# OPTIMIZING INTRUSION DETECTION ON THE INTERNET OF VEHICLES: A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS USING THE CICIOV2024 DATASET

**Nur 'Aliah Binti Amirudin[1,*], Said Jadid Abdulkadir[1]**

[1]Computer and Information Science Department, Universiti Teknologi PETRONAS, Malaysia

Email: nur_22009393@utp.edu.my:

## Abstract

Leveraging the CICIoV2024 dataset, we evaluate the performance of different algorithms in accurately identifying and classifying intrusions in vehicular networks specifically using LightGBM, XGBoost, CatBoost, and LCCDE algorithms. Our comparative analysis considers key performance metrics such as accuracy, precision, recall, and F1-score, shedding light on the strengths and limitations of each approach. Through rigorous experimentation and evaluation, we achieve remarkable results, demonstrating 100% accuracy, recall, precision, and F1-score across all models. These findings highlight the efficacy of employing advanced machine learning techniques for enhancing intrusion detection capabilities in IoV environments. The insights gained from this study can inform the development of highly accurate and reliable intrusion detection systems tailored to the unique challenges of vehicular networks.

***Keywords:*** *IoV. CICIoV2024, XGBoost, CatBoost, LightGBM, LCCDE*

## INTRODUCTION

The transportation sector, much like other industries, has undergone a profound shift due to technological advancements. Modern vehicles now integrate an array of sensors, software, and technology, enabling them to share filtered sensor data via the Internet cloud with a network of autonomous vehicles (AUVs) [1]. This evolution has led to the emergence of the Internet of Vehicles (IoV), where interconnected vehicles communicate with each other and infrastructure elements to enable advanced functionalities such as cooperative driving, real-time traffic management, and enhanced safety features [2]. This transformation is reflected in the substantial growth of the global connected vehicles market, projected to reach around USD 331.9 billion by 2032 [3]. However, alongside these advancements come new security challenges, with vehicles increasingly vulnerable to cyber threats [4].

Intrusion detection systems (IDSs) play a crucial role in ensuring the security of IoV systems by identifying and mitigating malicious activities. AUVs are particularly susceptible to network threats, which can have severe consequences for human safety if, for instance, critical vehicle systems like braking or steering are compromised [5]. The importance of IDSs was underscored in a notable incident in 2016, where hackers infiltrated a Jeep Cherokee's systems, demonstrating the potential risks of cyber-attacks on vehicles [6]. By continuously monitoring network traffic and detecting anomalous behavior, IDSs contribute to maintaining the integrity, confidentiality, and availability of IoV services. However, designing effective IDS for IoV environments is challenging due to the dynamic nature of vehicular networks, diverse cyber threats, and resource constraints of onboard electronic control units (ECUs) [7]. While there are existing IDS solutions, there is ongoing room for improvement.

Machine learning (ML) algorithms have emerged as promising tools for enhancing intrusion detection in IoV, leveraging the vast amount of data generated by interconnected vehicles to detect unauthorized infiltrations [8]. ML algorithms can be trained to identify abnormal patterns in network traffic indicative of cyber-attacks, thereby enhancing detection accuracy. However, the performance of these algorithms is influenced by various factors such as feature selection, model architecture, and dataset characteristics, necessitating careful parameter tuning for optimal results.

## RELATED WORKS

ML algorithms such as LightGBM, CatBoost, XGBoost, and LCCDE are widely recognized for their utility in detecting anomalies and identifying potential malicious activities within datasets pertinent to the Internet of Vehicles (IoV). To uphold network security, the ability to effectively detect common threats like DoS and spoofing attacks is imperative. A systematic review conducted by [17] on research papers focusing on ML methods applied to cybersecurity has revealed promising results regarding the efficacy of these algorithms in addressing various threats, including DoS and spoofing. For example, [4] proposed an IDS based on LightGBM for detecting DoS attacks in VANETs, with their study showcasing LightGBM's notable accuracy and low false positive rates in detecting DoS attacks within IoV scenarios. Similarly, [19] developed a CatBoost-based IDS for detecting anomalous behavior in connected vehicles, with their findings indicating that CatBoost exhibited superior accuracy and robustness, particularly when handling high-dimensional and imbalanced datasets. Furthermore, research by [20] presented an XGBoost-based IDS tailored for detecting DoS attacks in VANETs, which achieved commendable detection rates and minimized false positives, thus positioning it as a viable solution for intrusion detection in IoV environments. Additionally, [9] introduced an IDS based on LCCDE for detecting spoofing attacks in connected vehicles, noting the balanced performance of LCCDE across different attack classes and its adaptability to dynamic IoV environments. These studies collectively contribute valuable insights into the effectiveness of ML-driven IDSs in bolstering the security posture of interconnected vehicular networks against evolving cyber threats. The findings from these studies suggest that LightGBM, XGBoost, and CatBoost have been effectively utilized in detecting DoS and spoofing attacks with high accuracy rates and low false positive rates. Their effectiveness can be attributed to their ability to analyze large datasets, identify patterns, and adapt to emerging threats.

## METHODOLOGY

### System Overview

The models were implemented using Scikit-learn, XGBoost, LightGBM, CatBoost libraries from Python (v. 3.10) in Google Colab platform and T4 GPU hardware.

### Proposed Framework

The CICIoV2024 dataset originally employed machine learning methods including Logistic Regression (LR), Random Forest (RF), Adaboost (AD), and Deep Neural Network (DNN). While these algorithms have demonstrated satisfactory performance in accurately identifying attacks and non-attacks, their precision, recall, and F1-score are found to be suboptimal [7]. This may be attributed to the extensive size of the dataset. To address this limitation, various optimization techniques and newer machine learning algorithms are likely to yield improvements in overall performance metrics such as accuracy, precision, recall, and F1-score.

The purpose of this work is to develop an IDS that can enhance the effectiveness and efficiency of detecting Denial of Service (DoS) and spoofing attacks in the CICIoV2024 dataset which would ultimately benefit the automotive industry due to its enhanced reliability and security at safeguarding vehicle and passenger safety using advanced machine learning algorithms LightGBM, XGBoost, CatBoost and Leader Class and Confidence Decision Ensemble (LCCDE).

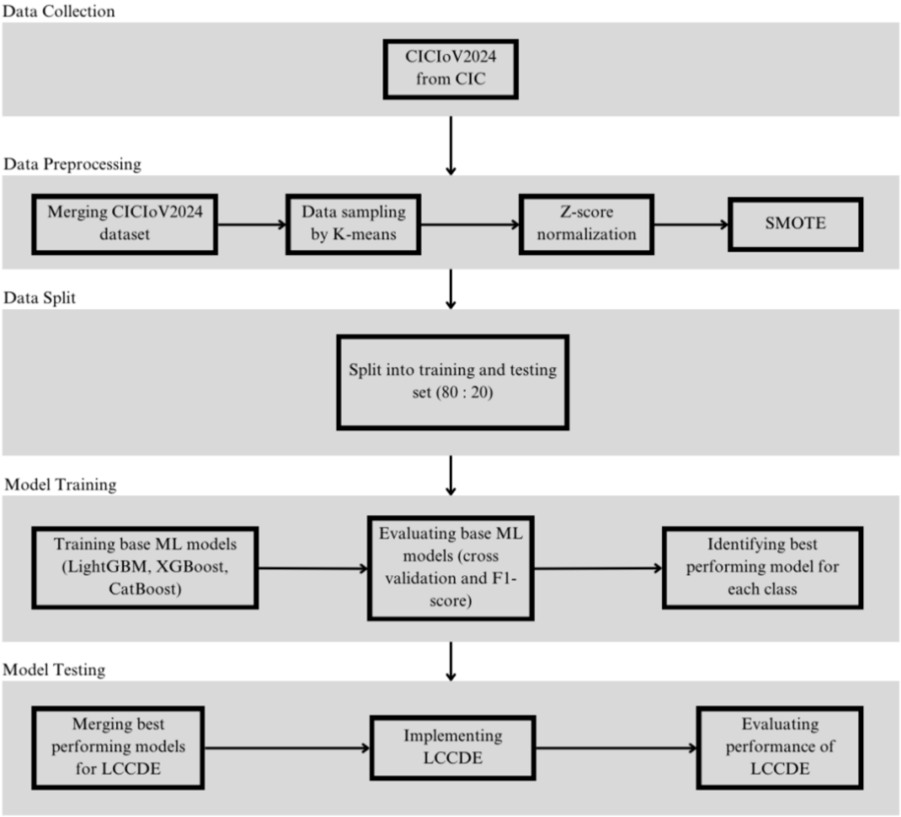
Data Collection
CICIoV2024 from CIC
Data Preprocessing
Merging CICIoV2024 dataset
Data sampling by K-means
Z-score normalization
SMOTE
Data Split
Split into training and testing set (80 : 20)
Model Training
Training base ML models (LightGBM, XGBoost, CatBoost)
Evaluating base ML models (cross validation and F1-score)
Identifying best performing model for each class
Model Testing
Merging best performing models for LCCDE
Implementing LCCDE
Evaluating performance of LCCDE

**Figure 1**: Proposed framework

## Dataset Characteristic

The CICIoV2024 dataset is the product of extensive experimentation conducted by the Canadian Institute for Cybersecurity (CIC) on the ECUs of a 2019 Ford vehicle. This dataset provides a view of intra-vehicle communications, shedding light on the intricate interactions within the vehicle's internal network. To ensure the safety of the vehicle's occupants, researchers meticulously orchestrated five distinct attacks on the intact inner structure of the 2019 Ford car. It's noteworthy that these attacks were carefully designed to immobilize the vehicle while posing no risk to the driver or passengers. Executed through the CAN bus protocol, these attacks fall within the categories of spoofing and DoS, showcasing the vulnerabilities inherent in vehicular communication systems.

The dataset includes 12 distinct features extracted from intra-vehicular communications, meticulously documented and outlined in Table 1. Each feature serves a specific purpose in characterizing the interactions within the vehicle's network and identifying potential anomalies or malicious activities. The CICIoV2024 dataset distribution is as shown in Table 2.

**Table 1**: Features of CICIoV2024 dataset

| Feature Name | Description |
|---|---|
| ID | Arbitration |

| | |
|---|---|
| Data_0 | Byte 0 of the data transmitted |
| Data_1 | Byte 1 of the data transmitted |
| Data_2 | Byte 2 of the data transmitted |
| Data_3 | Byte 3 of the data transmitted |
| Data_4 | Byte 4 of the data transmitted |
| Data_5 | Byte 5 of the data transmitted |
| Data_6 | Byte 6 of the data transmitted |
| Data_7 | Byte 7 of the data transmitted |
| label | The identification of benign or malicious traffic |
| category | The category to which traffic belongs |
| specific_class | The identification of the specific class of the traffic |

**Table 2**: Distribution of CICIoV2024 dataset

| Attack Type | Label | Count |
|---|---|---|
| - | Benign | 1223737 |
| DoS | DoS | 74663 |
| Spoofing | RPM | 54900 |
| | Speed | 24951 |
| | Steering Wheel | 19977 |
| | Gas | 9991 |

## Data Preprocessing

Firstly, to ensure that ML training is more efficient, the dataset undergoes Z-score normalization, where the mean of all values is 0 and the standard deviation is 1. Datasets with largely different features that do not undergo normalization may cause biased ML models that only emphasize large-scale features. The Z-score method implemented for each normalized feature value, $x_n$, is denoted by:

$$x_n = \frac{x - \mu}{\sigma}$$

where x is the original feature value, μ and σ are the mean and standard deviation of the feature values, respectively.

Secondly, due to the sheer size of the CICIoV2024 dataset, k-means clustering method was applied to the majority class to reduce the size of the dataset and is denoted by:

$$\sum_{i=0}^{n_k} \min_{\mu_k \in c_k} (x_i - u_j)^2$$

where $(x_1,\ldots, x_n)$ is the data matrix, $u_j$, also called the centroid of a cluster $C_k$, is the mean of all the samples in $C_k$; and $n_k$ is the total number of sample points in the cluster$C_k$. K-means has a linear time complexity of $O(nkt)$, where $n$, is the data size, $k$, is the number of cluster, and $t$ is the number of iterations. The number of optimal clusters was identified using the Elbow Method.

Lastly, to obtain a balanced representation of the imbalanced dataset and ensure that the model will not be biased to the majority class, the minority class is inflated to at least half of the number of the majority class. However, caution was taken to not cause oversampling of the minority class too aggressively because it will lead to poor generalization on unseen data. For each instance $X$ in the minority class, assuming $X_i$ is a sample randomly selected from the k nearest neighbours of $X$, a new synthetic instance $X_n$ can be denoted by:

$$X_n = X + rand(0,1) * (X_i - X), i = 1,2, \ldots, k$$

where $rand$(0,1)represeents a random number in ther ranger of (0,1).

## Model Training

For model training, the dataset is split into a training and testing set (80:20) [9]. LightGBM is based on the gradient boosting decision tree (GBDT) algorithm, which iteratively builds a series of decision trees to make predictions. However, what sets LightGBM apart is its novel approach to building these decision trees. Instead of using the traditional depth-wise approach, LightGBM employs a leaf-wise approach, which enables it to grow deeper trees by splitting nodes that yield the maximum reduction in loss. This approach results in more accurate models with fewer nodes, leading to faster training times and reduced memory consumption. Additionally, it implements efficient algorithms for handling categorical features and missing values, reducing the need for preprocessing and improving overall model accuracy. Overall, LightGBM is a powerful and versatile tool for a wide range of ML tasks, including classification, regression, and ranking. Its combination of speed, accuracy, and scalability makes it a popular choice among data scientists and practitioners seeking to build high-performance predictive models [10].

XGBoost, short for eXtreme Gradient Boosting, is a powerful and widely used ML algorithm known for its efficiency, effectiveness, and versatility. It belongs to the family of gradient boosting algorithms, which are ensemble learning techniques that combine the predictions of multiple individual models, typically decision trees, to produce a more accurate and robust final prediction. One of the key features of XGBoost is its innovative approach to building decision trees. Unlike traditional gradient boosting algorithms that build trees sequentially, XGBoost employs a more sophisticated optimization technique known as gradient boosting with parallel tree construction. This technique allows XGBoost to build trees in parallel, greatly accelerating the training process [11].

CatBoost is a powerful gradient boosting algorithm designed to handle categorical features seamlessly, making it particularly well-suited for datasets with a mix of numerical and categorical variables. CatBoost incorporates several advanced techniques to improve model performance and robustness. These include ordered boosting, which optimizes the order in which the trees are built to minimize overfitting, and a novel method for handling numerical features, which allows CatBoost to automatically detect and use the most appropriate scaling for each feature [12].

The LCCDE model employs leader models for each class to generate the ultimate prediction within the ensemble. In cases where multiple leader models are available for different classes, LCCDE prioritizes the one exhibiting the highest prediction confidence to determine the final decision. By incorporating LCCDE, the ensemble model achieves peak performance in detecting various types of attacks. LCCDE is specifically crafted to enhance the accuracy and robustness of classification models by capitalizing on the distances between class centers within the feature space. During the prediction phase, LCCDE consolidates the predictions of all base classifiers to formulate a final prediction for each instance. This consolidation process can manifest in various forms, including majority voting or weighted averaging, contingent upon the specific implementation employed. One of the key advantages of LCCDE is its ability to capture the local structure of the data, allowing it to make more accurate predictions, especially in regions where the class boundaries are non-linear or complex. Additionally, by dividing the feature space into local regions, LCCDE can handle datasets with imbalanced class distributions more effectively, as it can adapt its predictions to the local characteristics of each region [9].

## RESULT & DISCUSSION

The performance of the advanced ML models (LightGBM, XGBoost, CatBoost, LCCDE) was evaluated in terms of accuracy, precision, recall, and F1-score. Additionally, their performance with the previously used models in the original paper [7], was also compared to assess their effectiveness in handling the specific characteristics of the CICIoV2024 dataset. The results (Table 3) indicate that LightGBM, XGBoost, CatBoost and LCCDE all show superior results in terms of accuracy, precision, recall and F1-score in comparison to LR, AD, DNN and RF. LightGBM, XGBoost, CatBoost, and LCCDE all achieved an accuracy of 0.99, 1.0, 1.0, and 1.0, respectively. This indicates that all four algorithms have demonstrated exceptional performance in handling this dataset. Although the results are quite similar, each algorithm might have its unique advantages and disadvantages depending on the specific problem and dataset.

**Table 3**: Performance evaluation on CICIoV2024

| | Accuracy | Precision | Recall | F1 | Execution Time |
|---|---|---|---|---|---|
| LR | 0.95 | 0.74 | 0.68 | 0.63 | - |
| AD | 0.87 | 0.14 | 0.17 | 0.15 | - |
| DNN | 0.95 | 0.74 | 0.68 | 0.63 | - |
| RF | 0.95 | 0.60 | 0.68 | 0.62 | - |
| **LightGBM** | **0.99** | **0.99** | **0.99** | **0.99** | **5.39s** |
| **XGBoost** | **1.0** | **1.0** | **1.0** | **1.0** | **4.59s** |
| **CatBoost** | **1.0** | **1.0** | **1.0** | **1.0** | **1 min 11s** |
| **LCCDE** | **1.0** | **1.0** | **1.0** | **1.0** | **1 min 43s** |

LightGBM is highly accurate and efficient in handling large-scale datasets, such as network traffic. Its fast-training speed and low memory usage make it suitable for real-time applications [13]. However, its lack of interpretability and sensitivity to hyperparameters can present notable challenges. CatBoost offers compelling advantages in detecting DoS and spoofing attacks, including its ability in handling categorical features, robustness to overfitting, and built-in cross-validation. However, challenges such as computational resource requirements and limited interpretability should be taken into consideration [14]. XGBoost offers formidable capabilities in cyber threat detection, leveraging its scalability, regularization techniques, and flexibility to detect and mitigate DoS and spoofing attacks effectively. However, there are challenges such as sensitivity to hyperparameters and interpretability [15]. LCCDE adopts a sophisticated leader model selection strategy, where models with the highest prediction confidence for each class are designated as leaders. This selective approach ensures that the ensemble prioritizes the most reliable and accurate predictions for decision-making, amplifying its efficacy in detecting various types of attacks, including DoS and spoofing, with precision and confidence. However, it is important to address that there are challenges in implementation due to its complexity [16].

Overall, the LCCDE model was identified to have similar results to the other algorithms (LightGBM, XGBoost, CatBoost). This can be attributed to the fact that they share common objectives of maximizing predictive accuracy and generalization performance. Each model may excel in certain aspects, such as handling categorical features (CatBoost), scalability (XGBoost), or interpretability (LightGBM), but they ultimately strive to achieve optimal predictive performance. Therefore, if

LCCDE effectively combines diverse base models and leverages their complementary strengths, it is possible to achieve similar results to these individual algorithms.

## CONCLUSION

Overall, the comparative analysis underscores the efficacy of ML models in enhancing intrusion detection capabilities. While each model demonstrates remarkable performance on the CICIoV2024 dataset, the choice of model ultimately depends on factors such as computational efficiency, interpretability, and specific requirements of the IDS deployment. In summary, LightGBM, CatBoost, XGBoost, and LCCDE offer distinct strengths and weaknesses in detecting DoS and spoofing attacks. As cyber threats continue to evolve, ongoing research and development efforts in ML-based IDS hold promise for bolstering cybersecurity defenses and mitigating emerging threats effectively. Through the project results, LightGBM, XGBoost, CatBoost and LCCDE have been found to outperform LR, AD, DNN and RF, and are proven to be effective ML algorithms for detecting DoS and spoofing attacks in the CICIoV2024 dataset. The proposed LightGBM, XGBoost, CatBoost and LCCDE models achieved high accuracy, precision, recall and F1-score of 99%, 100%, 100% and 100% respectively.

In the future, it is suggested that other methods that can enhance robustness of ML models against adversarial attacks are explored to ensure that IDS systems remain effective in detecting more sophisticated and evolving threats in vehicular networks.

## ACKNOWLEDGMENT

The authors would like to acknowledge Universiti Teknologi PETRONAS for supporting the completion of this work.

## REFERENCES

[1] E.-K. Lee, M. Gerla, G. Pau, U. Lee, and J.-H. Lim, “Internet of Vehicles: From intelligent grid to autonomous cars and vehicular fogs,” International Journal of Distributed Sensor Networks, vol. 12, no. 9, p. 155014771666550, Sep. 2016, doi: https://doi.org/10.1177/1550147716665500.

[2] M. N. O. Sadiku, M. Tembely, and S. M. Musa, “INTERNET OF VEHICLES: AN INTRODUCTION,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, no. 1, p. 11, Feb. 2018, doi: https://doi.org/10.23956/ijarcsse.v8i1.512.

[3] “Connected Vehicles Market,” *market.us*, Oct. 2023. https://market.us/report/connected-vehicles-market/

[4] W. Gou, H. Zhang, and R. Zhang, “Multi-Classification and Tree-Based Ensemble Network for the Intrusion Detection System in the Internet of Vehicles,” *Sensors*, vol. 23, no. 21, pp. 8788–8788, Oct. 2023, doi: https://doi.org/10.3390/s23218788.

[5] L. Yang, A. Moubayed, and A. Shami, “MTH-IDS: A Multi-Tiered Hybrid Intrusion Detection System for Internet of Vehicles,” *IEEE Internet of Things Journal*, pp. 1–1, 2021, doi: https://doi.org/10.1109/jiot.2021.3084796.

[6] J. Golson, “Jeep hackers at it again, this time taking control of steering and braking systems,” *The Verge*, Aug. 02, 2016. https://www.theverge.com/2016/8/2/12353186/car-hack-jeep-cherokee-vulnerability-miller-valasek

[7] E. Carlos Pinto Neto *et al.*, “Ciciov2024:Advancing Realistic Ids Approaches Against Dos and Spoofing Attack in Iov Can Bus,” *Social Science Research Network*, Feb. 21, 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4733521 (accessed Mar. 26, 2024).

[8] E. Alalwany and I. Mahgoub, “Security and Trust Management in the Internet of Vehicles (IoV): Challenges and Machine Learning Solutions,” *Sensors*, vol. 24, no. 2, p. 368, Jan. 2024, doi: https://doi.org/10.3390/s24020368.

[9] L. Yang, A. Shami, G. Stevens, and Stephen de Rusett, “LCCDE: A Decision-Based Ensemble Framework for Intrusion Detection in The Internet of Vehicles,” GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Dec. 2022, doi: https://doi.org/10.1109/globecom48099.2022.10001280.

[10] Nissar Nabil, N. Najib, and Jamali Abdellah, “Leveraging Artificial Neural Networks and LightGBM for Enhanced Intrusion Detection in Automotive Systems,” Arabian journal for science and engineering, Feb. 2024, doi: https://doi.org/10.1007/s13369-024-08787-z.

[11] C. D. Kokane, G. Mohadikar, S. Khapekar, B. Jadhao, T. Waykole, and V. V. Deotare, “Machine Learning Approach for Intelligent Transport System in IOV-Based Vehicular Network Traffic for Smart Cities,” International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 11s, pp. 06-16, Sep. 2023, Available: https://ijisae.org/index.php/IJISAE/article/view/3430/2017

[12] M. S. Korium, M. Saber, A. Beattie, A. Narayanan, S. Sahoo, and P. H. J. Nardelli, “Intrusion detection system for cyberattacks in the Internet of Vehicles environment,” Ad Hoc Networks, vol. 153, p. 103330, Feb. 2024, doi: https://doi.org/10.1016/j.adhoc.2023.103330.

[13] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[14] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” Neural Information Processing Systems, 2018. https://papers.nips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

[15] T. Chen and C. Guestrin, “XGBoost: a Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, pp. 785–794, 2016, doi: https://doi.org/10.1145/2939672.2939785.

[16] Zhang. Y, Zhang. J, and Wu. J, “ Leader class and confidence decision ensemble for intrusion detection in vehicular networks,” IEEE Transactions on Vehicular Technology, 68(5), 4620-4633, 2019.

[17] I. D. Aiyanyo, H. Samuel, and H. Lim, “A Systematic Review of Defensive and Offensive Cybersecurity with Machine Learning,” Applied Sciences, vol. 10, no. 17, p. 5811, Aug. 2020, doi: https://doi.org/10.3390/app10175811.

[19] Nitesh Singh Bhati and Manju Khari, “A New Intrusion Detection Scheme Using CatBoost Classifier,” Springer eBooks, pp. 169–176, Jan. 2021, doi: https://doi.org/10.1007/978-3-030-69431-9_13.

[20] S. Dhaliwal, A.-A. Nahid, and R. Abbas, “Effective Intrusion Detection System Using XGBoost,” Information, vol. 9, no. 7, p. 149, Jun. 2018, doi: https://doi.org/10.3390/info9070149.