Vol. 2 Issue No. 3 (2024) pp. 1-37

Journal of Intelligent Systems and applied data science (JISADS)

ISSN (2974-9840) Online

We are pleased to publish the third issue of the Journal of Intelligent Systems and Applied Data Science (JISADS). JISADS is a multidisciplinary peer-reviewed journal that aims to publish high-quality research papers on Intelligent Systems and Applied Data Science. Published: **15-01-2025**.

Editor-In-Chief:
Dr. Wasim Ali
Journal of Intelligent Systems and Applied Data Science (JISADS)
Politecnico di Bari, Italy
Editor@jisads.com / editor.jisads@gmail.com

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

CLOSED-CIRCUIT TELEVISION TECHNOLOGY FOR PREVENTING EXAMINATION MALPRACTICES

Pius Uagbae Ejodamen

Department of Computing Sciences, Faculty of Science, Admiralty University of Nigeria, Ibusa.

piusejodamen@adun.edu.ng, +234-9135257551

ABSTRACT

This study examines Closed Circuit Television CCTV systems as part of a series of security infrastructures intended to prevent or detect examination malpractices in tertiary institutions. A structured questionnaire was designed and administered to Admiralty University of Nigeria students. Four research questions with multiple-item constructs were used to obtain the respondents' perceptions. A random online sampling technique was adopted to get the needed information on a five-point Likert scale. A 10-item questionnaire, with other questions on demographics, produced a high-reliability coefficient of 0.941. Exploratory analysis of respondents' perceptions indicated an agreement, with a statistical mean greater than 3, that the presence of CCTV will deter people from being involved in examination malpractices. This study provides statistical evidence that CCTV has great potential to improve the quality of supervision and student conduct during examinations. It is recommended that tertiary institutions should deplore CCTV cameras in examination venues for effective monitoring and to aid post-examination investigations. Future studies may investigate the perceptions of supervisors and management of tertiary institutions concerning CCTV deployment.

Keywords: Examination malpractice; CCTV; Closed Circuit Television; tertiary institution

1. INTRODUCTION

Education can either be formal or informal. Informal learning can take place in several settings, including markets, village squares, community centres, and after-school programs. Informal learning may not strictly be guided by a set curriculum and may arise unintentionally in connection with specific events or as a result of shifting practical needs. On the other hand, formal education takes place in a controlled setting with the specific goal of instructing students. It could be in a classroom setting with several students being taught by qualified/certified teacher(s). At the end of a learning period, the knowledge gained is evaluated and results are given as grades. In formal education, the desire for good grades has been prioritized by students more than actual learning. [1] opined that this may be due to pressure from parents, a decline in reading culture, as well as the availability and abuse of technology.

Examining students gives the trainer feedback on their degree of knowledge acquisition and gauges how well they retain that knowledge. This feedback mechanism is distorted by any wrongdoing or irregularity, which produces a misleading learning consequence. The violation of rules set up to ensure credible feedback is known as examination malpractice. This phenomenon is mostly carried out by examination candidates, who give the trainer feedback on their degree

of knowledge acquisition and their parents or sponsors, schools, parent-teacher associations, examination organizing bodies, and even rules enforcement agents, such as invigilators, policemen, security personnel, and proxy examination writers [2].

This study is focused on the use of CCTV in examination malpractice detection, prevention and post-offence investigation in universities, using Admiralty University of Nigeria (ADUN) as a case study. The university was selected because its students are from across Nigeria. The main limitation encountered is the potential fear of students that the result of this research may be adopted and implemented by the school authority. Some students who refused to fill out the form expressed fears that CCTV would do more harm than good because it would lead to mass failure by students. This could indicate an over-reliance on examination misconduct to pass exams. Another limitation was the passive interest of academic staff in responding to the questionnaire. This may be due to the significant workload on the staff during the survey period. There may also be staff indifference to the approved method of supervising examinations. A future study may investigate the perceptions of staff on the use of CCTV during examinations. However, these limitations did not reduce the quality of the work, because sufficient sample from the students was achieved...

2. THEORETICAL FRAMEWORK

Generally, the principal objective of examination malpractice is to enable its perpetrator(s) to achieve examination success and obtain higher grades without the corresponding knowledge, talent, or ability related to the course of study or program. In the course of this action, the authorities concerned may directly or indirectly cooperate with or allow examination malpractice to take place. Cheating in examinations is now being perceived as a norm partly due to the approach to its prevention and punishment of offenders. This has led to the escalation of examination misconduct in our school system at a scale that would result in unavoidable repercussions in future. There is an urgent need to curtail this ugly trend.

The use of Closed-circuit television (CCTV) to monitor examinations has been proposed [1][3]. This technology provides additional supervision capability by enabling real-time remote monitoring of an examination [4]. While CCTV plays a significant role in providing security intelligence, it has not been extensively adopted for preventing cheating during examinations in Nigerian universities. The apathy to study among young people may have resulted in some students depending on cheating during examinations to pass their courses. This is a serious problem which ranges from lack of self-confidence, lack of knowledge of relevant concepts, and the tendency to be a nuisance to the society in future. Students' awareness of the almost impossible chance of malpractices during examinations, may compel them to study. Thereby, knowledge is gained and self-confidence is improved.

Generally, this study aims to find out the perception of the academic community about the use of CCTV surveillance cameras in preventing examination malpractices. Specifically, the study seeks to ascertain the effectiveness of CCTV technology in preventing/detecting misconduct during examinations. It also provides statistical data on the perceptions of students about the use of CCTV to stop examination malpractices.

Therefore, the following research questions were formulated to guide this study.

Q1: Does the presence of CCTV increase focus and confidence in the examination process?

Q2: Will CCTV improve the invigilator's confidence to act professionally?

Q3: Is CCTV a deterrent against examination malpractice?

Q4: Does CCTV help in exposing examination malpractice?

3. LITERATURE REVIEW

Examination malpractice can be in various forms including impersonation, inscription on paper or parts of the body, copying from one another, bribery and/or intimidation of supervisors [2]. Also, [5] opines that adequate use of surveillance will serve as deterrence from examination malpractices. Studies have recommended ways to reduce examination misconduct, including enactment and implementation of laws, retraining of teachers, effective continuous assessment, as well as using CCTV to monitor examination venues.

3.1 Related works

The study by [1] addressed the pervasive issue

of examination malpractice in Nigeria, highlighting its causes, methods, and potential solutions through e-supervision. They identified factors such as a lack of reading culture, laziness, and overloaded syllabi as contributors to examination malpractice. The research further suggested that improving student reading habits, and teaching methods, and implementing technology like CCTV can help mitigate these issues, emphasizing the need for collective efforts to uphold examination integrity.

Another investigation by [6] revealed that many public tertiary institutions in Rivers State have adopted security measures like CCTV cameras, biometric systems, and signal jamming devices to combat exam cheating among students in advanced studies. The research examined how electronic invigilation could limit examination misconduct among postgraduate students in some public tertiary institutions in Rivers State. Of 400 students selected through multi-stage sampling techniques, 365 were completed and returned for analysis. The survey tool was validated by three experts and achieved a high-reliability score of 0.83 using the Cronbach Alpha Method. The study addressed research questions using mean and standard deviation, while testing formulated hypotheses with the z-test.

To simplify the complex and expensive traditional examination process, [3] proposed a framework for automating traditional invigilation of examination using biometric authentication and 360-degree CCTV surveillance. The method aims to eliminate student malpractices as well as reduce the number of invigilators in examination venues. The proposed system uses a biometric reader for authentication, allowing only registered students to enter the exam hall while an invigilator monitors from a distance through live CCTV feeds and communicates via microphones. The proposed model is reported to be cost-effective, and efficient, and enhances the integrity of an examination process.

An automated cheating detection system was developed by [7] using video surveillance to observe the behaviour of students during examinations. The system considered head movements, eye movements and hand movements to detect examination malpractice. Video input in real-time is captured, analysed, and classified as normal or abnormal behaviours. Misconduct would trigger an alarm to attract the proctor. In [8] a tracking application was developed that could limit misconduct during an online English academic writing examination. It was reported that this method is more cost-effective than traditional proctoring and it reflects the knowledge gained by the student.

A study by [9] have recommended combining artificial intelligence (AI), real-time CCTV coverage, and data analytics for an enhanced monitoring of examinations. Since AI systems can analyse human movements, it makes it suitable to detect unacceptable behaviour within the examination premises.

3.2 Research Gap and Problem Statement

The consequences of allowing examination malpractices to foster is unimaginable. It could lead to the collapse of society and systems. Previous studies made a case for the introduction of CCTV in examination processes in different cities and countries. There is little mention of students' perception concerning having CCTV in examination venues. Also, in a public university such as ADUN – the case study – where the population comprises of students from different parts of the country, it is suitable to understand how this technology will be received.

This study focuses on the perceptions of students in a tertiary institution. It provides statistical data to aid policy makers as evidence of the need for CCTV in monitoring examinations. Specific acts of misconduct such unauthorized movements, using smart devices, and post-examination investigation were all examine.

3.3 Methodology and Research design

This section presents the methods and procedures that were applied in this study. It describes the research design, area of the study, population of study, sample size and sampling technique, and method of data collection. The validation of the instrument, reliability of the instrument, and method of data analysis are also described in this section.

The study made use of a cross-sectional survey research design. Researchers collect detailed descriptions of existing phenomena to use the collected data to justify current conditions and practice or to make better plans for improving phenomena. Surveys are generally done to collect three kinds of information. First, data concerning existing conditions. Secondly, a comparison of the existing status of a situation and the required standard. Finally, data for improving existing conditions.

The cross-sectional survey design enables the researcher to collect his data at a particular point or period from a selected sample. This method was selected because it enabled the researcher to use a sample drawn to represent the various elements of the population under study. The item construct used in this study was derived from published works and interviews with members of the ADUN university community. This was necessary to measure responses that is peculiar to a Nigerian university.

3.4 Population and Sample

Admiralty University of Nigeria (ADUN) is a tertiary institution located in Ibusa town of Delta State, Nigeria. Focus was mainly on students who have been on campus for at least one semester and participated in an examination. As at the time of data collection, the student population from all departments in the school was 652, with male gender at 472 and 180 females.

The sampling technique applied in selecting the sample for the study is simple random sampling. This type of sampling involves a random selection of respondents from a larger sample or population, giving all individuals in the sample an equal chance to be chosen. In simple random sampling, individuals are chosen at random and not more than once to prevent biases that will negatively affect the validity of the result of the experiment. Therefore, an online survey using a Google form questionnaire and the link was shared with both students and staff. Out of 208 responses, 202 from students were accepted and 6 from staff were rejected. The responses from staff were rejected because it was insignificant compared to the number of academic staff in the university. The reason for low staff response may be due to the request for only academic staff who actively participate as examination proctors.

3.5 Method of Data Collection

A self–developed online survey questionnaire was designed with 10 items to measure students' perception of CCTV in preventing examination malpractices. In addition, 5 items were used to obtain the demographics of the students and a declaration of participation in examination malpractice. Notably, the respondents were assured that the instrument would be treated confidentially; hence names or private information was not requested.

The questionnaire was divided into two (2) sections. Section A contained demographic information about the respondent such as their gender, age range, staff/student, academic level, and declaration if they have been involved in examination malpractice. In Section B, a five-point Likert scale, similar to [10], was applied to obtain the respondents' position as regards each item. The five-point Likert scale was used for this study over the "Yes/No" and "four-point" Likert scale because the student is expected to respond to a set of close-ended questions/statements but with support for uncertainty. In this case, a respondent chooses from options such as 'Strongly Agree', 'Agree', 'Undecided/Neutral', 'Disagree', and 'Strongly Disagree'. Scores on this scale ranged from 1 ('Strongly Disagree') to 5 ('Strongly Agree') and the respondents checked the box that best reflected their view on the items stated.

3.6 Validity and Reliability of the Instrument

The validity of an instrument refers to the extent to which the questionnaire measures what it claims to measure [10]. Validity means the extent to which the scores and the conclusions based on these scores can be used for the intended purpose of the questionnaire. In other words, it is the degree to which results obtained from the analysis of the data represent the phenomena under the study. For this research, face validity and content validity of the instruments were carried out by two experts in the field of computer science for validation. Their contributions were considered in restructuring the questionnaire.

The test-retest method was used to ensure the reliability of the instrument. The instrument was trial-tested in the sample with few students, and their responses were collected. After about two months of the administration, the same test items were re-administered to the same group of respondents. Thereafter, the final questionnaire was administered to all students after the semester's examination. Cronbach's Alpha method was used to determine the reliability coefficient of the instrument which was established as 0.941, indicating very high reliability. The generally agreed-upon lower limit for Cronbach's Alpha is 0.7 [11], although it may decrease to 0.6 in exploratory research [11][12]. It was suggested that the score for each construct should be greater than 0.6 for it to be reliable [13]. Thus, a score of 0.6 and above was accepted in this study. The instrument used can, therefore, be said to be suitable for measuring the perception of students accepting CCTV as a technology that could prevent examination malpractices

4. DATA ANALYSIS

The analysis of data from the questionnaire was quantitative. Descriptive statistical method of analysis was employed in this analysis. The research questions were answered and summarized, while each

questionnaire item was analysed with the aid of a frequency percentage and/or charts. To answer the research questions, the decision was based on the instrument scale mean of 50%. Any item with a mean response above 50% was taken as "agreed" while any within 50% and below was considered as "disagreed". The hypotheses were tested at a significant level of 0.05. All analyses were computed using the International Business Machine Statistical Packages for the Social Sciences (IBM SPSS) version 27.

4.1 Demographic Analysis

Out of the 202 respondents, 153 (75.7%) of them were males and 49(24.3%) were females, as shown in Figure 1. This is a reflective estimate of the gender balance in the university as most students in ADUN are males. Hence, the selected sample used in this study is an adequate representation of the global population of ADUN concerning gender.

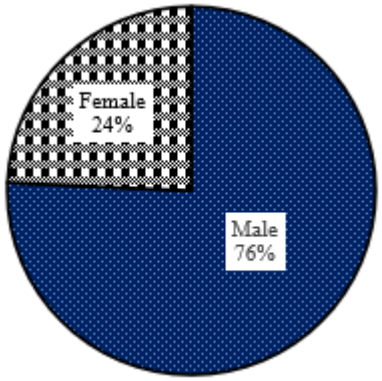

Figure 1: Gender of respondents

The age of the respondents ranged from 18 to 44 years as shown in Table 1. Most of the respondents – 58.9% – were aged between 18 and 29 years, while 37.1% of respondents were less than 18 years. Marginally, 2.5% were above 30 years and 1.5% were between 19 and 29 years. With none of the respondents being above 44 years of age, this indicates that the respondents to this survey are youth in their prime. It is expected that people in this age group are active and may want to consider any means possible to advance their studies/careers. So, their response is very relevant in determining the perceptions of students about the use of CCTV to prevent exam malpractices.

Table 1: Age (in years) of respondents

Age Range	N	%
18 – 29	119	58.9%
less than 18	75	37.1%
30 – 44	5	2.5%
19 – 29	3	1.5%

Frequency analysis of the academic status of respondents showed that 63.9% of the respondents were in the 100 level and 18.8% were in the 300 level, accounting for 82.7% of total respondents. As shown in Table 2, respondents at the 400 and 200 levels were 12.4% and 5.0% respectively. It is worthy of note that at the time of administering this instrument, there was an incremental surge in admission which is responsible for more 100 level respondents. Nevertheless, all participants had witnessed at least one examination in the institution.

Table 2: Academic Level of Respondents

Level	**N**	**%**
100	129	63.9%
300	38	18.8%
400	25	12.4%
200	10	5.0%

4.2 Descriptive Analysis

To analyse the past participation of respondents in examination malpractices, a cross-tabulation method was applied. Figure 2 data visualisation shows that 19.3% of respondents have either assisted or directly gotten involved in examination malpractice. However, 80.7% denied ever been involved in the act. With a clear majority of respondents who have never aided or been involved in examination malpractice, it is expected that an unbiased perception will be expressed about the use of CCTV to prevent examination misconduct.

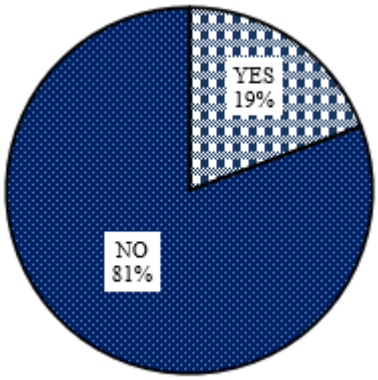

Figure 2: Involvement in examination malpractices

There are four research questions the data will be used to answer. Multiple constructs were used to obtain the perception of respondents about the use of

CCTV to prevent examination malpractices. The statistical mean for each construct is calculated to determine the perception of respondents for that question. Thereafter, the average of all the mean values for the various constructs is used to make decisions if the research questions are answered as "YES" or "NO". For example, from Table 3, Q1 has three constructs with their mean as shown. The decision is "YES" if the average of the three means is greater than 3.00, and "NO" if it is lesser.

Table 3: Answers to research questions

Code	Research Questions	Item Construct	Mean	Decision (Mean Average)
QI	Does the presence of CCTV increase focus and confidence in the examination process?	The presence of CCTV will create awareness that everyone is being watched	3.69	3.45 YES
		The presence of CCTV will prevent distractions from other students during examinations	3.05	
		Students will be afraid to swap seats with others	3.60	
Q2	Will CCTV improve invigilator's confidence to act professionally?	The presence of CCTV will prevent invigilators from assisting students	3.37	3.48 YES
		It will deter students from assaulting invigilators/supervisors during an examination	3.59	
Q3	Is CCTV a deterrent against examination malpractice?	It will deter students from going into the examination hall with implicating written materials (e.g. *expo*, *bomb* e.t.c.)	3.27	3.22 YES
		It will deter students from smuggling in digital devices such as smartphones and wristwatches	3.32	
		The presence of CCTV will discourage and deter students from talking/whispering during an examination	3.08	
Q4	Does CCTV help in exposing examination malpractice?	CCTV footage/videos will provide evidence during the investigation of an examination malpractice	3.75	3.40 YES
		It will prevent impersonation i.e students paying someone else to write their examination	3.05	

5. DISCUSSION ON THE RESULTS

5.1: Does the presence of CCTV increase focus and confidence in the examination process?

A mean response of 3.69 showed agreement that CCTV would create awareness that the examination is being monitored, thereby increasing the credibility of the process. Similarly, a mean response of 3.05 slightly agrees that distractions will be minimized while students' focus will improve. Also, respondents agreed that movements around the examination hall can make other students lose focus. Unapproved seat changes by students are regarded as misconduct, but a mean response of 3.06 indicates that students will be afraid to commit such offences in the presence of CCTV. Cumulatively, the average of these mean values resulted in 3.45. This indicates an agreement that the presence of CCTV technology when deployed in examination

venues will enhance students' focus and confidence in the examination process.

5.2: Will CCTV improve the invigilator's confidence to act professionally?

There have been reported cases where invigilators assist students in committing examination malpractice, which is unprofessional conduct. A mean of 3.37 implies that respondents believe that the installation of CCTV in examination halls will reduce such incidences. Furthermore, a 3.59 mean implies that some students who threaten supervisors and sometimes assault them will be deterred. Hence, with a mean average of 3.48, it is agreed that CCTV will improve invigilators' confidence to act professionally in an examination venue without fear or favour.

5.3: Is CCTV a deterrent against examination malpractice?

Respondents agree (average mean = 3.22) that the presence of CCTV will serve as a deterrent against those who might want to be involved in examination malpractice. It will deter students from entering the hall with unapproved materials and digital devices to aid them in cheating. When students and proctors are aware of the presence of CCTV in the examination hall, talking/whispering will be curbed.

5.4: Does CCTV help in the detection of examination malpractice?

Sometimes, when examination malpractice is reported, the accused would mostly deny it. There is also the possibility of false accusations. This makes it a challenge to determine if the student committed a punishable offence. This gap can be covered with the presence of CCTV in examination venues. Respondents with a mean of 3.75 affirmed that the recorded CCTV footage would provide reliable evidence to establish malpractice or otherwise. Similarly, a mean response of 3.05 shows that CCTV may prevent impersonation – a situation whereby someone else sits for an examination for another. If the impersonator is not caught at that moment, CCTV footage analysis will help detect the fraud. A cumulative mean average of 3.40 implies that CCTV technology can be deployed to expose examination misconduct.

6. ETHICAL CONSIDERATIONS

Despite the perceptions of students' acceptability, implementations should consider ethical requirements. It is appropriate for the students to be aware of the presence of CCTV and deploy the technology with fairness and equality according to necessity [14][15].

7. CONCLUSION

The discourse so far reveals that the menace of examination malpractices is growing rapidly in tertiary institutions. The use of CCTV technology for monitoring activities in the examination hall is gaining prominence because of its benefits which include providing evidence for investigative reasons and also serving as a deterrence to intending offenders since they are aware that they are being watched.

Based on the findings of this study, it is concluded that the place of CCTV technology in ensuring crime-free examination cannot be overemphasized. This means that the use of CCTV surveillance in place of human invigilation or in addition to human invigilation could go a long way in assisting universities to nib the issue of examination malpractices in the bud. Results from this study suggest that where CCTV technology is adopted during the conduct of examination, students' willingness to cheat will be reduced drastically and confidence in the examination process will be improved.

It is recommended that tertiary institutions should invest in and adopt the use of CCTV technology to monitor their examinations. Future studies may consider automating the use of CCTV systems for remote monitoring examinations. Furthermore, the scope of a similar study could be expanded to cover several universities across Nigeria.

REFERENCES

[1] S. O. Oso, P. O. Owoeye, and B. O. Amoran, "Conquering the Indomitability of Examination Malpractice in Nigeria in Gallantry Through E-Supervision," United Int. J. Res. Technol., vol. 5, no. 1, pp. 61–69, 2023.

[2] C. O. Onyibe, U. U. Uma, and E. Ibina, "Examination Malpractice in Nigeria: Causes and Effects on National Development", Journal of Education and Practice, vol 6, no 26, 12–17, 2015.

[3] J. Hoque, R. Ahmed, J. Uddin, and M. M. A. Faisal, "Automation of Traditional Exam Invigilation using CCTV and Bio-Metric", International Journal of Advanced Computer Science and Applications (IJACSA), vol 11, no 6, pp. 392–399, 2020.

[4] R. Smith and G. Wilson, "CCTV surveillance in examination halls: Effectiveness in detecting malpractice," Journal of Academic Integrity, vol. 19, no. 2, pp. 115-128, 2023.

[5] A. Mohammed, M. Adamu, and T. Johnson, "The role of surveillance in curbing cheating: Case studies from Nigerian universities," Educational Research Review, vol. 9, no. 4, pp. 120-133, 2024.

[6] E. P. Menyechi, and N. T. Kelechi, "Electronic Invigilation Inclusion in Curbing Examination Malpractices Among Postgraduate Students in Selected Public Tertiary Institutions in Rivers State", British Journal of Education, Learning and Development Psychology, vol 6, no 2, pp. 100 – 113, 2023, doi: 10.52589/bjeldp-y5jnzppc.

[7] R. M. Al_airaji., I. A. Aljazaery, H. T. Alrikabi, and A. H. M. Alaidi, "Automated Cheating Detection based on Video Surveillance in the Examination Classes", International Journal of Interactive Mobile Technologies (iJIM), vol 16, no 8, 2022, pp. 124–137, doi: 10.3991/ijim.v16i08.30157.

[8] P. Pipattarasakul, and R. Phoophuangpairoj, "Alleviation of Cheating in English Academic Writing Exams Using a Developed Tracking Application", Proceedings of the 6th UPI International Conference on TVET 2020 (TVET 2020), pp. 297–302, 2021.

[9] H. Nguyen and M. Lee, "Artificial intelligence in CCTV surveillance for exam security: A technological overview," Journal of Educational Technology and Innovation, vol. 6, no. 2, pp. 89-102, 2024.

[10] U. O. Ekong, V. E. Ekong, P. U. Ejodamen, and I. B. Nderiya, "Technology Acceptance Modelling of Bring Your Own Device (BYOD): A Confirmatory Factor Analysis", Computing, Information Systems, Development Informatics & Allied Research Journal, vol 13, no 2, pp. 15-26, 2023.

[11] P. B. Robinson, D. V. Stimpson, J. C. Huefner, and H. K. Hunt, "An attitude approach to the prediction of entrepreneurship", Entrepreneurship: Theory & Practice, vol 15, no 4, pp. 13–31, 1991.

[12] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis, 7th Edition, Pearson, New York, 2010.

[13] J. C. Nunnally, Psychometric Theory, New York: McGraw-Hill, 1967.

[14] D. J. Solove, Understanding Privacy. Cambridge, MA, USA: Harvard University Press, 2021.

[15] L. Sweeney, "Ethical considerations in surveillance for security," Journal of Information Ethics, vol. 31, no. 1, pp. 56-71, 2021.

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com
ISSN (2974-9840) Online

ADVANCING ORCHARD MANAGEMENT WITH RFID: SYSTEM INTEGRATION, CHALLENGES, AND SOLUTIONS IN MOBILE APPLICATIONS

Laila A. Wahab Abdullah Naji

University of Aden-Faculty of Aden, Yemen
Tefke2010@Gmail.Com

ABSTRACT

Radio Frequency Identification (RFID) technology employs wireless radio frequency signals for non-contact data transfer, facilitating automatic identification of targets. It is emerging as a pivotal technology in the advancement of modern agriculture. This research investigates the integration of RFID technology into mobile applications for orchard management, focusing on the system's hardware components, software structure, and data management approaches. The study highlights several challenges including usability of software, scalability of functions, cost management, signal interference, tag longevity, standardization, and compatibility issues, offering adaptable solutions for improvement. Lastly, this paper anticipates future applications of RFID in orchard management and discusses the significant potential of Android Studio in developing RFID applications.

Keywords: RFID, Orchard Management, Mobile Applications, System Integration, Agricultural Technology,

1. INTRODUCTION

The rapid advancement of information technology is driving significant changes in the agricultural sector, particularly through digital transformation. Radio Frequency Identification (RFID), a wireless and contactless automatic identification technology, has demonstrated impressive outcomes in enhancing agricultural product safety supervision. It has gradually emerged as one of the key technologies underpinning agricultural modernization [1]. By using radio frequency signals, RFID enables seamless information transfer and automatic identification of target objects without requiring manual intervention. When integrated with technologies such as the Internet and mobile communications [2], RFID facilitates global item tracking and information sharing, forming a foundational component of Internet of Things (IoT) infrastructure.

In orchard management, RFID technology offers substantial improvements in efficiency, precision, and data-driven decision-making. By attaching RFID tags to fruit trees and utilizing mobile terminal devices, real-time collection and monitoring of orchard environmental conditions and fruit tree growth become possible. While the adoption of RFID technology enhances orchard management accuracy and operational efficiency, several challenges persist:

- High Costs: Deployment and maintenance of RFID systems remain expensive.
- Signal Interference: Environmental factors such as moisture, soil, and vegetation can disrupt RFID signals, reducing reading accuracy.
- Network Coverage: Orchards located in remote regions often face poor network signal coverage, hampering real-time data transmission and remote monitoring.

- Technical Complexity: Implementing RFID systems demands technical expertise and support.
- Tag Durability: RFID tags must withstand harsh environmental conditions, including sunlight and rain, which increases replacement frequency and costs.
- Data Security and Privacy: Proper measures must be in place to prevent unauthorized access and data breaches.
- Compatibility and Standardization: Variations across different RFID systems and devices may hinder overall system performance and scalability.
- User Acceptance: Orchard managers accustomed to traditional methods may require time and training to adopt new technologies.
- Equipment Maintenance: Devices such as RFID readers and mobile terminals require regular upkeep, adding to long-term operational expenses.
- Software Integration: Seamlessly integrating RFID technology into existing orchard management systems often necessitates additional development work to ensure smooth data flow and processing.

Addressing these challenges is crucial for maximizing the potential of RFID technology in orchard management and achieving sustainable digital transformation in agriculture.

2. TECHNICAL OVERVIEW

2.1 RFID Technology and Its Application in Orchard Management Systems

Radio Frequency Identification (RFID) is a contactless automatic identification technology that utilizes radio frequency signals for object identification and data acquisition. It has become an essential tool in modern agricultural management, significantly enhancing operational efficiency and accuracy in orchard systems. RFID technology relies on wireless communication between a tag and a reader at its core. The tag, equipped with an electronic chip and antenna, can store and transmit specific data [3].

RFID tags are often attached to fruits, trees, or plants in orchard management, storing details such as variety, planting location, and maturity stage. The RFID reader can rapidly retrieve this information over long distances without direct physical contact, effectively minimizing errors and inefficiencies associated with manual data entry. Research indicates that RFID systems in orchard management can achieve asset-tracking accuracy rates exceeding 99%, compared to just 85% with traditional manual methods.

Beyond improving data collection accuracy, RFID technology enables real-time monitoring and informs decision-making. This allows orchard managers to allocate resources more effectively and plan production with enhanced precision, ultimately optimizing overall productivity.

2.2 Android Studio Development Technology

Android Studio serves as the primary integrated development environment (IDE) for developing RFID applications on Android devices. By 2023, Android Studio supports over 2.5 billion active devices, providing a robust foundation for scalable RFID application deployment.

To ensure seamless integration, developers must install the latest Android Software Development Kit (SDK) and ensure compatibility with the required API levels. Configuring virtual devices or connecting physical Android devices is essential for thorough testing and stability checks.

Android Studio's comprehensive toolset enables developers to design, build, and optimize RFID-integrated applications, focusing on technical performance and user experience. This environment ensures that RFID-based solutions operate reliably across diverse Android devices.

2.3 Connecting RFID Technology to Android Devices

Connecting RFID technology to Android devices begins with selecting an appropriate RFID reader, such as a USB-based or Bluetooth-enabled card reader. Key considerations during selection include range, compatibility, battery life, and connection methods.

Once a suitable RFID reader is chosen, specific applications (e.g., RFID Tag Tracker or RFID Reader Scanner) must be installed on the Android device. These apps are typically designed to support particular reader models.

After installing the necessary software:

Open the application and connect with the RFID reader via Bluetooth pairing or USB connection.

Configure the RFID reader settings, including frequency, reading range, and data format.

Initiate the tag-reading process. The RFID reader emits signals, and upon receiving them, the RFID tag returns its unique identifier, enabling the reader to process and interpret the data.

This seamless integration allows Android devices to function as efficient tools for RFID data collection, processing, and management in orchard environments.

3. PROBLEMS AND IMPROVEMENTS IN ORCHARD MANAGEMENT SYSTEMS

3.1 Traditional Orchard Management Methods and Their Limitations

Traditional orchard management primarily relies on manual recording and observation to track tree growth, monitor pests and diseases, and determine fruit maturity and harvest timing. For instance, fruit farmers often assess fruit ripeness based on personal experience, which lacks precision and can lead to mistimed harvesting. This inaccuracy frequently results in reduced fruit quality and market value, with studies indicating that losses from improper harvest timing can range between 10% and 20%.

Manual asset management also poses significant challenges. Tracking tools, machinery, and other resources through handwritten records is both time-consuming and error-prone, hindering real-time monitoring and efficient resource allocation. These limitations highlight the need for modern technological solutions to address the inefficiencies of traditional methods.

In today's fast-paced agricultural landscape, the demand for swift and accurate responses to changing conditions has driven the adoption of advanced technologies like RFID systems, which offer significant improvements in efficiency, precision, and data management.

3.2 Advantages of Mobile Orchard Management Systems Based on RFID Technology

Mobile orchard management systems powered by RFID technology enable real-time and accurate data collection and processing. By attaching RFID tags to fruit trees, essential data—such as growth status, fertilization schedules, and irrigation history—can be instantly accessed via mobile devices. This minimizes human error, ensures data accuracy, and allows managers to make informed decisions promptly.

RFID technology excels in fruit maturity monitoring. Tags affixed to fruits enable real-time monitoring of ripening stages, allowing managers to predict optimal harvest times accurately. This prevents economic losses associated with premature or delayed picking and ensures better fruit quality and market value.

Additionally, RFID supports end-to-end traceability, tracking fruits from harvesting through packaging and transportation. This guarantees quality control and enhances product safety, instilling greater confidence among consumers and stakeholders.

In terms of asset management, RFID tags attached to tools and machinery allow for real-time tracking of asset location and usage status. Mobile devices can instantly display this data, reducing the risks of asset loss or damage and improving resource utilization.

The mobile application interface further enhances convenience, enabling orchard staff to monitor and manage operations anytime and anywhere via smartphones. This flexibility significantly improves overall productivity and operational efficiency.

In summary, mobile orchard management systems leveraging RFID technology enhance data accuracy, streamline resource allocation, and provide valuable decision-making support. As RFID technology continues to evolve, its potential applications in orchard management will only expand.

3.3 Enhancements in RFID Technology for Mobile Orchard Management Systems

The ongoing improvement of RFID technology in mobile orchard management systems focuses on user interface optimization, functionality expansion, and system performance upgrades.

Key Improvements Include:

-Intuitive User Interface: The mobile application interface is designed for ease of use, allowing fruit farmers to quickly learn the system and access critical data and analytical reports effortlessly.

-Remote Troubleshooting and Diagnosis: Technical support personnel can now provide remote assistance to address equipment malfunctions, minimizing production downtime and operational disruptions.

-Integrated Weather Forecasting and Disaster Alerts: The system delivers timely weather updates and risk alerts, helping fruit farmers implement preventive measures to safeguard orchard production.

-AI Integration for Predictive Insights: Artificial intelligence algorithms are employed to predict market

trends and consumer demand, offering customized planting recommendations. This reduces reliance on manual decision-making and enhances the orchard's competitiveness in the market.

These enhancements not only improve the usability and reliability of mobile RFID systems but also empower orchard managers with data-driven insights for better decision-making. As these technologies continue to advance, they promise to revolutionize orchard management further, making it smarter, more efficient, and highly sustainable.

Performance Optimization of RFID Systems

1. Reducing Interference

Interference is a common challenge in RFID systems, stemming from metal objects, electromagnetic sources, signal reflections, label occlusion, and environmental factors. Below are strategies to address these issues:

Metal Interference: Tools, equipment, or metal orchard facilities can disrupt RFID signals.

Solutions: Use metal-shielded tags or anti-metal tags specifically designed to function near metallic surfaces. Adjust the antenna's position and angle to minimize direct interference.

Electromagnetic Interference (EMI): Electromagnetic sources can disrupt normal RFID operations.

Solutions: Reduce or shield interference sources near RFID installations. Use a spectrum analyzer to detect and avoid interference frequencies. Choose RFID devices and antennas with strong anti-interference capabilities.

Multi-path Interference: Reflections from surrounding surfaces cause radio frequency signals to propagate via multiple paths, leading to signal instability.

Solutions: Use circularly polarized antennas, optimize the position and orientation of the reader and antenna, and minimize reflection paths.

Label Occlusion: Physical obstructions between the RFID tag and reader can severely disrupt signal transmission and reception.

Solutions: Use anti-interference tags such as waterproof or anti-metal tags. Adjust the relative positioning of the reader and tag to avoid obstructions.

Environmental Interference: Temperature, humidity, and dust can impact RFID performance.

Solutions: Select environment-specific RFID equipment designed to withstand high/low temperatures and humidity. Implement protective measures such as dust covers and waterproof enclosures.

Signal Conflicts: Overlapping signals from multiple RFID systems operating in close proximity can cause communication issues.

Solutions: Use RFID systems operating on different frequency bands, avoid frequency conflicts, and coordinate system operating times. Employ anti-collision protocols to prevent data overlaps.

2. Enhancing Tag Durability

The lifespan and reliability of RFID tags depend on material quality, structural design, environmental conditions, and maintenance practices.

Material Quality: High-quality materials, such as ABS plastic, offer excellent wear resistance, chemical corrosion resistance, and impact resistance.

Structural Design: Tags should be structurally reinforced to withstand mechanical stress and daily wear and tear. Materials like PET foam substrates combined with aluminum etching antennas ensure flexibility and durability.

Anti-Metal Design: Use anti-metal RFID tags equipped with electromagnetic-absorbing materials on the back to minimize interference from metal surfaces.

Maintenance and Upkeep: Regularly clean and inspect RFID equipment, including antennas and connectors, to ensure secure and corrosion-free connections.

Temperature Management: Store tags at optimal temperatures between −20°C and 60°C to prevent performance degradation caused by extreme heat or cold.

3. Solving Standardization and Compatibility Issues

Standardization is critical to ensuring interoperability and compatibility across RFID systems. Adhering to international and national standards is essential for seamless integration.

International Standards: Follow ISO/IEC specifications, including:

ISO/IEC 18000 Series: Covers frequency bands and applications (e.g., LF, HF, UHF).

ISO/IEC 14443 Series: Focuses on high-frequency RFID systems.

National Standards: Comply with domestic standards, such as GB/T 33848.3-2017, to align with regional requirements.

Compatibility Testing: Regular compatibility tests ensure that RFID readers and tags can communicate effectively across different devices and manufacturers.

Technical Specifications: Harmonize technical specifications and testing methodologies to guarantee consistency and seamless integration between devices.

By addressing these key performance challenges—interference, durability, and standardization—RFID systems in orchard management can achieve higher reliability, efficiency, and scalability.

4. Application and Implementation of RFID Technology in Orchard Management Systems

4.1 Hardware Components of the Orchard RFID System

The hardware architecture of an RFID system in orchard management consists of four primary components: tags, readers, antennas, and middleware. These components work together to enable efficient data collection, transmission, and integration.

RFID Tags: These are attached to fruit trees or individual fruits and contain unique identification data. Using wireless communication technology, readers can access this information rapidly. For instance, a standard UHF RFID tag typically has a data storage capacity of 64 to 128 bits, sufficient to store key details such as fruit variety, maturity level, and planting location.

Readers: The primary function of RFID readers is to send query signals to tags, receive response signals, and transmit the collected data to the orchard management information system for further processing.

Antennas: The design and placement of antennas play a crucial role in signal coverage and reading efficiency. These are often customized based on the orchard's size, layout, and terrain characteristics to ensure optimal signal transmission and reception.

Middleware: Acting as an intermediary layer, middleware filters, integrates, and forwards data from readers to the backend database. This component ensures smooth communication and seamless data processing between the hardware and software systems.

When these hardware components operate in synergy, the RFID system facilitates real-time monitoring of orchard assets and fruit status, leading to enhanced management efficiency and improved fruit quality.

4.2 Software Architecture and Data Management of Orchard RFID System

The software architecture of an orchard RFID system is structured into three main layers: the data acquisition layer, the data processing layer, and the application layer. This architecture ensures seamless data collection, analysis, and visualization for informed orchard management.

Data Acquisition Layer: This layer is responsible for extracting information directly from RFID tags. It captures critical data such as fruit location, variety, and maturity stage and passes it to the next layer for processing.

Data Processing Layer: Advanced algorithms are employed to clean, integrate, and analyze the raw data collected from the acquisition layer [5]. This step ensures the accuracy, reliability, and consistency of the data before it moves to the final layer.

Application Layer: In this layer, processed data is converted into user-friendly visual information. Orchard managers can access dashboards, charts, and analytics reports to make informed decisions regarding fruit harvesting, irrigation, and pest control.

Data Management is a cornerstone of the software system, and it involves:

Large-scale Data Processing: The system must be capable of handling data from thousands of RFID tags efficiently while ensuring rapid responses to query requests.

Data Storage: Robust databases are required to store vast amounts of collected data securely.

Data Security and Privacy: Encryption technologies and access control protocols must be implemented to protect data during storage and transmission and prevent unauthorized access or leaks.

The combination of efficient software architecture and secure data management practices ensures the smooth operation of RFID systems in orchards, empowering managers with accurate, real-time insights for improved productivity and resource allocation.

5. Developing RFID Applications with Android Studio

5.1 Hardware Interface Selection and Software Development Environment Preparation

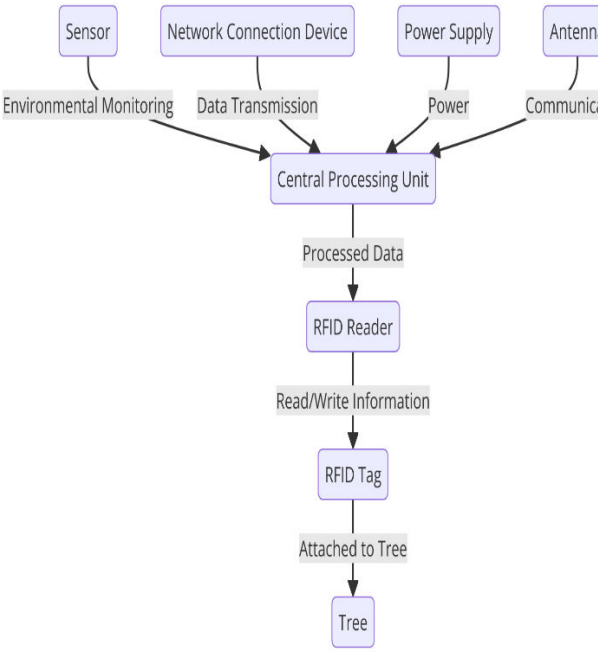

Figure 2. Flowchart of the RFID Reader Initialization

To enable RFID tag reading on the Android platform, developers typically use RFID readers/writers compatible with Android systems. These devices can connect to Android devices through three main interfaces:

USB: Provides a stable and high-speed connection for RFID readers.

Bluetooth: Offers wireless communication but requires pairing with Android devices.

Near-Field Communication (NFC): NFC is integrated into many modern smartphones, making it an accessible RFID solution. However, NFC has limited reading distance, which can constrain certain use cases.

For application development, Android Studio serves as the primary Integrated Development Environment (IDE). Developers need to:

Install Android Studio and configure it with the appropriate SDK versions.

Download and integrate the RFID reader's Software Development Kit (SDK) or Application Programming Interface (API).

Declare necessary permissions in the AndroidManifest.xml file, including access for USB, Bluetooth, and NFC.

Proper configuration ensures seamless communication between the Android device and the RFID hardware, laying the foundation for efficient application development.

5.2 Reader-Writer Connection and RFID Data Reading and Parsing

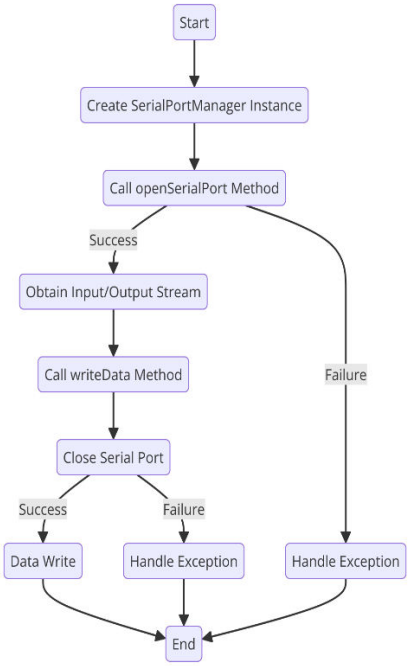

Figure 3. Flowchart of the Serial Port Initialization and Read/Write Operations

(Placeholder for diagram if needed)

The process for connecting RFID readers to Android devices depends on the chosen communication interface:

USB Interface: Use USB Host API to identify and communicate with RFID peripherals.

Bluetooth Interface: Utilize BluetoothAdapter and BluetoothSocket classes for pairing and data transfer.

NFC Interface: Leverage NfcAdapter for scanning and reading NFC tags.

Once the connection is established, the application calls the RFID API to extract data from the tag. This data undergoes parsing and analytical processing to retrieve meaningful information, such as fruit variety, location, and maturity status.

Key Steps in Data Reading and Parsing:

Establish a connection with the RFID device using the chosen interface.

Retrieve tag data using the API functions provided by the SDK.

Parse and clean raw data for accuracy.

Display processed information in a user-friendly format on the Android application.

5.3 Implementation Details

In Android environments, serial communication serves as the core method for exchanging data with RFID hardware. Below are the implementation details:

Hardware Requirements:

ACR122U RFID Reader

USB-to-Serial Module (e.g., CH340)
Android Device with OTG Support
Software Configuration:
Add the android-serialport-api library dependency in the build.gradle file.
Declare serial communication permissions in AndroidManifest.xml.
Serial Communication Process:
Instantiate the SerialPortManager class.

Open the serial port:
serialPort = new SerialPort(new File(path), baudRate, 0);

Set up input and output streams:
InputStream inputStream = serialPort.getInputStream();
OutputStream outputStream = serialPort.getOutputStream();

Write data to the RFID reader:
outputStream.write(data);

Read data from the RFID tag:
int bytesRead = inputStream.read(buffer);
Handle exceptions for any failures during communication.

Close the serial port when the process is complete:
serialPort.close();

ACR122U Integration Workflow:
Instantiate the RFIDManager class.

Open the reader connection:
reader.open("ACR122U");
Read tag data using the API functions.
Close the reader after completing operations:
reader.close();

Handle any exceptions during connection or disconnection processes.
Serial Communication Workflow:
Attempt to open the serial port and initialize streams.
Write data to the RFID tag.
Read data from the tag.
Close the connection properly.
This integration ensures robust communication between RFID hardware and Android devices, enabling smooth data exchange and processing.

6. Development direction of RFID technology in orchard management system- Discussion

With the continuous development of Internet of Things technology, the application of radio frequency identification (RFID) technology in the field of orchard management is becoming increasingly in-depth, and its future development trend mainly focuses on intelligent, refined, and sustainable management. For example, the application of RFID technology makes it possible to monitor each fruit tree in the orchard in real-time. It can accurately record the growth status of each tree, the occurrence of pests and diseases, and management activities such as fertilization and irrigation. Studies have shown that orchard management using RFID technology can reduce the error of fruit maturity detection to less than 1%, significantly improving fruit quality and yield. In addition, combined with big data analysis and machine learning algorithms, RFID systems can predict the best time to pick fruits, thereby reducing losses caused by picking too early or too late. RFID technology provides accurate data support for orchard managers, enabling them to measure and manage every aspect of the orchard better, thereby realizing intelligent and refined orchard management.

7. Concolusion

With the rapid development of IoT technology, the integration of RFID technology and Android Studio is opening a new chapter in smart applications. With its powerful functions and flexible customization, Android Studio provides unprecedented convenience for the development of RFID applications. Developers can use Android Studio's efficient code editor, rich debugging tools, and intuitive user interface design to quickly build stable and user-friendly RFID applications. In addition, Android Studio's Gradle build system supports modular development, making the maintenance and update of RFID applications more efficient. In terms of data security, Android Studio provides a powerful encryption library and security framework to help developers protect the security of RFID data transmission and prevent unauthorized access. In the future, as Android Studio continues to be updated and optimized, its role in RFID application development will become more important, providing developers with more powerful tools and resources to promote the in-depth application of RFID technology in various industries.

REFERENCES

[1] Othman, A. M. A., Ahmad, N. A., Ripin, N., Saadon, E. I. S., Ismail, M. F., Abd Rahman, N. H., ... & Ramli, N. (2024). Exploring the Opportunities of Applying RFID Technology in Smart Agriculture. Journal of Advanced Research in Applied Mechanics, 127(1), 144-154.

[2] Shen, X., Shi, G., Cheng, L., Gu, L., Rao, Y., & He, Y. (2023). Chipless RFID-inspired Sensing for Smart Agriculture: A Review. Sensors and Actuators A: Physical, 114725.

[3] Zhou, J., & Shi, J. (2009). RFID localization algorithms and applications—a review. Journal of intelligent manufacturing, 20, 695-707.

[4] Alwadi, A., Gawanmeh, A., Parvin, S., & Al-Karaki, J. N. (2017). Smart solutions for RFID based inventory management systems: A survey. Scalable Computing: Practice and Experience, 18(4), 347-360.

[5] Yeh, Sheng-Cheng, et al. "A Performance Improvement for Indoor Positioning Systems Using Earth's Magnetic Field." Sensors 23.16 (2023): 7108.

[6] Zayou, R., Besbe, M. A., & Hamam, H. (2014). Agricultural and environmental applications of RFID technology. International Journal of Agricultural and Environmental Information Systems (IJAEIS), 5(2), 50-65.

[7] Abu, N. S., Bukhari, W. M., Ong, C. H., Kassim, A. M., Izzuddin, T. A., Sukhaimie, M. N., ... & Rasid, A. F. A. (2022). Internet of things applications in precision agriculture: A review. Journal of Robotics and Control (JRC), 3(3), 338-347.

[8] Wu, C., Zhang, H., Zhang, J., Tian, W., & Cheng, H. (2013, June). An Orchard Management Systemwith RFID-Based Apple Tree Identify Detection. In 2013 Fourth International Conference on Digital Manufacturing & Automation (pp. 180-184). IEEE.

[9] Visconti, P., de Fazio, R., Velázquez, R., Del-Valle-Soto, C., & Giannoccaro, N. I. (2020). Development of sensors-based agri-food traceability system remotely managed by a software platform for optimized farm management. Sensors, 20(13), 3632.

[10] Bresilla, K. (2019). Sensors, Robotics and Artificial Intelligence in Precision Orchard Management (POM).

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

INTELLIGENT ANALYSIS OF SCIENTIFIC AND TECHNOLOGICAL LITERATURE: A NEW PARADIGM FOR RESEARCH EFFICIENCY AND INSIGHT DISCOVERY

ZAINAB KAREEM ABDULLAH[*1]

[1]*Ministry of Education, Iraq, almwswymhmd125@gmail.com*

ABSTRACT

This paper proposes a new scientific research paradigm, intelligent analysis of scientific and technological literature. By comparing traditional literature analysis methods, it emphasizes the significant advantages of intelligent analysis of scientific and technological literature in improving research efficiency and depth. The article elaborates on the concept of intelligent analysis of scientific and technological literature and its great role in scientific research, and looks forward to the theoretical basis of natural language processing, machine learning and other technologies in realizing intelligent analysis of scientific and technological literature. A proof-of-concept system is designed, and some core functions are tested and analyzed using some random paper data. Intelligent analysis of scientific and technological literature is expected to become an indispensable research tool for researchers and promote scientific research to a new level.

Keywords: Intelligent Literature Analysis, Natural Language Processing (NLP), Machine Learning in Research, Scientific Knowledge Discovery, Bibliometric Tools

1. INTRODUCTION

As an important means to extract valuable information from massive literature, understand and gain insights into research trends, and assist in scientific research decision-making, the development of scientific and technological literature analysis is closely related to the development of information technology. In the traditional period, it mainly relied on manual reading, manual sorting, and simple statistical analysis. It was inefficient and easily affected by subjective factors. Entering the computer-assisted stage, with the popularization of computers, various literature management tools such as EndNote and Zotero [1] emerged to assist researchers in collecting, sorting, and annotating literature. Then entered the data mining era. The introduction of data mining technology made it possible to conduct more in-depth mining of literature. Researchers began to use clustering, classification, association rules and other technologies to develop tools such as CiteSpace, VOSviewer, Bibliometrix and SciMAT [2] to extract implicit knowledge from literature. With the rise of artificial intelligence, we have gradually entered the era of artificial intelligence. Especially with the progress in the fields of big data and natural language processing, scientific and technological literature analysis has entered a new stage. Machine learning, deep learning and other technologies are widely used in tasks such as literature summarization, sentiment analysis, and topic modeling, which will greatly improve the efficiency and accuracy of analysis. Based on this, this paper proposes a new intelligent analysis of scientific and technological literature (Intelligent Insights) for scientific and technological literature analysis, and designs a corresponding proof of concept (Proof of Concept) system.

2. LITERATURE REVIEW

2.1 Functions and features of scientific literature quantitative analysis software

Existing bibliometric analysis tools [2] CiteSpace, VOSviewer, Bibliometrix and SciMAT are designed to help researchers deeply explore the huge scientific literature database and reveal the complex relationships therein. Their core function is to visualize the co-citation network. Through this intuitive way, researchers can clearly observe the relationship between different research fields, the evolution path of knowledge, and the distribution of key authors, papers and topics. The generated co-citation

network not only presents a static knowledge graph, but also provides some interactive functions. Researchers can delve into co-citation papers of interest and understand the development status of research frontiers in different fields. This function is valuable for conducting a comprehensive literature review and identifying key topics and influential works in a field. By analyzing the citation activity of a specific research topic, emerging research trends can be identified, helping researchers to gain insight into future research directions in advance. These quantitative software can also analyze the time development of citations and track the changes and development trends of literature, so as to have a deeper understanding of the knowledge evolution process in a certain research field. In addition, the software can also detect and visualize the cooperation relationship between authors, so as to understand the collaboration mode between different research teams and the impact of cooperation on research output and innovation. In addition, the software can track the emergence of keywords, conduct in-depth analysis of topics in the literature, and identify emerging research areas by tracking the frequency and changes of keywords.

These analytical software can help researchers keep up with the latest research trends and provide researchers with a diverse toolbox to help them better understand the generation, dissemination and evolution of scientific knowledge. Through visualization, interactive and predictive analysis functions, researchers can position their research in the current academic environment.

2.2. Deficiencies of existing scientific literature quantitative analysis software

In today's era of rapid development and widespread application of artificial intelligence, existing scientific literature analysis software has highlighted some functional deficiencies in the following aspects. The specific manifestations are as follows, such as the lack of real-time data integration. Existing software usually requires users to manually import paper data in a specified format, cannot import full-text files, and cannot achieve real-time connection with the literature database, which greatly reduces efficiency; without machine learning capabilities, the analysis results of existing software are often static and cannot be automatically updated as new literature emerges; lack of context analysis, the analysis of existing software on literature mostly stays at the level of keywords and abstracts, and cannot deeply understand the relationship between literature and citation context; weak natural language processing function, analysis software overly relies on article abstracts and cited literature, and cannot extract more valuable information from massive text. In addition, due to the lack of natural language processing function, analysis software cannot analyze subjective emotions in literature and cannot understand the author's attitude and views on research results. Faced with today's information explosion, scientific and technological literature has shown an exponential growth. The massive data generated by scientific research results worldwide requires us to no longer stay on the visualization function of existing software, but to deeply understand the data and content of literature. Obviously, traditional retrieval, export methods, and quantitative analysis software are no longer sufficient to meet the needs of researchers.

2.3. What is "intelligent analysis of scientific literature"?

Traditionally, researchers often complete such work by combing literature, including manual collection and combing, or using bibliometric analysis tools to process the collected data in a certain format. The intelligent analysis of scientific and technological literature proposed by us refers to the use of artificial intelligence technology to deeply mine, analyze and understand massive amounts of unstructured scientific and technological literature data, so as to quickly and accurately obtain key information and discover potential knowledge associations, thereby helping researchers, engineers, etc. to quickly obtain the required information and promote scientific research innovation. The text of scientific and technological literature contains rich implicit knowledge and potential value. The intelligent analysis of scientific and technological literature also represents a paradigm shift in research. From manual processing, importing data into quantitative analysis software to generate visualization, to extracting deeper and more usable knowledge information from complex data sets (unstructured data). Such a system integrates artificial intelligence, natural language processing, machine learning, data analysis, visualization and cognitive reasoning technology. It can unlock deeper information in the literature, predict future scientific research development trends, and thus promote the rapid development of scientific research. The intelligent analysis system of scientific and technological literature also uses machine learning functions to train local data models and automatically update and iterate, thereby providing forward-looking guidance for researchers' decision-making. The system is also expected to better achieve cross-disciplinary and cross-professional collaboration and complementarity, and improve scientific research prediction capabilities.

3. INTELLIGENT ANALYSIS SYSTEM OF SCIENTIFIC AND TECHNOLOGICAL LITERATURE AND ITS IMPLEMENTATION

3.1. Overview of the Intelligent Analysis System for Scientific and Technological Literature

Based on the above ideas, we designed a proof of concept system for intelligent analysis of scientific literature.

First, the support for real-time collection and input of unstructured data is a significant difference between the scientific and technological literature intelligent analysis system and traditional literature quantitative analysis tools. Traditional software often relies on pre-defined structured

data and requires input data to have clear formats and labels. However, there is a large amount of unstructured text data in the real world. For example, academic papers are often file data in various formats, such as Word, PDF, or HTML documents. The scientific and technological literature intelligent analysis system can support these documents without the need for user conversion or export.

As mentioned above, traditional quantitative analysis software can usually only provide objective information such as the frequency of citations and publication year of the literature, and cannot conduct in-depth semantic analysis of the text content. Therefore, it cannot accurately judge the author's attitude towards the research results. Traditional literature analysis software lacks semantic understanding and has limited ability to understand text. It cannot accurately identify key information in the text, nor can it provide users with an overview of key articles. At the same time, it cannot provide fully flexible customized visualization.

The intelligent analysis system of scientific and technological literature automatically discovers implicit topics in the literature and classifies the literature through concept extraction and nomenclature recognition. By analyzing the distribution of topics in different periods, we can understand the changes and development trends of hot spots in the research field. The fully customized visualization function can flexibly use the latest Python call function library to generate visualizations such as word clouds and theme maps to intuitively display the distribution and evolution of topics. Based on the function of opinion mining (also known as sentiment analysis) of the system, the system can analyze the opinions in the literature and judge whether the author's attitude towards the research results is supportive, opposed or neutral. The sentiment polarity analysis of the opinions can be performed to determine whether the opinions are positive, negative or neutral. Therefore, relevant high-quality literature and the latest research results can be recommended to users. Literature similar to the target literature can also be recommended based on the content similarity of the literature. With the help of the overview extraction function, the system can extract relevant wonderful paragraphs in the original text according to user needs, saving users time reading the original text, and also obtaining valuable paragraphs far higher in quantity and quality than the abstract. Based on intelligent question and answer of generative AI, users can obtain knowledge in specific fields from trained big data models by asking questions, such as "Who is the authoritative scholar in this field?", "Which institutions are most active in research in this field?", "What is the main contribution of this article?", etc.

Compared with the function set of the above-mentioned intelligent analysis of scientific and technological literature, the functional deficiencies of the existing software are obvious. Table 1 compares the function sets of the two.

3.2. Overview of the implementation methods of intelligent analysis of scientific literature

The user input, data collection and preprocessing module is the foundation of the system and also the input module. This module is responsible for importing unstructured text data such as PDF, Word, TXT, HTML and other files from various channels (such as user-uploaded documents, network, document library real-time download, etc.), and performing data cleaning, denoising, word segmentation and other preprocessing to provide high-quality data for subsequent analysis.

Table 1. Comparison of features between traditional tools and intelligent Insights system

Functional Classification	**Traditional Bibliometric Analysis Tools**	**Intelligent Analysis System for Scientific and Technological Literature**
Topic network generation	Support	Support
Author network generation	Support	Support
Co-citation network generation	Support	Support
Evolution timeline generation	Support	Support
Emergence Detection	Support	Support
Unstructured Data	Not supported	Support
Full-text-based nomenclature recognition	Not supported	Support
Concept extraction based on full text	Not supported	Support
Full-text based text classification	Not supported	Support
Opinion mining based on full text	Not supported	Support
Text profile based on the full text	Not supported	Support
Smart Question and Answer	Not supported	Support
Visual customization	Partial support	Support

The system’s large language model interface, LLM API (Large Language Model API), provides users with a means to quickly access the most advanced large language models [3] (such as GPT-4, Google Bard, Meta LLaMA, etc.) to achieve human-computer dialogue. The LLM API is highly flexible and can customize the model’s output style, content scope, and interaction method according to different application scenarios, and develop customized intelligent question and answer systems.

The NLP module is at the center of the scientific and technological literature intelligent analysis system, including the data mining engine (Text Mining Engine), machine learning and model training, and internal model submodules. Based on preprocessing, the NLP module extracts valuable features from the text, such as keywords,

word frequency, sentiment tendency, etc. These features will be used as input for machine learning and model training to build and train the internal model. This internal model is different from the LLM from the outside. It is the basis of the entire scientific and technological literature intelligent analysis system. The extracted features are used to build and train internal models based on deep learning, such as classification models, clustering models, and generation models. Based on the internal model that is continuously iteratively trained, users can extract named entities and concepts, classify them, and compress long texts to generate a refined overview of the core information. We will discuss these core functions of the scientific and technological literature intelligent analysis system in detail in the next chapter.

The report generation and visualization module of the system presents the analysis results to the user in a concise and visual manner, such as in the form of graphs, tables, text overviews, etc., to generate reports with clear structure and rich content.

In general, the scientific literature intelligent analysis system is a highly integrated system, with each module interdependent and mutually reinforcing. From data collection to report generation, the entire process is a continuous process of intelligent knowledge processing. Through such a system, we can more deeply explore the value contained in the literature and provide support for scientific research decision-making.

4. KEY FUNCTIONS AND IMPLEMENTATION OF INTELLIGENT ANALYSIS OF SCIENTIFIC LITERATURE

To realize the intelligent analysis system of scientific and technological literature, it is necessary to integrate the research results of multiple interdisciplinary fields, including computer science, artificial intelligence, natural language processing, data mining, etc. Its theoretical basis mainly comes from the following disciplines: knowledge graph, data mining, natural language processing, and machine learning. Among them, knowledge graph and data mining based on statistical results have been widely used in existing methods and will not be repeated here. How to solve the functional deficiencies of existing software tools in Table 1 technically? In theory, these problems can be answered by natural language processing (NLP) technology [4]. We discuss them in the following sections and use our proof-of-concept system to give practical applications and demonstration results.

4.1. Named Entity Recognition for Intelligent Analysis of Scientific Literature

4.1.1. Theoretical basis of named entity recognition

Named Entity Recognition (NER) in intelligent analysis of scientific and technological literature [5] is a subtask in NLP. Its goal is to identify entities with specific meanings in text and classify them into predefined categories, such as names of people, places, and names of organizations. In the application of intelligent analysis of scientific and technological literature, NER plays a vital role. It provides a basis for subsequent tasks such as text analysis, information extraction, and knowledge graph construction. The theoretical basis of NER mainly comes from the following aspects: statistics and probability models. NER is essentially a classification by calculating the probability that a word belongs to a certain named entity. Commonly used probability models include hidden Markov model (HMM) and conditional random field (CRF). The text features are converted into numerical features that can be processed by the model, such as part of speech, word frequency, context, etc. NER usually adopts supervised learning, that is, the model is trained through labeled training data. Some classification algorithms, such as support vector machine (SVM), decision tree, random forest and other traditional machine learning algorithms have been widely used in NER tasks.

4.1.2. Experiments and results on the named entity recognition function

In the proof-of-concept system, we input a photo of a recently published article in the Journal of Intelligent Learning Systems and Applications [6] for named entity recognition and uploaded the article's PDF file "jilsa2024164_59601667.pdf". The system extracted the named entities shown in Table 1. This information is not obtained from a fixed-format file imported by the user like the trauma software, but from unstructured text such as PDF files. Table 2 includes geographic locations, names, and names of organizations. In total, 13 organizational names, 28 names, and 5 geographic locations were detected in the paper. Due to space constraints, not all information can be listed. Table 3 gives the geographic locations, and Figure 3 gives a bar chart of the distribution of detected organizations.

Table2. Named entity recognition result

Named Body	**quantity**
Organization Name	13
Name	28
Location	5

Table 3. Detection results of named geographic location

Place Names	**Confidence**
Madina	56.87%
Saudi Arabia	92.56%
Los Angeles, California	51.03%
New Orleans, Louisiana	88.38%

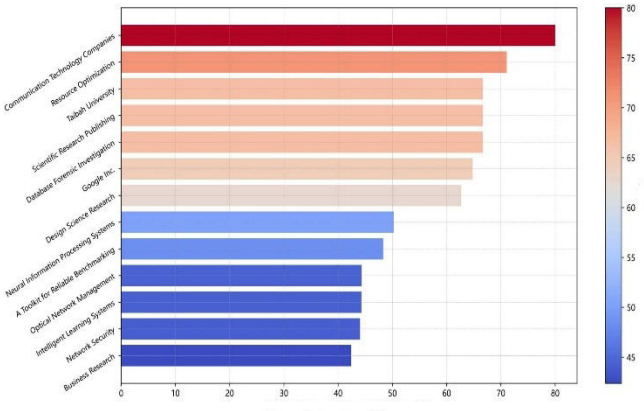

Figure 1. Named entity recognition demonstration and experimental results

As can be seen above, this proof-of-concept system can extract relevant real-name entities from randomly downloaded texts. Traditional software must import relevant data (such as abstracts and citations) in a certain format to extract relevant information such as research topics.

4.2. Concept extraction in intelligent analysis of scientific literature

4.2.1. Theoretical basis of concept extraction

Concept extraction [7] is an important branch of NLP. Different from named entity recognition, it aims to automatically identify and extract concepts with clear meanings from text. These concepts can be people, objects, organizations, events, and attributes, or more abstract concepts. Concept extraction provides a basis for many NLP tasks such as information extraction, knowledge graph construction, and text classification. The theoretical basis of its application is linguistic theory, statistics, and machine learning. Concept extraction is rooted in linguistic theory, especially semantics. Concepts in text can be identified by analyzing the semantic roles, semantic relationships, and contextual information of words. Statistical methods also provide powerful tools for concept extraction. For example, potential concepts can be discovered through word frequency statistics, co-occurrence analysis, and other methods. Machine learning algorithms, especially deep learning models, play an increasingly important role in concept extraction. By training a large amount of labeled data, the model can automatically learn complex feature representations and accurately identify concepts.

4.2.2. Functional Experiments and Results of Concept Extraction

We randomly selected an article from the open access journal “Open Journal of Social Sciences”, “Investigating the Impact of Conflict Management Approaches on Organizational Productivity in Healthcare Settings: A Qualitative Exploration” [8] as an example, and used our scientific literature intelligent analysis system to extract concepts, inputting the downloaded file jss20241211_231769282.pdf. The test results are shown in Figure 4. It is worth noting that the concepts extracted from the article are not limited to the information in the abstract, title, or keywords of the article, but the content of the full text of the paper is analyzed and processed. The system test results give the top ten complex concepts and simple concepts, and output the word cloud diagram shown in Figure 4. The system can generate different styles of word cloud diagrams according to different languages and fonts, and different mask images. This diagram uses a spherical mask as the mask image and the ERNHC.TTF font to generate a better visual effect. The system can also output detailed concept extraction results in the form of a bar heat map, as shown in Figure 5 .

Figure 2. Concept extraction word cloud graph

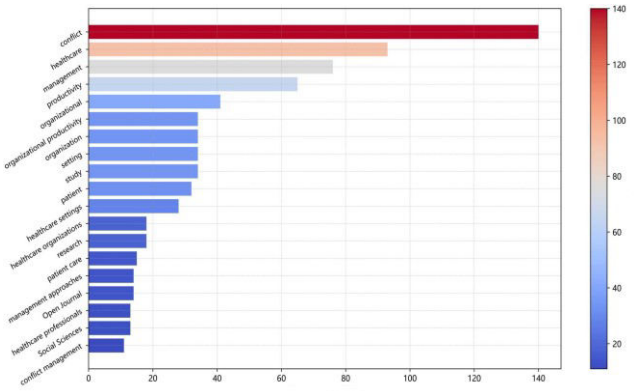

Figure 3. Concept extraction word frequency graph

4.3. Text Classification in Intelligent Analysis of Scientific and Technological Literature

4.3.1. Theoretical basis of text classification

Text classification in intelligent analysis of scientific and technological literature can be divided into different categories according to the content of the literature, such as research direction, theme, etc. As mentioned above, through NLP technology, we can analyze and classify any form of text, and are no longer limited to structured data. After text preprocessing and feature extraction, that is, through the bag-of-words model, TF-IDF, Word Embedding [9] and other technologies, the text is converted into a numerical feature vector for processing by the machine learning model. Then, the classification model uses machine learning

algorithms such as naive Bayes, support vector machine, deep learning model (such as RNN, CNN), etc. to classify the text. The advantage of NLP text classification is that it greatly improves the classification efficiency and reduces manual intervention. By training a large amount of labeled data or existing data models, a high classification accuracy can be achieved. It can classify various types of text and has strong adaptability. Moreover, with the continuous training and iteration of the model, the classification effect can be continuously improved.

4.3.2. Experiments and Results of Text Classification

Generally speaking, the dedicated local model in the intelligent analysis system of scientific and technological literature can accurately classify unstructured data. However, such a dedicated model requires a lot of machine training to improve the accuracy of classification. Due to time and space reasons, we adopted a shortcut mode to simplify the machine training process. In the test, we selected four journals from Hans Publishing House: "Material Science", "Frontiers in Sociology", "Statistics and Applications", and "Computer Science and Applications" for random downloads. The detailed file distribution in the test corpus is shown in Table 4 .

Table 3. Training corpus documents distribution

Journal Name	Number of papers	Text size
Materials Science	17	38,119,464
Frontiers of Social Sciences	19	11,375,037
Statistics and Applications	29	74,913,303
Computer Science and Applications	29	88,378,902

After learning the system machine learning, we selected another set of files for benchmarking. The file list is shown in Table 4 .

The results of the local model's benchmark proofreading are measured by the following indicators: Precision measures the accuracy of the model's prediction of the positive class, Recall measures the model's ability to identify all relevant instances, and F Score (F1 score) is the harmonic mean of Precision and Recall. Silence refers to the inability of the model to classify a text when doing benchmark verification. Silence = 1 − Precision. Noise refers to the interference information encountered by the model when processing the text, which causes the text to be incorrectly classified. Noise = 1 − Recall. It can be seen that the performance of the scientific and technological literature intelligent analysis proof of concept system can achieve a certain accuracy (~90%) even based on a small corpus. In actual operation, if you want to achieve higher precision and recall, you must expand the corpus and iterate machine learning.

Table 4. Journal paper file list for benchmarking

file name	Magazine	File size
ms20241000000_85110778.pdf	Materials Science	745,644
ms20241410_101281771.pdf	Materials Science	3,801,839
ms20241410_111281763.pdf	Materials Science	2,963,651
sa2024135_192581414.pdf	Statistics and Applications	2,365,446
sa2024135_202581421.pdf	Statistics and Applications	3,978,672
sa2024135_212581436.pdf	Statistics and Applications	438,700
ass20241311_392397511.pdf	Frontiers of Social Sciences	448,676
ass20241311_402397901.pdf	Frontiers of Social Sciences	533,220
ass20241311_412397940.pdf	Frontiers of Social Sciences	461,201
ms20240200000_45453600.pdf	Computer Science and Applications	5,023,856
ms20240200000_49175675.pdf	Computer Science and Applications	998,072
ms20240200000_71545278.pdf	Computer Science and Applications	7,702,578

4.4. Opinion Mining in Intelligent Analysis of Scientific Literature

4.4.1. Theoretical basis of opinion mining

Opinion mining, also known as sentiment analysis [10], aims to analyze subjective information such as emotions, opinions, and attitudes expressed in texts. For academic literature, sentiment analysis can help us understand the author's evaluation of the research results, whether it is positive affirmation or negative criticism, so as to have a deeper understanding of his academic views. Sentiment analysis technology can make up for the shortcomings of traditional methods by deeply understanding the semantics of the text. It can: For example, it can identify sentiment polarity: determine whether the text expresses positive, negative or neutral emotions, locate sentiment words to find keywords that express emotions, such as "good", "bad", "excellent", "bad", etc. Analyzing sentiment intensity can evaluate the intensity of emotions, such as "very satisfied", "average", "very dissatisfied". Understanding the reasons for emotions: analyzing the reasons for expressing emotions in the text, so as to have a deeper understanding of the author's views. The theoretical basis of sentiment analysis comes from statistics, and classification is performed by calculating the probability that the text belongs to different sentiment categories. Commonly used statistical models include naive Bayes and support vector machine discriminant models. Machine learning provides a large number of algorithms and models for learning rules from data and applying them to new data. For example, a classifier can be trained to map text features to sentiment

labels. Deep learning, especially neural network-based models, has achieved remarkable results in sentiment analysis tasks. Models such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and Transformers can automatically extract high-level features from text, thereby improving the accuracy of sentiment classification.

4.4.2. Experiments and results of opinion mining

Regarding the opinion mining test, we can enter two paragraphs of text in the scientific literature intelligent analysis system:

1. [John Mack] provides a thorough and insightful analysis of [1991], satisfactorily summarizing the existing problems and offering reasonable suggestions…
2. [Victor King]'s evaluation of [2] is too general and lacks in-depth analysis of the research details…

Table 6. Sentimental analysis test demonstration

Sentence	Positive Score	Negative Score	Overall Tone
John Mike conducted an in-depth and detailed analysis of the research in [1], and clearly sorted out the contributions and shortcomings of the research in this field. The author particularly emphasized the innovation of [1] and compared it with [related research] to highlight its uniqueness. In addition, the author also put forward constructive suggestions on the possible future development direction of [1], providing useful inspiration for subsequent research.	58.34	7.34%	Positive
Victor's evaluation of [2] is too general and lacks in-depth analysis of the research details. Although the author mentioned some of the advantages of [2], he did not fully discuss its limitations. In addition, when comparing [2] with other related studies, the author did not select clear comparison points, which affected the objectivity of the evaluation.	16.34%	53.8%	Negative

In traditional bibliometric software, as long as two documents are in the references, they will be included in the calculation of co-citation documents and have the same weight. But the actual situation is not so. Not all cited articles play a positive role in the author's research. Sometimes, some authors try to show the uniqueness and innovation of their research by citing some non-frontier documents. Through the intelligent analysis of scientific and technological literature, we can identify them. The second case in this example is actually a negative evaluation, which is of general significance to both the author and the paper reader. Through the intelligent analysis system of scientific and technological literature, we can analyze the tone between the lines of the dialogue and mark the negative comments. The test results are shown in Table 6. In this example, the first document is judged as a positive evaluation, while the second document is listed as a negative evaluation. With such automatic marking, scientific researchers can choose documents for further reading.

4.5. Text overview in intelligent analysis of scientific literature

4.5.1. Theoretical basis of text profiles

Text summarization in intelligent analysis of scientific and technological literature is different from the abstract of scientific and technological literature. The abstract is a short and coherent text generated by the author from an entire article or document, extracting its core ideas, main arguments and conclusions. NLP-based text summarization [11] focuses more on extracting key information from the text and generating a shorter version than the original text while retaining the core meaning of the original article. It can be an abbreviated version of the entire article or a summary of a specific paragraph. Compared with the abstract, the text summary pays more attention to the compression and retention of information to ensure that the generated text can accurately reflect the main ideas and meaning of the original text; it can generate summary content of different lengths and styles according to different needs. Text summarization can be achieved through the following methods: converting the text into a representation that can be processed by a computer, such as word vectors, sentence vectors, etc.; sorting the sentences or words in the text by importance in order to extract key information; using various compression algorithms to compress the text into a shorter version; and converting the compressed information into natural language text. Common methods for text overview include statistical methods, such as algorithms centered on high-frequency words; TF-IDF methods to measure the importance of words in documents and consider the prevalence of words; and graph-based methods, which represent text as graphs, with nodes representing words and edges representing the relationship between words, and extracting key information through graph algorithms. In addition, machine learning-based methods use labeled data to train classifiers to determine whether sentences are important. Through a reward mechanism, the training model generates high-quality overviews. Deep learning-based methods use, for example, the Seq2Seq model: text is encoded into vectors and then decoded to generate overviews.

4.5.2. Experiments and results on text overview

We used an article by the author of this paper ("Trust Construction in Commercial Transactions in the Mobile Internet Era—Based on an Investigation of WeChat Group

Buying Groups in the J Community") [12] as a sample and generated a summary of the article. In the test, we chose to generate 8% of the full text as the proportion parameter. In actual applications, this percentage parameter is user-controllable and can generate appropriate summary content according to actual needs. The results show that the important points in the paper are well extracted, far exceeding the amount of information provided by the paper abstract. Compared with reading tens of thousands of words of the original text, reading the abbreviated summary of the text can indeed help researchers save a lot of time and energy and focus on innovation and practice.

4.6. Intelligent Question Answering in Intelligent Analysis of Scientific and Technological Literature

4.6.1. Theoretical basis of intelligent question answering

Intelligent question answering is an important function of the system. It uses the large language model (LLM) to allow users to ask questions to the system and obtain accurate and relevant answers [13]. Users do not need to read the literature one by one, but only need to ask questions to quickly obtain the required information. The system can answer various questions raised by users, including factual questions, conceptual questions, comparative questions, etc. through its deep understanding of the text. Personalized service: The system can provide personalized question-and-answer services based on the user's question history and interest preferences. The technical implementation of intelligent question answering benefits from natural language understanding (NLU) and uses knowledge graphs to extract entities and relationships from a large number of documents to build a knowledge graph. The user's question is converted into a query statement on the knowledge graph, and the answer is obtained from the knowledge graph. The machine learning behind intelligent question answering uses a large number of question-and-answer dialogues to train the machine learning model and improve the accuracy of the model. The user's question is input into the model to obtain the model's prediction result. The currently well-known ChatGPT is a conversational large language model developed by OpenAI that is good at generating human-level text. It can be used to build a more natural and fluent dialogue system to improve the user experience.

4.6.2. Experiments and Results of Intelligent Question Answering

An important module in the intelligent analysis system of scientific literature is the application interface module of the LLM model, so as to call the API in the LLM. The latest Google Gemini is a large language model application developed by Google, which has powerful text generation, translation, code writing and information retrieval capabilities. In the intelligent question-answering system, Gemini can be used to understand users' questions more accurately and generate more comprehensive and logical answers. use [14] as the theme (the development trend of foreign Internet trust research under the guidance of technical logic - recognition and visualization analysis based on CiteSpace) to let Gemini provide relevant information. The test results show that the answers obtained have considerable accuracy. The quality of the answers is also very high, covering the highlights of the paper, giving the paper ideas, research methods, research characteristics, and refining the research and development trends of the overview in the paper. The advantages of the large model are well reflected.

5. CONCLUSION

By reviewing traditional literature analysis tools, this paper proposes intelligent analysis of scientific and technological literature and discusses in detail this emerging research paradigm assisted by artificial intelligence. It reveals the great potential of intelligent analysis of scientific and technological literature in improving research efficiency and depth. The proposal of intelligent analysis of scientific and technological literature provides a new perspective for quantifying the correlation between literature and provides researchers with more precise guidance.

From the technical implementation level, the use of technologies such as natural language processing and machine learning provides favorable support for the intelligent analysis of scientific and technological literature. We have designed a proof-of-concept system, conducted an in-depth discussion on some core functions, and provided test and demonstration effects. In the future, with the continuous advancement of technology, intelligent analysis of scientific and technological literature is expected to achieve more complex and refined literature analysis, bringing revolutionary changes to scientific research. We need to continuously improve relevant technologies, strengthen human-computer collaboration, and actively respond to possible challenges. We believe that intelligent analysis of scientific and technological literature is a major change in the paradigm of scientific research. Through continuous exploration and innovation, we have reason to believe that intelligent analysis of scientific and technological literature will become an indispensable research tool for researchers and promote scientific research to new heights.

REFERENCES

[1]. George, P. M., & Robbins, K. (1994). Reference accuracy in the dermatologic literature. Journal of the American Academy of Dermatology, 31(1), 61-64.

[2]. Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El Profesional de la Información*, 29(1), e290103.

[3]. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arXiv preprint arXiv:2402.06196.

[4]. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 21-25) . IEEE.

[5]. Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. Computer Science Review, 29, 21-43.

[6]. Wang, D., & Zhang, M. (2021). Artificial intelligence in optical communications: from machine learning to deep learning. Frontiers in Communications and Networks, 2, 656786..

[7]. Chatterjee, N., & Mohan, S. (2008, February). Discovering word senses from text using random indexing. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 299-310). Berlin, Heidelberg: Springer Berlin Heidelberg.

[8]. Leon-Perez, J. M., Notelaers, G., & Leon-Rubio, J. M. (2016). Assessing the effectiveness of conflict management training in a health sector organization: evidence from subjective and objective indicators. European Journal of Work and Organizational Psychology, 25(1), 1-12.

[9]. Brown, PF, de Souza, PV, Mercer, RL, Della Pietra, VJ and Lai, JC (1992) Class-Based n- Gram Models of Natural Language. Computational Linguistics , 18, 467-479.

[10]. Keith, B., Fuentes, E. and Meneses, C. (2017) A Hybrid Approach for Sentiment Analysis Applied to Paper. Proceedings of ACM SIGKDD Conference , Halifax, 13-17 August 2017, 1-1011. Gupta, L., Jain, R., & Agrawal, S. (2020). A survey on 5G network: Architecture and emerging technologies. IEEE Access, 7, 75415-75443.

[11]. Knight, K. and Marcu, D. (2002) Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. Artificial Intelligence , 139, 91-107.

[12]. Huang Xiaoye. Trust construction in commercial transactions in the mobile Internet era: Based on an investigation of the J community WeChat group buying group[J]. Journal of China University of Mining and Technology (Social Sciences Edition), 2023, 25(3): 91-104.

[13]. Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, 1- 5 November 2016, 2383-2392

[14]. Huang Xiaoye. The development trend of foreign Internet trust research under the guidance of technical logic: Identification and visualization analysis based on CiteSpace[J]. Journal of Jiangnan University (Humanities and Social Sciences Edition), 2023, 22(2): 52-65.

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

ENSEMBLE-BASED MODEL FOR MITIGATING FEATURE DISCREPANCIES FOR ENHANCED THREAT DETECTION USING DOMAIN ADAPTATION

Joshua J. Tom[1], Pius U. Ejodamen[2], Taiwo Fele[3]*

[1]*Department of Mathematics and Computer Science, Elizade University, Ilara Mokin, Nigeria,*
[2]*Department of Computing Sciences, Admiralty University of Nigeria, Ibusa, Nigeria,*
[3]*Computer Science Department, School of Science and Computer Studies, The Federal Polytechnic, Ado-Ekiti, Nigeria,*

drtomjoshua@gmail.com, piusejodamen@adun.edu.ng, fele_ta@fedpolyado.edu.ng

ABSTRACT

In today's highly interconnected digital world, there are varieties of threat actors and threat types which necessitate a deep and robust threat detection system. Algorithms for detecting threats rely on various features of security data to identify potential threats. However, some threats are feature-dependent making it nontrivial to detect all types of threats using the same set of features in the dataset. Discrepancy in security telemetry datasets can be a potential cause of threat misclassification and consequently low threat detection system performance. In this paper, we propose an ensemble technique (Ensemble-DAFE) that integrates two techniques for mitigating feature discrepancy in security data viz domain adaptation (DA) and feature engineering (FE) techniques leveraging the strengths of the two to improve threat detection accuracy. We conducted experiments to determine the impact of feature discrepancies on threat detection performance. We obtained a threat detection performance accuracy of 99.96%. when the combined DA and FE was implemented compared to performance accuracy 96.38% without DA. Our result for Ensemble-DAFE with DA combined with FE outperforms state-of-the-art methods without DA compared to ours in terms of detection accuracy. We evaluate the effectiveness of our Ensemble-DAFE threat detection model using a synthetic dataset of network traffic with real-world security features. Based on the result, we noticed a 3.58% improvement in detection performance due to the integration of DA in the threat detection process and demonstrate its ability to reduce false negatives and false positives compared to individual feature-based detection methods.

Keywords: Feature Discrepancy, Ensemble Model, Domain Adaptation, Threat Detection,

1. INTRODUCTION

The cybersecurity arena has witnessed organizations generating huge amount of security telemetry data including data collected from firewalls, routers and switches, endpoint logs (system logs, application logs, antivirus logs), data from user activity logs, feeds from threat intelligence, cloud infrastructure and services logs and data gathered from security information and events management (SIEM), intrusion detection system (IDS) and intrusion prevention system IPS). These data are collected, analyzed, and monitored for security operations, swift response to cybersecurity incidence, and good organizations' posture. Security data is fundamental in shaping an organization's security

strategy in detecting and mitigating threat. Several threat detection algorithms have utilized different features extracted from these data to detect existential threats. However, there may be some discrepancies in the set of extracted data in the process of threat detection as different features do not have the same effectiveness given all threat types. This can have far-reaching negative consequences on a threat detection model such as reduction in model accuracy due to threat misclassification (false positives), fragmented view of security posture, integration problems arising from the need for unified analysis of data from multiple sources, poor and inefficient incident response. These have been drawbacks of many existing threat detection models. Due to the huge amount of security data generated from different sources, the constantly changing and complex nature of threats, and the dynamic threat landscape, there is the need to be able to adapt to new patterns and the continuously learn from the data for improve real-time detection performance [1]. Machine learning algorithm (MLA) is suitable in this scenario as it can handle vast amount of data in real-time detection prevent breaches and prevent damage. Because of this, most detection systems are based on machine learning, AI, deep learning, etc. for threat detection [2]. The use of ML in threat detection can help improve threat detection compared to traditional approaches such as signature-based detection [3]. These ML algorithms can be integrated with specialized techniques that mitigate the performance degrading effect of discrepancies in dataset [4]. There exist many techniques for handling data feature discrepancy including domain adaptation, feature engineering, feature transformation, multi-view learning, etc. [5].

As our contribution in mitigating this problem, we propose an ensemble-based model comprising extra tree classifier, gradient boosting classifier, and random forest models that mitigates discrepancy in security data features by combining multiple feature-based detection algorithms to achieve efficient threat detection through improved accuracy. We integrate domain adaptation with innovative feature engineering techniques to significantly improve our threat detection model's performance in identifying evolving threats. To the best of our knowledge, our work is the first attempt to combine two mitigating approaches against data feature discrepancy. The first approach adopts the domain adaption framework proposed in [6]. The second approach employs feature engineering technique to determine signatures of traffic data and enable our model to distinguish between normal traffic and threats. In this paper, we have come up with a framework for developing a hybrid model capable of addressing data feature discrepancies in threat detection.

2. THEORETICAL FRAMEWORK

The fast-evolving cyber threats landscape has made it mandatory for organizations to pay specific attention to the development of efficient threat detection systems to identify and mitigate risks in real-time. This is because of increased sophistication in cyber threats dynamics and advancement in development tools [7]. The focus on ML techniques in threat detection has been intensified by the inadequacy of traditional rule-based detection systems. The adoption of Machine learning models in different fields have shown great results in terms of threat detection automation by learning patterns from vast amount security datasets. However, a significant problem to grapple with is data feature discrepancy during training [8]. Some key aspects of feature discrepancy in security telemetry data collected from various security events or logs include inconsistency in naming, differences in data types, missing features, varying feature availability, data granularity, temporal discrepancies, feature redundancy, etc. The presence of these feature discrepancies in datasets arises due to differences in the way in which features are distributed between source and target domains. As such, machine learning models must learn to adapt to a target domain different from a source domain in which it was trained [5].

Domain Adaptation (DA) is a desirable and an important component of a machine learning model whose major purpose is to improve the performance of models trained in one environment (source domain) and tested in a new environment (target domain) [9]. With regards to threat detection, domain adaptation is vital because of the rapidly evolving cyber threats, which changes rapidly and are characterized by varied behaviors. Different organizations require different approach to detecting threats in their digital life but the narrative can change when threat detection models are design to be able to adapt to changing environments such that a model trained on benign network traffic exhibits accurate threat detection capability and not struggle in a strange environment with different features like attack vectors, user behavior patterns, and network architecture. Therefore, to handle any disparity in security data features, domain adaptation techniques are used to bridge the gap between the source and target domains making threat detection models to perform optimally in the face of new and unfamiliar threats.

Domain adaptation mainly consists of a range of methodologies including instance re-weighting, feature transformation-based Maximum Mean Discrepancy, adversarial training-based Generative Adversarial Networks (GAN), and domain-invariant feature learning [10]. These techniques are often engaged to reduce the divergence between the source and target domain distributions. Furthermore, Generalized Adaptive Models (GMAs) have been recently been used in designing various machine learning based models to bring in flexibility to cater for variations in data features [11]. This approach is principally to adapt the feature engineering architecture and learning mechanisms to specific feature distribution characteristics. Figure 1 shows a general conceptual framework of discriminative cross domain adaption.

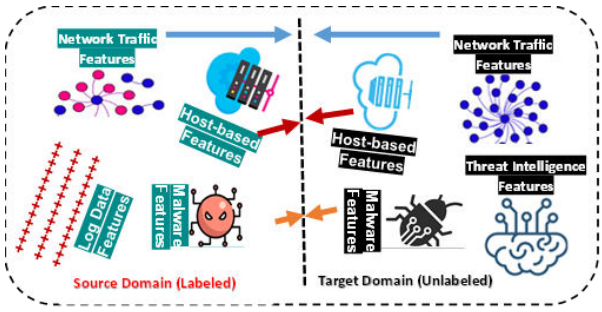

Figure 1: Illustrating Cross Domain Adaptation to Learn Discriminative Information by Mitigating Domain Discrepancy.

Feature engineering forms the foundation of machine learning models for effective threat detection [12]. In this approach, relevant features are selected or modified from the raw data that would enhance the model's understanding of the patterns that may indicate potential threats. Some important aspects of feature engineering in detecting threats are: eliciting specific domain for hunting threats e.g. network traffic or transaction data; the type of features e.g. raw features such as user IDs, timestamps, IP addresses, etc., behavioral features and contextual features; temporal features; statistical features; anomaly detection features, amongst others. For threat detection models to perform optimally, security data features must be carefully and correctly selected to ensure better performance. With the dynamic nature of threats, feature refinement and adaptation are key to combat emerging threats and contend the unpredictability in attacker behaviors [13]. Many research works have been conducted deploying other techniques for handling discrepancies in data features in dataset meant for machine learning analytics including feature alignment, data augmentation, bias correction, ensemble and multimodal approaches, etc.

Furthermore, the potentials of machine learning models in threat detection and domain adaptation cannot be overemphasized. Machine learning algorithms (neural network architectures or ensemble-based learning models) intrinsically providing mechanisms to take care of possible variances in domains can be a better option in handling feature discrepancies [14]. Studies have shown that advancements techniques like transfer learning and adversarial training are good candidates for improving the robustness of machine learning models against domain shifts [15]. Individual machine learning algorithm perform best given different aspects and scenarios. To leverage the diverse strengths of these model we need to bring them into a single framework, ensemble model. In the context of threat detection, the diversified characteristic of ensemble models can capture varying patterns and nuances in datasets leading to higher performance in threat detection [16]. Mostly use ensemble framework include those that use Decision Trees, Support Vector Machines, Neural Networks, Logistic Regression, K-Nearest Neighbor, Gradient Boosting Machines, etc. as the base learners for classification and regression tasks. From the foregoing, the deployment of domain adaptation, feature engineering, and machine learning models is expected to bring some level of innovation and novelty in the context of cybersecurity and can offer the much-envisaged real-time response and adaptability in tackling the rampaging and evolving threats. As organizations continuously grapple with different and sophisticated attack vectors, it is important to consider the design of robust and efficient real-time threat detection systems that can adapt to new environments and changing threat characteristics. This paper aims to model a threat detection system by intersecting domain adaptation, a subset of transfer learning, feature engineering technique and ensembled machine learning models for efficient threat detection. We carry out a systematic review of existing works, methodologies and frameworks, analyze the impact of feature discrepancy on threat detection model, and propose an ensemble-based model [17] integrated with innovative techniques described above to leverage the techniques' individual strengths.

3. LITERATURE REVIEW

3.1 Related Work

There have been several challenges posed by data feature discrepancies for machine learning models, prominent among them is low model performance. In recent times, extensive studies have been carried out by many researchers to solve the problem of feature

discrepancies in machine learning where numerous techniques including ensemble methods, domain adaptation, feature engineering, etc. have proposed multiple techniques, including to mitigate the adverse effects of such discrepancies. In this section, we review a handful of related works in this direction. One of the renowned techniques for improving the accuracy of machine learning models in the face of feature discrepancies in Ensemble Models. Here, techniques such as bagging and boosting [18] have been applied in effectively reducing feature variance and domain shift in datasets. Recently, some researchers have built models on different feature subsets using some sort machine learning model aggregation strategies to achieve robustness against feature discrepancies [19]. [20] proposed a hybrid feature selection with an ensemble classifier to select relevant features and provides consistent attack classification. To achieve effective threat detection, they used CfsSubsetEval, genetic search, and a rule-based engine to effectively select subsets of features with high correlation. They claimed that their model drastically reduces False Alarm Rate (FAR). However, there are specific limitations occasioned by individual techniques deployed by the authors which might degrade the model's performance. The CfsSubsetEval method is sensitive to data distribution, assumes linear relationships between features and ignores temporal variations in data. The genetic search component can introduce the risks of overfitting and parameter sensitivity following genetic algorithm's reliance on parameter tuning [21]. Rule-based feature selection is existentially based on heuristics to select features. This approach has notable drawbacks including lack of adaptability, domain knowledge dependent, limitation in caring for interactions, and there is the potential for bias [22]. [23] introduced a solution to the problem of botnet detection where they used a hybridized feature selection approach (consisting of Categorical Analysis, Mutual Information, and Principal Component Analysis) to enhance the detection capabilities of the ensemble learners. For the ensemble technique, Extra trees was used to help in adapting the detection model to new botnet threats. While the used of the comprehensive feature selection method offers some gains, there are individual feature selection disadvantages which might adversely affect the model's performance. For instance, Categorical Analysis may not be appropriate for continuous variables except the variables are discretized while for Principal Component Analysis method there is the problem of loss of interpretability, assumption of linear relationship among features, and is sensitive to scaling [24]. [25] presented a review on feature selection and ensemble techniques used in anomaly-based IDS research. The paper categorized feature selection techniques to determine individual technique's effectiveness on machine learning-based threat detection models during training and detection phases and concluded that selection of most relevant features in a dataset increases the efficiency of detection in terms of accuracy of the model. They also focused on ensemble techniques employed in anomaly-based IDS models and illustrated how this technique improves the performance of the anomaly-based IDS models. To offer significant improvements in existing intrusion detection systems (IDS) in the Internet of Things domain, [26] proposed an ensemble-based intrusion detection model using logistic regression, naive Bayes, and decision tree as the machine learning algorithms deployed with a voting classifier. The evaluated their model's performance with some existing state-of-the-art techniques using the CICIDS2017 dataset and the result showed significant improvement in terms of accuracy as compared to existing models in terms of both binary and multi-class classification scenarios. However, a stacked ensemble used in the model can make the training of the model very complex and slow since the individual models requires separate training. An existential ensemble model such as Extra Trees (Extremely Randomized Trees) could give a good balance in the bias-variance trade-off by introducing randomness offering robust performance across various datasets. In most cases, the selection of the base models to form stacked ensemble machine learning model in order to gain accuracy with neural models trained in detecting novel threats is a nontrivial problem. [27] presented a novel method named PANACEA to detect cyber-threat, by integrating ensemble learning with adversarial training. The main objective of their work was to enhance the accuracy of neural models addressing threat challenges by creating an ensemble consisting of different base models. The study focused significantly on the selection and pruning of these base models using eXplainable AI (XAI) to improve diversity and improve the accuracy of the ensemble model. They evaluated how different models respond to various input feature subspaces, and used the result of the evaluation to refine the training process, and targeted the models' performances in identifying diverse attack patterns. They conducted empirical validation on several benchmark datasets and results show that the combination of adversarial training, ensemble learning, and XAI techniques was effective in improving multiclass classification accuracy in the datasets. Since the proposed model was based on deep learning, there is

going to be issues of scalability with regard to computational resources for training the model. Wireless Sensor Networks (WSNs) is no doubt backbone of Internet of Things (IoT) and require adequate threat detection strategy. [28] presented a first of its kind method code named Weighted Score Selector (WSS) using ensemble-based machine learning (ML) techniques aimed at improving the detection of threats in WSNs. Their choice of a machine learning approach to the solution is a wise choice because of the recent adoption of this approach in threat detection especially for real-time threat monitoring. The WSS model uses a combination of machine learning classifiers utilizing the strengths of the prominent ones during threat detection in improving the overall performance of the model. Compared with traditional ensemble techniques such as Boosting, Bagging, and Stacking, WSS substantiates the position of the authors in terms of the model's effectiveness in threat detection. Recently, researches have been conducted which integrate ensemble learning techniques for the design of threat detection solutions to handle feature discrepancies. [29] propose a binary classifier approach developed from a machine learning ensemble method to filter and dump malicious traffic to prevent malicious actors from accessing the IoT network and its peripherals. They employed the gradient boosting machine (GBM) ensemble approach to train the binary classifier using pre-processed recorded data packets to detect the anomaly and prevent the IoT networks from zero-day attacks. [30] proposed a domain adaptive ensemble learning (DAEL) framework to address both unified domain adaptation (UDA) and domain generalization (DG) problems. The proposed framework consisted of one CNN feature extractor shared across domains and multiple classifiers with each classifier trained to specialize in a particular source domain thereby acting as an expert in its own domain. Overall, these experts in the DAEL framework collaboratively forms an ensemble and learns complementary information from each other to be efficient in an unfamiliar domain. Under this arrangement, one classifier's source domain becomes another classifier's target domain, which can actively check feature discrepancy across domains. The authors experimented their model on three multi-source UDA datasets and two DG datasets and the results show significant improvement in the state of the art on both problems. This approach leverages the diversity of predictions from multiple models while simultaneously addressing feature discrepancies during training. In [31], a novel approach to solve the problem of multisource domain adaptation (MDA) was proposed using Dual-Level Alignment Network with Ensemble Learning (DANE). In particular, the issue of intradomain and interdomain shifts that hinder knowledge transfer from multiple labeled source domains to an unlabeled target domain was addressed. The objective of the paper was to enhance the performance of the classifier(s) by effectively using knowledge present in the source domains. [32] presented a timely and important contribution to the field of malware detection by proposing an unsupervised domain optimization (UDA)-based malware detection method. It addressed the challenge of detecting of known and unknown malware, hence earning to reduce source (labeled) and target domain (unlabeled) using the distribution divergence between the source and target domain which is minimized with the help of symmetric adversarial learning. The traditional approach often falters in the face of rapidly changing cyber threats. The use of two public datasets in the evaluation enhances the reliability of the proposed method. They found that an accuracy rate of 95.63% was impressive in detecting unknown malicious code and indicates that UDA-based approaches are effective in a variety of situations. Although the proposed method shows promising results, the complexity of implementing symmetric adversarial learning poses challenges for practitioners.

3.2 Research Gap/Problem Statement

Despite the fact that existing threat detection systems are dependent on various algorithms that utilize specific features of security data to identify latent threats, there is significant unattended challenges posed by feature discrepancies across datasets. State-of-the-arts approaches fail to consider the non-homogeneity in the number and types of features present in different security datasets, which is capable of misclassifying threats and diminishing threat detection performance. Furthermore, a number existing methodologies do not adequately leverage the strength in combining domain adaptation (DA) and feature engineering (FE) techniques but often focus on either of these techniques in isolation. Hence, the complementary benefits of these approaches are never harnessed. While earlier results from studies in this direction suggest that ensemble techniques have the tendency to improve detection accuracy, an all-inclusive study that systematically assess the impact of feature discrepancies on the overall effectiveness of threat detection systems is lacking. Therefore, the purpose of this study is to fill the identified gap by proposing the Ensemble-DAFE method, which incorporates the DA and FE techniques to establish a more robust framework for threat detection in security.

3.3 Methodology

This section outlines the hybrid methodology employed to address data feature discrepancies and the ensemble machine learning approach to ensure efficient real-time threat detection. Our approach integrates ensemble methods, domain adaptation techniques, and a systematic process for feature engineering and selection.

We structure our methodology in five phases. Phase 1 handles data collection and preprocessing. In phase 2, we conduct feature engineering and selection. The other phases include domain adaptation phase, model development phase, and lastly the evaluation phase. We first all give the architecture of the proposed model as shown in figure 2.

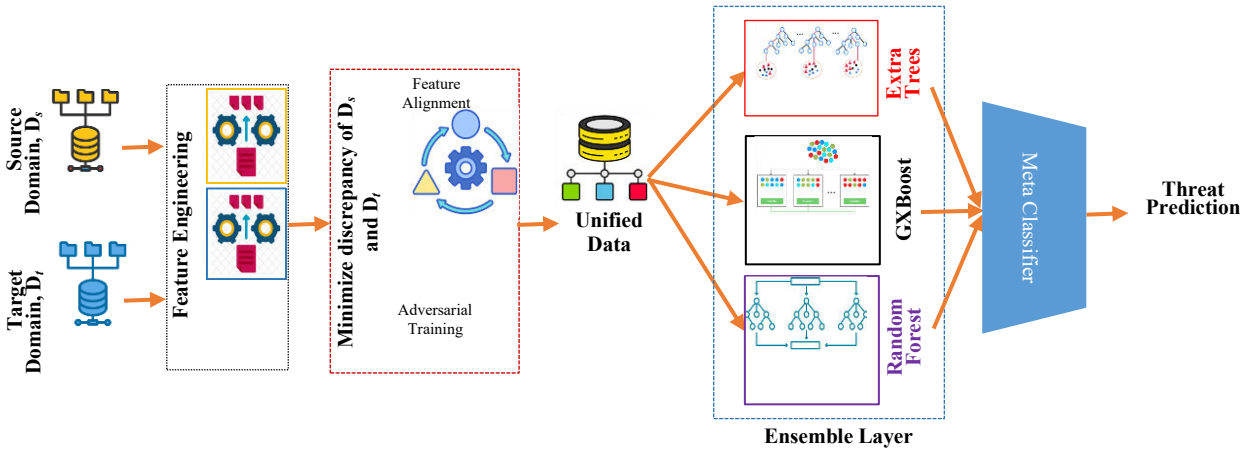

Figure 2: Architecture of Proposed Ensemble-DAFE model for with Feature Discrepancy Mitigation

i. Phase 1: Data Collection and Preprocessing

Dataset

For this experiment, we generated synthetic security dataset code named SYNCyberNet Dataset in Python environment using faker, pandas and NumPy libraries. A base structure of generated synthetic (SYNCyberNet) dataset is defined and comprised of 200,000 rows with 25 security features including Threat_ID, Timestamp, Destination_IP, Source_IP, Byte_Count, Protocol, Port, Attack_Vector, User_Agent, Session_Duration, and Malware_Detected, etc. For simplicity, we labelled 20% of the records as attack signifying that a threat is detected and 80% labeled as normal signifying no threat.

Though we understand that synthetic data cannot replace real world dataset and despite the limitations of synthetic datasets including possible domain gap, overfitting to synthetic data, lack of realism, etc., there are several reasons to justify our use of a synthetic dataset to validate the model proposed in this work. Some of the justifications include:

(i) Using synthetic data in validating our model offers us precise control over the characteristics of the dataset. With this, we are able to control how features are distributed, noise level, etc.

(ii) We also opted to use synthetic data to create data with subtle anomalies or with highly correlated feature combinations. This allows us to perform targeted testing to determine model robustness.

(iii) Using synthetic data ensures repeatable validation experiments due to the deterministic nature of the data generation. This allows varied model versions performance to be consistently compared.

(iv) Real world datasets do not address latest attacks and emerging threats. Synthetic data allows simulating threats not yet observed in real world data, enabling the model to be ready for evolving threats.

Defining Source and Target Domains

To create the source domain, we specified the SYNCyberNet Dataset as the source domain and generate the target domain from the source domain data by introducing domain specific variations in the datasets and simulate discrepancies in the target domain data using 6 of the features including Byte_count, Is_Vulnerable, Threat_Level, Protocol, User_role, and Attack_Vector to generate the target domain data with no labels. In this paper, we model the source domain data and target domain data using equation 1 and equation 2 respectively:

$$D_s = \{(\mathrm{x}_i^s, y_i^s)\}_{i=1}^{n_s} \quad (1)$$

$$D_t = \{(\mathrm{x}_j^t)\}_{j=1}^{n_t} \quad (2)$$

where D_s and D_t represent source domain with labels and target domain without labels and the two have different distributions, n_s and $n_t$denote the number of threats in D_s and D_t respectively, and y_i^s represents the label of sample x_i^s in D_s.

Feature Discrepancy Analysis

We specified a Python code with libraries such as seaborn, matplotlib with Numpy, and pandas to carry out correlation analysis on the source and target domain datasets using heatmaps to visualized the resulting correlation matrices to understand and access the complex relationship across the two domains. This step is playing a very prominent part in preemptive feature engineering and selection. Correlation matrices for source domain data and target domain data are shown in figures 3.

ii. Phase 2: Feature Engineering and Selection

We conducted feature engineering to create new features by combining the source and target domains based on domain knowledge resulting a unified domain saved in a csv files (unified_dataset.csv). to achieve this, we specified a Python code that import category_encoders to implement Target Categorical Encoding. The source and target domain files were loaded into pandas DataFrame and used to create combined features from selected features such as Byte_Count', 'Attack_Vector', 'Is_Vulnerable', 'Event_Type', 'Threat_Level', and 'User_Role'. We setup sets up a Target Encoder to convert categorical variables into numerical values. Finally, the DataFrame with the new features and encoded values were saved as unified_dataset.csv. Figure 4 captures the output of our Python code for feature engineering showing 5 rows of the combined features.

```
Python 3.12.4 | packaged by Anaconda, Inc. | (main, Jun 18 2024, 15:03:56) [MSC
v.1929 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 8.25.0 -- An enhanced Interactive Python.

Restarting kernel...

In [1]: runfile('C:/Users/Dr. Joshua Tom/combine_features.py', wdir='C:/Users/
Dr. Joshua Tom')
   Byte_Count      Attack_Vector  ...  High_Byte_Vulnerability User_Role_Threat
0        6083            Malware  ...                        1        Admin_Low
1        8286  Denial of Service  ...                        1     Admin_Medium
2        8252           Phishing  ...                        0      User_Medium
3        1529  Network Intrusion  ...                        0     Guest_Medium
4        8090  Network Intrusion  ...                        1     Guest_Medium

[5 rows x 9 columns]
Feature engineering and encoding completed successfully.

In [2]:
```

Fig. 4: Feature Engineering using Target Categorical Encoding using Category Encoders for Verification

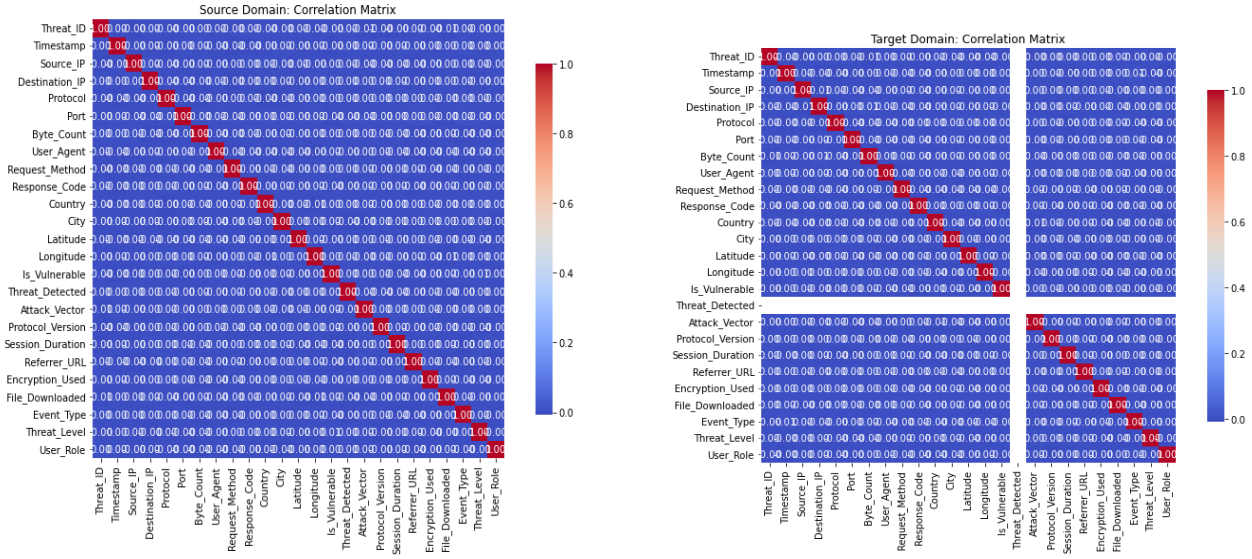

Figure 3: Correlation Matrices for Source and Target Domain

Fig. 3a: Correlation Matrice for Source Domain *Fig. 3b: Correlation Matrice for Target Domain*

We use two (2) methods for feature selection: filter and wrapper to select the associated features, and remove irrelevant or redundant features. We selected features using the filter method (ANOVA F-value as score function). Recursive Feature Elimination (RFE) was the Wrapper Method used.

3.4 Propose Ensemble Model

The ensemble technique creates a unified model by combining multiple feature-based detection methods that share the characteristics of voting. Every feature-based detection method is trained on different dataset and detects a specific threat class.

iii. Phase 3: Cross Domain Adaptation

There are four techniques for adapting a model to target domain. These include pretrained model fine-tuning, adversarial training, self-training, and feature alignment. To address discrepancies and improve model transferability across the source and target domains, we used Feature Alignment technique for domain adaptation known as Maximum Mean Discrepancy, which is based on feature transformation. Maximum mean discrepancy (MMD) is generally a class of nonparametric two-sample tests

aimed at maximizing the mean difference between samples from the source domain, X to the target domain, Y over all choices of data transformations living in some function space [33].

Detailed Implementation of MMD in a Domain Adaptation Setting

We implemented domain adaptation using MMD as follows:

We passed both source and target data through a CNN feature extractor to obtain feature representations for both domains. We choose the Gaussian (RBF) kernel function, $k(x, x')$ for MMD to map the data into a higher dimensional space as given below:

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (3)$$

where x and $x'$are feature vectors and σ represents the kernel bandwidth.

The MMD loss is calculated as the square of the distance between the mean embeddings of the source and target distributions in the reproducing kernel Hilbert space (RKHS) and is given as:

$$\mathrm{MMD}(\mathrm{P}, \mathrm{Q}) = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s} k(x_i, .) - \frac{1}{n_t}\sum_{i=1}^{n_s} k(y_j, .) \right\|_H^2 \quad (4)$$

where P and Q are the source and target distributions, x_i and x_i are samples from the source domain and target domain respectively, n_s and n_t are the number of source and target samples respectively and H is the Hilbert space. Equation 4 can be evaluated as:

$$\begin{aligned}\mathrm{MMD}(\mathrm{P}, \mathrm{Q}) = &\frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k(\mathrm{x}_i, \mathrm{x}_j) \\ &+ \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k(\mathrm{y}_i, \mathrm{y}_j) \\ &- \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(x_i, \mathrm{y}_j)\end{aligned} \quad (5)$$

where x_i's are the source domain data points and x_j's are the target domain data points. k is the *pairwise* radial basis function-based kernel provided in the scikit-learn library to calculate the radial basis function (RBF) kernel between pairs of samples in the source and target domain data. This allowed us to determine whether there is a serious feature discrepancy across the source and target data. In this experiment, the threshold for acceptable MMD value was set at 0.15. Firstly, our MMD value was 0.1811 suggesting that we have detected high discrepancy. Therefore, based on this value, domain adaptation using CCA is implemented which tries to overcome or minimize the computed discrepancy. The reduced MMD value of 0.0172, following the CCA minimization process confirms that there is considerable success in reducing initial feature discrepancy within data structure. Figure 5 presents a visualization of source and target domain (left) with transformed sources/targets after CCA minimization applied (right). Unsurprisingly, this decreased difference allowed the threat detection model to operate better. Finally, we minimize the MMD loss by adding the MMD loss as a regularization term to the overall loss function during training, which makes our proposed model to learn to minimize both the classification/regression loss on the labeled source data and the discrepancy between the source and target feature distributions.

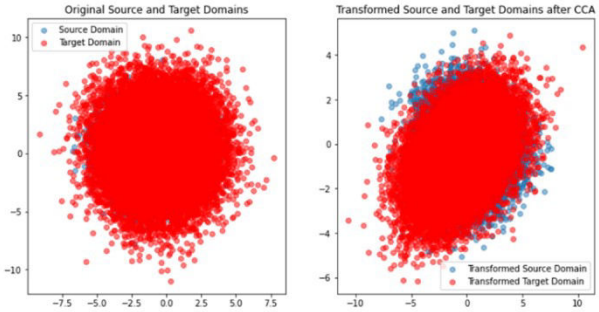

Figure 5: Visualization of the Original Source and Target Domain and the Transformed Source

iv. Phase 4: Model Development

Ensemble Learning Approach

An ensemble of multiple machine learning models made up of RandomForestClassifier, ExtraTreesClassifier, and XGBClassifier was constructed to leverage the strengths of different algorithms with Logistic Regression Classifier as the meta classifier. The base models were selected to form the ensemble model based on the respective strengths and capabilities. The Extra Trees were chosen because of its power of interpretability and effectiveness in capturing non-linear relationships. Random Forests was included to enhance robustness and prevent overfitting while offering improved performance across varying datasets. XGboost classifier was considered to provide for optimization of loss functions using its sequential learning ability. To build the ensemble model, we used StackingClassifier ensemble learning method from the scikit-learn library. We specified the StackingClassifier's initial estimators as the base models and the final estimator as the meta classifier.

Model Training and Tuning

For the training of an already developed model (Ensemble-DAFE), both the grid search and cross validation were used during hyperparameter

optimization so as to obtain the configurations of models to be incorporated in the final ensemble. Two of the source domains were set up for training and validation. The validation data set was 20% of the total source domain while training data set was 80%. In order to build a prediction model for the source domain data, each of the base models was trained and tested separately. All base models' predictions were implemented into the Logistic regression based stacked ensemble model, which was trained and tested as well. The final prognosis was based on the outputs of each model through a process known as majority voting to improve confidence and precision. Their performances in terms of model accuracy are given in table 1.

Table 1: Performances of the Base Models and the Stacked Ensemble Model before minimizing feature Discrepancy

Model Name	Model Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	93.18%	0.9441	0.9193	0.9316
Extra Trees Classifier	94.16%	0.9490	0.9245	0.9366
XGBoost Classifier	94.24%	0.9507	0.9340	0.9423
Proposed Ensemble Model	96.38%	0.9512	0.9368	0.9940

v. Phase 5: Evaluation

Evaluation Metrics

A detailed evaluation of model performance was carried out which included parameters like Accuracy, Precision, Recall, and F1 Score. The values obtained for the different metrics are as given in table 1. In this paper, we provided for model evaluation to be computed through the Accuracy, Precision, Recall and F1 Score as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where Accuracy is a measure of the ratio of correctly classified samples to the total number of samples. TP is true positives, TN true negatives, FP false positives and FN represents false negatives.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

where Precision is the ratio of positive samples to forecasted positive samples.

$$\text{Recall} = \frac{TP}{TP + FN} \quad 6)$$

where Recall refers to the ratio of predicted positive samples against actual positive samples.

$$\text{F1–score} = \frac{Precision \text{ x Recall}}{Precision + \text{Recall}} \quad (7)$$

where F1-score represents a weighted average of Precision and Recall to balance the effects of accuracy and recall, and to evaluate a classifier comprehensively. Targets were achieved after resolving the feature discrepancies through application of CCA as explained in the Domain Adaptation section above, we obtained the transformed data (figure 5b), and re-trained and re-evaluated the base and stacked ensemble models on the transformed dataset. The outcome of the re-training and re-evaluation is presented in table 2.

Table 2: Performances of the Base Models and the Stacked Ensemble Model on the feature Discrepancy Minimized Dataset

Model Name	Model Accuracy	Propose Ensemble model accuracy
Random Forest Classifier	96.87%	99.96%
Extra Trees Classifier	98.27%	
XGBoost Classifier	98.89%	

4. COMPARATIVE ANALYSIS

The performance of the proposed hybrid model was compared against baseline models trained solely on the target domain and models without domain adaptation. The study in [26] designed an ensemble-based intrusion detection model using logistic regression, naive Bayes, and decision tree as the machine learning algorithms and obtained an average accuracy of 88.94% without implementing domain adaptation for feature discrepancy handling. The proposed model integrated with domain adaptation for feature discrepancy handling outperforms theirs with 99.96 performance accuracy. [29] proposed an anomaly-based intrusion detection system for security against DDoS using gradient boosting machine ensemble classifier with no domain adaptation capability, which performed well for binary classification yielding accuracy of 98.27%. Compared to this, the proposed approach outperformed the one proposed in [29] even with high computational requirement of our model. The work in [32] used symmetric adversarial learning to minimize the distribution divergence between the source and target domain in their proposed unsupervised domain adaptation (UDA)-based malware detection method and obtained a performance accuracy of 95.63%,

which clearly shows that our method outperforms theirs in threat detection accuracy

5. DISCUSSION ON THE RESULTS

The results of the experiments conducted show that the ensemble model developed in this paper, Ensemble-DAFE outperforms individual base detection methods in terms of accuracy and reduced false negatives and false positives. We specifically noticed that our ensemble technique achieves an F1-score of 0.9940 against the highest F1-score of 0.9423 exhibited by the XGBoost classifier as the best individual model. Our ensemble model achieves a performance improvement of 2.51% in detection accuracy due to feature discrepancy minimization. As a result of this discrepancy handling, the developed ensemble technique reduced false negatives by 6% and false positives by 10% compared to individual models used in the ensemble.

6. CONCLUSION

Ensemble methods have been shown to effectively resolve the issue of feature disparity in threat detection algorithms. Improving more than one feature-based detection strategy concurrently allows us to increase the efficiency of detecting a threat while minimizing both false positives and negatives. Our method can be used for different ranges of threat detection systems such as network behavior analysis, system log analysis, or user actions analysis. The method described in this paper for improving threat detection performance by addressing feature differences, between source and target domains in security datasets involves creating a model that can manage data feature variations in threat detection effectively. Stacked ensemble, with domain adaptation and thorough feature engineering and selection processes can improve both detection accuracy and overall performance in a domain setting. The literature review conducted in this paper highlights the increasing importance of dealing with data feature differences to maintain the efficiency of ensemble models. Researchers have been working on improving the performance and reliability of machine learning systems by using domain adaptation techniques and experimenting with methods, across different applications and datasets successfully over the years. In a nutshell although there has been advancement in handling variations in data features using these methods, the intricate nature of real-world situations calls for ongoing exploration of combined and flexible approaches to further boost both model reliability and effectiveness. In the future, we intend to evaluate the performance of our stacked ensemble model on larger and real-world datasets. This paper suggests further investigation into the implications of increasing the number of base models in a given ensemble architecture on the detection performance of a threat detection model. It is also important to explore and determine the appropriateness of other ensemble methods such as bagging and boosting in future threat detection solutions.

REFERENCES

[1] J. Lan X. Liu B. Li et al. A Novel Hierarchical Attention-based Triplet Network with Unsupervised Domain Adaptation for Network Intrusion Detection. Appl Intell 53, 11705–11726, 2023.

[2] S. Sharma and N. S. Yadav. Ensemble-based Machine Learning Techniques for Attack Detection, 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, pp. 1-6, 2021.

[3] S. Dandyala and S. Banik Traditional Methods of Threat Detection. International Journal of Advanced Engineering Technologies and Innovations, 1(2), 161-177, 2021.

[4] L. L. Scientific Ensemble Machine Learning Algorithm Methods for Detecting the Attacks Using Intrusion Detection System. Journal of Theoretical and Applied Information Technology, 102(5), 2024.

[5] F. Khani and P. Liang. Feature Noise Induces Loss Discrepancy Across Groups. In International Conference on Machine Learning (pp. 5209-5219). PMLR, 2020.

[6] J. Yan R. Sun T. Liu S. Duan. Domain-adaptation-based active ensemble learning for improving chemical sensor array performance. Sensors and Actuators A: Physical, 357, 114411, 2023.

[7] T. Zaid and S. Garai. Emerging Trends in Cybersecurity: A Holistic View on Current Threats, Assessing Solutions, and Pioneering New Frontiers. Blockchain in Healthcare Today, 7, 2024.

[8] Dou Y, Yang H, Yang M, Xu Y, Ke D. Dynamically mitigating data discrepancy with balanced focal loss for replay attack detection. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4115-4122),

2021. IEEE.

[9] S. Zhao G. Wang S. Zhang Y. Gu Y. Li Z. Song K. Keutzer. Multi-Source Distilling Domain Adaptation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 12975-12983), 2020.

[10] P. Singhal R. Walambe S. Ramanna K. Kotecha. Domain Adaptation: Challenges, Methods, Datasets, and Applications. IEEE Access, 11, 6973-7020, 2023.

[11] I. Karna A. Madam C. Deokule R. Adhao V. Pachghare. Ensemble-based Filter Feature Selection Technique for Building Flow-based IDS. In 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS) (pp. 324-328), 2021. IEEE.

[12] P. R. Kothamali S. Banik S. V. Nadimpalli. Feature Engineering for Effective Threat Detection. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence, 12(1), 341-358, 2021.

[13] S. Zhou D. Chen J. Pan J. Shi J. Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2952-2963), 2024.

[14] H. Do Hoang T. B. Xuan T. N. N. Minh P. T. Duy V. H. Pham. DA-GAN: Domain Adaptation for Generative Adversarial Networks-assisted Cyber Threat Detection. In 2022 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 29-34), 2022. IEEE.

[15] A. S. Li, A. Iyengar A. Kundu E. Bertino. Transfer Learning for Security: Challenges and Future Directions. arXiv preprint arXiv:2403.00935, 2024.

[16] A. Odeh and A. Abu Taleb. Ensemble-Based Deep Learning Models for Enhancing IoT Intrusion Detection. Applied Sciences, 13(21), 11985, 2023.

[17] N.I. Haque, M.A Rahman, H. Shahriar. Ensemble-based efficient anomaly detection for smart building control systems. In 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 504-513), 2021. IEEE.

[18] A. Mosavi F. H. Sajedi, B. Choubin M. Goodarzi A. A. Dineva E. Rafiei Sardooi. Ensemble boosting and bagging based machine learning models for groundwater potential prediction. Water Resources Management, 35, 23-37, 2021.

[19] M. Moshawrab M. Adda A. Bouzouane H. Ibrahim A. Raad. Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. Electronics, 12(10), 2287, 2023.

[20] E. Jaw and X. Wang. Feature selection and ensemble-based intrusion detection system: an efficient and comprehensive approach. Symmetry, 13(10), 1764, 2021.

[21] M. Mosayebi and M. Sodhi. Tuning genetic algorithm parameters using design of experiments. In Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion (pp. 1937-1944), 2020.

[22] F. Chiroma M. Cocea H. Liu. Evaluation of rule-based learning and feature selection approaches for classification. In 7th Imperial College Computing Student Workshop (pp. 1-6). Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2019.

[23] M.A. Hossain, M.S. Islam. A novel hybrid feature selection and ensemble-based machine learning approach for botnet detection. Sci Rep 13, 21207. https://doi.org/10.1038/s41598-023-48230-1, 2023.

[24] C. Peng Y. Chen Z. Kang C. Chen Q. Cheng. Robust principal component analysis: A factorization-based approach with linear complexity. Information Sciences, 513, 581-599, 2020.

[25] M. Torabi N.I. Udzir M.T. Abdullah R. Yaakob. A Review on Feature Selection and Ensemble Techniques for Intrusion Detection System. International Journal of Advanced Computer Science and Applications, 12, 2021.

[26] A. Abbas M.A. Khan S. Latif et al. A New Ensemble-Based Intrusion Detection System for Internet of Things. Arab J Sci Eng 47, 1805–1819. https://doi.org/10.1007/s13369-021-06086-5, 2022.

[27] M. AL-Essa G. Andresini A. Appice et al. PANACEA: a neural model ensemble for cyber-threat detection. Mach Learn 113,

5379–5422. https://doi.org/10.1007/s10994-023-06470-2, 2024.

[28] S. Ismail Z. El Mrabet H. Reza. An Ensemble-Based Machine Learning Approach for Cyber-Attacks Detection in Wireless Sensor Networks. Applied Sciences. 13(1):30. 2023.

[29] P. Verma A. Dumka R. Singh A. Ashok A. Gehlot PK Malik M. Hedabou. A novel intrusion detection approach using machine learning ensemble for IoT environments. Applied Sciences, 11(21), 10268, 2021.

[30] K. Zhou Y. Yang Y. Qiao T. Xiang. Domain adaptive ensemble learning. IEEE Transactions on Image Processing, 30, 8008-8018, 2021.

[31] Y. Yang L. Wen P. Zeng B. Yan Y. Wang. DANE: A Dual-level Alignment Network with Ensemble Learning for Multi-Source Domain Adaptation. IEEE Transactions on Instrumentation and Measurement, 2024.

[32] F. Wang G. Chai Q. Li C. Wang. An efficient deep unsupervised domain adaptation for unknown malware detection. Symmetry, 14(2), 296, 2022.

[33] S. Paik, M. Celentano, A. Green, R. J. Tibshirani. Maximum mean discrepancy meets neural networks: The radon-kolmogorov-smirnov test. arXiv preprint arXiv:2309.02422, 2023.

[34]