Enhancing attack detection in IoT through integration of weighted emphasis formula with XGBoost

Januar Al Amien[1,2], Hadhrami Ab Ghani[2], Nurul Izrin Md Saleh[3], Soni[1,2]
[1]Department of Informatics Engineering, Faculty Computer Science, University Muhammadiyah Riau, Riau, Indonesia
[2]Department of Data Science, Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Kelantan, Malaysia
[3]Department of Software Engineering, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Article Info

Article history:

Received Jul 10, 2024
Revised Oct 8, 2024
Accepted Oct 30, 2024

Keywords:

Imbalanced ratio
Internet of things
Intrusion detection system
IoTID20 dataset
Machine learning
Weighted XGBoost

ABSTRACT

This research addresses the challenge of detecting attacks in the internet of things (IoT) environment, where minority classes often go unnoticed due to the dominance of majority classes. The primary objective is to introduce and integrate the imbalance ratio formula (IRF) into the XGBoost algorithm, aiming to provide greater emphasis on minority classes and ensure the model's focus on attack detection, particularly in binary and multiclass scenarios. Experimental validation using the IoTID20 dataset demonstrates the significant enhancement in attack detection accuracy achieved by integrating IRF into XGBoost. This enhancement contributes to the consistent improvement in distinguishing attacks from normal traffic, thereby resulting in a more reliable attack detection system in complex IoT environments. Moreover, the implementation of IRF enhances the robustness of the XGBoost model, enabling effective handling of imbalanced datasets commonly encountered in IoT security applications. This approach advances intrusion detection systems by addressing the challenge of class imbalance, leading to more accurate and efficient detection of malicious activities in IoT networks. The practical implications of these findings include the enhancement of cybersecurity measures in IoT deployments, potentially mitigating the risks associated with cyber threats in interconnected smart environments.

Corresponding Author:

Januar Al Amien
Department of Informatics Engineering, Faculty Computer Science, University Muhammadiyah Riau
Riau, Indonesia
Email: januaralamien@umri.ac.id

1. INTRODUCTION

Technology has become a basic necessity in society, bringing about significant changes in their lifestyles [1]. One of these technological advancements is the internet of things (IoT), which consists of interconnected everyday devices equipped with lightweight processors and network cards. These devices can be managed through web services and/or other types of interfaces [2]. A multitude of physical objects, such as temperature sensors, smartphones, air conditioners, and even smart power grids, are directly involved in the Internet, enabling environmental monitoring and collaborative task execution without human intervention [3].

The advancement of the internet of things (IoT) has played a significant role in everyday life. However, alongside the convenience and efficiency offered by IoT, the emergence of security challenges has become a primary concern. In 2018, Symantec's report noted that the total number of targeted attacks on IoT devices exceeded 57,000, with over 5,000 attacks recorded each month. Attackers employed various hacking techniques such as denial of service (DoS), distributed DoS (DDoS), ransomware, and other botnet attacks to exploit vulnerabilities in IoT systems and networks [4]. Nevertheless, they provide insights into traffic behavior and can help identify crucial information. One approach to recognizing changes in network behavior is through an intrusion detection system (IDS) that assists in detecting, assessing, and identifying unauthorized usage in information systems [5].

The IDS plays a crucial role in addressing potential network threats before exhibiting malicious behavior. IDS is responsible for identifying malicious activities on a host that can subsequently spread to other hosts within the network. Research utilizing IDS datasets has been conducted. Our innovative IDS model employs statistical pre-processing, Stack Denoising Auto Encoder (SDAE) for data reduction, and a transformer-enhanced classification approach, demonstrated on the NSL-KDD dataset [6]. In the study by sun *et al.* [7], the UNSW-NB15 dataset was employed for a Random Forest classification model. The ensemble model applied to the NSL-KDD, Kyoto, and CSE-CIC-IDS-2018 datasets yielded satisfactory results [8]. Experiments were conducted on the CSE-CIC-DS2018 dataset, combining convolutional neural network (CNN) and recurrent neural network (RNN) models [9]. Experimentation on the Bot-IoT dataset using the proposed method proved efficient and achieved an average accuracy exceeding 96% [10]. In this study, the authors refer to the IoTID20 dataset [11]. Asserts the existence of various types of attacks on the internet of things (IoT), including data exfiltration, DoS and DDoS attacks, Keylogging, as well as operating system (OS scan) and service scanning (service scan). Ullah in [12] introduced a dataset named IoTID20, which contains diverse types of IoT attacks and families. IoTID20 was developed for the detection of abnormal behavior in IoT, encompassing Mirai attacks, DoS, Scan, MITM ARP spoofing, scan host port, and Mirai-UDP [13].

However, an issue arises in the IoTID20 dataset, namely, imbalance. Imbalance is a novel concern in the field of machine learning, where imbalance occurs when the number of samples in one class is greater than the other in a dataset with two or multiple classes [14]. The consequence is that the model tends to learn less about minority classes, resulting in training bias towards the majority class [15]. To address the imbalance issue in the data, various sampling techniques have been proposed, such as oversampling, undersampling, random sampling, and others [14]. Several studies have investigated the imbalance problem in multi-class scenarios. For instance, utilized a combination of synthetic minority over-sampling technique (SMOTE) and undersampling based on gaussian mixture model (GMM) on the UNSW-NB15 and CICIDS2017 datasets [16]. Attack categories include common types such as DoS, DDoS, Botnet, PortScan, web attacks, and so on. Another study from Mqadi *et al.* [17] employed undersampling based on the near-miss algorithm with random forest. Therefore, a model is required to produce more optimal results, which can be achieved by utilizing a machine learning approach to address the imbalance in the IoTID20 dataset. In this research, the imbalance issue in multiclass is tackled by employing an imbalance ratio, referred to as imbalance ratio formula (IRF) [18], where each minority class is given weighted emphasis to ensure the model pays more attention to the minority classes.

Machine learning is a scientific exploration of algorithms and statistical models applied by computer systems to perform specific tasks without requiring direct programming [19]. Currently, anomaly detection techniques in networks generally rely on machine learning approaches, such as KNN and SVM [20]. According to research, some IDS use classification algorithms like decision trees, SVM, K-nearest, and others use feature selection [21]. In this study, the authors refer to the LightGBM approach. light gradient boosting machine (LightGBM) is one of the latest research findings in the gradient boosting framework that utilizes tree-based learning algorithms [22]. LightGBM, as a dominant ensemble method, utilizes the decision tree algorithm and is often applied to classification tasks due to its superiority [7]. For this research, the model used is XGBoost to provide better performance. The primary contributions of this study include:

- Introducing a novel formula termed IRF to apply weighted emphasis on minority classes, ensuring the model focuses on the minority classes, particularly in binary and multiclass scenarios.
- Integrating IRF into XGBoost to enhance performance in the detection of attacks within the IoT environment, thereby achieving improved accuracy and efficiency.

2. METHOD

In response to the prevalent security challenges in the internet of things (IoT) environment, we propose a specifically tailored intrusion detection methodology. The proposed approach includes the implementation of an IDS designed to address the unique characteristics of IoT. This strategic methodology is crafted to provide robust protection against evolving security threats in the dynamic IoT ecosystem. By addressing specific challenges in the IoT domain, our methodology aims to enhance the security posture and resilience of IoT devices and systems. Figure 1 illustrates the implementation of the analysis using the

IoTID20 dataset, involving a series of preprocessing steps such as label encoding, numerical transformation, and normalization. To tackle class imbalance, class weight techniques are applied to enhance the role of minority classes in the model. The data is then partitioned with an 80% allocation for training and 20% for testing, with the XGBoost model chosen as the primary algorithm. Evaluation of the analysis results using a confusion matrix provides in-depth insights into the model's ability to handle class imbalance in the IoTID20 dataset.

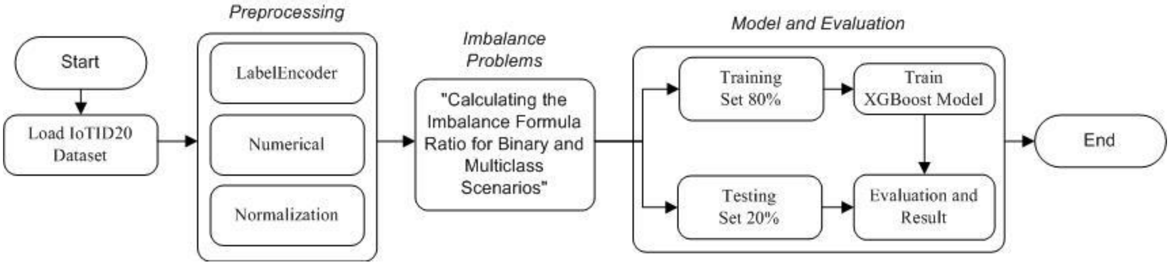

Figure 1. Diagram of model construction

2.1. Dataset IoTID20

The original IoTID20 dataset consists of 625,783 entries with 86 features per entry [23]. After eliminating duplicate data, the number of entries is reduced to 461,696 with 86 features. The purpose of the duplicate elimination process is to clean the dataset from potential redundant data, ensuring the accuracy of the analysis on the dataset. Table 1 shows the class distribution for both binary and multiclass classification: binary categorizes data into anomaly and normal, while multiclass provides finer categories such as Mirai, scan, DoS, and others, with a consistent total of 461,696 entries across classifications. The classification of class attributes involves the utilization of the 'Label' and 'Cat' attributes.

Table 1. Binary and multiclass class distribution IoTID20 dataset

	Label	Class number
Binary	Anomaly	423098
	Normal	38598
	Total	**461696**
	Cat	**Class number**
Multiclass	Mirai	281102
	Scan	59390
	DoS	56744
	Normal	38598
	MITM ARP spoofing	25862
	Total	**461696**

In the preprocessing phase, the code undergoes several crucial steps to enhance the dataset's suitability for machine learning tasks. Initially, it uses LabelEncoder to convert categorical features, namely \`Src_IP\`, and \`Dst_IP\`, into numerical representations. The 'Flow_ID' and 'Timestamp' feature will be removed. The code then addresses potential issues related to infinite values in certain columns by replacing them with large finite values. Furthermore, to ensure the robustness of the dataset, the data is scaled using RobustScaler, a technique designed to reduce sensitivity to outliers. These preprocessing steps collectively contribute to optimizing the dataset for subsequent machine learning models, enhancing robust performance.

2.2. Imbalance ratio formula

To enhance and advance the earlier method introduced in binary classification [12], this research presents a multi-class classification strategy incorporating a novel imbalance ratio approach. The antecedent investigation focused on binary classification challenges, primarily addressing distinctions between two classes. In the current study, we broaden the scope to investigate binary and multiclass classification concerns. The procedural steps for computing the IRF for a given dataset are as follows [24]:

- Find the number of samples in each class.

- For each class i, calculate the number of samples in the majority class *(N_i)* and the number of samples in the minority class *(n_i)*.
- Calculate the imbalance ratio *(IR_i)* for each class *i* as follows:

$$IR_i = N_i/n_i \tag{1}$$

- Calculate the IRF value for the dataset as the maximum imbalance ratio across all classes:

$$IFR = max(IR_1, IR_2, \dots, IR_k) \tag{2}$$

- Calculate the average of the values obtained in step 2.
- Return the result.

Where k is the total number of classes in the dataset.

2.3. XGBoost

According to Chen and Guestrin [25], These formulas represent key components of the decision tree model within the framework of gradient boosting. In the context of the gradient boosting algorithm, the IRFst formula $\mathcal{L}^{(t)}$ denotes the loss function at iteration (t), encompassing terms related to prediction errors, the current model's predictions, and a regularization component. The second formula w_j^* calculates the optimal weight for a specific node in the decision tree, considering the gradients and Hessians of the loss function, with an added regularization term.

The third formula $\mathcal{L}^{(t)}(q)$ defines the loss function for tree pruning at iteration (t), incorporating terms related to the sum of gradients, Hessians, a regularization parameter λ, and a pruning parameter γ. Lastly, the fourth formula $\mathcal{L}_{split}$ represents the loss function guiding the selection of a split at a tree node, involving sums of gradients and Hessians for both the left and right child nodes, regularization, and a pruning term γ. These formulations collectively contribute to the effective training and optimization of the gradient boosting algorithm. These formulas are part of the decision tree model used in gradient boosting methods. Here's a brief explanation for each formula:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} I\left(\hat{\mathrm{y}}_{\mathrm{i}}, y_{\mathrm{i}}{}^{(t-1)} + f_t(x_{\mathrm{i}})\right) + \Omega(f_t), \tag{3}$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} [l\left(\hat{y}, y_{\mathrm{i}}{}^{(t-1)}\right) + g_{\mathrm{i}}\, f_t(x_{\mathrm{i}}) + \frac{1}{2}\, h_i\, ft^2(x_{\mathrm{i}})] + \Omega(f_t), \tag{4}$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} (g_{\mathrm{i}} f_t(x_{\mathrm{i}}) + \frac{1}{2} h_{\mathrm{i}} f_t^2(x_{\mathrm{i}})] + \Omega(f_t), \tag{5}$$

$$w_j^* = -\frac{(\sum_{i\in I_j} g_{\mathrm{i}})^2}{\sum_{i\in I_j} h_i + \lambda}, \tag{6}$$

$$\mathcal{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i\in I_j} g_{\mathrm{i}})^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T, \tag{7}$$

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_{\mathrm{i}})^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_{\mathrm{i}})^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_{\mathrm{i}})^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma, \tag{8}$$

2.4. Evaluation

In the context of binary and multiclass classification, specifically for the Label Anda Cat task, evaluation metrics are generated similarly to the process followed in binary and multiclass classification. These metrics provide a quantitative assessment of the model's performance in handling multiple classes, offering insights into aspects such as precision, recall, F1-score, and accuracy for each class within the binary and multiclass classification problem. This provides a solid foundation for evaluating the effectiveness of the model in distinguishing between different classes in the binary and multiclass classification dataset [26].

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

$$F1 = 2 * \frac{Presisi*Recall}{Presisi+Recall} \tag{11}$$

$$FNR = \frac{FN}{TP+FN} \tag{12}$$

$$PR = \frac{FP}{FP+TN} \tag{13}$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{14}$$

3. RESULTS AND DISCUSSION

3.1. Imbalance ratio formula

The purpose of the results from Table 2 is to determine the class weights applied to each class within the dataset. These calculations yield the class weights for each class, expressed as the IRF. In binary classes, the anomaly class has an IRF of 0.54569846, while the normal class has an IRF of 5.97064434. The total IRF for binary classes is 10.94128860. Additionally, for multiclass classes, the Mirai class has an IRF of 1.554793736, the scan class has an IRF of 3.570458588, the DoS class has an IRF of 0.328490014, the normal class has an IRF of 2.392331209, and the MITM ARP Spoofing class has an IRF of 1.627294516. These results demonstrate the weights assigned to each class to address the imbalance within the dataset.

Table 2. Class weights based on IRF calculation

	Label	Class number	Class of number IRF
Binary	Anomaly	423098	0.54569846
	Normal	38598	5.97064434
	Total	**461696**	
Multiclass	**Cat**	**Class number**	**Class of number IRF**
	Mirai	281102	1.554793736
	Scan	59390	3.570458588
	DoS	56744	0.328490014
	Normal	38598	2.392331209
	MITM ARP Spoofing	25862	1.627294516
	Total	**461696**	

3.2. XGBoost model and evaluation

Table 3 presents a comprehensive evaluation of the IRF model's performance in both binary and multiclass scenarios. In the binary analysis, metrics including accuracy, precision, recall, and F1-score are detailed for the "anomaly" and "normal" classes. Notably, the "anomaly" class achieves exceptional performance with an accuracy of 0.999984, precision of 0.99975, recall of 1.00000, and an F1-score of 0.99988. Similarly, the "normal" class achieves perfect accuracy (1.00000) with high precision (0.99998) and recall (0.99999). Figure 2 provides a visual depiction of these metrics for clarity and enhanced interpretation.

In the multiclass assessment, classes such as "DoS," "MITM ARP Spoofing," "Mirai," "normal," and "scan" are evaluated. Key highlights include the "DoS" class with an accuracy of 0.999912, precision of 1.00000, recall of 0.99941, and an impressive F1-score of 0.99970. "MITM ARP Spoofing" also demonstrates exceptional accuracy and robust precision and recall metrics. Similarly, classes like "Mirai," "Normal," and "Scan" consistently exhibit strong performance across all evaluated metrics, underscoring the IRF model's efficacy in accurately classifying instances across diverse classes. A graphical representation of these multiclass evaluation metrics in Figure 2 complements the textual findings. Overall, IRF proves effective in improving the detection of minority classes across various scenarios. Although it requires more computational resources in multiclass models, this integration demonstrates IRF's potential for application in attack detection systems within IoT environments. It significantly enhances performance in threat detection.

Table 3. Evaluation results of IRF model performance in binary and multiclass scenarios

Label	Accuracy	Precision	Recall	F1-score	Time (Sec)
Anomaly	**0.999984**	0.99975	1.00000	0.99988	23.501898
Normal		1.00000	0.99998	0.99999	
Average		**0.999984**	**0.999984**	**0.999984**	
Cat	**Accuracy**	**Precision**	**Recall**	**F1-score**	
DoS		1.00000	0.99941	0.99970	
MITM ARP Spoofing		0.99957	1.00000	0.99978	
Mirai	**0.999912**	0.99995	0.99996	0.99996	50.832946
Normal		0.99975	1.00000	0.99988	
Scan		0.99987	0.99993	0.99990	
Weighted average		**0.999912**	**0.999912**	**0.999912**	

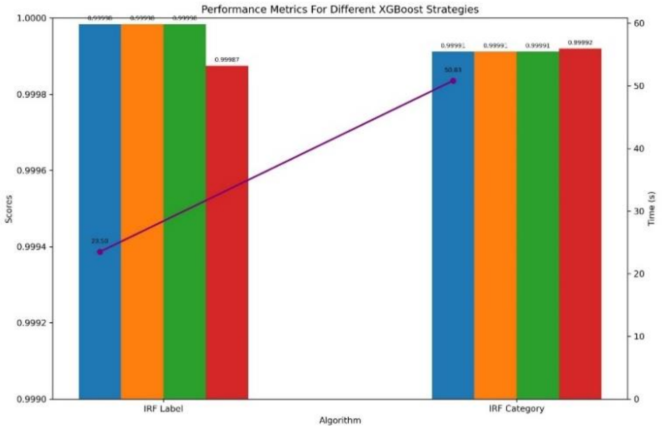

Figure 2. Performance metrics for different XGBoost label strategies

Table 4 summarizes the performance metrics of various algorithms for binary and multiclass classification. While methods like shallow neural networks (100% accuracy) and random forest (99.96%) demonstrate strong results, the proposed XGBoost Multiclass IRF achieves the highest accuracy of 99.99%, outperforming prior approaches. This highlights the effectiveness of XGBoost in handling multiclass classification tasks with exceptional precision.

Table 4. Performance metrics for different XGBoost binary and multiclass strategies

Algorithm	Accuracy	Precision
Ullah, Safi *et al.* [27]	DCNN	98%
Y. Song [28]	Deep Learning-MCC	94%
R. Qaddoura [29]	Single Hidden Layer Feed-Forward Neural Network (SLFN)	98%
A. A. Alsulami [30]	Shallow Neural Networks (SNNs)	100%
P. Maniriho [31]	Random Forest (DoS, MITM, Scan)	99,96%
K. Albulayhi [32]	Intersection Mathematical (IMF) and Union Mathematical (UMF)	99.7% and 99,7%
I. Ullah [12]	Decision Tree (Sub-Category)	88%
Proposed method	XGBoost Multiclass IRF	99,99%

4. CONCLUSION

This study successfully introduces and integrates the IRF into the XGBoost algorithm to enhance attack detection performance in internet of things (IoT) environments. Experimental results demonstrate that applying IRF effectively addresses class imbalance within datasets. In binary scenarios, IRF increased the recall metric from 0.988914 to 0.998635, with a negligible decrease in accuracy from 0.998971 to 0.998635. Similarly, in multiclass scenarios, IRF showed balanced performance with a slight decrease in accuracy from 0.993253 to 0.992733, though processing time increased from 37.63 seconds to 49.40 seconds. The implementation of IRF not only improves the detection of minority classes but also demonstrates significant potential for application in intrusion detection systems within IoT environments. Although IRF requires greater computational resources, the enhanced performance in detecting malicious activities substantiates its effectiveness and reliability as a promising solution to bolster cybersecurity measures in IoT settings. Future work should explore optimizing the IRF algorithm for computational efficiency and test its application on various IoT datasets to assess generalization and robustness.

Additionally, the development of adaptive detection systems with continuous learning capabilities should be investigated to improve responsiveness to emerging threats in dynamic IoT environments.

ACKNOWLEDGEMENTS

The authors express their profound appreciation for the support and assistance provided by Universitas Muhammadiyah Riau, Indonesia, and Universiti Malaysia Kelantan, Malaysia, which have enabled the completion of this article. Additionally, the authors extend their gratitude to fellow researchers who contributed, both formally and informally, to the preparation of this paper.

REFERENCES

[1] K. Jiang, W. Wang, A. Wang, and H. Wu, “Network intrusion detection combined hybrid sampling with deep hierarchical network,” *IEEE Access*, vol. 8, pp. 32464–32476, 2020, doi: 10.1109/access.2020.2973730.

[2] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset,” *Future Generation Computer Systems*, vol. 100, pp. 779-796, 2019, doi: 10.1016/j.future.2019.05.041.

[3] H. A. Abdul-Ghani, D. Konstantas, and M. Mahyoub, “A comprehensive IoT attacks survey based on a building-blocked reference model,” *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 9, no. 3, 2018, doi: 10.14569/IJACSA.2018.090349.

[4] N. Koroniotis, N. Moustafa, and E. Sitnikova, “A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework,” *Future Generation Computer Systems*, vol. 110, pp. 91–106, 2020, doi: 10.1016/j.future.2020.03.042.

[5] L. Vigoya, D. Fernandez, V. Carneiro, and F. J. Nóvoa, “IoT dataset validation using machine learning techniques for traffic anomaly detection,” *Electronics*, vol. 10, no. 22, 2021, doi: 10.3390/electronics10222857.

[6] H. Hanafi, A. Pranolo, Y. Mao, T. Hariguna, L. Hernandez, and N. F. Kurniawan, “IDSX-Attention: intrusion detection system (IDS) based hybrid MADE-SDAE and LSTM-attention mechanism,” *International Journal of Advances in Intelligent Informatics*, vol. 9, no. 1, pp. 121–135, 2023, doi: 10.26555/ijain.v9i1.942.

[7] X. Sun, M. Liu, and Z. Sima, “A novel cryptocurrency price trend forecasting model based on LightGBM,” *Finance Research Letters*, vol. 32, 2020, doi: 10.1016/j.frl.2018.12.032.

[8] M. Bakro *et al.*, “Efficient intrusion detection system in the cloud using fusion feature selection approaches and an ensemble classifier,” *Electronics*, vol. 12, no. 11, 2023, doi: 10.3390/electronics12112427.

[9] M. A. Khan, “HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system,” *Processes*, vol. 9, no. 5, 2021, doi: 10.3390/pr9050834.

[10] M. Shafiq, Z. Tian, A. K. Bashir, X. Du and M. Guizani, "CorrAUC: a malicious bot-iot traffic detection method in IoT network using machine-learning techniques," *IEEE Internet of Things Journal*, vol. 8, no. 5, 2021, doi: 10.1109/JIOT.2020.3002255.

[11] A. Churcher *et al.*, “An experimental analysis of attack classification using machine learning in IoT networks,” *Sensors (Switzerland)*, vol. 21, no. 2, pp. 1–32, 2021, doi: 10.3390/s21020446.

[12] I. Ullah and Q. H. Mahmoud, “A scheme for generating a dataset for anomalous activity detection in IoT networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12109 LNAI, pp. 508–520, 2020, doi: 10.1007/978-3-030-47358-7_52.

[13] K. Albulayhi, A. A. Smadi, F. T. Sheldon, and R. K. Abercrombie, “Iot intrusion detection taxonomy, reference architecture, and analyses,” *Sensors*, vol. 21, no. 19, 2021. doi: 10.3390/s21196432.

[14] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review,” *Journal of Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00349-y.

[15] M. Son, S. Jung, J. Moon and E. Hwang, "BCGAN-Based Over-Sampling Scheme for Imbalanced Data," *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Busan, Korea (South), 2020, pp. 155-160, doi: 10.1109/BigComp48618.2020.00-83.

[16] H. Zhang, L. Huang, C. Q. Wu, and Z. Li, “An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset,” *Computer Networks*, vol. 177, 2020, doi: 10.1016/j.comnet.2020.107315.

[17] N. M. Mqadi, N. Naicker, and T. Adeliyi, “Solving misclassification of the credit card imbalance problem using near miss,” *Mathematical Problems in Engineering*, vol. 2021, 2021, doi: 10.1155/2021/7194728.

[18] J. Al Amien, H. Ab Ghani, N. I. Md Saleh, E. Ismanto, and R. Gunawan, “Intrusion detection system for imbalance ratio class using weighted XGBoost classifier,” *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 21, no. 5, p. 1102, 2023, doi: 10.12928/telkomnika.v21i5.24735.

[19] B. Mahesh, “Machine learning algorithms-a review,” *International Journal of Science and Research (IJSR)*, 2018, doi: 10.21275/ART20203995.

[20] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, “BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset,” *IEEE Access*, vol. 8, pp. 29575–29585, 2020, doi: 10.1109/access.2020.2972627.

[21] F. E. Laghrissi, S. Douzi, K. Douzi, and B. Hssina, “Intrusion detection systems using long short-term memory (LSTM),” *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00448-4.

[22] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, “Diagnosis of diabetes mellitus using gradient boosting machine (Lightgbm),” *Diagnostics*, vol. 11, no. 9, 2021, doi: 10.3390/diagnostics11091714.

[23] N. Islam *et al.*, “Towards machine learning based intrusion detection in IoT networks,” *Computers, Materials & Continua*, vol. 69, no. 2, pp. 1801–1821, 2021, doi: 10.32604/cmc.2021.018466.

[24] J. Al Amien, H. Ab Ghani, N. Izrin Md Saleh, S. Soni, Y. Fatma, and R. Hayami, “A comprehensive evaluation of multiclass imbalance techniques with ensemble models in IoT environments,” *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 22, no. 3, pp. 690-701, 2024, doi: 10.12928/telkomnika.v22i3.25887.

[25] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 19, no. 6, pp. 785–794. doi: 10.1145/2939672.2939785.

[26] J. G. Almaraz-Rivera, J. A. Perez-Diaz, and J. A. Cantoral-Ceballos, “Transport and application layer DDoS attacks detection to IoT devices by using machine learning and deep learning models,” *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093367.

[27] S. Ullah *et al.*, “A New intrusion detection system for the internet of things via deep convolutional neural network and feature engineering,” *Sensors 2022*, vol. 22, no. 10, p. 3607, 2022, doi: 10.3390/S22103607.

[28] Y. Song, S. Hyun, and Y.-G. Cheong, “Analysis of autoencoders for network intrusion detection†,” *Sensors*, vol. 21, no. 13, 2021, doi: 10.3390/s21134294.

[29] R. Qaddoura, A. M. Al-Zoubi, I. Almomani, and H. Faris, “A multi-stage classification approach for iot intrusion detection based on clustering with oversampling,”*Applied Science*, vol. 11, no. 7, 2021, doi: 10.3390/app11073022.

[30] A. A. Alsulami, Q. Abu Al-Haija, A. Tayeb, and A. Alqahtani, “An intrusion detection and classification system for IoT traffic with improved data engineering,” *Applied* Science, vol. 12, no. 23, p. 12336, 2022, doi: 10.3390/APP122312336.

[31] P. Maniriho, E. Niyigaba, Z. Bizimana, V. Twiringiyimana, L. J. Mahoro and T. Ahmad, "Anomaly-based intrusion detection approach for IoT Networks using machine learning," *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, Surabaya, Indonesia, 2020, pp. 303-308, doi: 10.1109/CENIM51130.2020.9297958.

[32] K. Albulayhi, Q. A. Al-Haija, S. A. Alsuhibany, A. A. Jillepalli, M. Ashrafuzzaman, and F. T. Sheldon, “IoT intrusion detection using machine learning with a novel high performing feature selection method,” *Applied Science*, vol. 12, no. 10, 2022, doi: 10.3390/app12105015.

BIOGRAPHIES OF AUTHORS

Januar Al Amien completed education bachelor's degree in the Informatics Engineering Department, STMIK-AMIK Riau. And master's degree in Master of Information Technology at Putra Indonesia University Padang. Now working as a lecturer in the Department of Computer Science, University Muhammadiyah of Riau. With research interests in the field of Machine learning algorithms and AI. He can be contacted at email: januaralamien@umri.ac.id.

Hadhrami Ab Ghani received his bachelor degree in electronics engineering from Multimedia University Malaysia (MMU) in 2002. In 2004, he completed his master's degree in Telecommunication Engineering at The University of Melbourne. He then pursued his Ph.D. at Imperial College London in intelligent network systems and completed his Ph.D. in 2011. He can be contacted at email: hadhrami.ag@umk.edu.my.

Nurul Izrin Md Saleh obtained a bachelor's degree in information technology from Multimedia University Malaysia (MMU). He completed his master's degree in computer science at The University of Putra Malaysia. Then complete a Ph.D. at The University of Brunel London in the same field of study. She can be contacted at email: izrin@utem.edu.my.

Soni received the bachelor's degree in Informatics Engineering Department from STMIK AMIK Riau, Indonesia and the master’s degree in computer science Islamic University of Indonesia. now works as a lecturer at the Faculty of Computer Science, University of Muhammadiyah Riau. His current research interests include Machine learning algorithms and AI. He can be contacted at email: soni@umri.ac.id.