Machine Learning for Intrusion Detection System in IoT Environment with Permutation Importance

Vishav Pratap Singh[1,†], Raj Kumari[2,*,†] and Mandeep Kaur[3,†]
[1] *Panjab University, Chandigarh (U.T.)* 160014
[2] *Panjab University, Chandigarh (U.T.)* 160014
[3] *Panjab University, Chandigarh (U.T.)* 160014

Abstract

This study evaluates the efficacy of machine learning algorithms in enhancing intrusion detection systems (IDS) within IoT networks, focusing on logistic regression and deep neural network models. Initial findings reveal that without preprocessing, logistic regression performed poorly, underscoring the necessity of feature scaling and data balancing. Subsequent adjustments in these areas substantially improved the model's accuracy and F1-scores, demonstrating the critical importance of these preprocessing steps. Conversely, while a deep neural network achieved high accuracy, it struggled with a lower F1-score, highlighting challenges in achieving balance between precision and recall. The exploration of various preprocessing strategies, including feature importance, significantly contributed to refining the model's predictive capabilities. Future research directions include the development of advanced ensemble techniques to leverage diverse model strengths, optimization of deep learning models to better handle minority class predictions, and enhancement of real-time detection capabilities. Additionally, expanding the adaptability of these models across different IoT domains and configurations will be crucial for practical, real-world application. This study sets the groundwork for further advancements in IDS, aiming to bolster security measures across increasingly complex IoT environments.

Keywords

IoT security, machine learning, intrusion detection systems, real-time processing, federated learning, adversarial attacks

1. Introduction

The Internet of Things (IoT) networks, representing a cutting-edge frontier in technology, encompass interconnected devices and sensors that revolutionize various sectors by collecting and sharing vast amounts of data, facilitating smart homes, healthcare innovations, and industrial automation. Despite the transformative benefits, these networks face significant security challenges due to a vast number of devices, which create a large attack surface for cybercriminals. Vulnerabilities, often due to lax security standards, expose IoT to potential large-scale attacks, raising concerns about data privacy and integrity. To mitigate these risks, advanced encryption, robust authentication mechanisms, machine learning for anomaly detection, and blockchain for decentralized ledgers are being employed to enhance security and data traceability. Intrusion Detection Systems (IDS) play a crucial role by monitoring and safeguarding data streams, ensuring device integrity amidst diverse standards and resource constraints. Nonetheless, challenges in standardization and security implementation persist, emphasizing the need for a balance between innovation and security in the ever-evolving IoT landscape [1][2].

Proceedings of SNSFAIT 2024: International Symposium on Securing Next-Generation Systems using Future Artificial Intelligence Technologies, Delhi, India, August 08-09th, 2024
[1*] Corresponding author.
[†] These authors contributed equally.
pratapvishav92@gmail.com (Vishav Pratap Singh); rajkumari_bhatia5@yahoo.com (Raj Kumari); mandeep24@gmail.com (Mandeep Kaur)
(Vishav Pratap Singh); 0009-0006-2772-3067 (Raj Kumari); 0000-0003-3926-8007 (Mandeep Kaur)

In the domain of cybersecurity, the increasing number of network attacks and intrusions poses a grave threat to the availability and confidentiality of vital data in IOT. Intrusion detection systems, also known by their acronym IDS, are crucial to the security of IOT networks due to their capacity to identify and eliminate potential vulnerabilities. Most traditional IDS methods have primarily relied on singular classifier models to detect and categorize malicious activities. To effectively combat the ever-changing nature of today's cyber threats, we require protection measures that are both more intelligent and advanced.

Ensemble methods have recently emerged as a possible intrusion detection solution. These methods combine numerous classifiers to attain greater precision and robustness. The fundamental idea underlying ensemble approaches is that the combination of several different classifiers can, when employed collectively, produce results that are superior to those produced by a single classifier used alone. Ensemble approaches aim to improve detection rates, reduce the number of false positives and false negatives, and provide robust defences against adversarial attacks. This is achieved by utilizing the knowledge and experience of many classifiers.

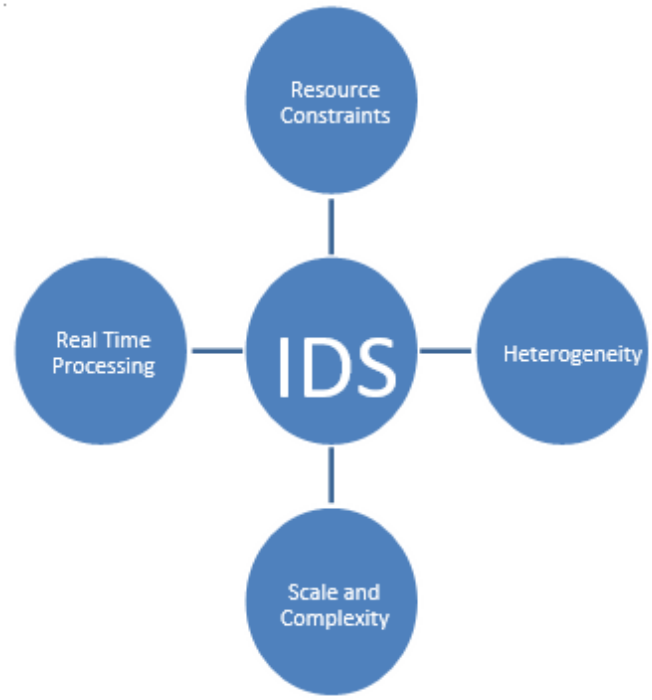

Figure 1: Different challenges faced in IDS

2. IOT Security Challenges and solutions

The Internet of Things (IoT) poses unique challenges due to its vast heterogeneity, scale, and resource constraints, necessitating specialized intrusion detection systems (IDS). IoT devices often have limited processing power, memory, and energy, which complicates the implementation of traditional, resource-intensive IDS. To address this, lightweight and energy-efficient detection techniques are being developed to balance effective security with minimal resource use [3, 4]. Additionally, the IoT ecosystem's diversity in communication protocols, data formats, and architecture makes it difficult to implement a standard IDS that integrates seamlessly across all devices [6]. The sheer volume and complexity of IoT networks, potentially encompassing billions of devices, demand scalable intrusion detection solutions that employ distributed and parallel processing [7]. Furthermore, to ensure timely security responses, IDS must optimize for low-latency, real-time data analysis to quickly address threats [8]. Innovative research and development are crucial to forge IDS that are both effective and tailored to the specific needs of the expansive and diverse IoT environments, ensuring robust protection across varied deployments.

A. Solutions and Techniques:

Intrusion Detection Systems (IDS) for the Internet of Things (IoT) must evolve to meet the distinct challenges of the expansive and diverse IoT landscape. To achieve this, lightweight IDS agents are deployed directly on IoT devices, designed to minimize CPU and memory usage while effectively monitoring local network behavior,

thus reducing data transmission to central systems and lessening the load on central processing, ideal for large-scale implementations [9]. Decentralization is furthered by edge computing, which processes data close to its origin, decreasing latency and bandwidth needs and enhancing the capacity for rapid response in real-time applications [10]. Additionally, flow-based analysis focuses on network flow data to detect anomalies efficiently [11], while blockchain technology secures IDS logs with a decentralized and immutable ledger [12]. Collaborative IDS systems enhance security through shared intelligence among IoT devices, fostering a proactive defence approach [13]. Together, these strategies form a comprehensive framework that addresses resource constraints, heterogeneity, and complexity, ensuring robust and real-time threat detection and response across IoT networks.

Figure 2: Different Types of network attacks faced by IoT

B. Types of Attacks on IoT Devices:

To enhance the security of Internet of Things (IoT) networks and mitigate potential cyberattacks, several key strategies are essential: Implementing robust authentication methods such as multi-factor authentication (MFA) and using strong cryptography can secure access to IoT devices and protect data exchanged [24]. Regularly updating software and firmware is crucial to address known vulnerabilities and counteract emerging threats, with automatic updates serving to reduce security gaps [25]. Network segmentation can effectively isolate critical systems and sensitive data, limiting the reach of potential attacks within the network [26]. Anomaly detection and traffic monitoring using IDS and machine learning can identify and respond to unusual activities swiftly, enhancing threat detection and network resilience [27]. Managing device identities ensures that only authorized devices can access the network, bolstering overall network security [28]. Additionally, ensuring the physical security of IoT devices through tamper-resistant hardware and controlled access prevents direct physical attacks [29]. Together, these measures provide a comprehensive approach to securing IoT ecosystems against a wide range of cyber threats.

C. Solutions to Prevent IoT Network Attacks:

To bolster the security of Internet of Things (IoT) networks against diverse cyber threats, several strategic measures can be implemented: Robust authentication techniques, such as multi-factor authentication (MFA) and strong encryption, ensure that only authorized users access IoT systems and protect data exchanges [24]. Regular software updates and patches mitigate known vulnerabilities, with automatic updates facilitating timely enhancements [25]. Network segmentation divides the network into manageable sections, isolating critical systems to minimize attack impact [26]. Continuous monitoring of network traffic through anomaly detection and IDS identifies and mitigates threats promptly, with machine learning enhancing detection capabilities [27]. Centralized device identity management controls network access by verifying each device's

unique identifier [28]. Additionally, physical security measures like tamper-resistant hardware and secure enclosures protect against physical tampering [29]. Collectively, these strategies form a comprehensive defence framework, safeguarding IoT networks from both digital and physical security risks.

D. Approaches to IoT Intrusion Detection through traffic monitoring and anomalies:

Intrusion detection systems employing Behavioral Anomaly Detection, Signature-Based Detection, and Machine Learning-Based Detection methods significantly bolster network security by proactively identifying and mitigating potential threats. Behavioral Anomaly Detection sets a baseline behavior for IoT devices and networks, flagging substantial deviations as potential anomalies, thus detecting new or unknown threats [30]. Signature-Based Detection utilizes known patterns of attacks to identify threats but may falter against zero-day attacks which are unforeseen hazards [31]. Machine Learning-Based Detection leverages supervised and unsupervised learning to discern attack patterns from historical data, offering robust defences against sophisticated attacks [32].

These methods contribute to network security by enabling early identification of abnormalities, which helps detect potential breaches swiftly [30]. Recognizing suspicious behavior through continuous monitoring aids in real-time threat identification and response, essential for mitigating attacks like DoS and brute-force attempts [31]. Additionally, predictive capabilities help in safeguarding against DDoS attacks through traffic pattern analysis, facilitating preemptive responses to unusual traffic spikes [32]. They also enhance the detection of zero-day attacks by identifying atypical patterns indicative of such vulnerabilities [33]. Traffic monitoring helps in spotting insider threats by analyzing unusual data access or transfers, while the focus on behavioral deviations minimizes false positives, enhancing the accuracy of threat detection [34]. Moreover, these methods strengthen incident response and network resilience by allowing security teams to act quickly and adapt security measures dynamically, thus maintaining network functionality and ensuring business continuity [36, 37]. Overall, integrating these advanced detection strategies into IDS ensures a comprehensive and adaptive security posture for networks, crucial for defending against a range of cyber threats [38].

E. Different Methods of Anomaly Detection

Anomaly detection systems are crucial for identifying patterns or behaviors that deviate significantly from established norms in datasets, potentially indicating errors, odd occurrences, or security threats. These systems utilize a variety of methodologies, each suited to specific types of data and applications [41]. Statistical anomaly detection has gained popularity across many industries, employing tests like Dixon's Q-test and Grubbs' test for small datasets, and z-score methods for data with a normal distribution [39]. Machine learning-based detection leverages techniques like the Isolation Forest for handling anomalies in large datasets and One-Class SVM for cybersecurity and fraud detection [40]. Time-series anomaly detection, essential for IoT applications, uses methods like STL and Prophet to monitor data over time [42]. Unsupervised detection methods are invaluable when labeled data is scarce, utilizing density and cluster-based approaches for scenarios like fraud and intrusion detection [43]. Ensemble anomaly detection methods improve detection accuracy and reduce false positives by combining multiple approaches [44, 91, 92]. Network anomaly detection, critical for IoT security, employs flow-based, packet-based, and behavior-based analyses to safeguard systems [41]. The choice of anomaly detection system depends on the specific use case, dataset characteristics, and required complexity, with researchers and practitioners often combining methods to optimize results [45, 91, 92].

3. Recent Works on IoT Intrusion Detection

In recent years, there has been a great deal of research on intrusion detection using ensemble deep learning, and it has demonstrated great promise. In the process of detecting and mitigating network intrusions, the combination of ensemble methods and deep learning approaches has improved accuracy, resilience, and generalization capabilities. The following is a list of major research advances in this field:

Researchers [46, 47, 91, 92] have investigated the possibility of combining multiple deep neural networks into a singular ensemble model to enhance intrusion detection. Ensemble models can capture multiple components of network traffic and identifying malicious activity. Either by training a variety of DNN architectures or by modifying the hyper parameters of specific networks, this is accomplished.

In the guise of adversarial attacks, intrusion detection systems face a significant barrier. Researchers [48, 49] have investigated how ensembles of deep learning models can be made more resistant to such assaults. As a result of combining the predictions of numerous deep neural networks, ensemble models can enhance adversarial sample detection and mitigation. This helps to assure resilience against increasingly sophisticated intrusion attempts. In cross-domain intrusion detection scenarios, ensembles of deep learning have also been utilized [50, 91]. Transfer learning techniques permit the transfer of information from one domain to another, from a source domain (such as a labelled dataset) to a target domain (such as a distinct network environment). It is becoming increasingly crucial for ensemble models to effectively adapt and generalize their detection skills across domains, thereby enhancing their performance in novel intrusion detection scenarios.

Recent research has placed a significant reliance on the interpretability of ensembles of deep-learning models employed in intrusion detection. Researchers [51] have investigated various methods for explaining the ensemble's decision-making process, allowing security analysts to comprehend and rely on the model's output. Various techniques, including attention processes, saliency maps, and rule extraction, have been implemented to enhance the explainability of ensemble models.

The increasing prevalence of distributed networks has led to the emergence of federated learning as a potentially fruitful solution to the intrusion detection problem. Federated learning protects the privacy of users while leveraging the collective intelligence of distributed networks to improve the precision of intrusion detection [53]. This is achieved through the collaborative training of ensemble models across multiple network nodes without sharing raw data.

Unbalanced datasets, in which the number of normal instances greatly outnumbers the number of invasions, pose a challenge for intrusion detection. Attempts have been made to address this issue by resolving class imbalance using ensemble deep-learning techniques. These methodologies may involve oversampling, under sampling, or hybrid techniques. Ensemble models have the potential to learn from the cases of the minority class, thereby enhancing their ability to detect intrusions with reliability [52].

These findings demonstrate the potential of ensemble deep-learning methods in the field of intrusion detection. By combining the characteristics of deep learning architectures and ensemble approaches, researchers aim to develop robust, accurate, and adaptable intrusion detection systems. These systems will be able to manage cyber threats that are both complex and dynamic. Historically, academicians have utilized a variety of machine learning-based intrusion detection strategies. This section explains some of the most recent and pertinent:

In a research paper (54), deep decision trees are proposed as a novel method for detecting network intrusion. The researchers construct a deep decision tree ensemble by combining numerous decision trees, each of which is trained on a distinct subset of characteristics. The ensemble's architecture enables it to capture intricate correlations between the characteristics of network traffic, resulting in improved detection precision. Experiments conducted on datasets obtained from real-world network environments demonstrated the effectiveness of the proposed method. Existing decision tree-based approaches were eclipsed by the new method in terms of detection rate and false alarm rate. The research demonstrates that deep decision trees have the potential to be an effective method for detecting intrusions in large and complex networks.

In research paper [55], an enhanced version of the random forest method specifically tailored for intrusion detection is presented. The proposed method enhances the performance of the random forest classifier by integrating strategies for feature selection and parameter tuning. Experiments conducted on a benchmark dataset for intrusion detection demonstrate that the enhanced random forest outperforms both the classic random forest and other intrusion detection methods considered to be state-of-the-art. The research emphasizes the significance of fine-tuning parameters and selecting essential features in intrusion detection applications in order to achieve greater accuracy.

The authors of a distinct study [56] propose an improved version of the Support Vector Machine (SVM) algorithm for intrusion detection. To improve the overall performance of the classifier, the modifications focus predominantly on fine-tuning the SVM parameters and optimizing the kernel functions. The evaluation performed on a real-world intrusion detection dataset demonstrates that the modified SVM model obtains a

higher detection accuracy than the conventional SVM and other baseline approaches. Additionally, other baseline approaches are evaluated. According to the findings of the study, optimizing the parameters of a support vector machine (SVM) can result in significant improvements in detecting network intrusions and enhancing network security in general.

An additional essential piece of research [57] describes an intrusion detection system based on deep learning and employing Convolutional Neural Networks (CNN) and k-means clustering to classify network traffic. The objective of the model is to discover complex representations and patterns concealed within the network traffic data. Experimental results obtained from a dataset containing real-world data demonstrate that the proposed CNN-based intrusion detection system outperforms extant machine learning techniques. This study emphasizes the potential of deep learning technologies, such as CNN, for efficient and accurate network intrusion detection in complex and dynamic environments.

Sarrar and Al-turjman present a distributed intrusion detection system based on the K-Nearest Neighbors (KNN) algorithm in their paper [58]. Numerous nodes are utilized by the proposed system in order to process network traffic data and effectively identify anomalies. This method is intended to reduce computational overhead while enhancing scalability; as a result, it is suitable for use in environments with limited resources and a large scale. Experimental evaluations performed on datasets derived from the real world demonstrate that the distributed KNN-based system obtains accuracy comparable to that of centralized approaches while drastically reducing the processing time required. This research demonstrates the effectiveness of distributed KNN algorithms for intrusion detection in the Internet of Things (IoT) and other distributed network environments.

Almarri et al. present a hybrid deep learning ensemble model for intrusion detection in their paper [59]. This model combines Naive Bayes with methods of deep learning. The hybrid model exploits the benefits that can be derived from both methods, such as the usability of Naive Bayes and the capacity for representation learning provided by deep learning. Experimental results on real-world datasets demonstrate that, in terms of detection accuracy, the hybrid model outperforms both conventional machine learning methods and distinct deep learning models. According to the findings of the study, the efficacy of intrusion detection can be improved by combining multiple methods.

Zhang et al. [60] present an adaptive gradient boosting-based effective ensemble learning approach for intrusion detection. This algorithm is highly effective. The objective of the proposed technique is to optimize the performance of the ensemble model by dynamically adjusting the weights and learning rates during the boosting process. Using experimental assessments conducted on real-world datasets, the Adaptive Gradient Boosting-based ensemble has been shown to achieve superior levels of accuracy and robustness compared to conventional ensemble approaches. The research demonstrates the importance of employing adaptive learning strategies when creating efficient ensemble models for intrusion detection systems.

In the landscape of Intrusion Detection Systems (IDS), recent studies have underscored the efficacy of various deep learning and machine learning models across diverse attack types, highlighting significant achievements in accuracy, recall, and precision metrics. In general, the cited studies cast light on the significance of sophisticated machine learning and ensemble methods for intrusion detection in a vast array of network contexts. The use of deep learning approaches, improved SVM and random forest algorithms, and hybrid models has yielded promising results, enabling the detection of network intrusions with greater precision and efficiency. These advancements contribute to the enhancement of the overall security posture of networked systems other than the Internet of Things. Researchers and practitioners in the field of intrusion detection can use these insights to create systems that are more efficient and reliable in their mission to safeguard vital network infrastructure. The following table I provides an exhaustive analysis of the numerous possible approaches:

Table 1. Results Obtained by different Preprocessing methods with Logistic Regression Classifier

Reference	Classifier	Methodology	Outcome	Limitations
[54] (2020)	Decision Trees	Deep decision tree architecture for intrusion detection	Improved detection accuracy and reduced false positive	-May suffer from overfitting -Less effective for high-dimensional
[55] (2021)	Random Forest	Improved random forest algorithm for intrusion detection	Enhanced accuracy and reduced detection time	- Ensemble size and complexity can impact computational resources
[56] (2021)	Support Vector Machines (SVM)	Improved SVM algorithm with feature selection and kernel modification	Improved detection accuracy and reduced false	- Sensitivity to hyperparameter tuning -Requires labeled training
[57] (2021)	Neural Networks	Convolutional neural network (CNN) with k-means clustering	Improved accuracy in network traffic	- Requires large amounts of labeled training data
[58] (2022)	K-Nearest Neighbors (KNN)	Distributed intrusion detection using the KNN algorithm	Improved detection accuracy and reduced false	- Memory-intensive for large datasets Computationally expensive
[59] (2022)	Naive Bayes	Hybrid deep learning ensemble with naive Bayes component	Enhanced detection accuracy and reduced false positives	- The assumption of feature independence may not hold true in complex datasets
[60] (2022)	Ensemble-Specific Classifiers	Adaptive Gradient Boosting algorithm for ensemble learning	Improved detection accuracy and reduced false positives	- Computationally intensive due to the sequential nature of boosting
[61] (2023)	Ensemble Methods	Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor with voting and stacking	High accuracy of more than 98.3%	-Only one dataset used, -Deep Learning is not used. -only binary classifier is used while multi classifiers required in real life situations.
[95] (2023)	Ensemble Methods	The paper proposes an intrusion detection system based on stacked ensemble learning for IoT networks, achieving a high average accuracy rate of 99.68%.	- Proposed IDS achieves high average accuracy rate of 99.68% -Potential to improve security of IoT devices	-Computational resource constraints of IoT devices - Production cost limitations of IoT Devices
[96](2023)	Ensemble Methods	- Recursive Feature Elimination (REE) with KDD 99 dataset - Hyper parameter tuning and ensemble learning	- Performance of IDS assessed using evaluation metrics. - Hyper parameter tuning and ensemble learning used for improvement.	- New large dataset such as IoTID23 need to be evaluate on the existing methods
[97](2023)	Ensemble Methods	- PCC-CNN model for anomaly detection - Traditional PCC-ML models for comparison	- Best accuracy achieved by KNN and CART models: 98%, 99%, and 98% - Proposed PCC-CNN model achieved detection accuracy of 99.89%	- New large dataset such as IoTID23 need to be evaluate on the existing methods
[98](2023)	Ensemble Methods	- Light Gradient Boosting Machine (LGBM) classifier - Multi-class classification with LGBM classifier	- Proposed method achieved the highest classification performance in the literature. - Proposed method effective in preventing cyber-attacks.	- New large datasets such as IoTID23 need to be evaluated on the existing methods

[99](2023)	Ensemble Methods	- Random Forest (RF) for dimensionality reduction - Ensemble learning method for intrusion detection and identification	- Proposed RF method outperformed other approaches - Achieved accuracy of 99% on IoTID20 dataset	- New large datasets such as IoTID23 need to be evaluated on the existing methods
[100](2023)	Ensemble Methods	- Hybrid CatBoost regression model - IDS2017 dataset	- The proposed system achieved 92.5% accuracy. - The system was compared with various state-of-the-art approaches.	-Old dataset
[101](2023)	Ensemble Methods	Copilot couldn't generate the response. Please try again after some time.	- Accuracy detection rate: 98% - F1-score: 92% (multi-class attacks)	- Existing IDS approaches are not suitable for IoT traffic. - Constrained nature of IoT devices.
[102](2023)	Ensemble Methods	- Deep Learning-based intrusion detection system - Four-layer deep Fully Connected network architecture	- Average accuracy of 93.74% -Precision, recall, and F1-score of 93.71%, 93.82%, and 93.47% respectively.	-old dataset
[103](2023)	Ensemble Methods	- IVD-IMT algorithm under Artificial Immune System (AIS) based Intrusion Detection System (IDS) - Tuning input parameters using synthetic datasets and NSL-KDD dataset	- Improved variable-sized detector generation algorithm for healthcare - Emphasis on lowering false alarm rate without compromising detection rate	-old dataset
[104](2023)	Ensemble Methods	- Distributed processing and feature selection on IoT data - Deep learning with Recurrent Neural Networks (Simple RNN and Bi-directional LSTM)	- Feature selection reduced dataset size by 90% - Models achieved higher recall rate with reduced feature space	- Communication overheads due to large volume of IoT data - Computation requirements for deep learning models

F. Research Gaps Identified from Literature Review

Based on our review of the pertinent literature, we have identified the following research gaps in current intrusion detection systems.

Resilience Against Adversarial Attacks: The IDS must be able to withstand attacks by adversaries attempting to deceive or manipulate the system. There are research voids in the development of defence mechanisms, such as adversarial training, robust feature representations, and anomaly detection techniques, to increase the IDS's resistance to adversarial attacks [62].

Unbalanced datasets, in which the number of normal instances vastly outnumbers the number of intrusion instances, pose a challenge for intrusion detection systems (IDS). To eliminate biases and improve the detection of rare incursions, it is necessary to conduct research on methods for managing imbalanced data, such as oversampling, under sampling, and hybrid approaches [63]. These methods consist of oversampling, under sampling, and hybrid techniques.

Methods for Protecting the Privacy of Users Intrusion detection systems frequently require access to sensitive network data, raising privacy concerns. There are research gaps in the development of privacy-preserving strategies, such as federated learning, secure multiparty computation, and differential privacy, in order to facilitate collaborative intrusion detection while protecting the privacy of the data sources [64]. Federated learning, secure multiparty computation, and differential privacy are some of these techniques.

Integration and Deployment in the Real World: Even though research has increased the capabilities of IDS, there is a deficiency in the deployment and incorporation of these systems in complex network environments. To effectively integrate IDS into operational networks, research should concentrate on the development of approaches and frameworks that account for the networks' heterogeneity, interoperability, and scalability [65].

These references provide a foundational understanding of current research efforts and pinpoint where further exploration and development are needed to fill the identified gaps in the field of IoT cyber-attack detection using machine learning. It will be possible to further enhance the capabilities of intrusion detection systems by addressing these research gaps and problems. This will enable the detection and mitigation of network intrusions to be conducted in a more resilient, accurate, and efficient manner. Although there are many research gaps, we have identified based on literature review but our focus in this study will be on robustness of the system (Resilience against the adversaries), unbalanced datasets, different feature engineering methods and techniques to improve the accuracy and robustness of the IDS targeting the first two research gaps only considering the wide scope of the research gaps.

4. The Proposed Methodology for IOT Threat Detection with Feature Engineering

Intrusion detection systems (IDS) are critical for detecting and mitigating malicious activities in computer networks, increasingly utilizing ensemble machine learning techniques to enhance detection accuracy. The proposed methodology encompasses several steps outlined in Figure 2, starting with data pre-processing and feature engineering. This involves collecting network traffic data, handling missing values, scaling numerical features, encoding categorical variables, and employing feature selection techniques like correlation analysis and recursive feature elimination to identify and engineer pertinent features [65]. The next step addresses data balancing in inherently imbalanced intrusion detection datasets, using methods like random oversampling, under-sampling, Synthetic Minority Over-Sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) to mitigate bias towards the majority class [66,67]. Ensemble techniques are then applied to improve overall performance by leveraging the strengths of various classifiers through methods such as voting, Bagging, Random Forest, boosting with AdaBoost and GBM, and stacking with a meta-classifier [66-70]. A deep learning ensemble approach is considered for future enhancement of system accuracy and resilience. Model evaluation is conducted via split datasets or cross-validation, utilizing metrics like accuracy, precision, recall, F1-score, and AUC-ROC to assess performance [79]. The methodology concludes with model optimization through hyper parameter tuning using grid search and Bayesian optimization, followed by real-time implementation and monitoring of the IDS in network environments.

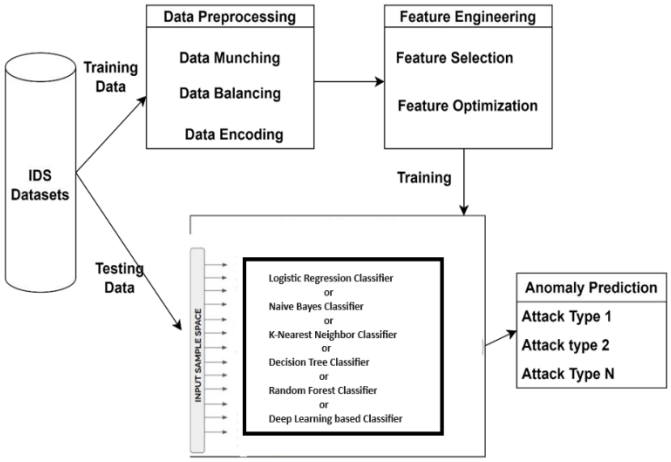

Figure 3: A model for intrusion detection with feature engineering

A. Feature Engineering for Selection and Optimization in IDS

Feature engineering techniques are invaluable for enhancing the functionality of Intrusion Detection Systems (IDS) by transforming input features into a lower-dimensional space while preserving essential information, thus optimizing the feature selection process and boosting the efficiency of IDS [93, 94]. Techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) reduce the

dimensionality of data, minimizing computation time and curbing overfitting in high-dimensional feature spaces. This dimensionality reduction not only retains critical information but also decreases noise and redundancy. Feature embedding also illuminates the significance of various features, aiding in discerning which characteristics are most indicative of normal versus anomalous activities, thereby guiding feature selection towards those that are most informative for detecting intrusions. Moreover, these methods enhance data representation by capturing complex inter-feature relationships that may not be visible in the original dataset, which can significantly improve machine learning model performance. Additionally, feature embedding can be integrated with optimization methods like grid search or Bayesian optimization to refine IDS classifier performance. It also addresses the challenge of imbalanced datasets common in intrusion detection by creating a more balanced feature space representation, facilitating more effective anomaly detection. Furthermore, embedding from related tasks or datasets can be applied to intrusion detection through transfer learning, especially useful when labeled data is scarce. Lastly, interpretable embedding methods like t-SNE enable security analysts to better understand and visualize the relationships within the data, offering valuable insights into intrusion patterns and supporting informed decision-making in cybersecurity operations.

B. Different Benchmarked Public Datasets for Machine Learning Based Threat Detection

This section describes three significant datasets used for intrusion detection system (IDS) evaluation, each with unique characteristics tailored to specific network environments. The UNSW-NB15 dataset [106], generated from real network traffic at a university, includes training and testing data that features labels indicating normal or attack states, making it a modern, large-scale dataset suitable for evaluating various IDS techniques. However, it may require feature engineering to optimize usage. The CIC-IDS2017 dataset [107], created from a real-world IoT environment, focuses on IoT network traffic and is presented across multiple CSV files. This dataset offers a glimpse into IoT network intrusion scenarios but demands significant data pre-processing and domain-specific expertise due to its diversity and recency. Lastly, the CCIoT2023 dataset [105] stands out as the largest in terms of the number of devices (105) used to establish the network topology and the variety of attacks (33), classified into seven categories, making it highly relevant for current IoT security research. It includes attacks carried out using IoT devices like Zigbee and Z-wave, emphasizing the complexity and evolving nature of network security threats.

5. Feature Selection using Feature Importance using

1. Dataset Selection

For our case study we have used the latest and most comprehensive and realistic dataset CCIoT2023. This dataset designed for evaluating the performance of machine learning algorithms in the context of IoT network security. It comprises a substantial volume of 1,048,575 records, enriched with 47 distinct features. This dataset is structured to include a wide variety of IoT network threats, featuring 34 different attacks that are categorized into 8 classes, illustrating a diverse range of security challenges against IoT devices. Unique to this dataset is its use of actual IoT devices both as attackers and victims within a meticulously constructed network topology that mirrors real-world IoT environments. This setup not only enhances the realism of the dataset but also aids in understanding the dynamics of IoT-specific vulnerabilities and attack vectors. The CCIoT2023 dataset is specifically designed for training and testing machine learning models to classify and detect IoT network traffic as either malicious or benign, making it an invaluable resource for researchers and practitioners working to fortify IoT networks against emerging security threats.

2. Preprocessing of the dataset

In addressing data preprocessing challenges such as handling missing values (NaNs), scaling features, and managing data imbalances, we observed a significant variance in records per threat class, ranging from 281 to 163,281, highlighting the dataset's severe imbalance. To mitigate this issue, we initially sampled each class to uniformly include only 10,000 records. The dataset was divided into training (80%) and validation (20%) sets to facilitate effective model training and evaluation. Using logistic regression on this balanced subset yielded a low Accuracy of 18.37% and an F1 Score of 10.87%. Recognizing the need for further preprocessing, we applied a standard scaler to normalize the feature values across a consistent range. Additionally, we employed the Synthetic Minority Over-Sampling Technique (SMOTE) to enhance the representation of minority classes, inflating each to 10,000 records. Consequently, the total number of records for our experiments increased to 340,000. These adjustments significantly improved model performance, with the revised logistic regression

model achieving an Accuracy of approximately 70.98% and an F1 Score of 70.75%. This illustrates the critical impact of comprehensive preprocessing and balancing strategies on the effectiveness of predictive models in handling highly imbalanced datasets.

3. Feature Engineering with Permutation importance

We then use the feature importance concept (permutation importance) which is a technique used to assess the significance of features in a predictive model, applicable across various model types due to its model agnostic nature. The process involves training a model and establishing a baseline performance using an appropriate metric like accuracy or mean squared error. Then, each feature's values are shuffled individually to disrupt their relationship with the target, and the model's performance is re-evaluated. The decrease in performance indicates the feature's importance, with a significant drop suggesting a high reliance by the model on that specific feature. This method is straightforward and helps in understanding feature interactions, but it can be influenced by randomness, especially in small datasets or when features are highly correlated, potentially leading to inflated importance of redundant features. Permutation importance thus offers a practical approach to determine which features most affect a model's predictions, providing valuable insights for feature selection and model refinement.

4. Results and Discussion

Utilizing the permutation importance technique with the `eli5` library, we identified the most influential features within our dataset, successfully reducing the number of features from 47 to 28. This reduction, focusing only on features impacting results by more than 1%, not only significantly decreased execution time but also improved model performance. As a result of this feature selection, we achieved a higher Accuracy of 75.18% and an F1 Score of 73.20%. Notably, this F1 Score surpasses the outcome reported in the main paper of CCIoT2023, which documented an F1 Score of only 67% using a deep neural network. While this represents a substantial improvement in model precision and recall, the accuracy, although improved, still remains an area for potential enhancement. This demonstrates the effectiveness of targeted feature reduction in enhancing model efficiency and performance in complex datasets.

Table 2. Results Obtained by different Preprocessing method with Logistic Regression Classifier

Algorithm	Accuracy	F1-Score
Logistic Regression without Scaling	18.37%	10.87%
Logistic Regression with Data Balancing	70.98%	70.75%
Logistic Regression with feature importance	75.18%	73.20%
Deep Neural network [CICIoT 2023 paper ref. [105]	98.61%	67.23%

The table II compares the performance of logistic regression under various configurations and a deep neural network on certain metrics. Initially, logistic regression without feature scaling showed poor performance with an Accuracy of 18.37% and an F1-Score of 10.87%, indicating substantial issues in handling raw data. Improvement was seen when data balancing techniques were applied, with the model's Accuracy and F1-Score rising to 70.98% and 70.75%, respectively, demonstrating the effectiveness of addressing class imbalance. Further enhancement was achieved through the use of feature importance to selectively reduce the number of features, which led to even better results with an Accuracy of 75.18% and an F1-Score of 73.20%, highlighting the benefits of focusing on relevant features. In contrast, the deep neural network, while achieving the highest Accuracy at 98.61%, had a lower F1-Score of 67.23% according to reference [105] from the CICIoT 2023 paper. This suggests that while the network was highly accurate overall, it struggled more with the balance between precision and

recall, especially in classifying minority classes effectively compared to some configurations of logistic regression.

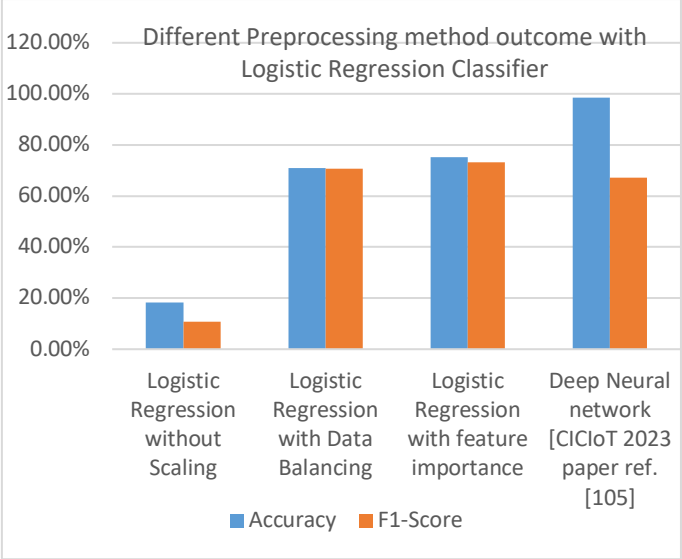

Figure 4: Different Pre-processing method outcome with Logistic Regression Classifier

6. Conclusion

This discussion underscores the significant impact that proper preprocessing and model configuration can have on the performance of machine learning algorithms for intrusion detection in IoT networks. Initial attempts using logistic regression without scaling demonstrated suboptimal performance, which was markedly improved through strategic adjustments such as feature scaling, data balancing, and the application of feature importance techniques. These modifications helped elevate the model's accuracy and F1-score significantly, reflecting the critical importance of feature normalization and balanced training sets in enhancing model efficacy. However, while logistic regression models showed notable improvements, the deep neural network presented a mixed outcome; it achieved high accuracy but a relatively lower F1-score, indicating an area where the model might be failing to effectively balance precision and recall, particularly in the context of minority class predictions.

7. Future Directions

Going forward, several avenues appear promising for further research and development:

1. **Advanced Ensemble Techniques:** Exploring more sophisticated ensemble methods that could combine the strengths of different underlying models might provide a pathway to both high accuracy and high F1-scores, ensuring robustness across various types of network intrusion scenarios.
2. **Deep Learning Optimization:** Given the high accuracy but lower F1-score of the deep neural network, there is a clear opportunity to refine these models, possibly by integrating techniques such as cost-sensitive learning or advanced oversampling methods tailored for deep learning to enhance minority class recognition.
3. **Feature Engineering and Selection:** Continued efforts in feature engineering and more dynamic feature selection methods could yield further improvements. Machine learning models can benefit from ongoing refinement of input features, possibly through automated feature engineering techniques that evolve based on emerging threat patterns.
4. **Real-Time Detection Capabilities:** Enhancing the real-time detection capabilities of IDS systems through the integration of streaming data models and incremental learning could help in effectively tackling the latest threats as they occur in IoT environments.
5. **Cross-Domain Adaptability:** Investigating the adaptability of the developed models across different IoT domains and network configurations would be valuable, ensuring that the solutions are not only effective in controlled test environments but also in diverse real-world settings.

By pursuing these directions, future research can continue to push the boundaries of what is possible in intrusion detection, ultimately leading to more secure and resilient IoT networks.

References

[1] Y. Chen, M. Patel, D. Wang, and P. Hui, ‘IoT Sentinel: Automated Device-Type Identification for Security Enforcement in IoT’, IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4203–4214, 2019.

[2] K. K. R. Choo, C. Liu, and K. K. W. Ho, ‘Intrusion detection in the Internet of Things (IoT) with minimal human intervention: A survey’, Journal of Network and Computer Applications, vol. 103, pp. 1–17, 2018.

[3] J. Luo, C. H. Liu, Y. Jin, J. Deng, and X. S. Shen, "Lightweight intrusion detection for Internet of Things with binary local linear embedding," IEEE Transactions on Information Forensics and Security, vol. 14, no. 1, pp. 132-147, Jan. 2019.

[4] R. Xu, Y. He, Q. Li, M. Guo, and J. Wang, "Real-time and lightweight anomaly detection for time series of counts," in 2018 24th IEEE International Conference on Parallel and Distributed Systems (ICPADS), pp. 705-714, 2018.

[5] L. Gao, J. An, F. Shang, W. F. A. Jia, R. O. Alaba, and M. O. Ayeni, ‘IoT-Botnet Attack Detection Based on Fuzzy Clustering and Extreme Learning Machine’, International Journal of Computer Applications, vol. 9, no. 9, pp. 13–18, 2017.

[6] F. A. Alaba, R. O. Ayeni, and M. O. Adigun, "An intrusion detection system model for Internet of Things," International Journal of Computer Applications, vol. 158, no. 9, pp. 13-18, Mar. 2017.

[7] N. M. Mahmood, M. S. Iqbal, A. Ahmed, H. Javaid, and M. Imran, "A scalable distributed intrusion detection system for the Internet of Things," Future Generation Computer Systems, vol. 80, pp. 408-422, Feb. 2018.

[8] P. K. Danso, E. C. P. Neto, S. Dadkhah, A. Zohourian, H. Molyneaux, and A. A. Ghorbani, ‘Ensemble-based intrusion detection for internet of things devices’, in 2022 IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), Marietta, GA, USA, 2022

[9] L. Zhang, Q. Wang, J. Wang, and W. Zhang, "A Lightweight Intrusion Detection System for the Internet of Things Based on Parallel K-means Clustering," Wireless Personal Communications, vol. 101, no. 3, pp. 1141-1155, Nov. 2018.

[10] M. F. Hasan, S. A. Alshehri, A. Alamri, and M. F. Alhamid, "A lightweight and distributed intrusion detection framework for Internet of Things," Computers & Security, vol. 83, pp. 108-127, Dec. 2019.

[11] J. Ghaleb, X. Wu, and Y. Yang, ‘A lightweight anomaly detection system for Internet of Things networks’, Journal of Network and Computer Applications, vol. 103, pp. 130–142, 2018.

[12] J. Yan, Y. Da Xu, H. Wang, S. Wang, and H. Hu, "Towards a blockchain-based framework for collaborative DDoS attack mitigation with smart contracts," Future Generation Computer Systems, vol. 88, pp. 173-180, Sep. 2018.

[13] L. Nie et al., ‘Intrusion detection for secure social internet of things based on collaborative edge computing: A generative adversarial network-based approach’, IEEE Trans. Comput. Soc. Syst., vol. 9, no. 1, pp. 134–145, Feb. 2022.

[14] G. S. Jangra and S. Kaur, ‘A Study of Various IoT Security Attacks and Their Detection Techniques’, in Proceedings of the 4th International Conference on Internet of Things and Connected Technologies (ICIoTCT), pp. 363–368, 2021

[15] M. Rashid and S. H. Hamid, ‘Security and Privacy Issues in IoT: A Comprehensive Survey’, Journal of Network and Computer Applications, vol. 126, pp. 11–31, 2019.

[16] R. D. Reddy and B. R. Lolla, ‘IoT Security: A Comprehensive Survey’, Journal of Computer Science and Technology, vol. 33, no. 3, pp. 531–563, 2018.

[17] Y. Xiao, D. Zhang, Y. Yang, and X. Wang, ‘A Survey of Security and Privacy Issues in Internet of Things’, Journal of Industrial Information Integration, vol. 10, pp. 1–11, 2018.

[18] M. Khan, R. Amin, M. Ali, H. Zhang, H. A. N. Akhtar, and S. U. Khan, "Dynamic threshold-based DoS attack detection in internet of things," IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7458-7466, Aug. 2020.

[19] J. Liu, X. Yang, S. Peng, C. Ma, and J. Han, "iToPSec: A lightweight intrusion detection system for Internet of Things in software-defined networks," Future Generation Computer Systems, vol. 78, pp. 1-11, Dec. 2017.

[20] S. Sami, S. S. Alwakeel, S. S. Alwakeel, T. W. D. Park, M. H. A. Almulla, and K. Salah, "Machine learning-based botnet detection approaches in the Internet of Things: A review," IEEE Access, vol. 8, pp. 209,730-209,755, 2020.

[21] Abdalqader, M. Hassan, S. A. Aljawarneh, and M. Alqatawna, "Secure boot for Internet of Things: Challenges and state of the art," in 2017 9th International Conference on Computer and Automation Engineering (ICCAE), pp. 114-118, 2017

[22] S. Singh, A. M. Prasad, V. Choudhary, S. K. Das, and S. Prasad, "A secure communication protocol for Internet of Things," in 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-5, 2015.
[23] M. Biswas and J. Gutierrez, "Internet of things device authentication techniques: A survey," Journal of Network and Computer Applications, vol. 98, pp. 18-29, Jan. 2018.
[24] Khan and I. H. Elhajj, "Securing Internet of Things: A Survey," IEEE Internet of Things Journal, vol. 7, no. 10, pp. 8778-8795, Oct. 2020.
[25] J. M. Kim, M. A. Hossain, and G. J. Lee, "Internet of Things (IoT) based security and privacy mechanisms for medical devices in healthcare systems," Computer Networks, vol. 165, pp. 106979, Nov. 2019.
[26] Z. Qin, S. Wang, D. Jiang, W. Jia, M. S. Hossain, and A. Ghoneim, "Security Threats and Solutions in Internet of Things: A Survey," IEEE Access, vol. 8, pp. 219-245, 2020.
[27] Marcelloni, L. Tanci, and S. R. Rajagopalan, "An Anomaly Detection System for Internet of Things Applications," IEEE Internet of Things Journal, vol. 7, no. 10, pp. 9516-9529, Oct. 2020.
[28] K. Salikhov, D. Kim, J. Song, D. Kim, and S. Rho, "Lightweight Trust-Based Device Identity Management for Internet of Things," Symmetry, vol. 13, no. 4, pp. 659, Apr. 2021.
[29] P. I. R. Grammatikis, P. G. Sarigiannidis, and I. D. Moscholios, 'Securing the Internet of Things: Challenges, threats and solutions', Internet of Things, vol. 5, pp. 41–70, 2019.
[30] N. Avgeriou, S. Papavassiliou, G. Apostolopoulos, and T. Karetsos, "Early Anomaly Detection for Network Security in SDN-based Infrastructures," IEEE Transactions on Network and Service Management, vol. 17, no. 1, pp. 124-139, Mar. 2020.
[31] Z. A. Baig, H. A. Alhumayzah, A. T. Alqarni, T. R. Sheltami, and T. N. Alotaibi, "An Adaptive Hybrid Network Intrusion Detection System using Traffic Behavior Analysis," in 2021 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 371-377, 2021.
[32] P. Sarkar, S. Misra, and H. V. Ramakrishnan, "DDoS Attack Detection and Mitigation Using Software Defined Networking," IEEE Transactions on Network and Service Management, vol. 17, no. 1, pp. 117-123, Mar. 2020.
[33] S. V. Pemmaraju and R. D. Reddy, "Zero-day Attack Detection using Machine Learning Algorithms in Cloud," in 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), pp. 1-6, 2019.
[34] J. Song, X. Chen, X. Yuan, and Z. Qin, "An Efficient Insider Threat Detection Scheme in Cloud Computing," in 2019 IEEE International Conference on Big Data (Big Data), pp. 576-581, 2019.
[35] M. A. Tahir, A. M. Abdalla, A. Al-Fuqaha, and S. H. Ahmed, "False Positive Reduction in Intrusion Detection Systems using a Combination of Decision Trees," in 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 1145-1150, 2019.
[36] Y. Li, J. X. Yu, C. Y. Zhang, and W. Sun, "Fast Detection of Cybersecurity Threats via Incremental Computation," IEEE Transactions on Dependable and Secure Computing, vol. 17, no. 2 , pp. 360-374, Mar. 2020.
[37] Z. Zheng, J. Chen, X. Hu, S. Chen, and M. W. Mutka, "Collaborative Anomaly Detection for IoT Security Using Federated Learning," in 2021 IEEE International Conference on Edge Computing (EDGE), pp. 187-194,), 2021.
[38] Z. S. Alawieh, R. Atani, and L. Alkotob, "Improving Network Resilience by Exploiting a Real Time DDoS Attack Detection Technique," in 2019 IEEE 4th Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp. 1-6. 2019
[39] V. Chandola, A. Banerjee, and V. Kumar, 'Anomaly detection: A survey', ACM Computing Surveys (CSUR), vol. 41, no. 3, 2009.
[40] S. Raymond and S. Sengupta, 'Deep Learning for IoT Intrusion Detection: A Survey', Sensors, vol. 21, no. 4, 2021.
[41] Y. Yu, X. Song, and X. Wang, 'Anomaly Detection and Prediction for IoT Data Streams: A Survey', IEEE Internet of Things Journal, vol. 7, no. 11, pp. 10856–10870, 2020.
[42] Y. Zheng, Q. Liu, E. Chen, J. Ge, and Z. Song, "Time Series Anomaly Detection: Algorithms, Applications, and Challenges," ACM Computing Surveys, vol. 53, no. 2, pp. 1-36, Mar. 2020.

[43] E. M. F. Skubch and J. Z. Kolter, "Variational Autoencoders for Anomalous Behavior Detection in Attributed Networks," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 737746, 2017.
[44] Y. Liu, G. Wang, J. Yang, and S. J. Maybank, "Towards good practice for braincomputer interface," in 2010 IEEE International Conference on Systems, Man and Cybernetics, pp. 3651-3658, 2010.
[45] L. Yin, S. Wang, and G. Gong, "Research on Network Intrusion Detection Method Based on Multi-Dimensional Data Analysis," in 2020 6th International Conference on Information Management (ICIM), pp. 277-282, 2020.
[46] M. A. Hassan, S. Iqbal, M. K. Khan, and A. K. Sangaiah, 'An Ensemble Deep Learning Approach for Intrusion Detection in IoT Networks', IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7170–7178, 2020.
[47] X. Zhang, C. Xie, Z. Li, and P. S. Yu, 'Ensemble deep learning for intrusion detection with heterogeneous unlabeled data', IEEE Transactions on Network Science and Engineering, vol. 7, no. 4, pp. 2589–2600, 2020.
[48] S. Oussama and T. Zouheir, 'An Ensemble Deep Learning Method Based on Stacking for Intrusion Detection Systems', in Proceedings of the 6th International Conference on Data Science, ACM, pp. 85–90, 2021.
[49] X. Wang, Z. Zhang, J. Zhang, X. Yang, and J. Ma, 'Ensemble deep learning for intrusion detection based on multi-feature fusion', Future Generation Computer Systems, vol. 121, pp. 402–413, 2021.
[50] Ouahman, A. Iqqou, H. Aboulhassan, and M. Rziza, 'Hybrid Ensemble of Deep Learning Classifiers for Intrusion Detection Systems', in Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, pp. 60–71, 2020.
[51] Chandra and P. N. Mahalle, 'An ensemble-based deep learning approach for intrusion detection using explainable AI', Journal of Information Security and Applications, vol. 60, 2021.
[52] Ghaffari, V. Balasubramanian, and V. Chang, 'A novel ensemble model using deep learning algorithms for intrusion detection systems', in Proceedings of the International Conference on Communication, Management and Information Technology, IEEE, pp. 283–288, 2020.
[53] X. Liu, L. Sun, X. Wang, and S. Yang, 'Federated ensemble deep learning for intrusion detection in IoT networks', IEEE Transactions on Industrial Informatics, vol. 17, no. 8, pp. 5537–5546, 2021.
[54] Y. Zhang and X. Gu, 'A deep decision tree-based network intrusion detection system', Future Generation Computer Systems, vol. 107, pp. 138–147, 2020.
[55] J. Zhou, Y. Qi, X. Xie, and X. Zhang, 'Intrusion detection method based on improved random forest algorithm', Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 2, pp. 1713–1721, 2021.
[56] X. Zhang, L. Xu, Y. Wang, and L. Huang, 'Intrusion detection model based on improved SVM algorithm', Security and Communication Networks, 2021.
[57] S. Sricharan and V. Tamarapalli, 'Deep learning based intrusion detection system using convolutional neural network with k-means clustering for network traffic classification', Computers & Security, 2021.
[58] N. H. Sarrar and F. Al-Turjman, 'An efficient distributed intrusion detection system based on k-nearest neighbors algorithm', IEEE Transactions on Industrial Informatics, vol. 18, no. 2, pp. 826–835, 2022.
[59] M. A. Almarri, M. K. Khan, and I. Ali, 'Hybrid deep learning ensemble model using naive Bayes for network intrusion detection system', Soft Computing, vol. 26, no. 5, pp. 3997–4013, 2022.
[60] H. Zhang, Y. Han, Q. Xie, W. Yang, and X. Li, 'Efficient ensemble learning algorithm for intrusion detection based on Adaptive Gradient Boosting', Computers & Electrical Engineering, vol. 100, 2022.
[61] Y. Alotaibi and M. Ilyas, 'Ensemble-learning framework for intrusion detection to enhance Internet of Things' devices security', Sensors (Basel), vol. 23, no. 12, Jun. 2023.
[62] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, 'Practical black-box attacks against machine learning', in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi United Arab Emirates, 2017.
[63] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 'SMOTE: Synthetic minority over-sampling technique', Journal of artificial intelligence research, 16, 321-357., 2002
[64] Q. Yang, Y. Liu, T. Chen, and Y. Tong, 'Federated Machine Learning: Concept and Applications', arXiv [cs.AI], 13-Feb-2019.
[65] S. Han, H. Mao, and W. J. Dally, 'Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding', arXiv [cs.CV], 01-Oct-2015
[66] T. Le and D. T. Nguyen, 'Intrusion detection using an ensemble deep learning approach with feature engineering and data balancing', Computers & Security, vol. 109, 2022.

[67] S. Barua, M. S. Rahman, and R. Islam, 'Ensemble intrusion detection system with adaptive boosting and random oversampling', Computers & Electrical Engineering, vol. 98, 2021.
[68] F. Tsai, C. F. Lai, and Y. M. Hsueh, 'Intrusion detection using stacking ensemble with feature engineering and ADASYN', Journal of Network and Computer Applications, vol. 190, 2022.
[69] J. Jiang, C. Cui, S. Yang, and Y. Zhang, 'Boosting ensemble machine learning approach for intrusion detection with performance evaluation of different data balancing techniques', Cluster Computing, pp. 1–16, 2022.
[70] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, 'A detailed analysis of the KDD CUP 99 data set', in 2009 IEEE Symposium on Computational Intelligence for Security and Defence Applications, Ottawa, ON, Canada, 2009.
[71] N. Moustafa, J. Slay, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, 'The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset', Proceedings of the 4th International Conference on, vol. 24, pp. 18–31, 2015.
[72] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on
[73] R. P. Lippmann et al., 'Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation', in Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00, Hilton Head, SC, USA, 2002.
[74] Gharib, M., Moattar, M. H., & Alqhtani, M. A. (2019). A comprehensive analysis of feature extraction techniques for the ISCX NSL-KDD intrusion detection dataset. PeerJ Computer Science, 5, e204.
[75] K. Kiryu, Y. Akira, and O. Yoshihiro, 'The Kyoto University network captures: A large-scale academic network traffic dataset for data-driven analysis', in Proceedings of the International Conference on Traffic Monitoring and Analysis, Springer, 2006, pp. 223–240.
[76] Dainotti, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and S. Antonio, 'Analysis of internet background radiation in Colombia', ACM SIGCOMM Computer Communication Review, vol. 42, no. 4, pp. 43–48, 2012.
[77] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.
[78] J. Davis and M. Goadrich, 'The relationship between Precision-Recall and ROC curves', in Proceedings of the 23rd international conference on Machine learning - ICML '06, Pittsburgh, Pennsylvania, 2006.
[79] M. Powers, 'Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation', Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.
[80] F. Provost and T. Fawcett, 'Robust classification for imprecise environments', arXiv [cs.LG], 13-Sep-2000.
[81] Provost, T. Fawcett, and R. Kohavi, 'The Case Against Accuracy Estimation for Comparing Induction Algorithms', in Proceedings of the 15th International Conference on Machine Learning (ICML), pp. 445–453, 1998.
[82] T. Saito and M. Rehmsmeier, 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', PLoS One, vol. 10, no. 3, p. e0118432, Mar. 2015.
[83] N. Moustafa and J. Slay, 'The significant features of the UNSWNB15 and the KDD99 data sets for network intrusion detection systems', in 2015 4th international workshop on building analysis datasets and gathering experience returns for security (BADGERS), IEEE, pp. 25–31, 2015.
[84] N. Moustafa and J. Slay, Eds., 'The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set', Information Security Journal: A Global Perspective, vol. 25, 1831.
[85] N. Koroniotis, N. Moustafa, E. Sitnikova, and J. Slay, 'Towards developing network forensic mechanism for botnet activities in the IoT based on machine learning techniques', in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Cham: Springer International Publishing, pp. 30–44, 2018.
[86] S. Meftah, T. Rachidi, and N. Assem, 'Network based intrusion detection using the UNSW-NB15 dataset', International Journal of Computing and Digital Systems, vol. 8, no. 5, pp. 477–487, 2019.
[87] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, 'An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset', Cluster Comput., vol. 23, no. 2, pp. 1397–1418, Jun. 2020.
[88] S. M. Kasongo and Y. Sun, 'Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset', J. Big Data, vol. 7, no. 1, Dec. 2020.

[89] V. Kumar, A. K. Das, and D. Sinha, 'UIDS: a unified intrusion detection system for IoT environment', Evol. Intell., vol. 14, no. 1, pp. 47–59, Mar. 2021.
[90] T. Nguyen, N. Nguyen, and T. Nguyen, 'Deep Learning-based Network Intrusion Detection System using Autoencoders and Deep Neural Networks', in Proceedings of the International Conference on Advanced Computational Intelligence (ICACI), 2021.
[91] T. T. Nguyen, G. H. Nguyen, and T. Q. Phan, 'Intrusion Detection Systems Using Deep Learning: A Review and Comparative Analysis', Symmetry, vol. 12, no. 6, 2020.
[92] S. Al-Muhtadi and S. D. Wolthusen, 'A Survey of Network-Based Intrusion Detection Data Sets', ACM Computing Surveys (CSUR), vol. 50, no. 3, pp. 1–36, 2017.
[93] Butun, S. Coleri Ergen, and A. Levi, 'A Survey of Common Intrusion Detection Techniques', International Journal of Computer Applications, vol. 137, no. 1, pp. 8–17, 2015.
[94] Y. Cao, Z. Wang, H. Ding, J. Zhang, and B. Li, 'An intrusion detection system based on stacked ensemble learning for IoT network', Comput. Electr. Eng., vol. 110, no. 108836, p. 108836, Sep. 2023.
[95] P., Ananthi., R., Janani. Ensemble based Intrusion Detection System for IoT Device. 1073-1078, 2023. doi: 10.1109/ICSCSS57650.2023.10169426
[96] M. Bhavsar., K. Roy., J. Kelly. Anomaly-based intrusion detection system for IoT application. Discover Internet of things, 3(1), 2023 doi: 10.1007/s43926-023-00034-5
[97] F. Kilinçer and O. Katar, 'A new Intrusion Detection System for Secured IoT/IIoT Networks based on LGBM', Gazi Üniv. Fen Bilim. Derg. C Tasar. ve Teknol., vol. 11, no. 2, pp. 321–328, Jun. 2023.
[98] S. Aamir and M. Faheem, 'Intrusion Detection System for IoT Environment using Ensemble Approaches', pp. 935–938, 2023.
[99] R. Latha and R. M. Bommi, 'Hybrid CatBoost Regression model based Intrusion Detection System in IoT-Enabled Networks', in 2023 9th International Conference on Electrical Energy Systems (ICEES), Chennai, India, 2023.
[100] Y. Al Sawafi, A. Touzene, and R. Hedjam, 'Hybrid deep learning-Based Intrusion Detection System for RPL IoT networks', J. Sens. Actuator Netw., vol. 12, no. 2, p. 21, Mar. 2023.
[101] A. Awajan, 'A novel Deep Learning-based intrusion detection system for IoT networks', Computers, vol. 12, no. 2, p. 34, Feb. 2023.
[102] P. Lakhotia, R. Dwivedi, D. K. Sharma, and N. Sharma, 'Intrusion Detection System for IoE-based medical networks', J. Database Manag., vol. 34, no. 2, pp. 1–18, Apr. 2023.
[103] N. F. Syed, M. Ge, and Z. Baig, 'Fog-cloud based intrusion detection system using Recurrent Neural Networks and feature selection for IoT networks', Comput. Netw., vol. 225, no. 109662, p. 109662, Apr. 2023.
[104] https://www.unb.ca/cic/datasets/iotdataset-2023.html
[105] Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. Sensors 23, 5941, 2023
[106] https://www.kaggle.com/datasets/kaggleprollc/nsl-kdd99-dataset
[107] https://www.kaggle.com/datasets/cicdataset/cicids2017