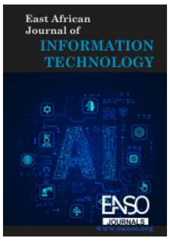


East African Journal of Information Technology
eajit.eanso.org
Volume 7, Issue 1, 2024
Print ISSN: 2707-5346 | Online ISSN: 2707-5354
Title DOI: https://doi.org/10.37284/2707-5354

EAST AFRICAN NATURE & SCIENCE ORGANIZATION

Original Article

Improving Network Security: An Intrusion Detection System (IDS) Dataset from Higher Learning Institutions, Mbeya University of Science and Technology (MUST), Tanzania

Daud M. Sindika[1], Dr. Mrindoko R. Nicholaus, PhD[1] & Dr. Nabahani B. Hamadi, PhD[1]*

[1] Mbeya University of Science and Technology, P. O. Box 131 Mbeya, Tanzania.
* Correspondence email: daudi.sindika@must.ac.tz.

Article DOI: https://doi.org/10.37284/eajit.7.1.1679

Date Published:

07 January 2024

Keywords:

HLIs,
IDS,
Cyber Threat,
ML,
Algorithm

ABSTRACT

Nowadays, Internet-driven culture securing computer networks in Higher Learning Institutions (HLIs) has become a major responsibility. Intrusion Detection Systems (IDS) are crucial for protecting networks from unauthorized activity and cyber threats. This paper examines the process of improving network security by creating a comprehensive IDS dataset using real traffic from HLIs, highlighting the importance of accurate and representative data in improving the system's ability to identify and mitigate future cyber-attacks. The IDS model was created using a variety of machine learning (ML) techniques. Metrics like accuracy, precision, recall, and F1-score were used to assess the performance of each model. The dataset used for training and testing was real-world network traffic data obtained from the institution's computer network. The results showed that the developed IDS obtained exceptional accuracy rates, with Random Forest, Gradient Boosting, and XGBoost models all achieving an accuracy of around 93%. Precision and recall values were likewise quite high across all algorithms. Furthermore, the study discovered that data quality has a substantial impact on IDS performance. Proper data preparation, feature engineering, and noise removal were found to be helpful in improving model accuracy and reducing false positives. While the IDS models performed well throughout validation and testing, implementing such systems in a production setting necessitates careful thought. As a result, the essay also examined the procedures for testing and deploying the IDS models in a real-world scenario. It underlined the significance of ongoing monitoring and maintenance in order to keep the model effective in identifying intrusions. The research aids in the progress of network security in HLI. Educational institutions can better protect their precious assets and sensitive information from cyberattacks by understanding the impact of data quality on IDS performance and implementing effective deployment techniques.

APA CITATION

Sindika, D. M., Nicholaus, M. R. & Hamadi, N. B. (2024). Improving Network Security: An Intrusion Detection System (IDS) Dataset from Higher Learning Institutions, Mbeya University of Science and Technology (MUST), Tanzania. *East African Journal of Information Technology*, *7*(1), 23-38. https://doi.org/10.37284/eajit.7.1.1679

CHICAGO CITATION

Sindika, Daud M., Mrindoko R. Nicholaus and Nabahani B. Hamadi. 2024. “Improving Network Security: An Intrusion Detection System (IDS) Dataset from Higher Learning Institutions, Mbeya University of Science and Technology (MUST), Tanzania”. *East African Journal of Information Technology* 7 (1), 23-38. https://doi.org/10.37284/eajit.7.1.1679.

HARVARD CITATION

Sindika, D. M., Nicholaus, M. R. & Hamadi, N. B. (2024) “Improving Network Security: An Intrusion Detection System (IDS) Dataset from Higher Learning Institutions, Mbeya University of Science and Technology (MUST), Tanzania”, *East African Journal of Information Technology*, 7(1), pp. 23-38. doi: 10.37284/eajit.7.1.1679.

IEEE CITATION

D. M. Sindika., M. R. Nicholaus & N. B. Hamadi “Improving Network Security: An Intrusion Detection System (IDS) Dataset from Higher Learning Institutions, Mbeya University of Science and Technology (MUST), Tanzania”, *EAJIT*, vol. 7, no. 1, pp. 23-38, Jan. 2024.

MLA CITATION

Sindika, Daud M., Mrindoko R. Nicholaus & Nabahani B. Hamadi “Improving Network Security: An Intrusion Detection System (IDS) Dataset from Higher Learning Institutions, Mbeya University of Science and Technology (MUST), Tanzania”. *East African Journal of Education Studies*, Vol. 7, no. 1, Jan. 2024, pp. 23-38, doi:10.37284/eajit.7.1.1679.

INTRODUCTION

A higher learning institution's (HLI) computer network is an essential component of institution life. The internet has impacted many facets of university core activities. Computer networks offer a range of services to the institution community, including internet access, teaching assistance, and learning processes, support for research projects, international conferences, workshops, lab space, and smart classrooms (Shanmugam & Malarkodi, 2019).

Recently, HLI has reportedly suffered an increase in security breaches due to handling vast volumes of valuable research and sensitive personal data, which, in academic settings, sums up the importance of information confidentiality, integrity, and availability (CIA) perfectly (Bongiovanni, 2019). Despite the deployment of safety precautions such as firewalls and encryption mechanisms, certain attackers manage to get past the HLI system's security defences. Therefore, it is crucial to detect them as soon as possible to reduce the possibility of damage to the important resources, and as a result, appropriate action can be taken into consideration to get rid of incursions (Abrar et al., 2020).

A variety of tools are being developed and deployed for various forms of network attacks. One of these instruments is the intrusion detection system (IDS). These tools can monitor a variety of network systems, cloud computing platforms, and information systems. Attacks may jeopardize the confidentiality, accessibility, and integrity of a system's security features, which the IDS can detect (Aljanabi et al., 2021a). The role of an intrusion detection system (IDS) involves acting as the first line of defence in identifying internet-based threats. Despite the use of secure network architecture, firewalls, passwords, encryptions, and personal screening, it is critical to have IDS techniques as the last line of defence against computer attacks in computer networks (Guezzaz et al., 2021a).

The demand for improved and more potent IDS is now even greater due to the rising rates of cyberattacks and cyberespionage. Researchers have begun to use cutting-edge Machine Learning (ML) techniques to efficiently detect hackers as their tactics become more sophisticated and challenging to detect, protecting internet users' information, and preserving general confidence in the security of the entire internet network (Ngueajio et al., 2023). To optimize the performance of an IDS, ensure that the data collected is accurate, complete, and relevant to the types of threats that are most common in the environment (Tran et al., 2022a).

This article discusses the process of creating a comprehensive IDS dataset from HLIs, highlighting the significance of accurate and representative data to improve the system's ability to detect and mitigate potential cyber threats.

Contribution

The significance of this study is that it addresses the topic of IDS datasets and their impact on IDS performance in higher education computer networks. Existing public datasets have become very large, with millions of records, necessitating the use of very efficient computers to find the right model for use. However, many educational institutions, particularly in developed countries, face a significant challenge with the presence of these devices, necessitating the challenge of developing these models. This study has provided an opportunity to strengthen the network of such organizations by developing a model that can work with computers that lack the power required for public data. The study's findings will also provide insights into how irrelevant datasets influence IDS performance in higher education institutions, allowing them to address these issues more proactively. This research will also be useful to IT administrators, system engineers, and security experts who are in charge of designing, implementing, and maintaining an IDS that is capable of accurately and effectively identifying security threats.

LITERATURE REVIEW

Theoretical Review

Intrusion detection is the process of identifying acts that attempt to compromise a resource's overall privacy and consistency (Samat, 2022). Due to the increase in the number of threats launched by attackers, an Intrusion Detection System (IDS) is critical nowadays, especially in an organization or industry (Lalduhsaka et al., 2021). IDSs are a type of security management system that can detect and prevent attacks on system security elements such as accessibility, integrity, and privacy (Samat, 2022).

IDSs can be either host-based (HIDS) or network-based (NIDS) (NIDS). NIDSs detect attacks by analyzing specific network events, whereas HIDSs detect intruders within individual hosts (Aljanabi et al., 2021a). A NIDS scans a packet sniffer, which is a program that reads raw packets from a local network segment. It can also track more network objectives in order to detect threats that HIDS may miss because HIDS cannot read packet headers and cannot detect certain types of attacks (Aljanabi et al., 2021a).

However, NIDSs do not rely on the operating system (OS) of the host as identification sources but HIDSs depend on the OS to function properly (Aljanabi et al., 2021a) . Hybrid IDSs that combine client and network-based technologies have also been developed for intrusion detection (ID). The ID techniques can also be in the form of misuse detection or anomalies detection. The detection mechanism used in IDS are three main types which are: statistical method, machine learning (ML), and data-mining methods

Machine learning is frequently useful when these techniques are combined to create a model as a solution to a problem. To develop effective solutions, a combination of different algorithms under different machine learning techniques can be used (Jadhav & Pellakuri, 2021). Researchers propose and develop numerous solutions to various problems using only this combination of machine learning techniques. Every technique in supervised and unsupervised learning has benefits and drawbacks. Despite this, when properly combined, these techniques can produce excellent results (Jadhav & Pellakuri, 2021).

There has been some prior research on the application of ML techniques to network intrusion detection systems. Each study's issue concerns vary, including those related to feature selection, data reduction, and classification technique optimization. The purpose of data reduction is to speed up and optimize the process, improving accuracy, precision, and specifications (Zhou et al., 2020.).

The requirement that ML approaches be taught for each network individually results in a significant issue. Different networks can impair the accuracy of ML techniques that are taught on open datasets. It appears that the difficulty raises the requirement for ML techniques that can label the data and regularly re-train for each network individually

(Devi & Kannan, 2021). There are some attempts to generate representative datasets. Therefore, exclusive datasets acquired from the target network and accurately labeled should be used to train ML techniques rather than public datasets (Devi & Kannan, 2021).

Empirical Review

One of the prevalent research problems in the domain of intrusion detection is the changing characteristics of both network traffic and the modern threat landscape (Komisarek et al., 2021a). The rate of change in the field is inextricably linked to the intensity of the cyber-arms race. The constant change in the threat landscape causes benchmark datasets to lose relevance. Due to privacy concerns and acquisition costs, telecom companies are hesitant to provide new, labeled data, resulting in a constant, high demand for new, relevant intrusion detection datasets (Komisarek et al., 2021a). There is a gap between the datasets available to intrusion detection researchers and the types of data that can be used in a real-world deployment of real-time ML-based network IDS (Komisarek et al., 2021a)

Anomaly-based techniques in IDS struggle with appropriate deployment, analysis, and evaluation due to a lack of adequate dataset. The researchers have employed a variety of such datasets, including DARPA98, KDD99, ISC2012, and ADFA13, to assess the effectiveness of their suggested methodologies for intrusion detection (Sharafaldin et al., 2018). The analysis of eleven publicly available datasets dating back to 1998 shows that several of these databases are out-of-date and untrustworthy for use. These datasets vary in terms of traffic types and quantities, the types of attacks they cover, the anonymized packet data and payload they use to reflect current trends, and the feature sets and metadata they contain (Khraisat et al., 2019a)

This study suggests creating an ML model for HLI computer network intrusion detection using actual domain datasets and considering all the crucial aspects of improving intrusion detection performance utilizing relevant HLI datasets. The model will improve computer network security and make it easier to implement actions to enhance IDS performance. The research will also support the creation of practical IDS in our nation's HLI as part of the effort to combat cybercrime.

Intrusion Detection System (IDS) in Higher Learning Institution (HLI)

Features such as Resource Sharing, a large number of users, the use of Virtual Local Area Networks (VLANs) and network segmentation, providing guest access to visitors (guest students, lecturers, and conference attendees), ensuring fair and efficient utilization of network resources, and holding sensitive and confidential information such as student records, research data, and intellectual property are important in differentiating the HLI computer network from other industries. Many academics have published about ways to improve IDS performance in diverse networks and, hence, enhance security by utilizing various models. There are also various models that have been presented by academics that aim to address issues with the dataset that those models employ.

Nowadays, higher education institutions (HLI) have built computer networks of varied sizes as an infrastructure platform for information construction. That network has influenced many aspects of HLI education, scientific research, management, etc., making it easier for teachers and students to work, study, and go about their everyday lives. (Wu & Wu, 2021). As technology has advanced, numerous dangers to data security have surfaced, which is not good at all for sensitive data transactions (Yang et al., 2019). Network security concerns have grown in importance, and security is now at the top of the list for network deployment and administration (Zheng et al., 2017).

Network security incidents like network worms, DoS attacks, network fraud, and so forth frequently happen. As a result of increasing innovation in network security incident tactics, data leakage, destruction, and other occurrences are now regular (Wang et al., 2019). It is impossible to completely defend our network, particularly during cyber-attacks. This is to advise HLI to prioritize cybersecurity mitigation plans within their organizations and train staff who will have the skills necessary as an effective security measure to secure their assets to fulfil the duty of a cybersecurity analyst in a threat-centric security operations centre. (Naagas et al., 2018). HLI networks' security needs to be strengthened in order to reduce security problems due to the complexity of the information system and the rapid development of new vulnerabilities and exploits (Nikoi et al., 2022).

The Importance of a Representative Dataset

The requirement that ML approaches be taught for each network individually results in a significant issue. Different networks can impair the accuracy of ML techniques that are taught on open datasets. It appears that the difficulty raises the requirement for ML techniques that can label the data and regularly retrain each network individually (Devi & Kannan, 2021). There have been some attempts to generate representative datasets. Therefore, exclusive datasets acquired from the target network and accurately labelled should be used to train ML techniques rather than public datasets (Devi & Kannan, 2021).

Realistic network traffic datasets are required to keep the defensive measures current and applicable (Komisarek et al., 2021b). Current datasets do not include all aspects of recent network traffic. Furthermore, NIDS is incapable of adapting to frequent network changes. Because networks are always changing, relying primarily on old datasets hinders the advancement of NIDS (Ghurab et al., 2021). To carry out the real-world application of machine learning-based intrusion detection, a set of labelled data from the end-user must be acquired (Komisarek et al., 2021b).

The analysis of eleven publicly available datasets dating back to 1998 shows that several of these databases are out-of-date and untrustworthy for use. These datasets vary in terms of traffic types and quantities, the types of attacks they cover, the anonymized packet data and payload they use to reflect current trends, and the feature sets and metadata they contain (Khraisat et al., 2019b).

A robust IDS requires a representative dataset that encompasses various network activities and traffic patterns commonly found in HLIs. By collecting data from different departments, campuses, and user groups, the IDS can better understand typical network behaviour and effectively distinguish between normal and malicious activities.

Algorithms for Machine Learning IDS

Random fore (RF) is an ensemble classifier used to increase accuracy. The random forest is made up of multiple decision trees. When compared to other standard classification techniques, random forest has a low classification error. RF Create forests that may be preserved for future reference, overcoming the problem of overfitting as well. In RF, accuracy and variable importance are generated automatically (Farnaaz & Jabbar, 2016). RF is an excellent choice for a tiny dataset since it can handle small quantities of data while maintaining high accuracy and resistance to overfitting. It is an ensemble approach that handles numerous classes efficiently and works well with both numerical and category information (Hasan et al., 2016).

Gradient Boosting (GB) classifier aggregates a number of weak models to build new models consecutively in an attempt to minimize the loss function. To compute the loss function, GB uses the gradient descent method. To reduce overfitting difficulties, boosting should be stopped on time using stopping criteria. A maximum number of models generated or a forecasted accuracy level could be used as a stopping criterion. GB should be investigated because it performs well even with small datasets and can handle multi-class classification workloads effectively. (Zulfiker et al., 2021).

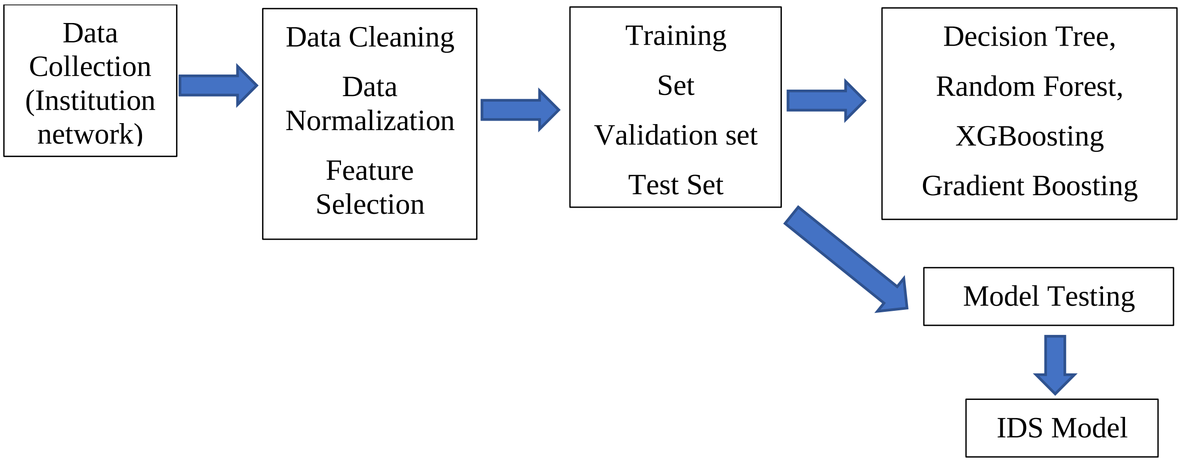
Classic Decision Tree: Is one of the supervised learning approaches that uses a tree-like approach for decisions and shows the probability of event outcomes. It has a conditional control statement that allows you to view and grasp the logic for the data, as well as decision rules that are simple to understand (Mahbooba et al., 2021). Classic decision trees in IDS have the advantage of being interpretable, which means that the decision-making process and the factors that contribute to the categorization can be understood (Lian et al., 2020).

XGBoost model: XGBoost is an ensemble learning method that builds multiple weak learners sequentially. Each new tree is constructed to correct the errors made by the previously added trees. This tool is so effective that it checks each and every data value in the database. This process gradually improves the model's accuracy, making it more effective at handling complex patterns in the data (Bhati et al., 2021).

CONCEPTUAL FRAMEWORK

The conceptual framework of the suggested model is depicted in the following figure: The model was created using a dataset gathered from institution network traffic, which was then cleaned up, normalized, and feature-selected as needed for training. The preprocessed dataset was then split into 70:30 training sets for validation and testing. The training set trained the Decision Tree, Random Forest, XGBoosting, and Gradient Boosting algorithms that perform better in the real world, and the best algorithm tested the model to create our final IDS model.

Figure 1: Conceptual framework of the proposed model for intrusion detection



METHODOLOGY

The study used a qualitative research design and data acquired from the MUST computer network using Firewall user logs. Firewall variables included time, log computer, log subtype, NAT rule name, source IP, destination IP, source port, destination port, protocol, bytes transferred, bytes received, and connection duration. The study compared the performance of the IDS model with data obtained from the firewall. The goal was to see if there was a link between the quality of the data in the dataset and the performance of the IDS.

A total of 610,600 network packets were gathered, pre-processed, trained, tested, and evaluated. To evaluate the model, the variables from the Firewall log were subjected to classification and grouping based on the attacks. This network traffic was captured on the MUST computer network using the Firewall program. The data was collected in May and June 2023. A Python library analyser was used to evaluate the data.

Materials and Methods

Python (Python Software Foundation, Wilmington, DE, USA) was used to clean the data, create visualizations, apply feature selection methods, and build the machine learning models within the Anaconda environment (Anaconda Software Distribution, Austin, TX, USA). Many Python packages were utilized. Pandas was used to clean and organize the data; Matplotlib was used to create the visualizations; Seaborn was used to create a heat map showing the correlation between features; Sci-Kit Learn was used for all machine learning and feature selection operations; and Numpy was used for general mathematical operations.

Data Collection

Data gathering is the initial stage in creating an IDS dataset. HLIs can use network monitoring tools and sensors to collect network logs, packet captures, and other relevant data sources. These sources may include firewall logs, network flow statistics, DNS logs, and authentication logs.

This study gathered the flow-based MD23 data set, which only includes DNS connections and was taken over a period of six days within a school network. A total of 610,600 unidirectional flows exists, of which 94,248 are malicious and the remainder represent typical user behaviour. Using logs from an intrusion prevention system, all flows were labelled. The authors did not intend to make the MD23 data set available to the general public due to privacy concerns.

Our generated dataset was named MD23. Four categories were used to categorize attacks in this dataset:

- DoS: A denial of service (DoS) attack prevents a legitimate user from accessing system and network resources. Email and online banking may be impacted (Alazzam et al., 2020). The SYN flood assault and the Smurf attack are examples of DoS attacks.
- Remote to Local (R2L): R2L attacks involve an attacker attempting to log into the target workstation without having an account (Alazzam et al., 2020).
- User to Root (U2R): This assault aims to give the perpetrator local access privileges on the victim’s machine (Alazzam et al., 2020).
- PROBE: In Probe, the attacker focuses on the host and seeks to learn more about it (Alazzam et al., 2020).

Table 1: Collected Data

Date	Number of Flows	Number of Attacks	Description
15/05/2023, Monday	155,495	26,790	Normal, DoS, and Probe
16/05/2023, Tuesday	145,201	22,452	Normal, Dos, and probe
17/05/2023, Wednesday	29,101	22,747	Normal, U2R, and R2L
18/05/2023, Thursday	63,901	20,998	Normal, U2R, and R2L
19/05/2023, Friday	98,801	51,454	Normal, Dos, Probe, U2R, and R2L
03/06/2023, Saturday	118,101	41,801	Normal, Dos, Probe, U2R, and R2L
Total	610,600	186242	Normal, Dos, Probe, U2R, and R2L

Data Preprocessing

After being collected, the data was pre-processed to verify its quality and uniformity. Data preprocessing is a crucial step in preparing the data for IDS training. It involves cleaning and transforming the data into an appropriate format.

Preprocessing procedures:

- There were numerous features in the gathered dataset's 23 feature-strong composition with missing values. We utilized Python programs to deal with missing values by deleting any rows or columns that had empty values.
- Both categorical and numerical data were collected for various features; however, the machine learning model only accepts numerical variables, so categorical variables must be transformed into numerical representations. To convert those characteristics with categorical variables to numerical ones, we utilized one-hot encoding and label encoding.
- To ensure that the numerical features we converted in the previous phase fit the algorithm I'm planning to employ in my IDS model, we scale and normalize them. Data scaling and normalization techniques are also

applied to bring all features to a similar scale, facilitating model training

- We determine the IDS's target variable by determining whether the traffic is normal or malicious and, if malicious, which type it is (there are four types of malicious: DoS, U2R, R2L, and probe).
- The dataset was then separated into subgroups for training, validation, and testing before models were trained.

Table 2: After data preprocessing

Number of Flows	Number of Attacks	Description	
559,736	92,000	Dos	23,250
		Probe	27,250
		U2R	21,250
		R2L	20,250
		Normal	407,167
		Abnormal/Unclassified	60,567

Feature Selection

The feature selection system seeks to remove irrelevant features as well as locate features that will aid in improving the detection rate based on the score each feature establishes during the selection process. In order to do that, a decision tree-based classifier and a recursive feature reduction procedure were both used, and the appropriate relevant features were later found. The IDS Dataset was subjected to this methodology, leading to the identification of pertinent features inside the dataset and an increase in accuracy. The findings of Herve Nkiama are consistent with the notion that feature selection considerably enhances classifier performance. A better intrusion detection system can be designed by knowing the criteria that identify significant aspects. (Nkiama et al., 2016)

Labelling

Each network activity must be tagged as normal or malicious in order to construct a supervised learning dataset. Data labels can be appropriately labelled by network administrators, cybersecurity specialists, and threat intelligence. The labelled data is used to train the model. We labelled normal as 1 and malicious as 2, 3, 4, 5, and 6 for Probe, R2L, U2R, and DoS, respectively, based on our data set.

Splitting the Dataset

Before model training, the dataset was subdivided into training, validation, and testing subsets. The training set was used to train the model, the validation set was used to fine-tune hyperparameters, and the testing set is used to assess the final model's performance.

The study employed a percentage-based technique in Python with the scikit-learn module to divide the dataset into training, testing, and validation subsets for evaluating the performance of the IDS model. Here is how it works:

Let:

The total number of samples (rows) in the dataset be 499,169; p_train be the proportion of the dataset that the study wants to dedicate to training (70%); P_test is the proportion of the dataset that the study wants to dedicate to testing (20%); P_val is the percentage of the dataset that the study wishes to dedicate to validation (10%).

The number of samples required for training (n_train), testing (n_test), and validation (n_val) can then be determined as a ratio of 70:20:10, respectively.

It should be noted that p_train + p_test + p_val should equal one, i.e., 100% of the dataset. After determining the number of samples for each subset, the data was splitted using the scikit-learn library's train_test_split function twice: Pandas' to_csv method was used to save the datasets created using Scikit-Learn's train_test_split function to CSV files. This will result into six CSV files with the respective data subsets (X_train.csv, X_test.csv, X_val.csv, y_train.csv, y_test.csv, and y_val.csv).

Table 3: Summary of MD23 Dataset

Attack Type	MD23	MD23-TRAIN-1	MD23-VAL	MD23-TEST
Dos, Probe, U2R, R2L	499,169	349,418	49,917	99,834

Model Selection and Training

A PC with an Intel (R) Core (TM) i5-10400 CPU running at 2.90 GHz and 8GB of RAM was used to carry out the research. The MD23 dataset's CSV files were used to test binary classification and multiclassification techniques.

HLIs can use a wide range of machine learning models, including Decision Trees, Random Forests, Support Vector Machines (SVM), Neural Networks, and Ensemble approaches such as AdaBoost. The nature of the dataset and the unique requirements of the IDS influence model selection. Labelled training data was used to train Classic Decision Tree, Random Forest, XG Boost, Light GBM, and SVM models, which are most popular and perform better in the field of IDS.

Hyperparameter Tuning

- If you did hyperparameter tuning, describe the method you employed (e.g., grid search, random search).
- Mention the tuned hyperparameters and the range of values considered.
- Describe the evaluation metrics used to choose the optimal hyperparameters.

Given that there was limited dataset of education data and the study wished to identify several types of attacks such as DoS, probing, U2R, and R2L on a standard desktop computer, the study recommended the following machine learning methods and a simple design for the IDS:

Algorithms for Machine Learning

Demonstration of the proposed model's results is shown in Table 4: accuracy (A), precision (P), recall (R), and F1-score (F1).

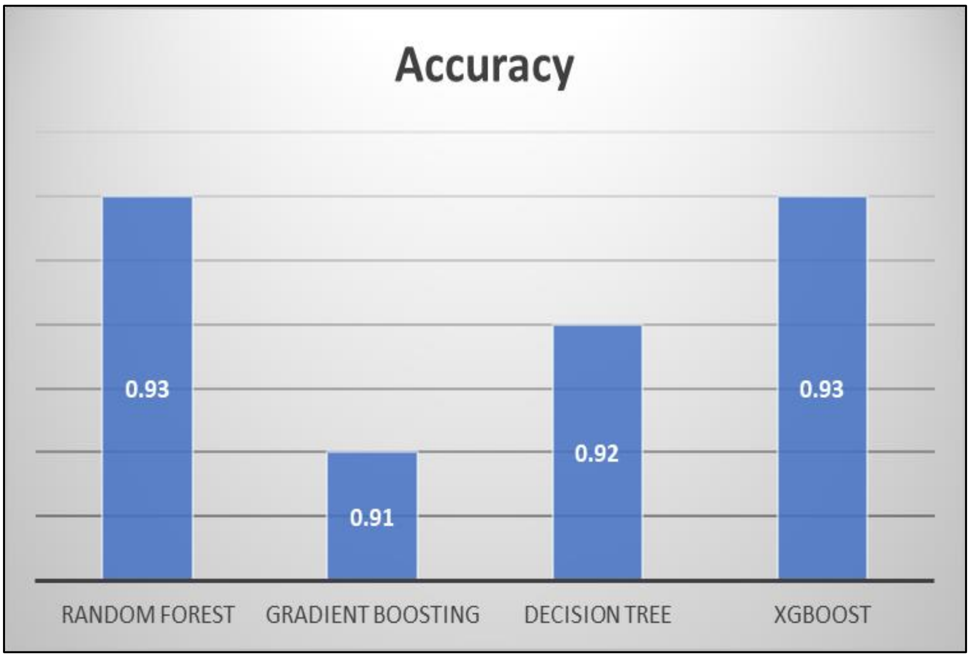
Table 4: Proposed model's results

Model	A%	P%	R%	F1%
DT	0.92	1	1	1
RF	0.93	1	0.95	0.88
XGBOOST	0.93	1	0.95	0.88
GB	0.91	0.99	0.96	0.99

A=Accuracy, P=Precision, R=Recall and F1=F1-score.

Matplotlib, Python's popular data visualization toolkit, was used. The bar plot *Figure 2* depicts some sample accuracy values for various models (Random Forest, Gradient Boosting, Decision Tree, and XGBoost) for easy comparison of their performance

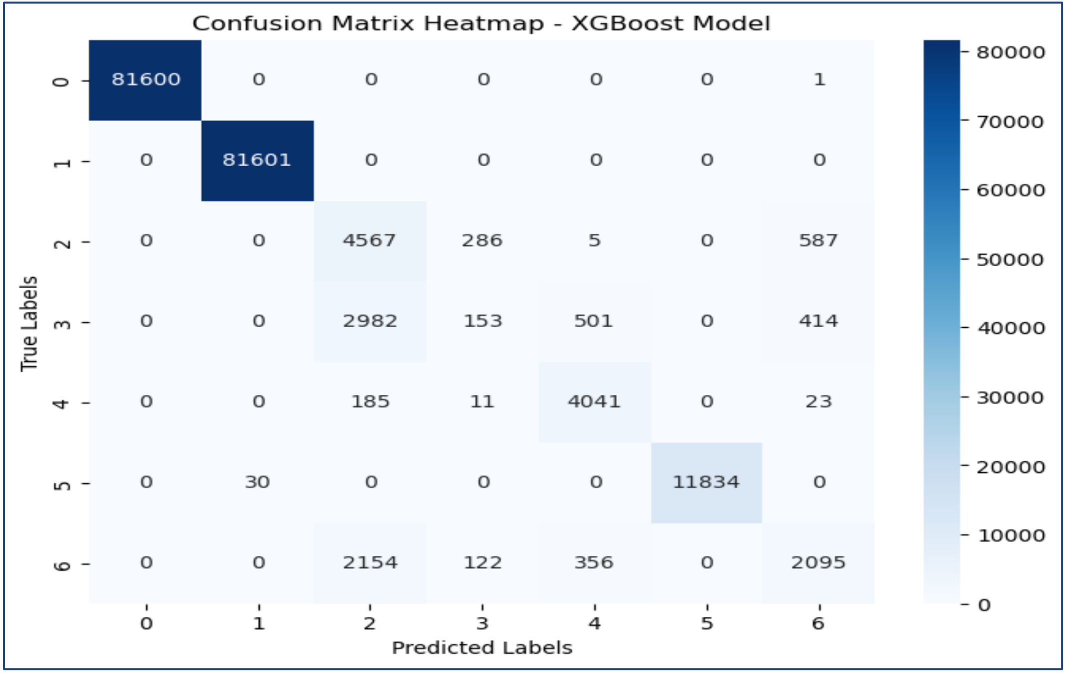
Figure 2: Results for model comparison - accuracy



The heatmap of the confusion matrix for the best-performing model XGBoost

In Python, the study utilized the Seaborn module, which works nicely with Matplotlib and provides simple functions for building visualizations like heatmaps. The confusion matrix heatmap for the best-performing model XGBoost is shown in the illustration *Figure 3*.

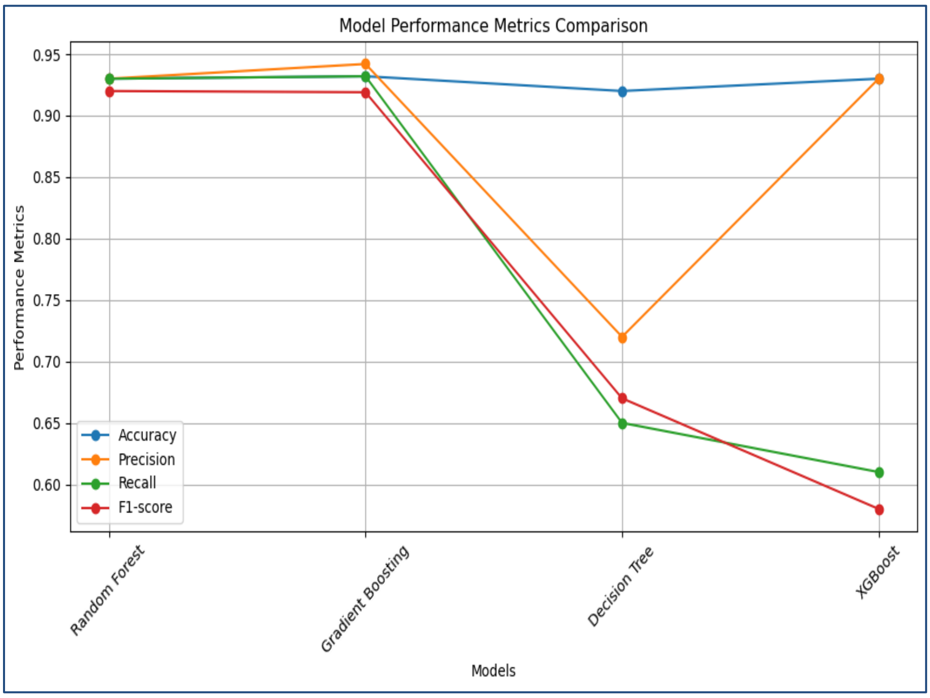
Figure 3: Heatmap of the confusion matrix



Line plot showing the variation of performance metrics across different models

The line plot below depicts the fluctuation of performance measures across different models using Matplotlib to visualize the data. Each model's performance metrics, such as accuracy, precision, recall, and F1-score, were displayed. The illustration is shown in Figure 4.

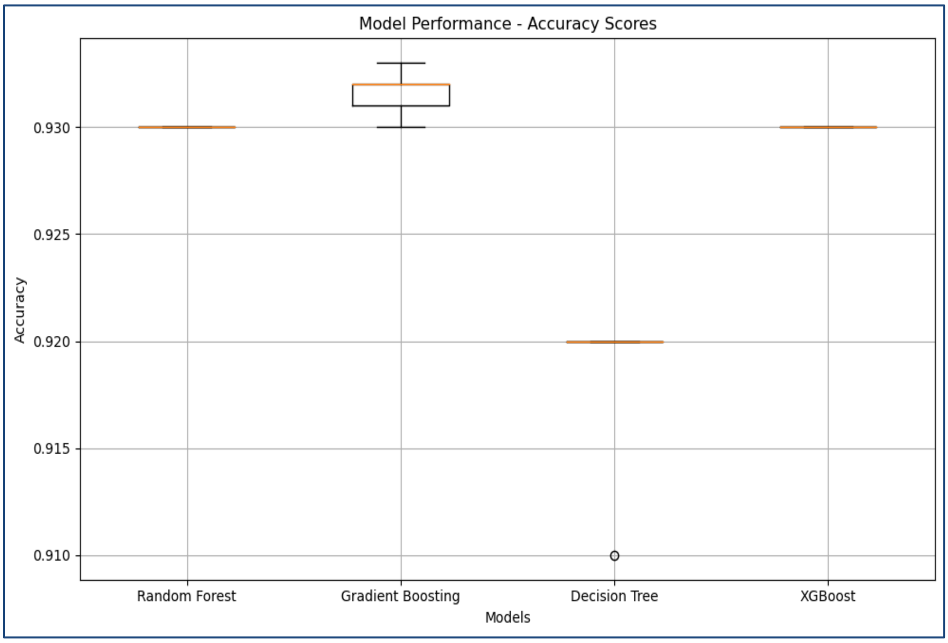
Figure 4: Model Performance Metrics Comparison



Box Plot - Model Performance

The box plot below compares the distribution of accuracy ratings for each model. The box plot will display the median, quartiles, and any outliers, offering insight into the model's variability.

Figure 5: Box Plot - Model Performance



RESULTS AND DISCUSSION

The study developed an Intrusion Detection System (IDS) by mixing machine learning methods such as Random Forest, Gradient Boosting, Decision Tree, and XGBoost. The goal was to increase the efficacy of these models in detecting intrusions and safeguarding computer networks at Higher Learning Institutions (HLIs).

Model Performance Evaluation

Metrics like accuracy, precision, recall, and F1-score were used to assess the performance of each IDS model. These metrics offered useful information about the models' ability to appropriately classify network traffic and detect intrusions.

The results showed that all four IDS models performed well, with Random Forest, Gradient Boosting, and XGBoost models obtaining approximately 93% accuracy (as seen on Table 4). All algorithms have good precision and recall values, showing that the models can efficiently distinguish between regular and malicious network data.

Table 4: Proposed model's results

Model	A%	P%	R%	F1%
DT	0.92	1	1	1
RF	0.93	1	0.95	0.88
XGBOOST	0.93	1	0.95	0.88
GB	0.91	0.99	0.96	0.99

A=Accuracy, P=Precision, R=Recall and F1=F1-score.

Impact of Data Quality on IDS Performance

Here are some examples of how data quality can impact IDS performance:

- False positives: If the IDS collect a lot of noise or irrelevant information, the system may generate a lot of false positive alerts, which can be time-consuming for security staff to examine and lower the IDS's efficacy (Tran et al., 2022b)
- Missed assaults: If the IDS do not gather all relevant data, it may miss some attacks or anomalies that may indicate a security problem. This can lead to security breaches and degrade the IDS's overall efficacy (Tran et al., 2022b)
- Alert accuracy: If the data acquired by the IDS is inaccurate or incomplete, the alarms issued by the system may be useless or ineffective. This can result in security personnel disregarding alerts or failing to respond to them in a timely manner (Guezzaz et al., 2021b)
- System performance: Collecting and analyzing huge amounts of data can be resource-intensive, and poor data quality can result in performance difficulties such as delayed response times or system breakdowns (Guezzaz et al., 2021b)

To improve the performance of an IDS, it's important to ensure that the data collected is accurate, complete, and relevant to the types of threats that are most common in the environment (Tran et al., 2022b). This may involve configuring the IDS to filter out noise and irrelevant information, tuning the system to collect the right types of data, or integrating the IDS with other security systems to improve the accuracy and relevance of alerts. Regular maintenance and monitoring of the IDS can also help ensure that it is performing optimally and detecting potential security threats effectively (Tran et al., 2022b)

The study emphasized the need for reliable and representative data to increase the effectiveness of the IDS system to detect and neutralize potential cyberattacks. Data preparation, feature engineering, and noise removal were discovered to be critical in improving model accuracy and lowering false positives.

As observed, the performance of the IDS model was shown to be highly dependent on the quality of the dataset used for training and testing, as measured by parameters such as accuracy, precision, and recall, F1-score, and confusion matrix. High-quality data guarantees that the model learns from meaningful patterns, resulting in increased detection skills.

We stressed the significance of rigorous testing and continued monitoring when deploying the

IDS models in a real-world setting. Before deployment, the models were thoroughly tested using a separate test dataset to confirm their generalization to new and previously encountered data.

Real-time monitoring was identified as a vital feature of IDS deployment. Continuous monitoring enables timely updates and tweaks to the model, ensuring its effectiveness in spotting new and emerging risks.

MD23 vs Public Dataset

For the model's performance to be more than 90% effective, it means that the model is good; however, the models of many previous studies that used public datasets appear to have higher results than our model. This does not make our model ineffective; rather, our model conforms to factors that lead to efficient datasets, such as the following:

Data Bias and Generalization: Most commonly used public datasets have various disadvantages, including obsolete attack versions that may not be reflective of real network attacks, insufficient information, and a large amount of redundant records that contribute bias to frequency records while training NID (Maseer et al., 2021).

Data Quality Control: ML models train and perform differently depending on data quality, regardless of model complexity (Tran et al., 2022a). Despite the widespread use of Intrusion Detection Systems (IDS) in higher learning institutions, the issue of data quality continues to affect their performance (Komisarek et al., 2021b). Poor data quality can lead to missed security threats or inefficient use of resources, which can result in security breaches (Devi & Kannan, 2021).

Real-world Constraints and Contextual Relevance: The academic community is increasingly looking for a trustworthy and real-world assaults dataset. Current Anomaly Intrusion detection systems perform poorly in terms of accuracy. This is due to the fact that no valid tests are performed and datasets are not checked (Abdulraheem & Ibraheem, 2019). The dataset should be an accurate representation of the network or system environment in which the IDS will be implemented. It should comprise realistic network traffic captures or system logs from different protocols, traffic quantities, and data types (Komisarek et al., 2021b).

These factors appear to be significantly better in our dataset than in previous research's public dataset, showing that our model is closer to real-world scenarios and captures the true domain of higher learning institutions.

CONCLUSION

This study is significant for readers because it expands the domain's ability to develop custom network datasets with the goal of enhancing institutional network security. As a result, real data will be available to develop IDS models for the appropriate institutions rather than using publicly available datasets that are incompatible with the institution's network infrastructure, users' behaviour patterns, and types of network devices.

The Intrusion Detection System (IDS) is critical for securing computer networks in Higher Education Institutions. Among the tested models, Random Forest and Gradient Boosting had the best accuracy of 93% on both validation and test datasets. They outperformed XGBoost and Decision Tree, which had somewhat lower accuracy and F1 scores.

The Random Forest and Gradient Boosting models displayed strong precision, recall, and F1-score across different classes, suggesting their ability to detect both normal and harmful behaviours successfully. However, due to its simplistic structure and inability to capture complicated patterns, the decision tree had certain limitations in distinguishing instances from specific classes. The findings of this study increase network security in HLIs and provide critical insights towards designing powerful and successful IDS systems.

Recommendations

Continued research and developments in intrusion detection algorithms and methodologies will be crucial in securing computer networks and creating a secure digital environment for HLIs and their stakeholders as cyber threats grow.

Random Forest and Gradient Boosting models are recommended for implementation in the IDS for Higher Education Institutions computer networks because of their higher performance and capacity to handle complex datasets with multiple classes. More study could be conducted to fine-tune hyperparameters and investigate other factors in order to improve model performance.

Finally, the study's conclusions pave the way for future research in HLI network security. We uncovered opportunities to improve IDS efficacy by studying more complicated algorithms, merging different datasets, and utilizing novel approaches to combating expanding cyber threats.

ACKNOWLEDGEMENT

I am grateful to Dr. Mrindoko R. Nicholaus and Dr. Nabahani B. Hamadi for assisting with this job in terms of writing abilities and suggestions on English expression in this article.

Conflict of Interest

I have no conflicts of interest.

REFERENCES

Abdulraheem, M. H., & Badie Ibraheem, N. (2019). A detailed analysis of new intrusion detection dataset. *Journal of Theoretical and Applied Information Technology*, *15*, 17. www.jatit.org

Abrar, I., Ayub, Z., Masoodi, F., & Bamhdi, A. M. (2020). A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset. *Proceedings - International Conference on Smart Electronics and Communication, ICOSEC 2020*, 919–924. https://doi.org/10.1109/ICOSEC49089.2020.9215232

A Detailed Analysis of Benchmark Datasets for Network Intrusion Detection System by Mossa Ghurab, Ghaleb Gaphari, Faisal Alshami, Reem Alshamy, Suad Othman: SSRN. (n.d.). Retrieved May 15, 2023, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3834787

Alazzam, H., Sharieh, A., & Sabri, K. E. (2020). A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer. *Expert Systems with Applications*, *148*, 113249. https://doi.org/10.1016/J.ESWA.2020.113249

Aljanabi, M., Ismail, M. A., & Ali, A. H. (2021a). Intrusion Detection Systems, Issues, Challenges, and Needs. *International Journal of Computational Intelligence Systems*, *14*(1), 560– 571. https://doi.org/10.2991/IJCIS.D.210105.001

Aljanabi, M., Ismail, M. A., & Ali, A. H. (2021b). Intrusion detection systems, issues, challenges, and needs. *International Journal of Computational Intelligence Systems*, *14*(1), 560– 571. https://doi.org/10.2991/IJCIS.D.210105.001

Bhati, B. S., Chugh, G., Al-Turjman, F., & Bhati, N. S. (2021). An improved ensemble based intrusion detection technique using XGBoost. *Transactions on Emerging Telecommunications Technologies*, *32*(6), e4076. https://doi.org/10.1002/ETT.4076

Bongiovanni, I. (2019). The least secure places in the universe? A systematic literature review on information security management in higher education. *Computers & Security*, *86*, 350–357. https://doi.org/10.1016/J.COSE.2019.07.003

Devi, P. P., & Kannan, S. (2021). Performance analysis of machine learning models for threats and attacks in network security traffic model. 48(12).

Farnaaz, N., & Jabbar, M. A. (2016). Random Forest Modeling for Network Intrusion Detection System. *Procedia Computer*

Science, *89*, 213– 217. https://doi.org/10.1016/J.PROCS.2016.06.047

Guezzaz, A., Benkirane, S., Azrour, M., & Khurram, S. (2021a). A Reliable Network Intrusion Detection Approach Using Decision Tree with Enhanced Data Quality. *Security and Communication Networks*, *2021*. https://doi.org/10.1155/2021/1230593

Guezzaz, A., Benkirane, S., Azrour, M., & Khurram, S. (2021b). A Reliable Network Intrusion Detection Approach Using Decision Tree with Enhanced Data Quality. *Security and Communication Networks*, *2021*. https://doi.org/10.1155/2021/1230593

Hasan, Md. A. M., Nasser, M., Ahmad, S., Molla, K. I., Hasan, Md. A. M., Nasser, M., Ahmad, S., & Molla, K. I. (2016). Feature Selection for Intrusion Detection Using Random Forest. *Journal of Information Security*, *7*(3), 129–140. https://doi.org/10.4236/JIS.2016.73009

Jadhav, A. D., & Pellakuri, V. (2021). Highly accurate and efficient two phase-intrusion detection system (TP-IDS) using distributed processing of HADOOP and machine learning techniques. *Journal of Big Data*, *8*(1), 1–22. https://doi.org/10.1186/S40537-021-00521-Y/FIGURES/7

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019a). Survey of intrusion detection systems: techniques, datasets, and challenges. *Cybersecurity*, *2*(1), 1–22. https://doi.org/10.1186/S42400-019-0038-7/FIGURES/8

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019b). Survey of intrusion detection systems: techniques, datasets, and challenges. *Cybersecurity*, *2*(1), 1–22. https://doi.org/10.1186/S42400-019-0038-7/FIGURES/8

Komisarek, M., Pawlicki, M., Kozik, R., Hołubowicz, W., & Choraś, M. (2021a). How to Effectively Collect and Process Network Data for Intrusion Detection? *Entropy 2021, Vol. 23, Page 1532*, *23*(11), 1532. https://doi.org/10.3390/E23111532

Komisarek, M., Pawlicki, M., Kozik, R., Hołubowicz, W., & Choraś, M. (2021b). How to Effectively Collect and Process Network Data for Intrusion Detection? *Entropy 2021, Vol. 23, Page 1532*, *23*(11), 1532. https://doi.org/10.3390/E23111532

Lalduhsaka, R., Khan, A. K., & Roy, A. K. (2021). Issues and Challenges in Building a Model for Intrusion Detection System. *2021 5th International Conference on Information Systems and Computer Networks, ISCON 2021*. https://doi.org/10.1109/ISCON52037.2021.9702322

Lian, W., Nie, G., Jia, B., Shi, D., Fan, Q., & Liang, Y. (2020). *An Intrusion Detection Method Based on Decision Tree-Recursive Feature Elimination in Ensemble Learning*. https://doi.org/10.1155/2020/2835023

Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model. *Complexity*, *2021*. https://doi.org/10.1155/2021/6634811

Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset. *IEEE Access*, *9*, 22351–22370. https://doi.org/10.1109/ACCESS.2021.3056614

Naagas, M., Jr, E. M., … T. P.-B. of E., & 2018, undefined. (2018). Defense-through-deception network security model: Securing university campus network from DOS/DDOS attack. *Beei.Org*, *7*(4), 593–600. https://doi.org/10.11591/eei.v7i4.1349

Ngueajio, M. K., Washington, G., Rawat, D. B., & Ngueabou, Y. (2023). Intrusion Detection Systems Using Support Vector Machines on the KDDCUP'99 and NSL-KDD Datasets: A

Comprehensive Survey. *Lecture Notes in Networks and Systems*, *543 LNNS*, 609–629. https://doi.org/10.1007/978-3-031-16078-3_42/COVER

Nikoi, S. N., Nsiah-Konadu, A., Adu-Boahene, C., & Nsiah-Konandu, A. (2022). Enhancing the Design of a Secured Campus Network using Demilitarized Zone and Honeypot at Uew-kumasi Campus Enhancing the Design of a Secured Campus Network using Demilitarized Zone and Honeypot at Uew-Kumasi View project Enhancing the Design of a Secured Campus Network using Demilitarized Zone and Honeypot at Uew-kumasi Campus. *Asian Journal of Research in Computer Science*, *13*(1), 14–28. https://doi.org/10.9734/AJRCOS/2022/v13i130304

Nkiama, H., Zainudeen, S., Said, M., & Saidu, M. (2016). A Subset Feature Elimination Mechanism for Intrusion Detection System. *International Journal of Advanced Computer Science and Applications*, *7*(4). https://doi.org/10.14569/IJACSA.2016.070419

Samat, N. A. (2022). Intrusion Detection System: Challenges in Network Security and Machine Learning. EasyChair.

Shanmugam, T., & Malarkodi, B. (2019). Analysis of campus network management challenges and solutions. Proceedings of the 2019 TEQIP - III Sponsored International Conference on Microwave Integrated Circuits, Photonics and Wireless Networks, IMICPW 2019, 312– 316. https://doi.org/10.1109/IMICPW.2019.8933236

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). *Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization*. https://doi.org/10.5220/0006639801080116

Tran, N., Chen, H., Bhuyan, J., & Ding, J. (2022a). Data Curation and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection. *IEEE Access*, *10*, 121900– 121923. https://doi.org/10.1109/ACCESS.2022.3211313

Tran, N., Chen, H., Bhuyan, J., & Ding, J. (2022b). Data Curation and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection. *IEEE Access*, *10*, 121900– 121923. https://doi.org/10.1109/ACCESS.2022.3211313

Wang, P., Wang, S., Chi, L., Huang, M., Luo, W., & Wan, X. (2019). Research on Network Security of Campus Network. *Journal of Physics: Conference Series*, *1187*(4), 042113. https://doi.org/10.1088/1742-6596/1187/4/042113

Wu, H., & Wu, H. (2021). The Construction and Implementation of the Security Defense System of University Campus Network. *Advances in Intelligent Systems and Computing*, *1282*, 691– 696. https://doi.org/10.1007/978-3-030-62743-0_99/COVER

Yang, X., Nan Zhu, A., Zhao, J., Li, X., Cao, Y., Huang, M., Luo, W., & Wan, X. (2019). Research on Network Security of Campus Network. *Journal of Physics: Conference Series*, *1187*(4), 042113. https://doi.org/10.1088/1742-6596/1187/4/042113

Zheng, S., Li, Z., & Li, B. (2017). *Campus Network Security Defense Strategy*. 356–359. https://doi.org/10.2991/MECAE-17.2017.67

Zhou, Y., Cheng, G., Jiang, S., networks, M. D.-C., & 2020, undefined. (n.d.). Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Elsevier*. Retrieved January 22, 2023, from https://www.sciencedirect.com/science/article/pii/S1389128619314203

Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T., & Uddin, M. S. (2021). An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Sciences*, *2*, 100044. https://doi.org/10.1016/J.CRBEHA.2021.100044