ISSN 2581 – 3781

AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

Available Online at www.ajcse.info
Asian Journal of Computer Science Engineering 2023;8(4)

RESEARCH ARTICLE

BalanceNet: Addressing Class Imbalance in AI-Powered Intrusion Detection Through Adaptive Sampling

Gaurav Sarraf

Independent researcher

sarrafgsarraf@gmail.com

Received on: 11-10-2023; Revised on: 19-11-2023; Accepted on: 12-12-2023

***Abstract*—The constantly increasing cases of computer-attacks in the modern digitally connected world leader to the necessity of the most efficient intrusion detection systems (IDSs). Since innocuous traffic flow greatly outweighs the occurrence of attacks, one of the most crucial difficulties in intrusion detection systems is investigating the class imbalance of data flow from networks. Since this is the case, it impacts the accuracy with which machine learning algorithms detect dangers to minority classes. The research study introduces an intrusion detection system that uses adaptive sampling techniques to tackle the issue of network traffic class imbalance. It uses the UNSW-NB15 dataset, Extreme Gradient Boosting (XGBoost), and oversampling based on ADASYN, and it promises to improve the capacity to detect intrusions that impact minority classes. The model's 99.59% accuracy, 99.8% precision, 99.5% recall, and 99.6% F1-score indicate that it is very good at detecting harmful activity with few false alarms. In comparison to LR, NB, and LSTM, XGBoost performs better across the board when it comes to critical metrics. The combination of adaptive data balancing with a robust ensemble classifier provides a scalable and robust solution to real-time network anomaly detection in complex and unbalanced network settings, which can be used to further develop intelligent cybersecurity systems.**

Keywords*—*Cyberattack, Internet of Things (IoT), Intrusion detection system, Network traffic, UNSW-NB15, Machine Learning.

INTRODUCTION

The increasing number of sensor-based data streams in the era of the IoT has brought about new possibilities and threats in the field of cybersecurity [1][2]. New studies have shown that there is a growing number of cybersecurity risks to sensor-based systems, including autonomous systems and the IoT networks [3][4]. One example is the IoT infrastructure, which exposes autonomous systems to the risk of distributed denial of service (DDoS) and data manipulation attacks because of its weak processing capacity and absence of security measures. Network resources must be kept available, private, and secure at all times; intrusion detection systems (IDSs) help with this by setting up protections for when danger strikes. They fall into two main categories: signature-based detection, which looks for previously identified patterns of threats, and anomaly-based detection, which uses a normalised pattern of network activity to identify potentially dangerous ones. Nonetheless, a major problem experienced when applying IDS is that of uneven training data [5][6].

Machine Learning (ML) is becoming a promising solution to the limitations of traditional IDS as it has attracted the attention of the cybersecurity community [7]. ML based IDS utilizes the behaviour analysis to identify anomalies and threats and provides the possibility of much greater accuracy and shorter detection times [8][9]. This is a paradigm change in the field of IDS which promises to not only enhance security, but also transform the privacy scene [10]. The effectiveness of ML algorithms is that they can detect threats, but this usually requires sensitive information [11][12]. ML in cybersecurity can be used as an effective tool to enhance the capacity of systems to interpret various patterns as well as predict possible data threats .

Motivation and Contribution

Cyberattacks on vital network infrastructures are becoming more sophisticated and common, necessitating the development of reliable IDS. Conventional detection techniques are generally ineffective with high-dimensional data, class imbalance and changing attack patterns, resulting in decreased accuracy and slower threat response. This project aims to provide a robust framework to support real-time network security monitors, enhance detection rates, decrease false alarms, and apply state-of-the-art ML models for efficient data preparation, feature selection, and class balancing. This study has a number of important contributions as follows:

- Created a full pipeline of pre-processing, consisting of cleaning, encoding, normalization, and class balancing with ADASYN on the UNSW-NB15 data.
- Applied chi-square statistical techniques to choose the most pertinent features, which minimizes the complexity of the computation and maximization of the performance of the model.
- Enhanced attack traffic categorisation using XGBoost, a hybrid of adaptive sampling and feature selection.
- The model's performance was evaluated using ROC curve analysis, F1, REC, ACC, and PRE, among other tools.

The proposed model also deals with an important problem of IDS, which is the issue of class imbalance, by combining adaptive sampling with an ensemble classifier with high performance. This guarantees enhancement in detecting minority-class attacks which are usually ignored by the traditional models. It is novel in the sense that it integrates ADASYN with XGBoost to achieve the best learning based on thin threat patterns and high accuracy and low false alarms. The solution does not only enhance detection reliability, but also adds to the modern cybersecurity systems with a scalable and data-sensitive solution.

Organization of the Paper

The structure of the paper is as follows: Study on IDS methods that is relevant to this topic is reviewed in Section II. Section III details the method that is being suggested. In Section IV, shows the experimental findings and compare how well the models performed. Conclusions and suggestions for further research are provided in Section V, which also summarises the study's main findings.

LITERATURE REVIEW

The construction of this study was guided and strengthened by a comprehensive assessment and analysis of significant research works on IDS.

Kabir et al. (2022) develop an intrusion detection system and intrusion prevention system model for an entire network. Using the ET Classifier and Mutual Information Gain feature selection techniques, this work presents two independent stacking ML models to increase the NIDS's ACC. One of the suggested models outperforms all other competing models in terms of ACC (96.24%), according to the comparison data [13].

Gupta and Saxena (2022) Despite advancements, the majority of commercial IDS that are currently available rely on signatures to identify intruders. Recently, anomaly detection has seen a rise in the use of ML-based classification algorithms. Results, recall, and ACC for the majority of ML methods on this dataset were 90% or higher. On the other hand, Radial Basis Function is the best of the seven algorithms when looking at the area under the ROC [14].

Umamaheshwari, Kumar and Sasikala (2021) employs a WSN-DS dataset that is open to the public to assess the system's efficiency. All of the suggested feature selection methods are tested with important performance indicators. Train duration, ACC, sensitivity, and specificity are 15.12 seconds, 98.58%, 92.81%, and 98.46%, respectively, while using MRMR feature selection. Thereby, a solid IDS in a WSN might be predicated on this research [15].

Das et al. (2020) offer a non-linear learning PIDS that integrates ML and NLP ensembles. A number of supervised and ensemble-based ML models are trained using the language-based vectors converted by the proposed NLPIDS from HTTP requests. With a lower number of false alarms (0.007) and a higher F1-score (0.999), the NLPIDS clearly outperforms competing methods. The NLPIDS is independent of attack vectors and tactics [16].

Srivastava, Agarwal and Kaur (2019) helped identify suspicious activity in the data pertaining to the traffic on the network. Much study has focused on the use of ML algorithms for anomaly identification in network data. The public repositories now accommodate additional datasets. Using innovative feature reduction based ML algorithms, the authors of this paper were able to spot suspicious patterns in the newly supplied dataset. A level of 86.15% ACC has been maintained [17].

Singh and Mathai (2019) Used the NSL KDD dataset for ML classification and compared the SPELM approach to its DBN counterpart. Computer time (90.8 vs. 102 seconds), accuracy (93.20 vs. 52.8%), and precision (69.492 vs. 66.836) are three areas where SPELM excels beyond the DBN method [18].

Table I provides an overview of current studies on adaptive sampling for IDS, including the models suggested, datasets used, important results, and problems encountered. There are still a number of unanswered questions about IDS, even though these technologies have made great strides in recent years. Most studies depend on popular datasets like UNSW-NB15, NSL-KDD, and Kyoto 2006+, which may not reflect the dynamic nature of zero-day threats and complex multi-stage invasions. This is a problem in the current state of cyberattack research. The ACC of detection has been enhanced by ensemble methods and feature selection strategies; nonetheless, numerous systems continue to face challenges when dealing with high-dimensional data, processing in real-time, and minimising false positives. Additionally, limited research has addressed adaptive or hybrid models that can dynamically adjust to new attack patterns without frequent retraining. There is also a lack of comprehensive studies integrating anomaly-based and signature-based detection to balance detection speed, ACC, and robustness across heterogeneous network environments. Because of these shortcomings, IDS require to be more flexible, scalable, and proven in the real world

TABLE I. RECENT STUDIES ON INTRUSION DETECTION SYSTEMS USING MACHINE AND DEEP LEARNING TECHNIQUES

Author	Proposed Work	Results	Key Findings	Limitations & Future Work
Kabir et al. (2022)	ML NIDS algorithms utilising ET classifiers and Mutual Information Gain	Testing ACC of stacking models: 96.24%	Stacking models outperform individual models; enhanced detection ACC on UNSW-NB15 dataset	Further optimization could improve performance on emerging attack types
Gupta & Saxena (2022)	Applied seven ML techniques for anomaly detection on Kyoto 2006+ dataset using information entropy	Most ML models achieved ~90% ACC, with performing best (AUC)	ML-based approaches are more effective than signature-based methods for anomaly detection	Extend to real-time detection and newer datasets
Umamaheshwari, Kumar & Sasikala (2021)	Built IDS for WSN using ML; feature selection via Correlation Score, Fisher Score, KW test, MRMR, and Relief	ACC 98.58%, Sensitivity 92.81%, Specificity 98.46%, PRE 93.86%, Training time 15.12s	Feature selection reduces detection time and improves IDS performance	Apply to larger WSN datasets and real-time deployment
Das et al. (2020)	The proposed NLPIDS uses ensemble ML and natural language processing to identify HTTP requests.	Using the CSIC 2010 dataset, the results demonstrate an F1-score of 0.999 and a false alarm rate of 0.007.	NLPIDS is attack-independent and achieves high detection performance	Explore application to other protocols and network types
Srivastava, Agarwal & Kaur (2019)	Used feature reduction-based ML algorithms to detect anomalies in network traffic	ACC 86.15%	Novel feature reduction techniques improve detection on recent datasets	Improve ACC and handle evolving attack types
Singh & Mathai (2019)	Proposed SPELM algorithm and compared with DBN using NSL-KDD dataset	SPELM: 93.20 percent vs. 52.8 percent for ACC; PRE:	SPELM outperforms DBN in accuracy and efficiency	Explore application to larger, more complex datasets and hybrid ML models

		69.49 percent vs. 66.736 percent for DBN.		

RESEARCH METHODOLOGY

This study employs the UNSW-NB15 dataset, applying pre-processing steps like cleaning, encoding, normalization, and ADASYN-based class balancing. Utilising chi-square feature selection allows for the preservation of critical attributes while simultaneously improving the model's performance. Following this, the cleaned-up dataset was split in half: half to be used for model training and half for model testing. For classification, employ an XGBoost model and assess its efficacy via ROC curve analysis, F1, REC, ACC, and PRE. shows the suggested IDS flowchart in Figure 1.

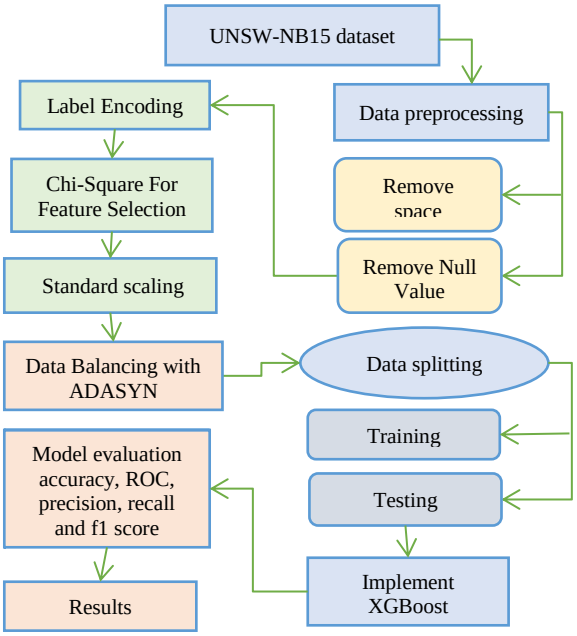

Fig. 1. Proposed Flowchart for Intrusion Detection system

The whole steps of implementation are explained in next section.

Data Gathering and Analysis

The UNSW-NB15 dataset, a new dataset, is referenced in this study. There are a total of 49 attributes in this dataset, with a class label and 25,40, 044 tagged occurrences that are categorised as either normal or attack. Data visualizations such as bar plots and heatmaps were used to examine attack distribution, feature correlations etc., are given below:

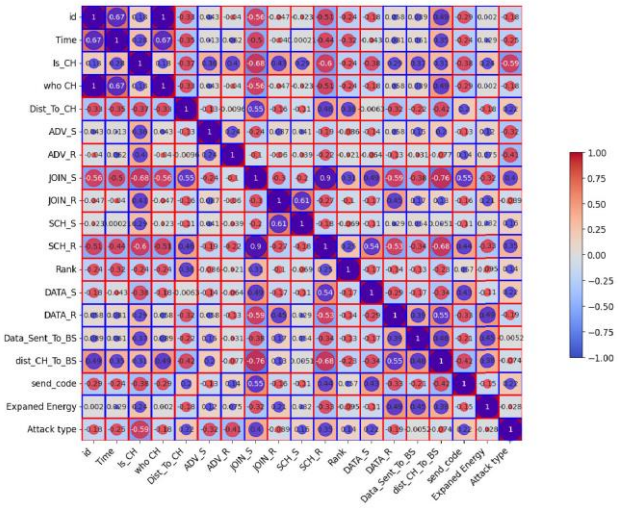

Fig. 2. Sample of Correlation Matrix on UNSW-NB15

Figure 2 provide comprehensive visual overview of inter-feature relationships, highlighting both positive and negative associations among variables such as Time, Dist_To_CH, ADV_S, JOIN_R, Expanded Energy, and Attack type. Each cell encodes the correlation coefficient using a color gradient from blue (strong negative) to red (strong positive), with white indicating near-zero correlation. The circular markers within cells further emphasize the magnitude of these relationships, aiding in intuitive pattern recognition. This matrix is instrumental for feature selection and model refinement, revealing potential redundancies and dependencies critical to cybersecurity analytics.

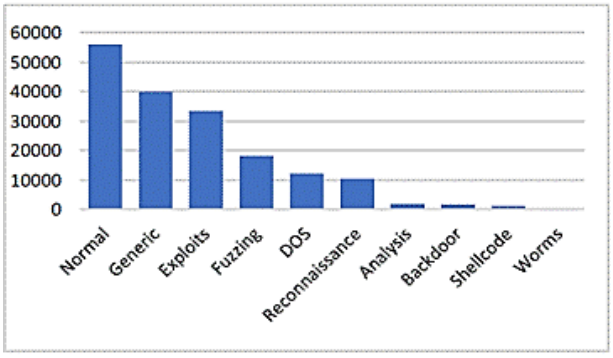

Fig. 3. Number of Records that Represent Normal Traffic and Malicious Types of Attacks in the UNSW-NB15 Dataset.

The UNSW-NB15 dataset includes a wide range of damaging attacks and traffic types, as illustrated in Figure 3. Normal traffic is the dominant type of data, containing more than 50,000 records and the next most prevalent data is the Generic traffic, which has a total of more than 30,000 records and the final and the most prevalent data is the Exploits, with the total of more than 30, 000 records. Fuzzing exhibits a significantly smaller, although still significant, number of 18,000 records and DOS and Reconnaissance attacks take their places, with counts ranging between 10,000 and 12,000. All other attack types Analysis, Backdoor, Shellcode, and Worms occupy a relatively small percentage of the dataset with only fewer than 2,000 records each, which suggests a

very skewed distribution centered around normal traffic, generic detection and attempts to exploit.

Data Pre-processing

Data preparation used the UNSW-NB15 dataset and entailed concatenation, cleaning and feature engineering. Its pre-processing steps involved handling of missing values, duplication, noise removing, encoding, feature selection, normalization and balancing. The most important steps of pre-processing are as follows:

- **Remove Space:** Remove spaces from column names for simpler manipulation, and keep only the first row and remove all others to eliminate duplicate rows from the dataset.
- **Remove Null values:** In order to improve the study's ACC, the wrong values of the attributes ct_flw_http_mthd, is_ftp_login, and attack_cat are removed.

Label Encoding For Data Encoding

Label encoding converts categorical data into numbers, allowing ML algorithms to handle the categorical data. Each distinct category is given an integer in the range 0 to (n -1), n being the number of distinct classes. As an example, using 11 categories, the number 0 through 10 is used.

Feature Selection Using Chi-Square

The term "feature selection" describes the steps used to determine which dataset characteristics are most relevant for building and training a ML model. In order to make AI models more compact and easier to work with, features are included. To find out which attributes are most essential to the target group, and compare the actual and expected frequencies of the categorical data using a statistical filter like chi-square. Features with high chi-square scores or low p-values are retained for improved model ACC.

Standard scalar for Normalization

A normal distribution with a mean of 0 and a standard deviation of 1 was generated by standardising the dataset using the StandardAero() function. Here observe that the standard deviation is divided by the mean of each observation and then subtracted once to achieve this transformation Equation (1)

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

The translated feature value (z), original descriptor values (x), mean (μ), and standard deviation (σ) are some of the variables found in this dataset.

Data Balancing using ADASYN

Data balancing strategies fix the problem of unequal class distributions and stop the model from happening. One adaptable oversampling approach that uses samples from minority classes is adaptive synthetic sampling, or ADASYN. To enhance classifier focus and decision boundaries, ADASYN generates synthetic data around harder-to-learn examples, prioritises samples from minority classes in low-density regions, and estimates the density of those classes.

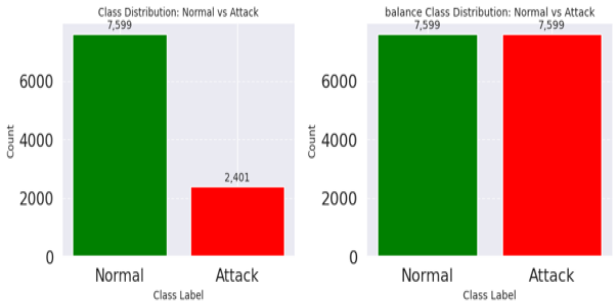

Fig. 4. Before and After Applying Adasyn for Class Blanacing

Figure 4 illustrates the impact of ADASYN on class balancing by comparing the original and resampled distributions of "Normal" and "Attack" instances. The dataset is initially unbalanced, which could lead to biassed model performance. With 7,599 samples in each class, the "Attack" minority class is synthetically extended to have the same size as the majority class after ADASYN is applied. Anomaly detection tasks in particular benefit from this tweak, since it increases the model's robustness for classification and its capacity to learn from patterns that are under-represented.

Data Splitting

The efficacy of the dataset was assessed by dividing it into training and testing subsets. 80% of the dataset was allocated for model development and parameter refining, while the remaining 20% was reserved for performance evaluation and testing.

Proposed Extreme Gradient Boosting (XGBoost)

XGBoost uses DT to generate predictions; it is an ensemble based learning method. Regression issues can be tackled in a few different ways: one is by minimising a loss function that measures the difference between actual and forecasted values. Two possible representations of the XGBoost regression model exist in mathematics Equation (2):

$$y = f(x) \quad (2)$$

where y represents the predicted price of the property, x represents the input feature (i.e., square footage, the number of bedrooms, etc.), and $f(x)$ represents the XGBoost model that predicts y as a result of x. XGBoost creates a sequence of decision trees to compute the $f(x)$ by training them to reach a minimum MSE loss function. The model uses the combined predictions of several DT to arrive at a final forecast. A simplified version of the XGBoost regression model is Equation (3):

$$y = \sum(k = 1\ to\ K)\ fk(x)) \quad (3)$$

$fk(x)$ is the forecast of the kth decision tree and K is the number of DT in the ensemble. Each tree is predicted as a weighted sum of the leaf values of the tree which are trained during the training process. The XGBoost model prediction of the input x is calculated by adding the prediction of all decision trees of the ensemble.

Evaluation Metrics

The suggested design was tested using several metrics to measure its performance. To summarise the results of the classification, a confusion matrix was created. The total number of correct and wrong predictions for each class is displayed in this matrix. Extracting useful metrics from this matrix included TP, FP, TN, and FN. Following the

formulation in (4) to (7), these values were utilised to calculate crucial performance indicators, such as ACC, PRE, REC, and F1:

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

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

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

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

A model's ACC can be defined as the percentage of cases for which it made a correct prediction relative to all instances in the dataset. PRE is the proportion of positive events that the model accurately anticipated as a percentage of all positive occurrences forecasted. The REC ratio is the number of positive events predicted out of all the possible positive instances. The F1 aids in remembering information and accurately recalling it since it is a harmonic mean of the two. With the help of the ROC curve, show how the percentage of FP and the percentage of TP for various decision criteria relate to one another schematically.

Results and Discussion

This section offerings the performance of the suggested model and describes the experimental setup. The experiments were conducted on a robust PC with an Intel Core (TM) i3-1005G1 CPU clocking in at 1.20 GHz, 4 GB of RAM, with Python installed. With 64 GB of RAM, the system can handle applications that require a lot of memory, and it comes with a substantial 40 GB of disc space for data storage. In Table II, show the proposed model's performance summarised. With a PRE of 99.59%, the suggested XGBoost model successfully categorised almost all network activities. The ACC of 99.5% in detecting real incursions and the PRE of 99.8% in minimising false positives demonstrate the model's usefulness. An F1 of 99.6% shows that the model is very reliable and robust for effective IDS in complicated network environments, since it strikes a great balance between REC and PRE.

TABLE II. Results of the Proposed Model for Intrusion Detection

Performance Matrix	Extreme Gradient Boosting (XGBoost)
Accuracy	99.59
Precision	99.8
Recall	99.5
F1-score	99.6

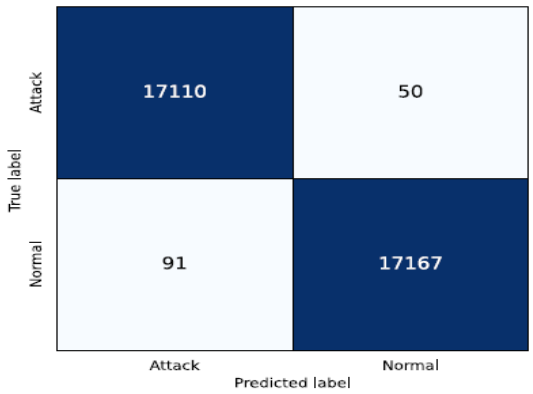

Fig. 5. Confusion Matrix for the XGBoost Model

A confusion matrix showing the results of a classification model is shown in Figure 5. The results of a model that classifies incoming data as "Attack" or "Normal" are shown in this array. Here, the rows show the actual labels and the columns show the expected ones. Matrix data shows that the model properly classified 17,110 occurrences as "Attack" and 17,167 as "Normal." False negatives totalling 50 and false positives totalling 91 occurred when it incorrectly classified 50 "Attack" instances as "Normal" and 91 "Normal" instances as "Attack" respectively. The model seems to be very accurate in general, with few misclassifications in comparison to the overall number of occurrences that were correctly detected.

Fig. 6. ROC Curve for XGBoost Model

In Figure 6, Shows how the TPR and the FPR intersect. Here can see the model's performance illustrated by the orange curve. The fact that the curve remains near the diagonal indicates that the model outperforms random guessing by a little margin. In spite of this, the reported AUC of 99.98 seems at odds with the curve's visual trend; after all, a top-notch classifier would have a ROC curve that is much higher than the diagonal. This discrepancy may indicate either a plotting or evaluation error in the results.

Comparative Analysis

Table III provides a comparison of the proposed XGBoost model's accuracy with that of other current models in order to evaluate its usefulness. Among the traditional ML models, LR achieved moderate performance with an accuracy of 70.5%, NB performed better in terms of PRE at 99%. The DL model, LSTM, showed significant improvement with an accuracy of 91.2%, balanced PRE and recall respectively. XGBoost demonstrated its exceptional capacity to accurately and reliably detect intrusions while minimising false positives by reaching virtually flawless metrics, outperforming all other models by a considerable margin.

TABLE III. Comparison of Different ML and NL Models for Intrusion Detection on UNSW-NB15 Dataset

Model	Accuracy	Precision	Recall	F1-score
LR[19]	70.5	65.9	96.1	78.2
NB[20]	76.5	99	69	82
LSTM[21]	91.2	87.3	80.6	83.8
XGBoost	99.59	99.8	99.5	99.6

The proposed IDS model has several interesting strengths that make it more effective in cybersecurity. Utilising adaptive sampling techniques, it can address the problem of class imbalance by reducing bias in favour of majority classes and improving the detection of unusual attack patterns. XGBoost is appropriate in complex and dynamic network environments

in which the predictive accuracy, robustness and scalability are required to be high. Its high performance in the major metrics proves that it has good classification with few false positives and negatives. These capabilities make them more balanced and smart IDS that can assist in real-time monitoring of threats and decision-making in current digital infrastructures.

Conclusion and Future Study

IDS are an important part of safeguarding digital infrastructure against more advanced cyber-attacks. This paper presented a new AI-based platform that would increase the IDS through the reduction of class imbalance. With the UNSW-NB15 data set and Extreme Gradient Boosting (XGBoost), the highest ACC of 99.59%, a PRE of 99.8%, a REC of 99.5% and an F1 of 99.6% were obtained, which proves the effectiveness of the method in detecting the frequent and rare attack patterns. Conventional methods such as Logistic Regression (70.5%) and the Naive Bayes (76.5%) demonstrated weak results, whereas DL based LSTM had a significant accuracy of 91.2%. XGBoost performs well in IDS but the evaluation on one dataset restricts its usefulness in a generalized setting in most network environment. ROC curve inconsistencies suggest potential issues in metric interpretation, and the computational cost of ADASYN and XGBoost may challenge deployment on low-resource systems. Future work will explore multi-dataset validation, real-time and edge optimization, and integration of explainable AI to enhance scalability, transparency, and practical applicability in dynamic cybersecurity settings.

References

[1] J. Thomas, K. V. Vedi, and S. Gupta, "The Effect and Challenges of the Internet of Things (IoT) on the Management of Supply Chains," *Int. J. Res. Anal. Rev.*, vol. 8, no. 3, 2021.

[2] N. Patel, "Sustainable Smart Cities: Leveraging Iot And Data Analytics For Energy Efficiency And Urban Development," *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 3, pp. 313–219, 2021.

[3] M. Abdur, S. Habib, M. Ali, and S. Ullah, "Security Issues in the Internet of Things (IoT): A Comprehensive Study," *Int. J. Adv. Comput. Sci. Appl.*, 2017, doi: 10.14569/ijacsa.2017.080650.

[4] V. M. L. G. Nerella, "Architecting Secure, Automated Multi-Cloud Database Platforms Strategies for Scalable Compliance.," *Int. J. Intell. Syst. Appl. Eng.*, vol. 9, pp. 128–138, 2021.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[6] A. Jamalipour and S. Murali, "A Taxonomy of Machine-Learning-Based Intrusion Detection Systems for the Internet of Things: A Survey," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9444–9466, Jun. 2022, doi: 10.1109/JIOT.2021.3126811.

[7] S. Thangavel, K. C. Sunkara, and S. Srinivasan, "Software-Defined Networking (SDN) in Cloud Data Centers: Optimizing Traffic Management for Hyper-Scale Infrastructure," *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.*, vol. 3, no. 1, pp. 29–42, 2022, doi: 10.63282/3050-9246.IJETCSIT-V3I3P104.

[8] D. Preuveneers and W. Joosen, "Sharing Machine Learning Models as Indicators of Compromise for Cyber Threat Intelligence," *J. Cybersecurity Priv.*, vol. 1, no. 1, pp. 140–163, Feb. 2021, doi: 10.3390/jcp1010008.

[9] A. S. Sohal, R. Sandhu, S. K. Sood, and V. Chang, "A cybersecurity framework to identify malicious edge device in fog computing and cloud-of-things environments," *Comput. Secur.*, vol. 74, pp. 340–354, May 2018, doi: 10.1016/j.cose.2017.08.016.

[10] S. B. Venkata Naga, K. C. Sunkara, S. Thangavel, and R. Sundaram, "Secure and Scalable Data Replication Strategies in Distributed Storage Networks," *Int. J. AI, BigData, Comput. Manag. Stud.*, vol. 2, no. 2, pp. 18–27, 2021, doi: 10.63282/3050-9416.IJAIBDCMS-V2I2P103.

[11] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," *J. Inf. Secur. Appl.*, vol. 44, pp. 80–88, Feb. 2019, doi: 10.1016/j.jisa.2018.11.007.

[12] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *J. Big Data*, vol. 7, no. 1, p. 41, Dec. 2020, doi: 10.1186/s40537-020-00318-5.

[13] M. H. Kabir, M. S. Rajib, A. S. M. T. Rahman, M. M. Rahman, and S. K. Dey, "Network Intrusion Detection Using UNSW-NB15 Dataset: Stacking Machine Learning Based Approach," in *2022 International Conference on Advancement in Electrical and Electronic Engineering, ICAEEE 2022*, 2022. doi: 10.1109/ICAEEE54957.2022.9836404.

[14] D. Gupta and A. K. Saxena, "Using Machine Learning for Network Intrusion Detection," in *2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, 2022, pp. 1–5. doi: 10.1109/ICATIECE56365.2022.10047400.

[15] S. Umamaheshwari, S. A. Kumar, and S. Sasikala, "Towards Building Robust Intrusion Detection System in Wireless Sensor Networks using Machine Learning and Feature Selection," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, ICAECA 2021*, 2021. doi: 10.1109/ICAECA52838.2021.9675609.

[16] S. Das, M. Ashrafuzzaman, F. T. Sheldon, and S. Shiva, "Network Intrusion Detection using Natural Language Processing and Ensemble Machine Learning," in *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020*, 2020. doi: 10.1109/SSCI47803.2020.9308268.

[17] A. Srivastava, A. Agarwal, and G. Kaur, "Novel Machine Learning Technique for Intrusion Detection in Recent Network-based Attacks," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, Nov. 2019, pp. 524–528. doi: 10.1109/ISCON47742.2019.9036172.

[18] K. Singh and K. J. Mathai, "Performance Comparison of Intrusion Detection System Between Deep Belief Network (DBN)Algorithm and State Preserving Extreme Learning Machine (SPELM) Algorithm," in *Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019*, 2019. doi: 10.1109/ICECCT.2019.8869492.

[19] M. Shushlevska, D. Efnusheva, G. Jakimovski, and Z. Todorov, "Anomaly Detection with Various Machine Learning Classification Techniques over UNSW-NB15 Dataset," in *Proceedings of International Conference on Applied Innovation in IT*, 2022.

[20] G. Kocher and G. Kumar, "Analysis of Machine Learning Algorithms with Feature Selection for Intrusion Detection using UNSW-NB15 Dataset," *Int. J. Netw. Secur. Its Appl.*, 2021, doi: 10.5121/ijnsa.2021.13102.

[21] L. Van Duong, "Network anomaly detection technique based on LSTM network and UNSW-NB15 dataset," *Int. J. Adv. Trends Comput. Sci. Eng.*, 2020, doi: 10.30534/ijatcse/2020/340942020.