup language, scalable vector graphics, and cascading style sheet [50]. The D3.js enables users to control outlined education datasets with clicks and scrolls in web browsers to improve in real-time teaching and learning process for tertiary institutes [50]. Menon et al. [52], in 2021 proposed a declarative statistical visualization library within a high-level, general-purpose programming language along with GitHub for visualizing smart healthcare datasets and applied a concurrent neutral network for extrapolation. Remotely enables practitioner and patients to collaborate, control, and evaluates effectively visualized health-related datasets' hidden patterns [52]. Allaymoun et al. [53] proposed an online tool for the visualization of sales big datasets. Visualizing customizable informative and statistical reports of sales datasets over Google Data Studio as the means of cost-cutting enables businesses to determine which product or service needs more attention in the supply and demand market chain [53]. Akhir et al. [54] 2018 proposed an E-Latihan System for data visualization for evaluation form management system [54] along with a psychometric response scale. The e-Latihan system visualizes operational organizational datasets for making a wise decision in the continuous process of business transaction which brings more fecundity besides brief statements of reports [54]. Ji and Gan [55], in 2020 reviewed on visualization of scholarly publications that report original experimental and academic work in the ordinary and community real-life consequences. Ji and Gan, highlighted Citespace, CitNetExplorer, Gephi, ScienceScape, SciMAT, and VOSviewer as the visualization tools to detect the relationship between scientific literature and to forecast upcoming applications and user behavior based on previous study patterns [55]. Desai et al. [56] highlighted computational expert mechanisms to visualize educational datasets [56]. The rule-based inferences system extracts large scales of educational datasets in a highly declarative way with control approaches to make extrapolation of informed content [56]. According to Chandrasekar et al. [58], selecting the perfect fit of visualization tool like ggplot2 that best meet needs as required is a challenging task for organizations and individuals [58]. Nevertheless, Chandrasekar et al. [58] declared an open-source data visualization package that is free, easy to use, and displays general visual properties to make inferences [58].

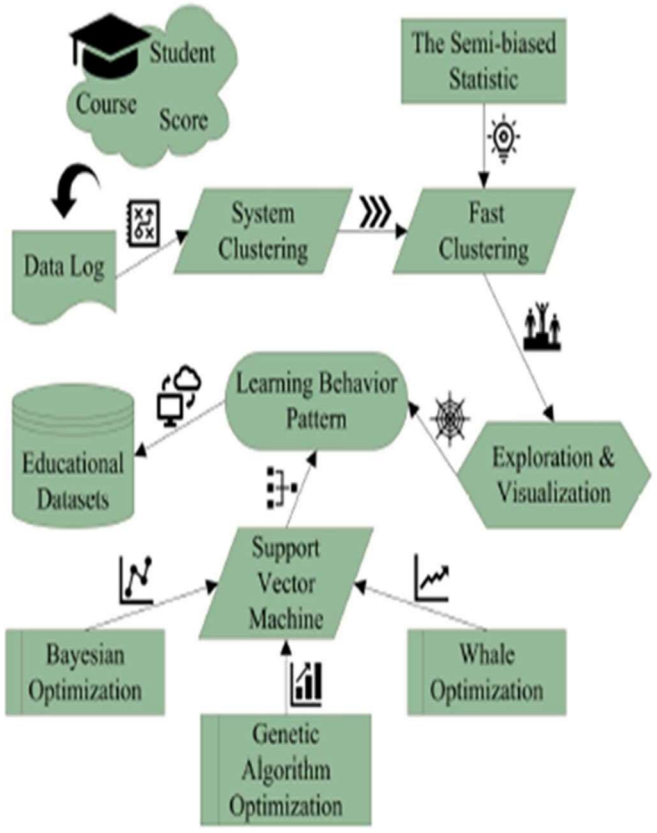



Figure 3: Educational process mining framework

According to a study by [59] visualizing datasets that are being generated over the Internet of Things is very crucial where cybercriminals can have unauthorized access to datasets. Khalid et al. [59] highlighted Decanter AI to evaluate Internet of Things datasets within semi-supervised machine learning techniques to maintain clean transactions of visualization [59]. Ordonez-Ante et al. [60] addressed on visualization of cyberbullying, cyberattacks, and cyber security threat datasets over online social media platforms [60]. Further on [60] indicated strategies of using structured design software tools within cluster computing with multiple processors and using an unbounded stream of events processing with computational application to maintain contextual state in visualizing misuse of

social media datasets at minimal delays [60]. Barik et al. [61] encompassed all aspects of proliferated moving information visualization. To match the correlation of objects, events, time, allocation, and other factors, Barkit et al. [61] suggested and reviewed White box GAT, ArcMap, GeoMesa, HadoopViz, and GRASS GIS, as the visualization tool for geospatial respective datasets [61].

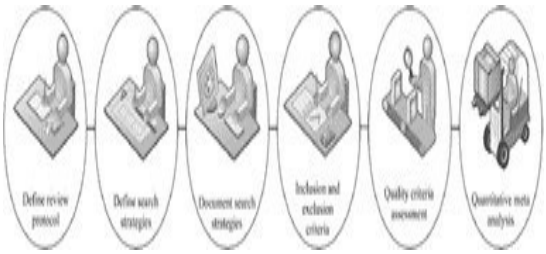



Figure 4: The figure shows the steps followed in the SLR.

## 8 F. Data Extraction

The relevant data were extracted from each of the included papers based on the review, assessment, and defined research questions. The significant data extracted is shown in different figures and tables and are briefly given as follows; • Table 3 shows all of the finally selected papers, with their titles, reference, and year of publication. • Table 4 provides a year-wise breakup of publications selected in which the number of publications is mentioned against each year. • Table 5 shows the answers to the research questions. This table shows show the details of each question and their answer with brief descriptions.

## 9 Data Synthesis

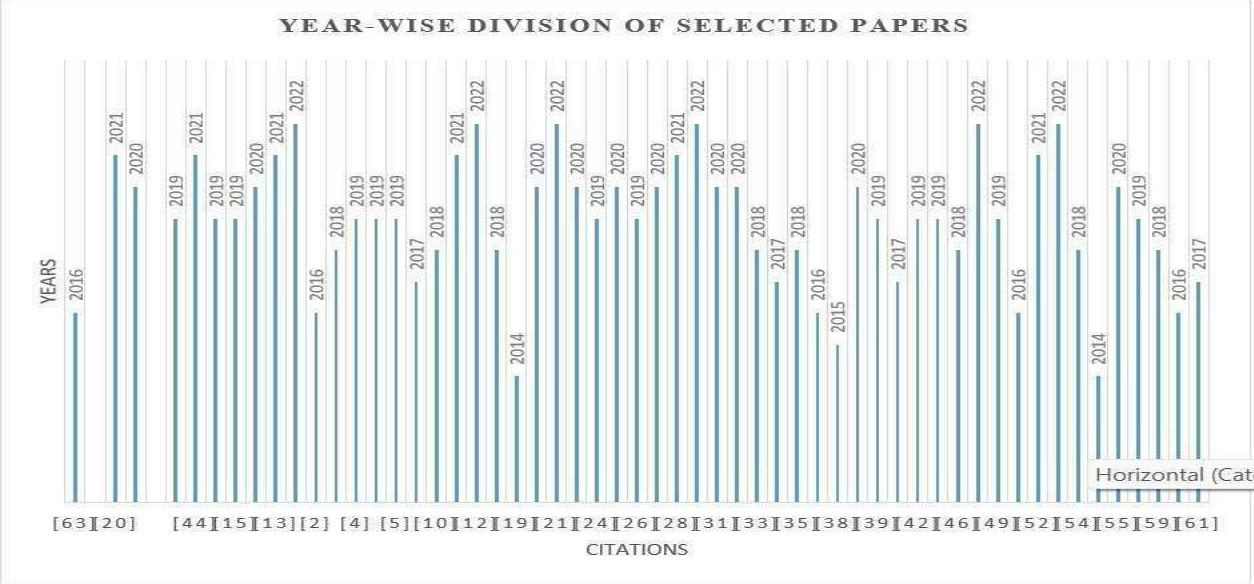
The main reviewer assisted the secondary reviewer in the data synthesis process. As a consequence, the 63-paper sample was used to produce the data extraction. These were read by the primary reviewer to compile a list of categories into which to group the success elements. The inclusion and exclusion criteria as well as the filtering procedure based on the keywords for articles are shown in Figure 2. Their names appropriately titled all of the articles and each library folder was thoroughly examined at the beginning. The duplicate papers were eliminated by examining the titles of the papers in each folder. The first selection and filtering procedure was done manually for all of the libraries, and 157 articles were found. The publications that were retrieved were then manually reviewed by abstract, and 135 articles in all were included. Finally, 63 articles were chosen after these articles underwent another manual content filtering. Each document had to be individually verified during the inclusion and exclusion

Table 3. Details of Selected Papers are listed in this table.

| S.N0 | Citation | Title | Year |
|---|---|---|---|
| 1 | [63] | PyramidViz: Visual Analytics and Big Data Visualization of Frequent Patterns | 2016 |
| 2 | [20] | Straightforward working principles behind modern data visualization approaches | 2021 |
| 3 | [31] | Data Visualization System based on Big Data Analysis", International Conference on Robots and Intelligent System | 2020 |
| 4 | [43] | An Overview of Data Visualization | 2019 |
| 5 | [44] | Big Data Analysis for Data Visualization A Review | 2021 |
| 7 | [15] | Visual analytics: A comprehensive overview | 2019 |
| 8 | [7] | Big Educational Data &; Analytics: Survey, architecture, and Challenges | 2020 |
| 9 | [13] | A comparative study of Chinese and foreign research on the Internet of Things in education: Bibliometric Analysis and visualization | 2021 |
| 10 | [16] | Exploration and visualization of learning behavior patterns from the perspective of educational process mining | 2022 |
| 11 | [2] | A New MI-Based Visualization Aided Validation Index for Mining Big Longitudinal Web Trial Data | 2016 |
| 12 | [3] | A Survey of Scholarly Data Visualization | 2018 |
| 13 | [4] | An IoT-Based Framework of Webvr Visualization for Digital Marketplaces Big Data in Connected Health | 2019 |
| 14 | [6] | Big Data Features, Applications, and Analytics in Digital Marketplaces—A Systematic Literature Review | 2019 |
| 15 | [5] | Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data | 2019 |
| 16 | [8] | Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges | 2017 |
| 17 | [10] | Data Lake Lambda Architecture for Smart Grids Big Data Analytics | 2018 |
| 18 | [11] | Smart Grid Big Data Analytics: Survey of Technologies, Techniques, and Applications | 2021 |
| 19 | [12] | Big data analysis on Urban mobility: Case of Bangkok | 2022 |
| 20 | [17] | Exploring the characteristics of issue-related behaviors in GitHub using visualization techniques | 2018 |
| 21 | [19] | Source code revision history visualization tools: Do they work and what would it take to put them to work? | 2014 |
| 22 | [18] | Mixing autoencoder with a classifier: Conceptual Data Visualization | 2020 |
| 21 | [21] | Systematic literature review on cyber situational awareness visualizations | 2022 |
| 22 | [23] | UVIS: A Formula-based end-user tool for Data Visualization | 2020 |
| 23 | [24] | Variable-based spatiotemporal trajectory data visualization illustrated | 2019 |
| 24 | [25] | Visual analytics of anomalous user behaviors: A survey | 2020 |
| 25 | [26] | Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges | 2019 |
| 26 | [27] | Visualization of Generic Utility of Sequential Patterns | 2020 |
| 27 | [28] | Visualization of Digital marketplaces Volume Data Based on Improved K-Means Clustering and Segmentation Rules | 2021 |
| 28 | [29] | Application of Big Data Visualization in Passenger Flow Analysis of Shanghai Metro Network | 2022 |
| 29 | [31] | Data Visualization System Based on Big Data Analysis | 2020 |
| 30 | [32] | Film Big Data Visualization Based on D3.js | 2020 |
| 31 | [33] | Generic Data Visualization Platform | 2018 |
| 32 | [34] | Overview of Data Visualization and Film Expert Evaluation System | 2017 |
| 33 | [35] | Research on 3D Electronic Power Big Data Visualization | 2018 |
| 34 | [36] | Spatial-Crowd: A Big Data Framework for Efficient Data Visualization | 2016 |
| 35 | [38] | Web-Application For Real-Time Big Data Visualization Of Complex Physical Experiments | 2015 |
| 36 | [37] | The Trend, Hotspots, Frontier, and Path of Big Data Visualization Research in China: Based on the Knowledge Graph Analysis of Citespace5.5.R2 | 2020 |
| 37 | [39] | A Big Data visualization layer meta-model proposition | 2019 |
| 38 | [41] | An approach toward Data Visualization based on AR principles | 2017 |
| 39 | [42] | An innovative methodology for big data visualization for telemedicine | 2019 |
| 40 | [45] | Mobile phone data visualization using Python QGIS API | 2019 |

| | | | |
|---|---|---|---|
| 41 | [46] | Big Data Visualization: Allotting by R and Python with GUI tools | 2018 |
| 42 | [47] | Big Data Visualization and analysis of various factors contributing to airline delays in the United States | 2022 |
| 43 | [49] | Big Data Visualization in Digital Marketplaces—a systematic review and future directions | 2019 |
| 44 | [50] | Big Data Visualization in Smart Cyber University," 2016 IEEE East-West Design & Test Symposium (EWDTS) | 2016 |
| 45 | [52] | Data Visualization and predictive analysis for Smart Healthcare: Tool for a Hospital | 2021 |
| 46 | [53] | Data Visualization and statistical graphics in big data analysis by Google Data Studio – sales case study | 2022 |
| 47 | [54] | Data Visualization for evaluation from the management system | 2018 |
| 48 | [56] | Data Visualization in educational datasets using a rule-based inference system | 2014 |
| 49 | [55] | Data visualization for making sense of scientific literature | 2020 |
| 50 | [58] | Deriving big data insights using data visualization techniques | 2019 |
| 51 | [59] | Exploratory Study for Data Visualization in the Internet of Things | 2018 |
| 52 | [60] | Interactive querying and data visualization for abuse detection in social network sites | 2016 |
| 53 | [61] | Investigation into the efficacy of geospatial big data visualization tools | 2017 |

Table 4: Year-wise division of selected papers




| Table 5: Description of the answers to the questions defined in | Method | Description | RQ1 | RQ2 | RQ3 |
|---|---|---|---|---|---|

| the proposed study itation | | | | | |
|---|---|---|---|---|---|
| [1] | hybrid information infrastructure, business intelligence. Educational data mining | The authors proposed visualizing, optimizing, creating, distributing, and scaling, the company's daily transactional and related resources datasets. Applied data mining algorithms to study the patterns of datasets in visualization processes. | □ | □ | □ |
| [2] | MI-based Visualization | The papers emphasize on visualization of clinical datasets which is being generated over the network. Applied MI-based Visualization aided validation index (MIVOOS) to study and evaluate the patterns of clinical report datasets. Future work aims to improve the response rate of clinical datasets' visualization. | □ | □ | □ |
| [3] | scientific visualization, information visualization, and visual analytics | The paper highlights aspects of visualization and its respective tools within a wide variety of literature generated over digital platforms. | □ | □ | □ |
| [4] | lightweight progressive transmission algorithm | The paper outlined on visualization of body scan datasets reports within digital transformation. | □ | □ | □ |
| [5] | Big Data Analytics and Mining | The authors focus on the visualization of crime datasets generated over a web-based platform. Applied state of art machine learning ad neutral algorithms in visualizing crime datasets. | □ | □ | □ |
| [6] | Analytics in Digital Marketplaces | The author emphasizes on visualization of Digital Marketplaces datasets. Applied data mining algorithms in visualizing and diagnosing Digital Marketplaces datasets in digital platforms. | □ | □ | □ |
| [7] | predict analytics and learning analytics | The authors focus on systematics analysis to visualize educational datasets and henceforth proposed advanced approaches to conclusions within specialized software systems. The paper also outlined future research on confidentiality and ethical issues of educational big datasets. | □ | □ | □ |
| [8] | Big IoT Data Analytics | The authors highlight on visualization of IoTs generated datasets. Applied interactive data visualization tools in visualizing IoT datasets. | □ | □ | □ |
| [10] | Tableau and Matlab software | The authors reviewed on visualization of datasets that are being generated from smart grid applications. Applied data mining clustering K-means algorithms in visualizing smart grid applications generated datasets. | □ | □ | □ |
| [11] | Smart grid data analytics | The paper focuses on visualizing the consumption of energy datasets that are being generated over the online platform. | □ | □ | □ |
| [12] | statistical analysis | The authors reviewed on visualization of movement datasets that are being generated over a satellite-based radio navigation system. Applied visual curve fitting for predicting the next trip of dedicated customers to Bangkok, China. | □ | □ | □ |
| | | visualization techniques along with K-mean clustering algorithms. | | | |
| [29] | Echart.js and D3.js | The authors highlighted on visualization of the pace of movement datasets that are being generated from an automatic operational system. | □ | □ | □ |
| [30] | Construction strategy of data visualization | The authors revised on series of construction strategies for data visualization over the online platform. | □ | □ | |
| [31] | high-dimensional data visualization | The authors proposed visual observation visualization techniques of specific datasets along with vector quantization. | □ | □ | □ |
| [32] | D3.js visualization techniques | The paper focuses on highlighting film datasets visualization techniques with general-purpose programming language. | □ | □ | □ |
| [33] | Web-based visualization platform | The authors intend to highlight on visualization of web-generated datasets with pre-written Javascript code, HTML, and CSS. | □ | □ | □ |
| [34] | film expert evaluation system | The authors intend to highlight on visualization of film datasets generated over a cloud-based platform within the film expert evaluation system. | □ | □ | □ |
| [35] | 3D electric power datasets visualization | The authors highlighted on visualization of energy consumption datasets using 3D electric power datasets. | □ | □ | □ |
| [36] | real-time analysis | The authors highlighted on visualization of social media-generated datasets using logic and mathematics to visualize and understand datasets patterns. | □ | □ | □ |

processes, which was exceedingly difficult. These 63 papers were managed in MS Word 2016 together with their citations. The procedure of creating the references was carried out manually because, in most cases, information is lost while getting citations from the internet. These details could take the form of the author's name, the year, the article's title, the location of publication, the page number, etc. The stated research questions in the suggested procedure for the literature review process were then applied to the chosen articles.

## 10 Results and Discussions

SLR is an established protocol used to study specific research systematically. The current research is an endeavor to study the visualization of big data in Digital Marketplaces. This section briefly describes the answers to the research questions defined below:

RQ 1. What research has been done from January 2010 till January 2022 in the visualization of big data in Digital Marketplaces? RQ 2. What techniques are used for the visualization of data in Digital Marketplaces? RQ 3. What are the benefits of visualizations in Digital Marketplaces?

Table 5 shows the description of answers for the questions defined in the section above.

In real life, processing and analyzing big data presents several difficulties. The fact that computers now represent all data visually makes it challenging to extract, see, and understand data. These activities take time, and the outcomes are not always accurate or satisfactory. Understanding human perception and finite cognition challenges is crucial for solving the visualization problems mentioned in this article. The area of design may then offer more effective and practical methods to exploit big data. Conclusion: By taking into account basic cognitive psychology principles and executing the most natural interaction with displayed virtual objects, the data visualization technique may be enhanced. Expanding it with features to eliminate blind spots and vision-reduced areas will significantly improve recognition times. Visualizing the data can considerably improve the average user's comprehension of the preselected information. The progress of visual data representation and imagery perception in the modern world is evident. Additionally, visualization software has been widely used and accessible to the general population. The authors have emphasized the extra complexity that data inherently adds to the design process based on existing literature. Adapting to late-stage data changes, foreseeing edge situations, articulating data-dependent interactions, conveying data mappings, and maintaining data mapping integrity are just a few of the data-related issues that might arise. These indicate several chances to develop tools with unique data-related capabilities

| | | | | | |
|---|---|---|---|---|---|
| [37] | 3D visualization, information visualization, and research visualization | The authors highlighted on visualization of specific datasets that are being generated over digital platforms using graphic content patterns, abstraction visualization, and applying exploratory analysis process. | | | |
| [38] | Advanced-Data Extraction Infrastructure | The authors intend to outline on visualization of real-time generated datasets using a dynamic web interface in the visualization process. | | | |
| [39] | big data metamodel visualizations layers | The authors intend to outline on visualization of relative information and SQL databases using metamodel visualization layers. | | | |
| [41] | conventional visualization techniques | The papers focus on highlighting on visualization of real-world datasets that are being generated on digital platforms using pie charts, line charts, bar charts, area charts, graphs, maps, heat maps, etc. | | | |
| [42] | Geo-JSON | The authors highlighted on visualization of heath-related datasets over the internet using an open standard geospatial data interchange format with a NoSQL database program. | | | |
| [43] | Tableau, Infogram, ChartBlocks, Datawrapper, Google Charts | The authors highlighted the Pros and Cons of data visualization tools, which would assist in selecting the perfect tool for visualizing the datasets in the digital marketplace. | | | |
| [44] | Treemap, circle packing, parallel coordinates, and stream graph method | The paper reviewed various visualization tools to visualize real-world generated datasets. | | | |
| [46] | R and python | The authors intend to highlight on visualization of quantitative datasets and qualitative datasets using R and Python programming languages. | | | |
| [47] | Big data Visualization and analysis | The authors studied on visualization of flight transaction datasets using Python programming language. | | | |
| [49] | Systematic Literature Review on cardiovascular system dataset visualization | The authors highlighted on visualization of heart-related datasets within a healthcare information system. Systematically analyzed 53 scientific-related literature. | | | |
| [50] | D3.js | The authors highlight on visualization of educational datasets along with HTML, SVG, and CSS. | | | |
| [52] | The declarative statistical visualization library | The authors proposed visualization of health-related datasets using a declarative statistical visualization library with Python programming language. | | | |
| [53] | visual analytics tool | The authors proposed visualization of sales datasets using interactive visual analytical tools within Google data studio. | | | |
| [54] | E-Latihan System | The authors proposed visualization of organizational operational datasets using the E-Latihan System. | | | |
| [55] | Big data Visualization tools | The authors proposed visualization of scholarly datasets using respective visualized tools. | | | |
| [56] | computational expert mechanism | The authors highlight on visualization of educational datasets using the computational expert mechanism. Applied rule-based inferences system to visualize educational datasets. | | | |
| [58] | open source data visualization package | The authors highlighted on visualization of datasets using an open-source data visualization package. | | | |
| [59] | Decanter AI | The authors highlighted on visualization of Internet of Things datasets along with machine learning technology algorithms. | | | |
| [60] | visual analytics tool | The authors intend to outline the visualization of cyber threat, and bullying datasets using visual analytics tools. | | | |
| [61] | Geospatial Data Visualization | The authors highlighted on visualization of geospatial datasets that are being generated over digital platforms for better use. | | | |
| [63] | PyramidViz | The authors intend to highlight an interactive visual interface tool to support analytical reasoning. Applied a set of trapezoidal and triangular pattern blocks within machine learning models to evaluate informative datasets patterns for deep understanding in visualization. | | | |

that directly assist the visualization design process. The development of these more potent tools could improve the robustness, effectiveness, and accessibility of the design process for individuals in various design positions.

Heterogeneous data exacerbate problems with data integration and big data processes. Since they demand a lot of data processing and storage space, both of them are crucial and challenging to display and analyze in large databases. The study on large data analysis for data visualization is reviewed in this publication. Additionally, it contrasts the outcomes based on various algorithms and techniques. As a result, the difficulties and techniques of the suggested approaches in related studies employing virtual reality based on big data visualization found a way to observe and analyze a variety of complicated data structures. Although several visualization strategies have been put out, some particular scholarly visualization methods are urged to be enhanced. For instance, the depiction of academic institutions has received very little attention. Another issue is that scholarly data can include a lot of information. It's still important to figure out how to use visualization tools to harvest meaningful information. The complexity of the network structure is also becoming more complicated; thus the efficacy has to be improved. How to effectively mix visualization methods with academic data analysis is another difficulty.

Visualization ideas and methodologies for academic data visual analysis are not effectively integrated into practice. These visualization techniques' data processing abilities must also be upgraded. Generally, academics obtain scholarly datasets from many online scholarly data portals. The dataset may be quite large, and researchers must pre-process this heterogeneous data (data merging, data partitioning, deleting unknown characteristics in the dataset, etc.) to match the data-input criteria of various visualization approaches.

Scholarly big data opens up new possibilities and problems for scholarly data analysis. Researchers have understood the need of using visualization tools on various datasets to better understand science. Thus, academic data visualization is

critical in resolving the issues that arise from large-volume, multivariate, and high-value data. It seems logical to focus more on this subject. In this survey work, we explore the developing field of academic data visualization to bring fresh insights. We showcase cutting-edge scholarly data visualization methodologies, with an emphasis on visualization tools and analytic systems.

Meanwhile, academic data analytic methods are being created to compile visual analyses of multivariate data from multiple perspectives (e.g., citation relationship, co-citation relationship, and co-authorship). As a result, by presenting details of the issue, this work offers a significant addition to research on academic data visualization. Despite becoming the focus of current research, several technologies still require improvement. One of the primary issues is figuring out how to efficiently combine information from complicated scholarly sources. Another problem is determining how to best mix various display approaches with analytical processing. Future research looking into these issues would be very important.

Through the findings of this systematic review, future researchers will be able to better understand current data visualization techniques and Visualize data using data science tools to Identify seasonal trends, correlation, and forecasting of customer behaviors.

## 11 IV. CONCLUSIONS

Creating graphics, diagrams, or animations to convey a message from the insight observed depends heavily on data visualization. Data visualization is the process of extracting crucial information from the data and plotting it to make decision-making simpler. This research provides a thorough report of the available literature on data visualization in the context of digital marketplaces to aid decision-makers. This research uses the SLR protocol and the data was collected from the research published from January 2010 to January 2022. Initially, a total of 1412 titles were found. Separate folders were maintained for the libraries. Each folder of the library was checked manually and their titles properly named all of the articles. The duplication of these publications was done by checking the titles in each folder. The inclusion and exclusion process was performed manually for all of the libraries by the titles and 157 articles were included. These 157 articles were then reviewed manually by abstract, and 135 articles were included. Finally, these 135 articles were reviewed by content, and 63 articles were selected. The process of exclusion and inclusion was very tricky as each of the papers was checked manually. These 63 papers along with their references were managed in MS Word. The literature on big data visualization in digital marketplaces that have been published is compiled and organized in this study. A restriction on this research was the use of only three of the libraries that were often cited. To prevent the hassle of duplicate entries and access to all of the publications, it was also decided not to use Google Scholar's keyword search function. Furthermore, because there is a lot of published research in the field, a small set of keywords was used in the search, primarily (”Digital) AND (”Visualizing marketplace big data” ”OR” data visualization ”OR” big data visualization ”OR” data visualization ”OR” Digital marketplaces data visualization” to get only related results. The suggested research will aid in the researchers' understanding of the existing research studies on the topic of big data visualization in digital marketplaces. It may eventually be utilized as support in further investigations. The findings of the suggested study indicate that there has been an increase in publications every year on big data visualization in digital marketplaces.

**Acknowledgment** The authors would like to express their sincere gratitude to Mrs. Anupriya Narayan of Fiji National University for assisting with paper collection and analysis.

**Conflict of Interest** The authors declare that no conflict of interest exists regarding this publication.

## REFERENCES

[1] O. Moscoso-Zea, J. Castro, J. Paredes-Gualtor and S. Lujan-Mora, ”A Hybrid Infrastructure of Enterprise Architecture and Business Intelligence & Analytics for Knowledge Management in Education”, IEEE Access, vol. 7, pp. 38778-38788I, 2019. Available: 10.1109/access.2019.2906343.

[2] Z. Zhang, H. Fang, and H. Wang, ”A New MI-Based Visualization Aided Validation Index for Mining Big Longitudinal Web Trial Data”, IEEE Access, vol. 4, pp. 2272-2280, 2016. Available: 10.1109/access.2016.2569074.

[3] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong, and F. Xia, ”A Survey of Scholarly Data Visualization”, IEEE Access, vol. 6, pp. 19205-19221, 2018. Available: 10.1109/access.2018.2815030.

[4] G. Xu et al., ”An IoT-Based Framework of Webvr Visualization for Digital Marketplaces Big Data in Connected Health”, IEEE Access, vol. 7, pp. 173866-173874, 2019. Available: 10.1109/access.2019.2957149.

[5] M. Feng et al., “Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data,” IEEE Access, vol. 7, pp. 106111–106123, 2019, doi 10.1109/access.2019.2930410.

[6] S. Nazir, M. Nawaz, A. Adnan, S. Shahzad, and S. Asadi, “Big Data Features, Applications, and Analytics in Digital Marketplaces—A Systematic Literature Review,” IEEE Access, vol. 7, pp. 143742–143771, 2019, doi 10.1109/access.2019.2941898.

[7] K. L.-M. Ang, F. L. Ge, and K. P. Seng, "Big Educational Data & Analytics: Survey, architecture, and Challenges IEEE Access, vol. 8, pp. 116392–116414, 2020.

[8] M. Marjani; F. Nasaruddin; A. Gani; A. Karim; I. A. T. Hashem; A. Siddiqa; I. Yaqoob "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," IEEE Access, vol. 5, pp. 5247–5261, 2017, doi: 10.1109/access.2017.2689040.

[9] X. Li et al., "Curriculum Reform in Big Data Education at Applied Technical Colleges and Universities in China," IEEE Access, vol. 7, pp. 125511–125521, 2019, doi: 10.1109/access.2019.2939196.

[10] A. A. Munshi and Y. A.-R. I. Mohamed, "Data Lake Lambda Architecture for Smart Grids Big Data Analytics," IEEE Access, vol. 6, pp. 40463–40471, 2018, doi: 10.1109/access.2018.2858256.

[11] D. Syed, A. Zainab, A. Ghrayeb, S. S. Refaat, H. Abu-Rub, and O. Bouhali, "Smart Grid Big Data Analytics: Survey of Technologies, Techniques, and Applications," IEEE Access, vol. 9, pp. 59564–59585, 2021.

[12] S. Panichpapiboon and K. Khunsri, "A big data analysis on Urban mobility: Case of Bangkok," IEEE Access, vol. 10, pp. 44400–44412, 2022.

[13] Z. Dai, Q. Zhang, X. Zhu, and L. Zhao, "A comparative study of Chinese and foreign research on the Internet of things in education: Bibliometric Analysis and visualization," IEEE Access, vol. 9, pp. 130127–130140, 2021.

[14] L. Wang, "Research on Data Visualization Information Processing Based on Computer Big Data", 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), DOI: 10.1109/TOCS53301.2021.9688643, 2021.

[15] W. Cui, "Visual analytics: A comprehensive overview," IEEE Access, vol. 7, pp. 81555–81573, 2019.

[16] G. Feng, M. Fan, and C. Ao, "Exploration and visualization of learning behavior patterns from the perspective of educational process mining," IEEE Access, vol. 10, pp. 65271–65283, 2022.

[17] Z. Liao, D. He, Z. Chen, X. Fan, Y. Zhang, and S. Liu, "Exploring the characteristics of issue-related behaviors in Git Hub using visualization techniques," IEEE Access, vol. 6, pp. 24003–24015, 2018.

[18] P. Hartono, "Mixing autoencoder with a classifier: Conceptual Data Visualization," IEEE Access, vol. 8, pp. 105301–105310, 2020.

[19] Chang Liu, Xin Ye, and En Ye, "Source code revision history visualization tools: Do they work and what would it take to put them to work?" IEEE Access, vol. 2, pp. 404–426, 2014.

[20] J. Montalvao, L. Miranda, and B. Dorizzi, "Straightforward working principles behind modern data visualization approaches," IEEE Access, vol. 9, pp. 4242–4252, 2021.

[21] L. Jiang, A. Jayatilaka, M. Nasim, M. Grobler, M. Zahedi, and M. A. Babar, "Systematic literature review on cyber situational awareness visualizations," IEEE Access, vol. 10, pp. 57525–57554, 2022.

[22] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: Visualization of intersecting sets," IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1983–1992, 2014.

[23] M. A. Kuhail and S. Lauesen, "UVIS: A Formula-based end-user tool for Data Visualization," IEEE Access, vol. 8, pp. 110264–110278, 2020.

[24] J. He, H. Chen, Y. Chen, X. Tang, and Y. Zou, "Variable-based spatiotemporal trajectory data visualization illustrated," IEEE Access, vol. 7, pp. 143646–143672, 2019.

[25] Y. Shi, Y. Liu, H. Tong, J. He, G. Yan, and N. Cao, "Visual analytics of anomalous user behaviors: A survey," IEEE Transactions on Big Data, pp. 1–1, 2020.

[26] F. Windhager et al., "Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges," IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 6, pp. 2311–2330, Jun. 2019, doi 10.1109/tvcg.2018.2830759.

[27] T. Witkowski, A. Krolak, K. Rosinska, P. Strumillo, and J. C.-W. Lin, "Visualization of Generic Utility of Sequential Patterns," IEEE Access, vol. 8, pp. 78004–78014, 2020, doi: 10.1109/access.2020.2989165.

[28] J. Ma, Y. A. Muad, and J. Chen, "Visualization of Digital Marketplaces Volume Data Based on Improved K-Means Clustering and Segmentation Rules," IEEE Access, vol. 9, pp. 100498–100512, 2021, doi: 10.1109/access.2021.3096790.

[29] H. Zhiyuan, Z. Liang, X. Liang, X. Ruithua, Z. Feng, "Application of Big Data Visualization in Passenger Flow Analysis of Shanghai Metro Network," IEEE Access, vol 2, 2017.

[30] L. T. MOHAMMED, A. A. AlHabshy, and K. A. ElDahshan, "Big Data Visualization: A Survey", International Congress on Human-Computer

Interaction, Optimization and Robotic Applications, DOI: 10.1109/HORA55278.2022.9799819, 2022.

[31] T. Liang, S. Lu, and Q. Liu, "Data Visualization System based on Big Data Analysis", International Conference on Robots and Intelligent Systems, DOI: 10.1109/ICRIS52159.2020.00027, 2020.

[32] A. Bing and A. L. Gu, "Film Big Data Visualization Based on D3.js", International Conference on Big Data and Social Sciences (ICBDSS), DOI: 10.1109/ICBDSS51270.2020.00019, 2020.

[33] A. R. Nada, S. M. Saad, and S. Abdennadher, "Generic Data Visualization Platform", International Conference Information Visualisation, DOI 10.1109/iV.2018.00020, 2018.

[34] S. Hu and J. Song, "Overview of Data Visualization and Film Expert Evaluation System", International Conference on Computer Technology Electronic and Communication (ICCTEC), DOI 10.1109/ICCTEC.2017.000, 2017.

[35] Y. Wei, "Research on 3D Electronic Power Big Data Visualization", IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), 2018.

[36] S. Atta, B. Sadiq, A. Ahmed, S. N. Saeed, E. Felemban, " Spatial-Crowd: A Big Data Framework for Efficient Data Visualization", IEEE International Conference on Big Data (Big Data), 2016.

[37] G. Min, M, Lin, Z. Li, and Y. Du, "The Trend, Hotspots, Frontier and Path of Big Data Visualization Research in China: Based on the Knowledge Graph Analysis of Citespace5.5.R2", International Conference on Culture-oriented Science & Technology (ICCST), 2020.

[38] K. A. Sergeevich, S. I. Petrovich, and A. M. Ovseevna, "Web-Application For Real-Time Big Data Visualization Of Complex Physical Experiments", International Siberian Conference on Control and Communications (SIBCON), 2015.

[39] A. Erraissi, B. Mouad, and A. Belangour, "A Big Data visualization layer meta-model proposition", 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), 2019.

[40] R.S. Raghav, S. Pothula, T. Vengattaraman, and D. Ponnurangam, "A Survey of Data Visualization Tools for Analyzing Large Volume of Data in Big Data Platform", IEEE Access, DOI: 10.1109/CESYS.2016.7889976, 2016.

[41] S. A. Hirve, A. Kunjir, B. Shaikh and K. Shah, "An approach towards Data Visualization based on AR principles", International Conference on Big Data Analytics and Computational Intelligence (ICBDACI), DOI: 10.1109/ICBDACI.2017.8070822, 2017.

[42] A. Galletta, L. Carnevale, A. Bramanti, and M. Fazio, "An innovative methodology for big data visualization for telemedicine," IEEE Transactions on Industrial Informatics, vol. 15, no. 1, pp. 490–497, 2019.

[43] M. Islam and S. Jin, "An Overview of Data Visualization", IEEE Access, International Conference on Information Science and Communications Technologies (ICISCT), DOI: 10.1109/ICISCT47635.2019.9012031, 2019.

[44] Ishika and N. Mittal, "Big Data Analysis for Data Visualization A Review", IEEE Access, 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), DOI: 10.1109/ICRITO51393.2021.9596423, 2021

[45] N. Grujic, O. Novovic, S. Brdar, V. Crnojevic, and M. Govedarica, "Mobile phone data visualization using Python QGIS API," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 2019.

[46] S. K. A. Fahad and A. E. Yahya, "Big Data Visualization: Allotting by R and Python with GUI tools," 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), 2018.

[47] H. Wang, "Big Data Visualization and Analysis of various factors contributing to airline delay in the United States," 2022 International Conference on Big Data, Information and Computer Network (BDICN), 2022.

[48] C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, Y. Wen, and A. Cuzzocrea, "Big Data Visualization and visual analytics of COVID-19 Data," 2020 24th International Conference Information Visualisation (IV), 2020.

[49] S. Nazir, M. Nawaz Khan, S. Anwar, A. Adnan, S. Asadi, S. Shahzad, and S. Ali, "Big Data Visualization in Digital Marketplaces—a systematic review and future directions," IEEE Access, vol. 7, pp. 115945–115958, 2019.

[50] T. Soklakova, A. Ziarmand, and S. Osadchyieva, "Big Data Visualization in Smart Cyber University," 2016 IEEE East-West Design & Test Symposium (EWDTS), 2016.

[51] S. M. Ali, N. Gupta, G. K. Nayak, and R. K. Lenka, "Big Data Visualization: Tools and challenges," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), 2016.

[52] A. Menon, A. M. S, A. Maria Joykutty, A. Y. Av, and A. Y. Av, "Data Visualization and predictive analysis for

Smart Healthcare: Tool for a Hospital," 2021 IEEE Region 10 Symposium (TENSYMP), 2021.

[53] M. H. Allaymoun, M. Khaled, F. Saleh, and F. Merza, "Data Visualization and statistical graphics in big data analysis by Google Data Studio – sales case study," 2022 IEEE Technology and Engineering Management Conference (TEMSCON EUROPE), 2022.

[54] E. A. Akhir, N. S. Aziz, and A. F. Roslin, "Data Visualization for evaluation form management system," 2018 IEEE Conference on Big Data and Analytics (ICBDA), 2018.

[55] H. Ji and W. Gan, "Data visualization for making sense of scientific literature," 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2020.

[56] A. Desai, M. Mian, D. Hazel, A. Teredesai, and G. Benner, "Data Visualization in educational datasets using a rule-based inference system," 2014 IEEE International Congress on Big Data, 2014.

[57] D. Zhu, Y. Wang, B. Wei, Z. Guo and F. Wan, "Data Visualization Overview", 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), DOI:101109/ICCASIT53235.2021.9633610, 2021.

[58] J. B. Chandrasekar, S. Murugesh, and V. R. Prasadula, "Deriving big data insights using data visualization techniques," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019.

[59] M. Y. Khalid, P. H. H. Then and V. Raman, "Exploratory Study for Data Visualization in the Internet of Things", 2018 42nd IEEE International Conference on Computer Software & Applications, DOI 10.1109/COMPSAC.2018.1028, 2018.

[60] L. Ordonez-Ante, T. Vanhove, G. V. Seghbroeck, T. Wauters and F. D. Turck, "Interactive querying and data visualization for abuse detection in social network sites", The 11th International Conference for Internet Technology and Secured Transactions (ICITST-2016), 2016.

[61] R. K. Barik, R. K. Lenka, S. M. Ali, N. Gupta, A. Satpathy and A. Raj "Investigation into the efficacy of geospatial big data visualization tools", International Conference on Computing, Communication, and Automation (ICCCA2017), 2017.

[62] M. P. Cota, M. D. Rodr´ıguez; M. R. Gonza´lez-Castro, R. M. M. Gonc¸alves, "Massive Data Visualization Analysis: Analysis of current visualization techniques and main challenges for the future", 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), DOI: 10.23919/CISTI.2017.7975704, 2017.

[63] C. K. Leung, V. V. Kononov, A.G.M. Pazdor, and F. Jiang, "PyramidViz: Visual Analytics and Big Data Visualization of Frequent Patterns", 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.158, 2016.

[64]. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big Data: the next frontier for Innovation, competition, and Productivity. June Progress Report. McKinsey Global Institute; 2011.

[65]. A. Bigelow, S. Drucker, D. Fisher, and M. Meyer. Reflections on How Designers Design with Data. In Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, AVI '14. ACM, 2014. doi: 10.1145/2598153.2598175

[66]. Agrawal D, Das S, El Abbadi A. Big Data and cloud computing: current state and future opportunities. In: Proceedings of the 14th International Conference on Extending Database Technology, ACM; 2011. pp 530–3 (2011).

[67]. Kaur M. Challenges and issues during visualization of Big Data. Int J Technol Res Eng. 2013;1:174–6.

[68] Top 4 Popular Big Data Visualization Tools. Accessed: Oct. 3, 2018. [Online]. Available: https://towardsdatascience.com/top-4-popular-big-data-visualization-tools-4ee945fe207d

[69] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Softw. Eng. Group School Comput. Sci., Mathematics Keele Univ., Keele, U.K., Tech. Rep. EBSE 2007-001, 2007.

[70]. Beyer MA, Laney D. The importance of "Big Data": a definition. Stamford: Gartner; 2012.

[71] (2018). Thomson Scienti_c Releases EndNote X1 for Windows. [Online]. Available: http://endnote.com/

[72] T. Dyba and T. Dingsoyr, "Empirical studies of agile software development: A systematic review," Inf. Softw. Technol., vol. 50, nos. 9_10, pp. 833 859, 2008.

[73]. Manicassamy J, Kumar SS, Rangan M, Ananth V, Venkataraman T, Dhavachelvan P. Gene suppressor: an added phase towards solving large scale optimization problems in genetic algorithm. Appl Soft Comp; 2015.

[74]. Akerkar R. Big Data computing. Boca Raton, FL: CRC Press, Taylor Francis Group; 2013.

[75]. Sethi IK, Jain AK. Artificial neural networks and statistical pattern recognition: old and new connections, vol. 1. New York: Elsevier; 2014.

[76]. Larose DT. Discovering knowledge in data: an introduction to data mining. Hoboken, NJ: John Wiley & Sons; 2014.

[77]. Maren AJ, Harston CT, Pap RM. Handbook of Neural Computing Applications. Academic Press; 2014.

[78]. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.

[79]. Cressie N. Statistics for spatial data. Hoboken, NJ: John Wiley Sons; 2015.

[80]. Lehnert WG, Ringle MH. Strategies for natural language processing. Hove, United Kingdom: Psychology Press; 2014.

[81]. Chu WW, editor. Data mining and knowledge discovery for Big Data. Studies in Big Data, vol. 1. Heidelberg: Springer; 2014.

[82]. Berry MJ, Linoff G. Data mining techniques: for marketing, sales, and customer support. New York: John Wiley & Sons; 1997. streams. Procedia Technol. 2014;12:255–63. Cambridge: Cambridge University Press; 2014.

[83]. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems; 2014. 3104–12.

[84]. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning. Adaptive computation and machine learning series: MIT Press; 2012.

[85]. Murphy KP. Machine learning: a probabilistic perspective. Adaptive computation and machine learning series. MIT Press; 2012.2014.

[86]. Xhafa F, Barolli L, Barolli A, Papajorgji P. Modeling and Processing for Next-Generation Big-Data Technologies: With Applications and Case Studies. Modeling and Optimization in Science and Technologies: Springer; 2014.

[87]. Giannakis GB, Bach F, Cendrillon R, Mahoney M, Neville J. Signal processing for Big Data. Signal Process Mag IEEE. 2014;31(5):15–6.

[88]. Shneiderman B. The big picture for Big Data: visualization. Science. 2014;343:730. analytic trends for today's businesses. Wiley CIO: Wiley; 2012.

[89]. Poli R, Rowe JE, Stephens CR, Wright AH. Allele diffusion in linear genetic programming and variable-length genetic algorithms with subtree crossover. Springer; 2002.

[90]. Langdon WB. Genetic programming and data structures: genetic programming + data structures = Automatic Programming!, vol. 1. Springer; 2012.

[91]. Kothari DP. Power system optimization. In: Proceedings of 2nd National Conference on Computational Intelligence and Signal Processing (CISP), IEEE; 2012; pp 18–21.

[92]. Moradi M, Abedini M. A combination of genetic algorithm and particle swarm optimization for optimal DG location and sizing in distribution systems. Int J Elect Power Energy Syst. 2012;34(1):66–74.

[93]. Melanie M. An introduction to genetic algorithms. Cambridge, Massachusetts London, England, Fifth printing; 1999. p 3.

**Authors**

**Mr. Anal Kumar** is a distinguished academician with a strong foundation in Information Technology. Graduating with a Bachelor's degree from the University of Fiji in 2009, he further excelled, acquiring a Master of Science in Information Technology in 2016. Presently, Mr. Kumar assumes the dual roles of Lecturer and Head of Department within the Department of Computing Sciences and Information Systems at Fiji National University. Concurrently, he is engaged in doctoral pursuits at the University of Fiji, focusing his research on Digital Marketplace data visualization using Machine learning algorithms. For inquiries, Mr. Kumar can be reached at anal.kumar@fnu.ac.fj.

**ABM Shawkat Ali** is a Bangladeshi-origin-Australian author, computer scientist, and data analyst. He is the author of several books in the area of Data Mining, Computational Intelligence, and Smart Grid. He is a newspaper columnist. He is an academic and well-known researcher in the areas of

Machine Learning and Data Science. He is also the founder of a research center and international conferences in Data Science and Engineering. He is now a Professor in Data Science at the University of Fiji. E-mail: abm.shawkat.ali@gmail.com