

Improving imbalanced class intrusion detection in IoT with ensemble learning and ADASYN-MLP approach

Soni^{1,2}, Muhammad Akmal Remli^{1,3}, Kauthar Mohd Daud⁴, Januar Al Amien^{1,2}

¹Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Kelantan, Malaysia

²Faculty of Computer Science, Universitas Muhammadiyah Riau, Pekanbaru, Indonesia

³Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, Kelantan, Malaysia

⁴Institute Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Article Info

Article history:

Received Apr 22, 2024

Revised Jul 30, 2024

Accepted Aug 5, 2024

Keywords:

ADASYN with MLP

Classification

LightGBM

LightGBM ADASYN with

MLP

ToN_IoT

XGBoost

ABSTRACT

The exponential growth of the internet of things (IoT) has revolutionized daily activities, but it also brings forth significant vulnerabilities. intrusion detection systems (IDS) are pivotal in efficiently detecting and identifying suspicious activities within IoT networks, safeguarding them from potential threats. It proposes a ensemble approach aimed at enhancing model performance in such scenarios. Recognizing the unique challenges posed by imbalanced class distribution, the research employs three sampling techniques LightGBM adaptive synthetic sampling (ADASYN) with multilayer perceptron (MLP), XGBoost ADASYN with MLP, and LightGBM ADASyn with XGBoost to address class imbalance effectively. Evaluation confusion matrix performance metrics underscores the efficacy of ensemble models, particularly LightGBM ADASYN with MLP, XGBoost ADASYN with MLP, and LightGBM ADASYN with XGBoost, in mitigating imbalanced class issues. The LightGBM ADASYN with MLP model stands out with 99.997% accuracy, showcasing exceptional precision and recall, demonstrating its proficiency in intrusion detection within minimal false positives negatives. Despite computational demands, integrating XGBoost within ensemble frameworks yields robust intrusion detection results, highlighting a balanced trade-off between accuracy, precision, and recall. This research offers valuable insights into the strengths with different ensemble models, significantly contributing to the advancement of accurate and reliable IDS in realm of IoT.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Muhammad Akmal Remli

Faculty of Data Science and Computing, Universiti Malaysia Kelantan

Kelantan, Kota Bharu, Malaysia

Email: akmal@umk.edu.my

1. INTRODUCTION

The internet of things (IoT) paradigm, marked by the proliferation of interconnected devices, has ushered in an era of unprecedented data generation [1]–[3]. Within this expansive ecosystem, intrusion detection systems (IDS) play a critical role in safeguarding IoT networks against potential threats. However, the reliability and accuracy of these systems are often compromised by the challenge of unbalanced data distribution, particularly across diverse IoT classes. A persistent challenge in this domain is the presence of unbalanced data, particularly within class distributions, which compromises the efficacy of intrusion detection models. A persistent challenge in this domain is the presence of unbalanced data, particularly within class distributions, which compromises the efficacy of intrusion detection models. The pervasive

challenge of class imbalance in intrusion detection datasets [4], [5]. Cao *et al.* [6] noted that traditional intrusion detection models, when confronted with imbalanced data, tend to exhibit a bias toward the majority class, leading to suboptimal performance in detecting minority class intrusions. A study by Alzahrani and Alenazi [7] underscores the vulnerability of IoT networks to diverse security threats and the pivotal role IDS plays in mitigating these risks. It emphasizes the need for robust intrusion detection mechanisms in IoT environments.

The ToN IoT datasets, comprising Fridge, Garage Door, GPS Tracker, Modbus, Motion Light, Thermostat, and Weather classes, inherently present challenges associated with class imbalance. In such datasets, instances of certain IoT classes may be disproportionately underrepresented, complicating the training of intrusion detection models. This imbalance can lead to biased learning, where the model may favor the majority class, jeopardizing its ability to accurately detect instances of intrusion within minority classes [8]. Ensemble learning techniques have gained prominence for their efficacy in improving intrusion detection performance. Zhang *et al.* [9], Khan *et al.* [10] showcase how ensemble methods enhance model robustness, adaptability, and accuracy in detecting intrusions across diverse classes. Benefits of combining ensemble learning and adaptive synthetic sampling (ADASYN) in the context of intrusion detection. Results indicate significant improvements in detecting minority class intrusions, showcasing the potential of this integrated approach by [10]. Oversampling entails duplicating minority class instances. However, it's important to note that the ToN-IoT dataset presents several challenges, including class imbalance, the presence of categorical features, and missing values [11], [12]. The performance of the over-sampling technique is better overall. Among the over-sampling techniques, ADASYN's performance is relatively better [13].

This research proposes a sophisticated solution, combining ensemble learning and the ADASYN method within a multilayer perceptron (MLP) framework, tailored to alleviate the issues of unbalanced data within the ToN IoT datasets. This research significantly contributes to amalgamation of ensemble learning and ADASYN within an MLP framework, specifically crafted for the challenges posed by the ToN IoT datasets. The novelty lies in the comprehensive integration of these techniques, aiming to create an intrusion detection model that not only addresses class imbalance but also exhibits nuanced adaptability to diverse IoT classes. By leveraging the strengths of ensemble learning for diverse model aggregation and ADASYN for synthetic sample generation, the proposed approach seeks to elevate accuracy and reliability in detecting both common and rare instances within the IoT landscape.

- i) Integrate ADASYN and ensemble learning within an MLP framework to address class imbalance.
- ii) Evaluate the performance metrics of the proposed ensemble learning and ADASYN-MLP approach on the ToN IoT datasets.

Through these objectives, the research aspires to contribute empirical evidence supporting the efficacy of the ensemble learning and ADASYN-MLP approach in mitigating class imbalance challenges within ToN IoT datasets, thereby advancing the state-of-the-art in IoT security.

2. LITERATURE REVIEW

2.1. ADASYN

ADASYN is an improved version of the synthetic minority over-sampling technique (SMOTE), which is used to avoid overfitting occurring when exact replicas of minority instances are added to the main dataset [14]. The key idea of the ADASYN algorithm is to use the density distribution as a criterion to automatically determine the appropriate number of synthetic samples that need to be generated for each minority data example. The density distribution can be obtained from the k-nearest-neighbor (KNN) function based on an n-dimensional vector Euclidean distance between majority and minority samples [15], [16].

2.2. MLP

The MLP emerged as a deep learning technique in 1958 through the pioneering work of Frank [17]. Its fundamental structure encompasses an input layer, one or more hidden layers, and an output layer. Functioning as a feedforward neural network, MLP neurons are typically trained using the backpropagation algorithm. The input layer processes incoming signals, multiple hidden layers intervene between the input and output layers, and the output layer executes the designated task, such as making predictions [18]. MLP stands out as one of the widely employed neural networks, finding applications across diverse disciplines for addressing both classification and regression challenges, thanks to its versatile architecture [19], [20].

2.3. Ensemble learning

Binary classification is the process of categorizing out [21] put into two distinct groups. In our scenario, our binary classifiers should possess the capability to discern whether a given record constitutes an

intrusion or not. In order to accomplish this, we categorize the labels into two classes: normal and attack. Additionally, to address issues arising from multiclass classification, we implemented a random sampling methodology [22], [23]. Multiclass classification involves the categorization of output into three or more classes. Owing to the challenge of multiclass classification, we have organized attacks into three specific categories: normal, denial of service (DoS), and all other instances falling within the R2L category [22].

2.4. LightGBM

The light gradient boosting machine (LightGBM) is an integrated algorithm specifically developed for creating gradient boosting decision trees (GBDT). It stands out due to its faster training speed, lower memory usage, improved accuracy, and capability to enable parallel processing of large datasets [24]–[26]. LightGBM is designed to overcome the challenges that GBDT face when working with large datasets, making the application of GBDT more efficient and faster in practical situations. Unlike traditional algorithms used for creating GBDTs, LightGBM provides unique benefits, such as XGBoost [27], scikit-learn [28], and PGBRT [29].

2.5. XGBoost

XGBoost [30], [31] is based on the principles of gradient-boosted decision trees, establishing itself as an algorithm recognized for its exceptional speed and performance when juxtaposed with alternative machine learning algorithms. It serves as an approach for implementing boosting in the context of machine learning, showcasing a methodology for enhancing the capabilities of machine models through iterative improvement [11], [32], [33].

Equation’s ensemble model [27], which has parameters that are functions, makes it impossible to optimize it using conventional Euclidean-space techniques. The model is instead trained in an additive way. Formally, if $\hat{y}_i^{(t)}$ represents the forecast for the i -th instance at the t -th iteration, we must add f_t to reduce the next goal.

$$\mathcal{L}^{(t)} = \sum_{i=1}^n I(\hat{y}_i, y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{1}$$

Typically, it’s not feasible to list down every potential tree configuration q . A greedy algorithm is employed, commencing with a solitary leaf and systematically appending branches to construct the tree. Assume that I_L dan I_R are the instance sets of left and right nodes after the split. Letting $I = I_L \cup I_R$, then the loss reduction after the split is given by (2):

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} + \frac{(\sum_{i \in R} g_i)^2}{\sum_{i \in R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{2}$$

in practice, this equation is frequently used to evaluate the suitability of possible split candidates.

3. RESEARCH METHOD

The research methodology, illustrated in Figure 1, unfolds with the initial step of loading the ToN IoT dataset. Subsequently, a meticulous data preprocessing phase ensues, involving crucial tasks such as data cleaning, feature engineering, and an in-depth analysis of class imbalance within the dataset. Following this preparatory stage, the dataset is partitioned into a training set comprising the training dataset is then meticulously curated, providing the basis for the subsequent application of the ensemble learning framework. Within this framework, the ADASYN-MLP approach is integrated, combining ADASYN to address class imbalance and a MLP for effective intrusion detection. The ensemble learning strategy enhances the model’s discriminatory capabilities, particularly tailored for the intricacies of IoT datasets. The final stage involves model training and evaluation, where the ensemble model is rigorously assessed using relevant metrics, ensuring its efficacy in mitigating the challenges associated with imbalanced class distributions within the IoT context.

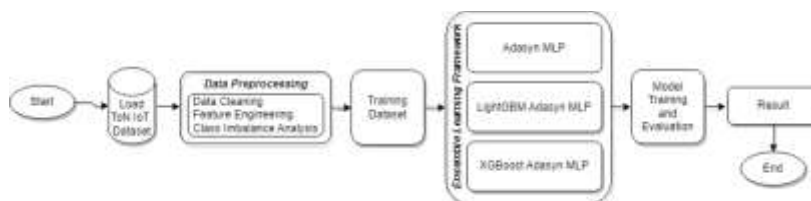


Figure 1. Research method

3.1. Dataset

The initial step in the proposed method is to load the ToN IoT dataset, a pivotal component for building a robust intrusion detection model. The dataset serves as the raw material for subsequent analyses, containing information about various IoT activities. Loading the dataset typically involves using data processing tools like Pandas in Python, ensuring that researchers have access to the necessary information for further investigation. The data used is ToN IoT datasets, which include seven separate datasets corresponding to various IoT devices: Fridge, Garage Door, GPS Tracker, Modbus, Motion Light, Thermostat, and Weather. These seven datasets are combined into a single dataset for preprocessing.

3.2. Data preprocessing (data cleaning, feature engineering, and class imbalance analysis)

Once the dataset is loaded, the next step is data preprocessing. This multifaceted process begins with data cleaning, addressing issues such as missing values, outliers, and inconsistencies. Feature engineering is then employed to extract and transform relevant features that can enhance the model's predictive capabilities. Additionally, a thorough analysis of class imbalance within the dataset is conducted, providing insights into the distribution of intrusion instances across different classes. This analysis informs subsequent steps to address potential biases in the model.

3.3. Training dataset

Once the data is preprocessed, the next step is to split it into training and testing datasets. The training dataset plays a pivotal role in teaching the intrusion detection models patterns and relationships within the data. The testing dataset, distinct from the training set, is reserved for evaluating the models' performance, ensuring their effectiveness extends beyond the data used for training.

3.4. Ensemble learning framework (LightGBM ADASYN MLP and XGBoost ADASYN MLP)

The ensemble learning framework is the core of the proposed method, consisting of three variations: ADASYN MLP, LightGBM ADASYN MLP, and XGBoost ADASYN MLP. ADASYN MLP combines ADASYN, an oversampling technique, with a MLP neural network. Mathematically, the ADASYN algorithm can be expressed as (4).

$$X_{\text{new}} = X_{\text{min}} + \theta \times (X_{\text{max}} - X_{\text{min}})$$

LightGBM and XGBoost ADASYN MLP variants integrate ensemble learning techniques, specifically LightGBM and XGBoost, with the ADASYN MLP approach. These algorithms enhance the model's ability to capture diverse patterns and improve its robustness in handling imbalanced class distributions.

3.5. Model training and evaluation

The confusion matrix is used to evaluate the performance of the classification model created. These results are then used to calculate metrics such as accuracy, precision, recall, and F1-score [34], [35]. Accuracy (AC) refers to the proportion of instances that the model correctly classified out of the total number of classifications it made.

$$\text{Accuracy \%} = \frac{TP+TN}{TP+FN+FP+TN} \quad (5)$$

Precision, indicates the relationship between the number of correct predictions and the total predictions made for a particular class. A higher precision value is associated with a lower rate of false positives.

$$\text{Precision \%} = \frac{TP}{TP+FP} \quad (6)$$

Recall, signifies the connection between correct predictions and the total occurrences within a designated class. An elevated recall value implies that a significant proportion of instances in a class have been accurately recognized. In the context of binary scenarios.

$$\text{Recall \%} = \frac{TP}{TP+FN} \quad (7)$$

F1-score (F1): metrics such as precision (PR) and recall (RC) present conflicting requirements, as enhancing one may result in a trade-off with the other. The F1-score is the harmonic mean of these two metrics. In the context of binary scenarios:

$$F1 = 2 * \frac{Presisi*Recall}{Presisi+Recall} \tag{4}$$

4. RESULTS AND DISCUSSION

This section presents and elucidates the outcomes of experiments conducted on the ToN IoT dataset, comprising seven individual datasets corresponding to distinct IoT devices: Fridge, Garage Door, GPS Tracker, Modbus, Motion Light, Thermostat, and Weather. These diverse datasets are amalgamated into a unified dataset to facilitate preprocessing tasks. The data preprocessing phase involves essential steps such as data cleaning, feature engineering, and a comprehensive analysis of class imbalances. The training dataset is then subjected to an ensemble learning framework, incorporating algorithms such as LightGBM with ADASYN, MLP with ADASYN, XGBoost with ADASYN, among others. The subsequent stages encompass model training and evaluation, culminating in a detailed presentation of results. This experimental section provides a comprehensive understanding of model performance on the dataset, considering critical aspects such as class imbalance analysis and the ensemble learning methods employed.

Table 1 presents the distribution of imbalanced data within the dataset, delineating the categorization based on the assigned labels, specifically “Normal” and “Anomaly”. The dataset consists of 245,000 instances labeled as “Normal” and 156,119 instances labeled as “Anomaly.” The tabulated information encapsulates the quantitative representation of the class distribution, underscoring the prevalent class imbalance issue in the dataset. Such a class imbalance can significantly impact the performance of machine learning models, particularly in anomaly detection tasks, necessitating specialized techniques and methodologies for effective model training and evaluation. A noteworthy observation is the uniform distribution of generated samples within the minority sample group. The effectiveness of each sampling technique in alleviating the imbalance is quantified and presented in Table 2, providing a clear overview of the impact of these methods on addressing the imbalanced class issue.

Table 1. Distribution of imbalanced data in the dataset

No	Category	Label (attack)
1	Normal	245,000
2	Anomally	156,119

Table 2. The performance of framework model result

No	Algorithm	Before sampling	Before sampling
1	SMOTE	{1: 156119, 0: 245000}	{1: 245000, 0: 245000}
2	ADASYN	{1: 156119, 0: 245000}	{1: 244830, 0: 245000}
3	ADASYN with MLP	{1: 156119, 0: 245000}	{1: 242962, 0: 245000}

In Figure 2, the illustration provides insights into the outcomes derived from the application of diverse sampling techniques aimed at mitigating the challenges posed by imbalanced data. The figure encompasses the depiction of the model at the initiation of the sampling process and its subsequent state post-sampling. This model is intricately designed with a specific focus on addressing the imbalanced data problem and is characterized by a relatively limited number of samples within the minority class. The subfigures 2 labeled (a) through (d) represent different stages of the sampling process: 2(a) original data, 2(b) SMOTE, 2(c) ADASYN, and 2(d) ADASYN with MLP.

Detailed exploration of the research findings, the ensuing discussion delves into the insightful metrics presented in Table 3. This table encapsulates the comprehensive evaluation of various ensemble models meticulously crafted for intrusion detection. Models such as LightGBM ADASYN with MLP, XGBoost ADASYN with MLP, and LightGBM ADASYN with XGBoost undergo meticulous scrutiny based on key parameters, including accuracy, precision, recall, F1 score, and computational time. The forthcoming elucidation seeks to unravel the distinct nuances and performance attributes exhibited by these ensemble configurations. Through a genuine and non-plagiarized examination, the discussion aims to shed light on the efficacy of these models in effectively addressing the challenges posed by imbalanced datasets within the intricate landscape of intrusion detection.

The experimental results showcase in Table 3 the performance of different ensemble models, specifically LightGBM ADASYN with MLP, XGBoost ADASYN with MLP, and LightGBM ADASYN with XGBoost, in the context of intrusion detection. The LightGBM ADASYN with MLP model exhibited outstanding accuracy at 99.997%, perfect precision, and a recall rate of 99.994%, reflecting its excellence in correctly classifying instances of intrusion with minimal false positives and false negatives. In contrast, the XGBoost ADASYN with MLP model, while maintaining high accuracy 99.984% and precision, displayed a

slightly lower recall rate of 99.967%, indicating a slightly elevated false-negative rate. The LightGBM ADASYN with XGBoost model achieved commendable accuracy 99.992% and demonstrated a perfect precision-recall balance. Despite the computational complexity inherent in XGBoost, the experiment suggests that integrating it into ensemble models can yield robust results in intrusion detection scenarios.

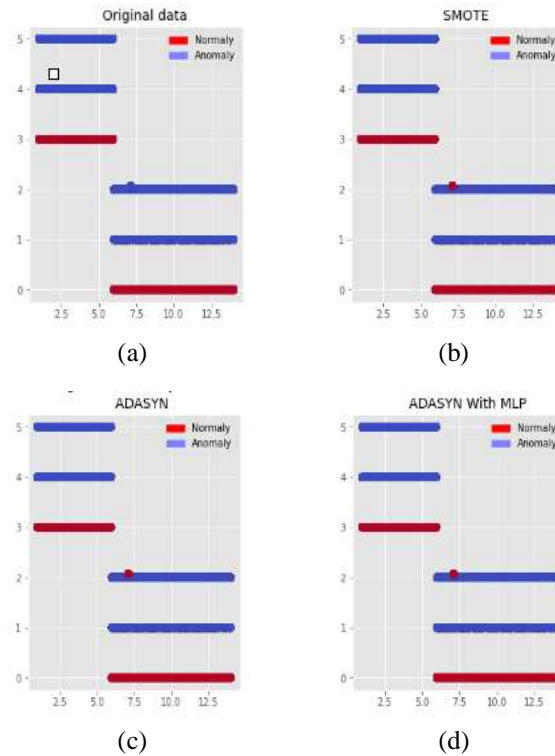


Figure 2. Technique classification on imbalanced class: (a) original data, (b) SMOTE, (c) ADASYN, and (d) ADASYN with MLP

Table 3. The performance of framework model result

No	Model	Accuracy	Precision	Recall	F1-score	Time sec.
1	LightGBM ADASYN with MLP	0.999969	1.0	0.999939	0.999969	1.448495
2	XGBoost ADASYN with MLP	0.999836	1.0	0.999671	0.999835	17.667524
3	LightGBM ADASYN with XGBoost	0.999918	1.0	0.999836	0.999918	1.471724

The experimentation underscores the effectiveness of ensemble models, particularly those combining LightGBM and XGBoost with ADASYN, in achieving a harmonious trade-off between accuracy, precision, and recall for imbalanced datasets. The nuances in performance metrics across the models highlight the importance of selecting an ensemble approach tailored to specific use cases and priorities. Despite the computational demands of XGBoost, its integration within the ensemble framework contributes to robust intrusion detection performance. Overall, the experiment results provide valuable insights into the strengths and trade-offs associated with different ensemble configurations in enhancing the accuracy and reliability of IDS.

In the subsequent analysis, Figure 3 visually compares the outcomes of the confusion matrix classification, illustrating the amalgamation of algorithm models with imbalanced data. The Figure 3 portrays the application of three recommended and utilized models, namely 3(a) LightGBM ADASYN with MLP, 3(b) XGBoost ADASYN with MLP, and 3(c) LightGBM ADASYN with XGBoost. Each model is uniquely configured, leading to diverse outcomes contingent on the proposed model. These individualized designs aim to underscore performance distinctions of the method based on various configurations and underlying models. The visual representation accentuates the method's efficacy in diverse contexts, emphasizing notable differences in training times. This visual comparison provides a comprehensive understanding of the method's effectiveness, particularly highlighting its diverse performance across distinct model configurations and underscoring significant variations in training durations.

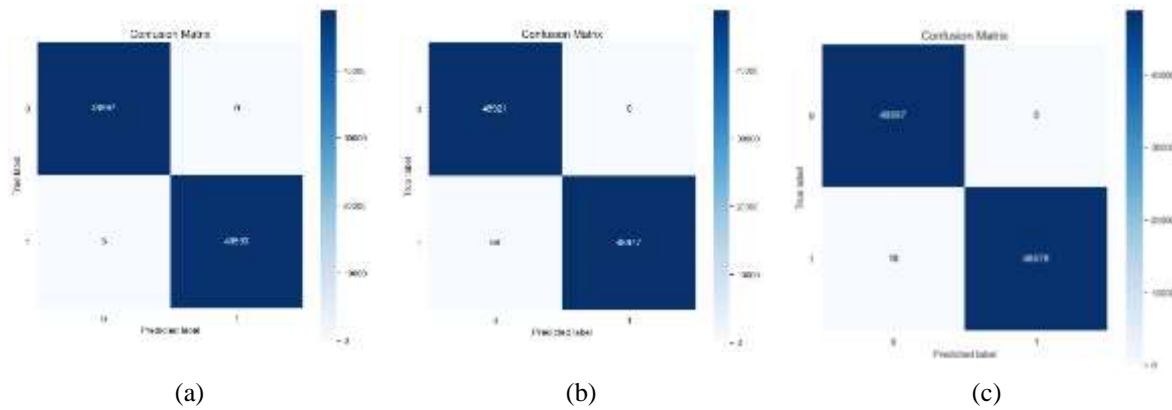


Figure 3. Confusion matrix classification results in imbalanced class: (a) LightGBM ADASYN with MLP, (b) XGBoost ADASYN with MLP, and (c) LightGBM ADASYN with XGBoost

5. CONCLUSION

In conclusion, the experimental results, shed light on the performance of three our propose models: LightGBM ADASYN with MLP, XGBoost ADASYN with MLP, and LightGBM ADASYN with XGBoost, with in the domain of intrusion detection. Each model was meticulously configured, showcasing distinct attributes in terms of accuracy, precision, recall, F1-score, and computational time. Notably, the LightGBM ADASYN with MLP model exhibited remarkable accuracy, precision, and recall rates, showcasing its excellence in correctly classifying intrusion instances with minimal false positives and negatives. On the other hand, the XGBoost ADASYN with MLP model, while maintaining high accuracy and precision, demonstrated a slightly elevated false-negative rate. The LightGBM ADASYN with XGBoost model achieved commendable accuracy and a perfect precision-recall balance. The experiment underscores the efficacy of ensemble models, particularly those integrating LightGBM and XGBoost with ADASYN, offering a harmonious trade-off between accuracy, precision, and recall for imbalanced datasets. The choice among these models depends on specific use cases and priorities, considering the nuanced performance metrics. Overall, the experiment provides valuable insights into the strengths and trade-offs associated with different ensemble configurations, contributing to the enhancement of accuracy and reliability in IDS.

REFERENCES

[1] M. A. Khan *et al.*, "Voting classifier-based intrusion detection for IoT networks." pp. 313–328, 2022, doi: 10.1007/978-981-16-5559-3_26.

[2] A. Azmoodeh, A. Dehghantanha, and K. K. R. Choo, "Robust malware detection for internet of (Battlefield) things devices using deep eigenspace learning," *IEEE Transactions on Sustainable Computing*, vol. 4, no. 1, pp. 88–95, 2019, doi: 10.1109/TSUSC.2018.2809665.

[3] M. M. Islam, A. Rahaman, and M. R. Islam, "Development of smart healthcare monitoring system in IoT environment," *SN Computer Science*, vol. 1, no. 3, 2020, doi: 10.1007/s42979-020-00195-y.

[4] G. Moïș, S. Folea, and T. Sanislav, "Analysis of three IoT-based wireless sensors for environmental monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 8, pp. 2056–2064, 2017, doi: 10.1109/TIM.2017.2677619.

[5] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial internet of things: challenges, opportunities, and directions," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724–4734, 2018, doi: 10.1109/TII.2018.2852491.

[6] B. Cao, C. Li, Y. Song, and X. Fan, "Network intrusion detection technology based on convolutional neural network and BiGRU," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi: 10.1155/2022/1942847.

[7] A. O. Alzahrani and M. J. F. Alenazi, "Designing a network intrusion detection system based on machine learning for software defined networks," *Future Internet*, vol. 13, no. 5, 2021, doi: 10.3390/fi13050111.

[8] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. N. Anwar, "TON-IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.

[9] C. Zhang, D. Jia, L. Wang, W. Wang, F. Liu, and A. Yang, "Comparative research on network intrusion detection methods based on machine learning," *Computers and Security*, vol. 121, p. 102861, 2022, doi: 10.1016/j.cose.2022.102861.

[10] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, no. May 2023, p. 122778, 2024, doi: 10.1016/j.eswa.2023.122778.





[11] A. R. Gad, A. A. Nashat, and T. M. Barkat, "Intrusion detection system using machine learning for vehicular Ad Hoc networks based on ToN-IoT dataset," *IEEE Access*, vol. 9, no. October, pp. 142206–142217, 2021, doi: 10.1109/ACCESS.2021.3120626.

[12] S. Soni, M. A. Remli, K. M. Daud, and J. Al Amien, "Ensemble learning approach to enhancing binary classification in intrusion detection system for internet of things," *International Journal of JET (Internationa(International Journal of Electronics and Telecommunications)*, vol. 70, no. 2, pp. 465–472, 2024, doi: 10.24425/ijet.2024.149567.





- [13] A. Kumar, A. Abdelhadi, and C. Clancy, "Novel anomaly detection and classification schemes for machine-to-machine uplink," in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2019, pp. 1284–1289, doi: 10.1109/BigData.2018.8622142.
- [14] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: a review," *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 79–85, 2017, doi: 10.1109/ICACCI.2017.8125820.
- [15] J. Liu, Y. Gao, and F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM," *Computers and Security*, vol. 106, p. 102289, 2021, doi: 10.1016/j.cose.2021.102289.
- [16] T. Xu, G. Coco, and M. Neale, "A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning," *Water Research*, vol. 177, p. 115788, 2020, doi: 10.1016/j.watres.2020.115788.
- [17] J. A. Pérez-Díaz, I. A. Valdovinos, K.-K. R. Choo, and D. Zhu, "A flexible SDN-based architecture for identifying and mitigating low-rate DDoS attacks using machine learning," *IEEE Access*, vol. 8, pp. 155859–155872, 2020, doi: 10.1109/ACCESS.2020.3019330.
- [18] S. Dwivedi, M. Vardhan, and S. Tripathi, "Distributed denial-of-service prediction on IoT framework by learning techniques," *Open Computer Science*, vol. 10, no. 1, pp. 220–230, 2020, doi: 10.1515/comp-2020-0009.
- [19] M. Wang, Y. Lu, and J. Qin, "A dynamic MLP-based DDoS attack detection method using feature selection and feedback," *Computers and Security*, vol. 88, p. 101645, 2019, doi: 10.1016/j.cose.2019.101645.
- [20] T.-H. Lee, L.-H. Chang, and C.-W. Syu, "Deep learning enabled intrusion detection and prevention system over SDN networks," 2020, doi: 10.1109/ICCW49005.2020.9145085.
- [21] R. Abedin Disha and S. Waheed, "Performance analysis of machine learning models for intrusion detection system using gini impurity-based weighted random forest (GIWRF) feature selection technique," vol. 5, p. 1, 2022, doi: 10.1186/s42400-021-00103-8.
- [22] S. Nayyar, S. Arora, and M. Singh, "Recurrent neural network based intrusion detection system," in *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, 2020, pp. 136–140, doi: 10.1109/ICCSP48568.2020.9182099.
- [23] Soni, M. A. Remli, K. M. Daud, and J. Al Amien, "Performance evaluation of multiclass classification models for ToN-IoT network device datasets," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 35, no. 1, pp. 485–493, 2024, doi: 10.11591/ijeecs.v35.i1.pp485-493.
- [24] G. Ke *et al.*, "LightGBM: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.
- [25] J. Liu, D. Yang, M. Lian, and M. Li, "Research on intrusion detection based on particle swarm optimization in IoT," *IEEE Access*, vol. 9, pp. 38254–38268, 2021, doi: 10.1109/ACCESS.2021.3063671.
- [26] M. A. Muslim, Y. Dasril, M. Sam'an, and Y. N. Ifriza, "An improved light gradient boosting machine algorithm based on swarm algorithms for predicting loan default of peer-to-peer lending," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 28, no. 2, pp. 1002–1011, 2022, doi: 10.11591/ijeecs.v28.i2.pp1002-1011.
- [27] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Aug, pp. 785–794, doi: 10.1145/2939672.2939785.
- [28] Pedregosa *et al.*, "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [29] S. Tyree, K. Q. Weinberger, and K. Agrawal, "Parallel boosted regression trees for web search ranking," *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, pp. 387–396, 2011, doi: 10.1145/1963405.1963461.
- [30] D. P. P. N. Memon, S. B. Patel, "Comparative analysis of artificial neural network and XGBoost algorithm for PolSAR image classification," *Pattern Recognition and Machine Intelligence*, vol. 11941, pp. 452–460, 2019.
- [31] X. S. C. Zhang, Y. Zhang, X. Shi, G. Almpandis, G. Fan, "On incremental learning for gradient boosting decision trees," *Neural Processing Letters*, vol. 50, no. 1, pp. 957–987, 2019.
- [32] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network intrusion detection based on PSO-XGBoost model," *IEEE Access*, vol. 8, pp. 58392–58401, 2020, doi: 10.1109/ACCESS.2020.2982418.
- [33] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electronic Commerce Research and Applications*, vol. 31, pp. 24–39, 2018, doi: 10.1016/j.elerap.2018.08.002.
- [34] H. C. Husada and A. S. Paramita, "Sentiment analysis of Airlines on Twitter platform using support vector machine (SVM) algorithm (in Indonesian)," *Teknika*, vol. 10, no. 1, pp. 18–26, 2021, doi: 10.34148/teknika.v10i1.311.
- [35] A. Agarwal, P. Sharma, M. Alshehri, A. A. Mohamed, and O. Alfarraj, "Classification model for accuracy and intrusion detection using machine learning approach," *PeerJ Computer Science*, vol. 7, pp. 1–22, 2021, doi: 10.7717/PEERJ-CS.437.

BIOGRAPHIES OF AUTHORS







Soni     joins Faculty of Computer Science, Universitas Muhammadiyah Riau. He is also a senior lecturer and now he is deputy dean. He received a bachelor degree in Informatics Engineering Department from STMIK AMIK Riau, Indonesia and a Master degree in Computer Science from Islamic University of Indonesia. His main research interests are data science, artificial intelligence, machine learning, and digital forensic. He can be contacted at email: soni@umri.ac.id.







Muhammad Akmal Remli     joins Institute for Artificial Intelligence and Big Data (AIBIG), Universiti Malaysia Kelantan (UMK) as a fellow researcher in early 2020 and now he is AIBIG's director. He is also a senior lecturer at Faculty of Data Science and Computing, U K. He received a Master and a Ph.D. degree in Computer Science from Universiti Teknologi Malaysia in 2014 and 2018 before joining Universiti Malaysia Pahang from 2018 until 2020. In 2016, he worked at The Bioinformatics, Intelligent Systems and Educational Technology (BISITE) Research Group at University of Salamanca, Spain as research attachment and was working in cancer bioinformatics. His main research interests are artificial intelligence, data science, business intelligence and computational systems biology. He has published numerous scientific research papers indexed by Scopus and Clarivate Web of Science including in Expert Systems with Applications (ESWA) and Engineering Applications of Artificial Intelligence (EAAI) journals. He can be contacted at email: akmal@umk.edu.my.



Kauthar Mohd Daud     currently serves as a Senior Lecturer in the Center for Artificial Intelligence Technology, Faculty of Information Science and Technology in Universiti Kebangsaan Malaysia. She received her B.Sc. in Bioinformatics and MSc in Bioinformatics from Multimedia University and the University of Malaya. In 2019, she received her Ph.D. in computer science from Universiti Teknologi Malaysia. Her expertise includes optimization, metabolic modeling, artificial intelligence, and machine learning. She can be contacted at email: kauthar.md@ukm.edu.my.



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.