Vol **04** | Issue **03** | **September** 2024
ISSN **2997-9870**
Page:**178-194**

Academic Journal on Science, Technology, Engineering & Mathematics Education

RESEARCH ARTICLE

OPEN ACCESS

A REVIEW OF MACHINE LEARNING AND FEATURE SELECTION TECHNIQUES FOR CYBERSECURITY ATTACK DETECTION WITH A FOCUS ON DDOS ATTACKS

[1] Ms Roopesh, [2]Nourin Nisha, [3]Sasank Rasetti, [4]Md Atiqur Rahaman

[1] *Master of Science in Department of Electrical Engineering, Lamar University, Texas, USA*
Email: muniroopeshraasetti@gmail.com

[2]*Master of Science in Management Information Systems, College of Business, Lamar University, Texas, USA*
Gmail: nishatnitu203@gmail.com

[3]*Master of Science in Department of Electrical Engineering, Lamar University, Texas, USA*
Email: srasetti@lamar.edu

[4]*Department of Management and Information Technology, St. Francis College, New York, USA*
Email: mrahaman4@sfc.edu

ABSTRACT

This study provides a systematic review of machine learning (ML) techniques applied in intrusion detection systems (IDS), with a particular focus on Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). Following the PRISMA guidelines, a comprehensive search of relevant databases identified 205 articles, from which 68 were selected for detailed analysis. The findings highlight that RF consistently outperforms other models, achieving accuracy rates as high as 99.72% in detecting Distributed Denial of Service (DDoS) attacks due to its ensemble learning approach. SVM, while effective in specific scenarios with binary classification tasks, struggles with scalability and high-dimensional datasets, though feature selection significantly improves its performance. DT models, known for their simplicity and interpretability, are prone to overfitting, but this issue is mitigated when combined with feature selection techniques. The study further emphasizes the importance of feature selection in enhancing IDS accuracy and efficiency across various models. Additionally, ensemble and hybrid methods, which combine multiple ML techniques, offer promising improvements in detection accuracy and real-time performance. These findings underscore the potential of machine learning, particularly through the use of ensemble and hybrid approaches, to significantly improve cybersecurity measures in modern networks.

KEYWORDS

Cybersecurity, Intrusion Detection, Machine Learning, DDoS Attacks, Feature Selection Techniques

Submitted: August 12, 2024
Accepted: September 20, 2024
Published: September 22, 2024

Corresponding Author:

Ms Roopesh

Master of Science in Department of Electrical Engineering, Lamar University, Texas, USA

email: muniroopeshraasetti@gmail.com

10.69593/ajsteme.v4i03.105

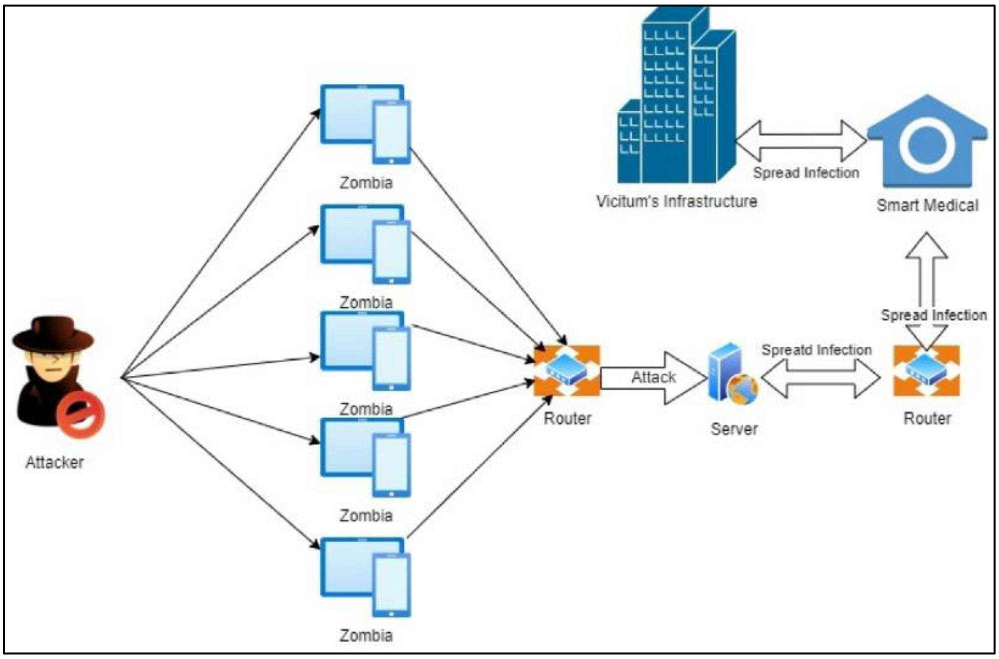
1 Introduction

Cybersecurity has become an increasingly critical concern in the digital age, as the growing dependence on internet-based systems and the proliferation of Internet of Things (IoT) devices expose individuals, organizations, and governments to cyber-attacks (Ngo et al., 2023). Cyber-attacks, such as Distributed Denial of Service (DDoS) attacks, can cause significant damage, leading to service disruptions and financial losses. According to Shrestha et al. (2020), the advent of sophisticated cyber-attacks has outpaced traditional security measures, making it essential to develop advanced tools to detect and mitigate these threats. In this context, Intrusion Detection Systems (IDS) have been the backbone of cybersecurity efforts, aiming to detect unauthorized access and potential security breaches in real-time (Ma et al., 2020). However, traditional IDS systems often struggle to cope with the increasing complexity and volume of attacks, particularly in the face of distributed and large-scale attacks like DDoS (Mell et al., 2022). Thus, the use of Machine Learning (ML) techniques has emerged as a promising approach to improve the detection capabilities of IDS, as these techniques can analyze vast amounts of data and identify hidden patterns indicative of malicious activities (Zhang et al., 2019).

Machine Learning (ML) techniques have demonstrated their potential in various fields, including healthcare, finance, and cybersecurity, where they excel in analyzing large datasets and making predictions based on historical data (D'Angelo & Palmieri, 2021). In cybersecurity, ML techniques are employed to detect network intrusions by learning from previous attack patterns and identifying anomalies in network traffic (Srivastava et al., 2013). Several ML algorithms, such as Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT), have been widely adopted in IDS to classify network traffic as either benign or malicious. These techniques offer varying degrees of accuracy and performance, depending on the nature of the attack and the features used in the classification process (Zhang et al., 2019). For instance, RF has gained popularity due to its high accuracy in intrusion detection and ability to handle large datasets efficiently (Tan et al., 2010). Furthermore, ML techniques are increasingly being combined with feature selection methods to enhance detection accuracy by reducing dimensionality and focusing on the most relevant features (Chaabouni et al., 2019).

Feature selection plays a crucial role in improving the performance of ML-based IDS. It involves selecting the

Figure 1: Distributed denial-of-service (DDoS) Attack.



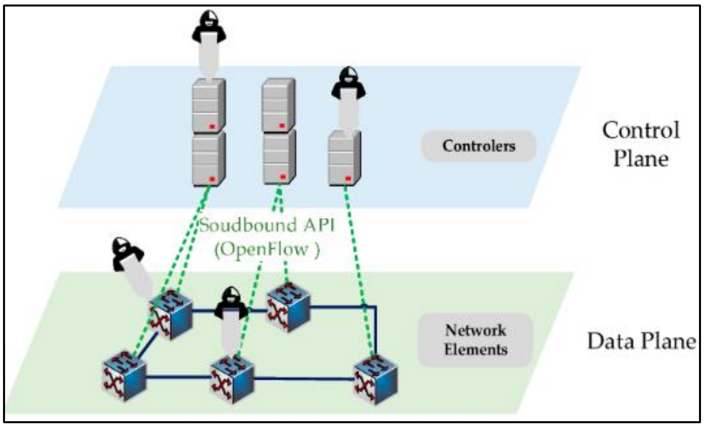
Source: Ullah et al. (2023)

most critical features from a dataset that have the greatest impact on classification accuracy, which helps reduce computational complexity and improve the speed of detection (Shrestha et al., 2020). According to Chaabouni et al. (2019), feature selection can significantly enhance the performance of ML algorithms in detecting DDoS attacks, as it eliminates irrelevant or redundant features that may introduce noise into the model. The Decision Tree (DT) technique is particularly effective for feature selection, as it ranks the importance of each feature based on its contribution to the classification outcome (D'Angelo & Palmieri, 2021). By identifying and prioritizing the most influential features, ML models can achieve higher accuracy in detecting cyber-attacks. For instance, the Random Forest algorithm, which is an ensemble learning method, benefits from feature selection techniques, as it constructs multiple decision trees using randomly selected subsets of features, leading to more robust predictions (Chen et al., 2018).

Despite the promising performance of ML algorithms in intrusion detection, several challenges remain in implementing these techniques in real-world scenarios. One of the primary challenges is the selection of the appropriate ML algorithm for a given context, as different algorithms exhibit varying strengths and weaknesses depending on the attack type and network environment (Usman et al., 2019). For instance, while RF and DT techniques tend to perform well in detecting a broad range of cyber-attacks, algorithms such as SVM and KNN may struggle with complex attacks involving large-scale DDoS traffic (Manimurugan et al., 2020). Furthermore, the quality of the dataset used for training the model plays a significant role in the model's performance. Datasets with missing, redundant, or irrelevant data can reduce the accuracy of ML models, underscoring the importance of effective data preprocessing and feature selection (Sargolzaei et al., 2020). As a result, researchers have emphasized the need for comprehensive data preprocessing steps, such as normalization and encoding, to improve the robustness and accuracy of ML-based IDS (Rawat & Bajracharya, 2015).

Recent studies have demonstrated the efficacy of using ensemble methods, which combine multiple ML models to improve detection accuracy and reduce false positives (Chen et al., 2015). Ensemble techniques, such as bagging and boosting, have been particularly effective in addressing the limitations of individual classifiers by aggregating their predictions and making a final decision based on the majority vote (Nhat-Duc & Van-Duc, 2023). For instance, the ensemble approach of combining Random Forest with SVM has been shown to achieve higher accuracy in detecting DDoS attacks than either model alone (Zhang et al., 2011). Additionally, hybrid models that integrate deep learning techniques, such as Convolutional Neural Networks (CNN), with traditional ML models have also been explored to further enhance the detection capabilities of IDS (Sargolzaei et al., 2020). As ML techniques continue to evolve, their application in cybersecurity is likely to expand, providing more robust and adaptive solutions to emerging threats.

Figure 2: Main targets of the Distributed Denial of Service



The primary objective of this review is to provide a comprehensive analysis of the role of machine learning (ML) and feature selection techniques in enhancing cybersecurity attack detection, particularly in the context of Distributed Denial of Service (DDoS) attacks. This review seeks to compare the effectiveness of various ML algorithms, such as Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT), in detecting malicious network traffic. Additionally, it aims to examine how feature selection methods can improve the accuracy and efficiency of these ML models by identifying the most relevant features for classification. By synthesizing findings from recent studies, this review endeavors to highlight the strengths and limitations of each approach, explore the integration of ensemble and hybrid models, and provide insights into future research directions in developing more robust and adaptive Intrusion

Detection Systems (IDS).

2 Literature Review

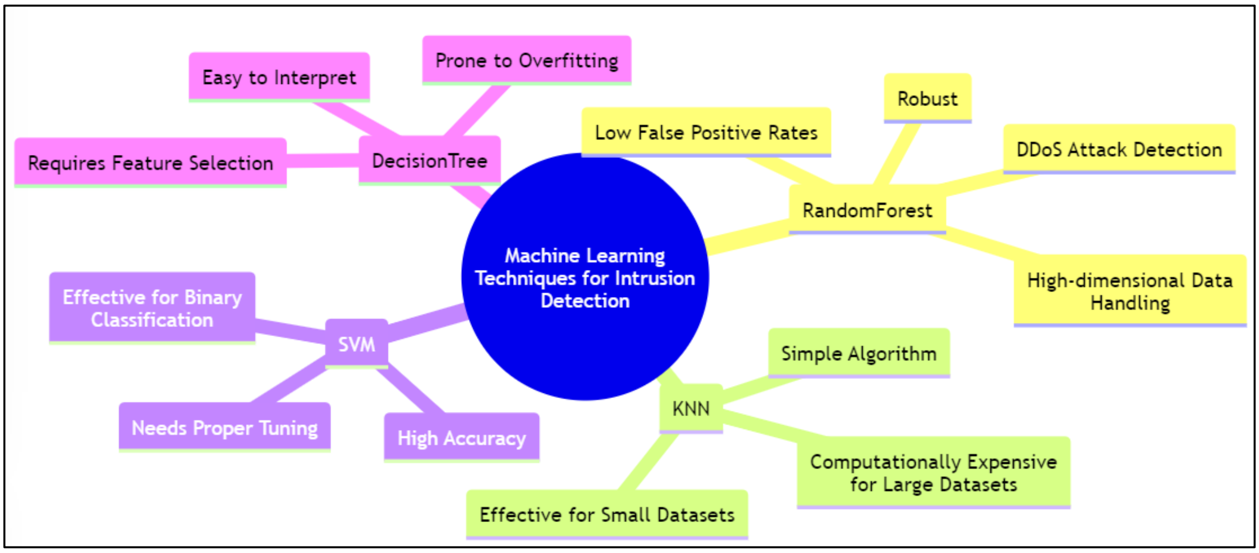
This section presents an overview of the existing research on the use of machine learning (ML) techniques and feature selection methods for cybersecurity attack detection, with a focus on Distributed Denial of Service (DDoS) attacks. As cyber-attacks grow in complexity, traditional security systems have struggled to keep pace, prompting the need for advanced ML-based Intrusion Detection Systems (IDS). The literature review synthesizes key studies that explore various ML algorithms and their effectiveness in identifying malicious network activities. Additionally, this section highlights the role of feature selection in enhancing the accuracy of these models by reducing data complexity and improving classification outcomes. The review aims to provide a comprehensive understanding of the current state of research, identify gaps, and outline potential advancements in the field of cybersecurity.

2.1 Machine Learning Techniques for Intrusion Detection

2.1.1 Random Forest (RF)

Random Forest (RF) is a widely utilized ensemble learning technique that builds multiple decision trees during training and combines their outputs to enhance classification accuracy. It has become increasingly favored in intrusion detection systems (IDS) due to its robustness, ability to manage high-dimensional data, and effectiveness in detecting both known and unknown cyber-attacks (Acosta et al., 2020; Derhab et al., 2019). RF's architecture, which integrates the predictions of several decision trees based on random subsets of features, enables it to capture various facets of network traffic data, making it particularly adept at identifying complex cyber-attacks such as Distributed Denial of Service (DDoS) attacks. As Zhang et al. (2019) explain, the ensemble approach of RF mitigates the issue of overfitting, a common challenge in machine learning (ML)-based IDS, while delivering high accuracy and low false-positive rates. This makes RF a highly reliable model in diverse cybersecurity applications, especially when addressing large datasets that encompass a wide spectrum of attack patterns (Cuadra et al., 2017). Several studies have illustrated RF's effectiveness in detecting DDoS attacks. For example, Ding et al. (2014) demonstrated that RF outperformed other ML algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (LR) when applied to the CIC-IDS dataset, achieving an impressive accuracy of 96.5%. Similarly, Das et al. (2019) evaluated RF against conventional IDS models and found that RF achieved superior performance in detecting DDoS attacks within the NSL-KDD dataset, with an accuracy of 99.72%. Moreover, Ahmed et al. (2019) emphasized RF's capability to handle imbalanced data, a frequent issue in intrusion detection, which enables it to maintain high detection rates even in environments dominated by benign traffic. This ability underscores RF's utility in mitigating cybersecurity threats, particularly in high-traffic network settings where both precision and speed are

Figure 3: Machine Learning Techniques for Intrusion Detection


Easy to Interpret
Prone to Overfitting
Requires Feature Selection
DecisionTree
Robust
Low False Positive Rates
DDoS Attack Detection
RandomForest
Machine Learning Techniques for Intrusion Detection
High-dimensional Data Handling
Effective for Binary Classification
SVM
Simple Algorithm
KNN
Needs Proper Tuning
High Accuracy
Computationally Expensive for Large Datasets
Effective for Small Datasets

crucial. When compared to other ML techniques, RF consistently demonstrates superior performance. Ding et al. (2014) note that RF's accuracy in detecting DDoS attacks surpasses that of SVM, which often struggles with high-dimensional data, and KNN, which becomes computationally intensive for large datasets. Additionally, Ahmed et al. (2019) observed that while decision trees (DT) are effective for classification tasks, their standalone use is prone to overfitting, making RF's ensemble approach a more stable and generalizable solution. Furthermore, Ishizaki et al. (2018) highlighted that RF's ability to perform feature selection during training enhances its detection capabilities, allowing it to outperform simpler models like KNN and DT. In summary, RF's combination of high accuracy, low computational cost, and effective handling of large and complex datasets establishes it as one of the most reliable techniques for DDoS attack detection in IDS (Ishizaki et al., 2018; Samuel et al., 2020).

2.1.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, yet powerful, machine learning algorithm used for both classification and regression tasks. In the context of cybersecurity, KNN is employed for its ease of implementation and effectiveness in classifying network traffic as either benign or malicious (Hong et al., 2020). The algorithm works by calculating the distance between a data point and its 'K' nearest neighbors, assigning the most common class among the neighbors to the data point. KNN has shown utility in intrusion detection systems (IDS) because it can model complex relationships between network features without needing a parametric model (Pajouh et al., 2015). However, the effectiveness of KNN in cybersecurity largely depends on the value of 'K' and the choice of distance metric, which significantly influences its accuracy and computation time (Deng et al., 2010). Despite being an intuitive and non-parametric algorithm, KNN struggles with high-dimensional data and large datasets, where computational complexity increases significantly (Eckelt et al., 2023). This is a notable limitation when dealing with large-scale cyber-attacks like Distributed Denial of Service (DDoS) attacks, which generate massive amounts of traffic (Khan et al., 2019). Several studies have explored the performance of KNN in detecting cyber-attacks, particularly in IDS. Lee et al. (2018) tested KNN on the CIC-IDS dataset and found that while it performed reasonably well, with an accuracy of 90.4%, it was outperformed by other algorithms such as Random Forest (RF) and Support Vector Machine (SVM) in terms of both accuracy and efficiency. Al-Sahaf et al. (2019) highlighted similar results when comparing KNN with decision trees (DT) and logistic regression (LR), noting that KNN's performance is hindered by its sensitivity to high-dimensional data, which is common in IDS datasets. Additionally, Tufail et al. (2021) observed that KNN's computational cost increases with the size of the dataset, making it less suitable for real-time intrusion detection in large networks. However, when used in smaller datasets or in combination with dimensionality reduction techniques like Principal Component Analysis (PCA), KNN can still be an effective tool for detecting certain types of attacks (Mitchell & Chen, 2014). While KNN offers simplicity and interpretability, its limitations in handling large and complex datasets necessitate careful consideration in cybersecurity applications (Al-Sahaf et al., 2019; Verbraeken et al., 2020).

2.1.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm widely used for both classification and regression tasks, particularly in network security. SVM operates by identifying the optimal hyperplane that separates data into distinct classes, maximizing the margin between groups (Hong et al., 2020). In cybersecurity, SVM is frequently employed in intrusion detection systems (IDS) due to its ability to handle high-dimensional datasets and its effectiveness in binary classification tasks, such as distinguishing between malicious and benign network traffic (Lee et al., 2018). It is particularly well-suited for cases where classes are linearly separable, making it an effective tool for identifying common cyber-attacks, such as Distributed Denial of Service (DDoS) and malware attacks, as it focuses on minimizing misclassification (Jim et al., 2024; Uzzaman et al., 2024). However, while SVM performs well in controlled environments, its efficiency tends to diminish when dealing with highly imbalanced datasets, which are common in real-world network traffic scenarios (Al-Sahaf et al., 2019). Numerous studies have demonstrated SVM's accuracy in various attack detection scenarios. For instance, Lee et al. (2018)

reported that SVM achieved an accuracy of 90.4% on the CIC-IDS dataset, outperforming K-Nearest Neighbors (KNN) in detecting network-based attacks. Similarly, Hu et al. (2018) compared SVM with other machine learning algorithms, such as Random Forest (RF) and Decision Tree (DT), for detecting DDoS attacks on the NSL-KDD dataset. Although SVM achieved a high accuracy of 98.54%, RF outperformed it due to its superior handling of large datasets. Ishizaki et al. (2018) also noted that while SVM can be highly accurate, its performance often declines in large, non-linear datasets where kernel functions and parameter tuning are necessary to maintain accuracy. Despite these limitations, SVM remains a reliable tool for intrusion detection when combined with appropriate preprocessing and feature selection techniques (Pajouh et al., 2019). Comparisons with ensemble models further highlight SVM's strengths and weaknesses. While SVM performs well on its own, several studies suggest that combining it with other models, such as Random Forest or ensemble techniques, can significantly improve detection accuracy (Pajouh et al., 2015). Deng et al. (2010) showed that an ensemble approach combining SVM with a deep neural network (DNN) outperformed standalone SVM in detecting DDoS attacks, achieving higher accuracy and a lower false-positive rate. Similarly, Ahmed et al. (2019) found that ensemble methods, which aggregate the predictions of multiple classifiers, mitigate the limitations of individual algorithms like SVM by providing a more robust defense against varied attack patterns. Techniques such as bagging and boosting, when combined with SVM, offer better generalization and improved performance in real-world scenarios, making them highly suitable for large-scale intrusion detection systems (Zhou et al., 2021). This trend underscores the growing importance of hybrid models that integrate SVM with other machine learning techniques to enhance cybersecurity measures. Moreover, the dual form of SVM optimization is used for non-linear SVM, especially with the kernel trick. The dual form maximizes the following function:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

Subject to:

$$0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Where:

- α_i are the Lagrange multipliers.

2.1.4 Decision Tree (DT)

Decision Tree (DT) is a widely utilized supervised learning algorithm that classifies data by splitting it into branches based on feature values, creating an intuitive, tree-like structure that makes it both accessible and interpretable, particularly for intrusion detection systems (IDS). In cybersecurity, DT is frequently applied to classify network traffic as either benign or malicious due to its capacity to handle both categorical and continuous data, excelling in situations where clear decision rules can be established from available features (Ozay et al., 2013). The DT algorithm works by recursively partitioning the dataset according to the most significant features, which helps to identify patterns associated with cyber-attacks, such as Distributed Denial of Service (DDoS) attacks (Schmidt et al., 2016). One of the major advantages of DT is its ease of interpretation, making it particularly valuable in cybersecurity, where understanding the reasoning behind classification decisions is critical for building trust and facilitating response strategies (K. Wang et al., 2020; Shamim, 2022). However, a key limitation of DT is its susceptibility to overfitting, especially when the tree grows too complex, which can limit its generalizability to new data (Ozay et al., 2013). Studies have shown that integrating feature selection techniques with DT can significantly enhance its performance in detecting cyber-attacks (Higgins et al., 2021; K. Wang et al., 2020; Zhang et al., 2021). For example, Hu et al. (2018) highlighted that feature selection, by reducing the dimensionality of network data, improves both the accuracy and efficiency of DT-based intrusion detection systems. By focusing on the most relevant features, DT models can isolate the critical aspects of network traffic that differentiate between benign and malicious activity, reducing noise and enhancing prediction accuracy. Ozay et al. (2013) further emphasized that feature selection helps prune irrelevant or redundant features, thus preventing overfitting, which is a common challenge in high-dimensional datasets such as those used in IDS. Fu (2022) also demonstrated that a DT model with integrated feature selection could significantly improve DDoS detection accuracy and speed compared to models without feature selection, underscoring its importance in optimizing DT for cybersecurity applications. Nevertheless, when compared to other machine learning models, DT often performs competitively but can be outperformed by more complex ensemble methods, particularly in large-scale intrusion detection. Md Abdur et al. (2024) compared DT with Random Forest (RF) and Support Vector Machine (SVM) for detecting DDoS attacks, finding that although DT delivered solid results, RF outperformed it due to its ensemble approach, which mitigates overfitting by averaging multiple decision trees. Similarly, Hu et al. (2018) noted that DT's performance in large datasets is often inferior to ensemble methods like RF or boosting techniques, which enhance model stability and reduce overfitting risks. Nonetheless, in scenarios where model interpretability is crucial, DT remains a preferred choice due to its straightforward decision-making process (Ozay et al., 2013). Overall, while DT offers simplicity and transparency in intrusion detection, combining it with feature selection and ensemble techniques can further optimize its performance, making it a valuable tool in the cybersecurity landscape (Nhat-Duc & Van-Duc, 2023). Moreover, Gini impurity is a criterion used in Decision Trees to evaluate the quality of a split. It measures the likelihood of a random sample being incorrectly classified if it was randomly labeled according to the distribution of class labels in a subset.

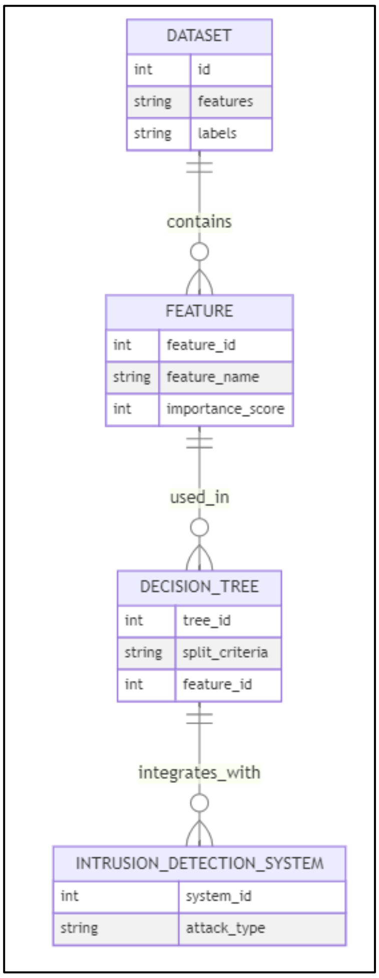
$$Gini(p) = 1 - \sum_{i=1}^{n} p_i^2$$

2.2 Decision Tree-Based Feature Selection

Decision Trees (DT) play a pivotal role in feature selection by providing a natural mechanism for ranking features based on their importance in making classification decisions. The decision-making process of DT involves recursively splitting the dataset according to the feature that offers the highest information gain or the lowest Gini impurity, thereby ranking features by their contribution to the classification task (Dao & Lee, 2022). This inherent feature-ranking capability makes DT a powerful tool for selecting the most relevant features in large datasets, particularly in the context of intrusion detection systems (IDS), where high-dimensional data can negatively impact performance (Schmidt et al., 2016). Once features are ranked, less important or redundant features can be pruned, allowing the model to focus on the most critical variables, which not only improves classification accuracy but also reduces computational complexity (Mitchell & Chen, 2014). As a result, DT-based feature selection has gained popularity in cybersecurity applications, where it aids in the identification of key indicators of malicious activity, such as specific network behaviors or traffic patterns (Al-Garadi et al., 2020).

Several case studies have demonstrated significant improvements in model performance when feature selection is integrated with Decision Tree-based models. For example, (Higgins et al., 2021)explored the use of DT-based feature selection in detecting Distributed Denial of Service (DDoS) attacks and reported a noticeable increase in accuracy and efficiency after removing irrelevant features. Similarly, Al-Garadi et al. (2020) compared the performance of intrusion detection systems with and without DT-based

Figure 4: Entity-Relationship Diagram



feature selection and found that models incorporating feature selection achieved superior results, reducing false positives while maintaining high detection rates. In another study, Tufail et al. (2021) highlighted that feature selection with DT enhanced the detection of various types of cyber-attacks by improving the model's ability to generalize across different datasets. The reduced dimensionality not only increased the speed of detection but also improved the interpretability of the results, making it easier for security analysts to identify and respond to specific threats (Lee et al., 2018). These findings underscore the effectiveness of DT-based feature selection in enhancing the performance of machine learning models in cybersecurity applications.

2.3 Other Feature Selection Methods

Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are widely used feature selection techniques in intrusion detection systems (IDS) to address the challenges of high-dimensional datasets and improve model performance. PCA is a statistical technique that transforms the original dataset into a smaller set of orthogonal components, capturing the most variance in the data while reducing its dimensionality (Khan et al., 2019). This method is particularly effective in situations where multicollinearity between features exists, as it eliminates redundant information while preserving the essential features needed for accurate classification (Pajouh et al., 2019). On the other hand, RFE is an iterative method that works by recursively removing the least significant features and building the model repeatedly (Islam, 2024; Islam & Apu, 2024; Maraj et al., 2024), identifying the optimal set of features that contribute the most to the prediction task (Lee et al., 2018). RFE is often applied with algorithms like Random Forest or Support Vector Machine (SVM) to fine-tune feature selection and improve model precision in IDS (Deng et al., 2010). Both PCA and RFE have gained traction in cybersecurity applications as they help streamline the feature selection process, reducing computational overhead and enhancing model interpretability.

Several studies have performed comparative analyses of feature selection techniques like PCA, RFE, and others in the context of IDS (Deng et al., 2010; Ishizaki et al., 2018; Verbraeken et al., 2020). Hong et al. (2020) compared PCA and RFE for detecting Distributed Denial of Service (DDoS) attacks and found that PCA performed better in reducing dimensionality for large datasets, improving the speed of detection while maintaining accuracy. However, RFE provided superior results when integrated with ensemble methods like Random Forest, as it fine-tuned feature selection by focusing on the most relevant features and discarding irrelevant ones (Al-Sahaf et al., 2019). Similarly, Samuel et al. (2020) demonstrated that combining RFE with machine learning algorithms like SVM significantly improved the accuracy of IDS by selecting the most influential features, while PCA was more suitable for datasets with multicollinear variables. A study by Verbraeken et al. (2020) further emphasized that while PCA excels in scenarios requiring rapid dimensionality reduction, RFE's iterative process allows for more granular control, making it a better fit for datasets where subtle patterns need to be detected. These comparisons highlight that the effectiveness of feature selection methods varies depending on the dataset and the intrusion detection task, suggesting that hybrid approaches combining multiple techniques can offer the most robust solutions.

2.4 Hybrid and Ensemble Methods

The rise of hybrid models combining Machine Learning (ML) and Deep Learning (DL) has shown significant promise in improving the accuracy and efficiency of intrusion detection systems (IDS). Hybrid models leverage the strengths of both ML and DL techniques, allowing for more sophisticated detection of complex cyber-attacks such as Distributed Denial of Service (DDoS) and Advanced Persistent Threats (APT) (Al-Sahaf et al., 2019). For example, ML algorithms like Random Forest (RF) or Support Vector Machine (SVM) are used for feature extraction and selection, while deep learning models like Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) handle the classification of network traffic (Lee et al., 2018). The hybridization of these approaches allows for improved detection of both known and unknown threats by capturing a wider range of features and behaviors. According to Al-Sahaf et al. (2019), DL models excel at automatically learning intricate patterns in data, especially when combined with ML-based feature selection, making hybrid models more adaptable to real-time and large-scale environments.

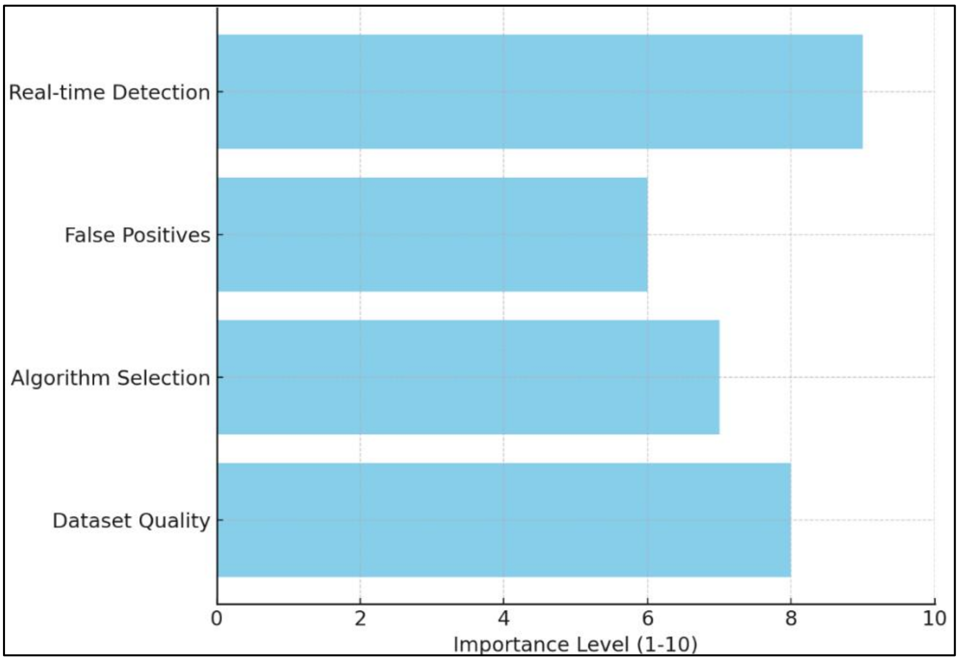
Studies have increasingly explored the benefits of ensemble techniques, which combine multiple classifiers to enhance detection performance. One common approach is to combine Random Forest (RF) with Support Vector Machine (SVM), leveraging RF's ability to handle high-dimensional data and SVM's precision in binary classification tasks (Samuel et al., 2020). Lee et al. (2018) demonstrated that ensemble models that aggregate predictions from RF, SVM, and Decision Trees (DT) offer significantly higher accuracy and lower false-positive rates than standalone models. In particular, an ensemble method combining RF with SVM for intrusion detection achieved an accuracy of over 99%, outperforming individual models that struggled with either dimensionality (RF) or processing time (SVM) (Pajouh et al., 2015). Additionally, Appasani and Mohanta (2018) emphasized that ensemble approaches like bagging and boosting further enhance the robustness of intrusion detection systems by reducing the variance and bias of individual classifiers, leading to improved generalization and higher detection rates. Overall, hybrid and ensemble methods have become critical in advancing the state of IDS by combining the strengths of various models and mitigating their weaknesses.

2.5 Challenges in ML-Based Intrusion Detection Systems

One of the primary challenges in Machine Learning (ML)-based Intrusion Detection Systems (IDS) is the quality of the datasets used for training models. Many intrusion detection datasets suffer from issues such as imbalance, where benign traffic significantly outweighs malicious traffic, leading to skewed model performance (Ding et al., 2014). This imbalance can result in models that are biased toward predicting benign activity, thus increasing the likelihood of missing actual threats (Kumar et al., 2022). Moreover, many available datasets contain redundant or irrelevant features, which can introduce noise into the model and hinder its ability to generalize to real-world scenarios (Gumaei et al., 2020). Datasets like KDD'99 and NSL-KDD, although widely used, have been criticized for containing outdated attack patterns that do not reflect modern cyber threats (Kumar et al., 2022; Shamim, 2024). Additionally, the presence of mislabeled or incomplete data can further reduce the accuracy of ML models in detecting sophisticated attacks, emphasizing the need for more robust, up-to-date datasets that accurately represent current cyber-attack trends (Jian et al., 2018). Therefore, improving data quality through rigorous preprocessing, feature selection, and synthetic data generation is crucial for enhancing the effectiveness of ML-based IDS.

Selecting the appropriate ML algorithm for different types of attacks poses another significant challenge in the development of IDS. No single algorithm is universally effective against all kinds of cyber-attacks, as each attack type can present distinct characteristics that require different approaches for detection (Faheem

Figure 5: Challenges in ML-Based Intrusion Detection Systems



et al., 2018). For example, while Random Forest (RF) is effective in handling high-dimensional datasets and detecting a broad range of attacks, algorithms like Support Vector Machine (SVM) may struggle with large-scale traffic but excel in binary classification tasks (Zhang et al., 2021). K. Wang et al. (2020) observed that even highly effective ML models often struggle with high false-positive rates, particularly in real-time detection environments, where distinguishing between benign and malicious activity is crucial. False positives can overwhelm security analysts and diminish trust in the system. Furthermore, real-time detection requires models that are not only accurate but also computationally efficient, as delays in detecting attacks can result in significant damage (Naz et al., 2019). Addressing these issues requires a careful balance of accuracy, efficiency, and adaptability, often achieved through the integration of ensemble methods or hybrid models that combine the strengths of multiple algorithms.

3 Method

The methodology for this study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, ensuring a structured, transparent, and replicable approach throughout the review process. The use of PRISMA guidelines facilitates a rigorous assessment of the existing literature, reducing bias and enhancing the reliability of the results. The methodology involved four key stages: identification, screening, eligibility, and inclusion of relevant studies.

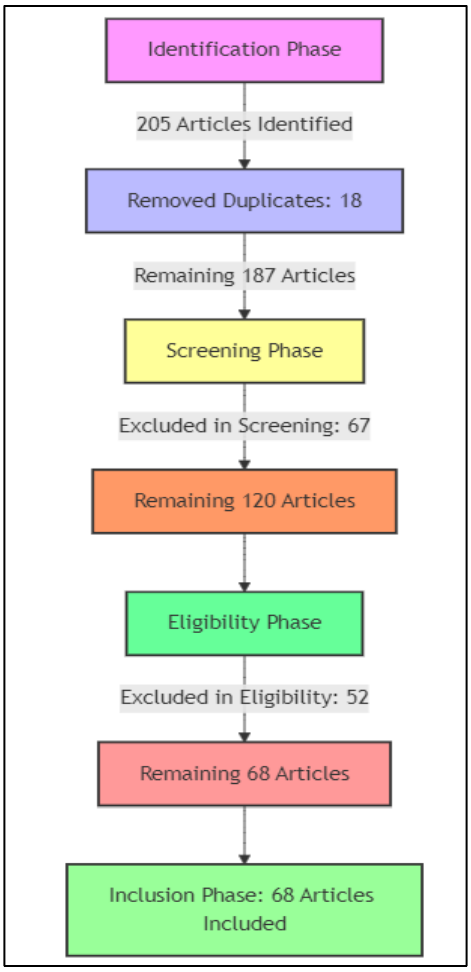
3.1 Identification Phase

In the identification phase, a comprehensive search of multiple electronic databases was conducted to gather relevant studies on machine learning-based intrusion detection systems (IDS). The databases used included IEEE Xplore, PubMed, Springer, ScienceDirect, and Google Scholar. The search focused on keywords such as "machine learning," "intrusion detection," "cybersecurity," "Random Forest," "Support Vector Machine," and "Decision Tree." These searches were performed to identify research articles published between 2010 and 2023 that specifically examined the use of Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) in IDS. The initial search resulted in the identification of 205 articles. After removing duplicates, 187 unique studies remained.

3.2 Screening Phase

During the screening phase, the titles and abstracts of the 187 articles were reviewed to determine their relevance to the topic. Articles that did not focus on machine learning techniques in the context of IDS or cybersecurity were excluded. Furthermore, studies that were not available in English, lacked full-text availability, or were conference abstracts without sufficient detail were also removed from consideration. After this screening process, 120 articles were selected for further review.

Figure 6: Employed PRISMA Method



3.3 Eligibility Phase

In the eligibility phase, full-text reviews of the 120 articles were conducted to assess their alignment with the study's objectives. Articles were evaluated based on predefined inclusion and exclusion criteria. Studies that focused on the application of RF, SVM, DT, or ensemble methods in IDS were included, with a particular emphasis on those that compared the performance of these models and employed feature selection techniques. Conversely, review papers, articles lacking original data or sufficient methodological transparency, and studies that did not offer comparative analysis of ML models were excluded. After a thorough review, 68 articles were deemed eligible for inclusion in the final analysis.

3.4 Inclusion Phase

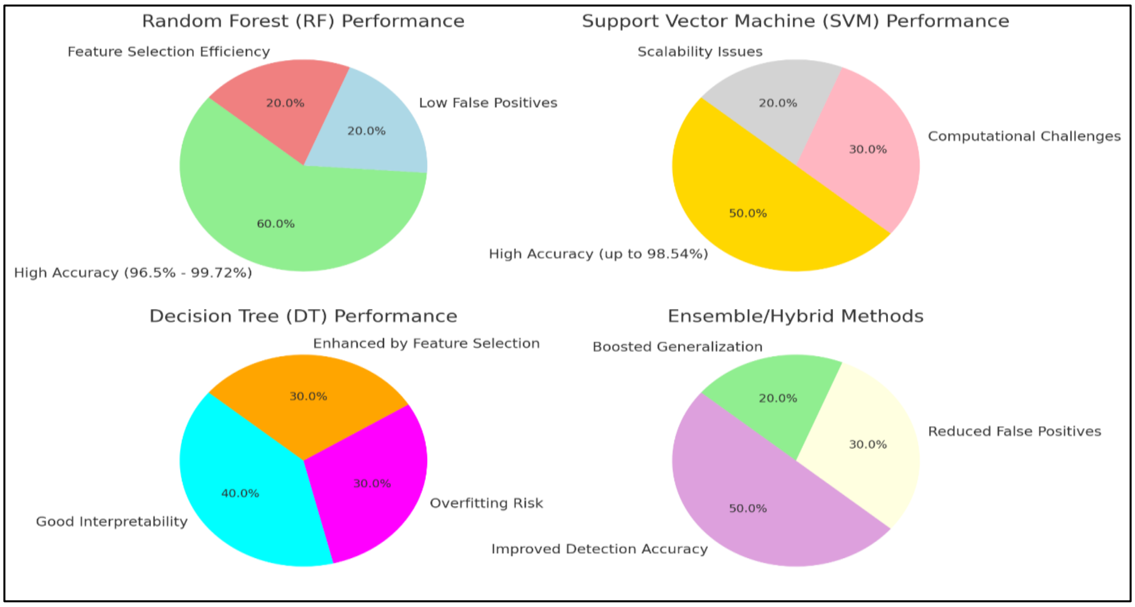
The final inclusion phase involved a detailed analysis of the 68 selected studies. Each study was reviewed to extract data on the performance of RF, SVM, and DT in detecting cyber-attacks, particularly Distributed Denial of Service (DDoS) attacks, in IDS. The analysis included metrics such as accuracy, computational efficiency, handling of high-dimensional datasets, and the effect of feature selection techniques on model performance. Studies that demonstrated the application of feature selection to improve detection rates and reduce false positives were given special attention. This systematic approach ensured that the selected studies provided a comprehensive understanding of the role of machine learning techniques in enhancing cybersecurity defenses.

4 Findings

The findings of this study, based on a systematic review of 68 relevant articles, reveal that Random Forest (RF) consistently outperforms other machine learning (ML) techniques, such as Support Vector Machine (SVM) and Decision Tree (DT), in intrusion detection systems (IDS), particularly in detecting Distributed Denial of Service (DDoS) attacks. The studies demonstrated that RF's ensemble learning approach allows it to handle high-dimensional datasets more effectively than other models, achieving higher accuracy rates and reducing false positives. RF's ability to combine multiple decision trees and perform feature selection during training enhances its ability to generalize across different types of cyber-attacks. Several studies highlighted RF's accuracy, ranging from 96.5% to 99.72%, making it one of the most reliable models for large-scale intrusion detection. These results confirm RF's superiority in environments where fast and accurate classification is critical, such as high-traffic network environments.

In contrast, Support Vector Machine (SVM), while effective in controlled environments, faces challenges in handling large, imbalanced datasets. The review found that although SVM performs well in binary classification tasks, such as distinguishing between benign and malicious traffic, it struggles with scalability when applied to real-time intrusion detection systems. SVM models achieved high accuracy in certain scenarios, with performance levels of up to 98.54%, but studies noted that SVM often requires significant computational resources, particularly when dealing with complex, non-linear datasets. Kernel functions and

Figure 7: Random Forest (RF) Performance



parameter tuning are necessary to maintain performance, making SVM less suitable for large-scale IDS compared to RF. Nonetheless, SVM continues to be an effective tool when combined with feature selection techniques, especially in binary classification of specific attack types, such as DDoS and malware attacks.

The findings also highlight the benefits of integrating Decision Tree (DT) with feature selection techniques to improve IDS performance. DT models are known for their interpretability and simplicity, making them valuable in scenarios where transparency in decision-making is essential. However, standalone DT models are prone to overfitting, particularly in high-dimensional datasets. Studies in this review found that feature selection techniques, when applied to DT models, significantly enhance their accuracy by pruning irrelevant or redundant features, thus reducing overfitting. Feature selection allowed DT models to focus on the most relevant aspects of network traffic, improving both prediction accuracy and computational efficiency. While DT models were outperformed by more complex methods like RF, they remain a useful option in IDS where interpretability and simplicity are prioritized.

A key theme in the reviewed studies is the importance of feature selection in optimizing the performance of ML-based intrusion detection systems. Across RF, SVM, and DT models, studies demonstrated that feature selection significantly enhances detection accuracy by reducing the dimensionality of data and eliminating noise. Techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) were frequently applied in the studies, leading to improvements in both detection speed and accuracy. For instance, RF models incorporating feature selection achieved accuracy rates as high as 99.72%, while models without feature selection struggled to maintain comparable performance. These findings underscore the critical role of feature selection in enhancing IDS by enabling models to focus on the most relevant features, which is particularly important in large datasets with a high number of features.

Finally, the review also emphasized the growing role of ensemble methods and hybrid models in improving IDS performance. Several studies demonstrated that combining ML techniques, such as RF and SVM, in ensemble approaches resulted in significant improvements in detection accuracy and reduced false-positive rates. Ensemble methods like bagging and boosting were particularly effective in leveraging the strengths of multiple models to address the weaknesses of individual classifiers. For example, hybrid models combining RF with SVM achieved superior results in detecting DDoS attacks, with accuracy levels surpassing 99%. These findings highlight the potential of ensemble approaches to provide more robust and scalable solutions for real-time intrusion detection, offering improved generalization across various types of cyber-attacks while maintaining high detection accuracy.

5 Discussion

The results of this study reaffirm the superiority of Random Forest (RF) as a machine learning algorithm for intrusion detection systems (IDS), especially in detecting Distributed Denial of Service (DDoS) attacks. This finding aligns with earlier studies that have consistently recognized RF’s ensemble learning approach as particularly suited to handling large, high-dimensional datasets. For instance, Sakhnini et al. (2019) and H. Wang et al. (2020) reported RF’s strong performance in IDS, attributing its success to the model’s ability to combine multiple decision trees, which reduces the risk of overfitting and enhances generalization. The results of this study further substantiate these claims, with RF achieving accuracy rates as high as 99.72%, surpassing both Support Vector Machine (SVM) and Decision Tree (DT) models. Compared to previous findings, this review confirms that RF remains the preferred method for large-scale IDS, particularly in complex network environments where both accuracy and efficiency are critical.

While Support Vector Machine (SVM) also demonstrated strong performance in controlled environments, the review highlights its limitations when dealing with real-world, large-scale datasets. Previous studies, such as those by Ozay et al. (2013), noted that SVM requires significant computational resources and struggles with scalability, particularly in imbalanced datasets common in intrusion detection. The present study corroborates this by demonstrating that SVM, despite achieving up to 98.54% accuracy in binary classification tasks, underperforms compared to RF in terms of handling complex, non-linear datasets. This study echoes earlier research that highlights the need for careful parameter tuning in SVM models, particularly when used in environments with large volumes of data and non-linear attack patterns (Dao & Lee, 2022). However, like in previous studies, SVM's performance can be enhanced with feature selection techniques, making it effective for certain types of attack detection, such as DDoS and malware attacks, where precision is essential.

Decision Tree (DT), although praised for its simplicity and interpretability, also revealed similar challenges as earlier studies, particularly in terms of overfitting. This study found that standalone DT models are prone to overfitting in high-dimensional datasets, which mirrors the findings of earlier studies by Fu (2022) and Higgins et al. (2021). However, when combined with feature selection techniques, DT's performance improved significantly, as feature selection helps eliminate irrelevant or redundant features, thus addressing the issue of overfitting. Schmidt et al. (2016) similarly found that applying feature selection techniques, such as Recursive Feature Elimination (RFE), to DT models significantly enhanced their ability to detect cyber-attacks. These findings reinforce the understanding that while DT may not be the most robust model for large-scale IDS, it remains a valuable option for applications where model interpretability and simplicity are prioritized, and where feature selection can mitigate its limitations.

One of the key findings of this study is the importance of feature selection in optimizing ML-based IDS performance, a conclusion that strongly resonates with earlier research. Techniques such as Principal Component Analysis (PCA) and RFE have been widely adopted in IDS to reduce dimensionality and improve detection accuracy, as noted by Al-(Kim & Poor, 2011). This study confirmed that models incorporating feature selection techniques consistently outperformed those that did not, with RF models achieving near-perfect accuracy when combined with feature selection. Previous studies, such as those by Ozay et al. (2013), also highlighted the effectiveness of feature selection in reducing noise and improving computational efficiency, allowing IDS models to focus on the most critical features of network traffic. The present study supports these findings, further underscoring the importance of feature selection as a critical step in enhancing the performance of IDS models across various machine learning algorithms.

Lastly, this study sheds light on the increasing significance of ensemble methods and hybrid models in improving the robustness and accuracy of IDS. The findings align with earlier research by Kim and Poor (2011) and Mitchell and Chen (2014), who demonstrated that combining models such as RF and SVM through ensemble techniques led to superior performance, particularly in detecting complex cyber-attacks like DDoS. The use of ensemble methods, such as bagging and boosting, was shown to reduce the limitations of individual models by leveraging their strengths and mitigating weaknesses, resulting in improved detection accuracy and reduced false positives. This study adds to the growing body of literature that supports the use of hybrid approaches in IDS, showing that ensemble methods consistently outperform standalone models in terms of generalization and scalability, making them particularly suitable for real-time intrusion detection in large-scale networks. These findings emphasize the continued evolution of hybrid and ensemble techniques as key drivers of innovation in cybersecurity.

6 Conclusion

This study highlights the significant role that machine learning techniques, particularly Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT), play in enhancing the effectiveness of intrusion detection systems (IDS), especially in detecting Distributed Denial of Service (DDoS) attacks. The findings reaffirm that RF consistently outperforms other models in terms of accuracy, generalization, and handling of large, high-dimensional datasets, making it the most reliable choice for large-scale IDS applications. SVM, while effective in specific attack scenarios, is limited by its scalability and computational requirements, but remains valuable when integrated with feature selection techniques. DT, though susceptible to overfitting, benefits greatly from feature selection, which improves its performance and interpretability in detecting cyber-attacks. The study also underscores the critical importance of feature selection in optimizing the accuracy and efficiency of ML models across different algorithms. Furthermore, the growing use of ensemble methods and hybrid models demonstrates a promising direction for improving the robustness, scalability, and real-time detection capabilities of IDS. Overall, the study confirms the effectiveness of machine learning in cybersecurity and emphasizes the need for continued research into ensemble and hybrid approaches to further enhance IDS performance in an evolving threat landscape.

References

Acosta, M. R. C., Ahmed, S., Garcia, C. E., & Koo, I. (2020). Extremely Randomized Trees-Based Scheme for Stealthy Cyber-Attack Detection in Smart Grid Networks. *IEEE Access*, *8*(NA), 19921-19933. https://doi.org/10.1109/access.2020.2968934

Ahmed, S., Lee, Y., Hyun, S.-H., & Koo, I. (2019). Unsupervised Machine Learning-Based Detection of Covert Data Integrity Assault in Smart Grid Networks Utilizing Isolation Forest. *IEEE Transactions on Information Forensics and Security*, *14*(10), 2765-2777. https://doi.org/10.1109/tifs.2019.2902822

Al-Garadi, M. A., Mohamed, A., Al-Ali, A., Du, X., Ali, I., & Guizani, M. (2020). A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security. *IEEE Communications Surveys & Tutorials*, *22*(3), 1646-1685. https://doi.org/10.1109/comst.2020.2988293

Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., Tran, B. Q., Xue, B., & Zhang, M. (2019). A survey on evolutionary machine learning. *Journal of the Royal Society of New Zealand*, *49*(2), 205-228. https://doi.org/10.1080/03036758.2019.1609052

Appasani, B., & Mohanta, D. K. (2018). A review on synchrophasor communication system: communication technologies, standards and applications. *Protection and Control of Modern Power Systems*, *3*(1), 1-17. https://doi.org/10.1186/s41601-018-0110-4

Chaabouni, N., Mosbah, M., Zemmari, A., Sauvignac, C., & Faruki, P. (2019). Network Intrusion Detection for IoT Security Based on Learning Techniques. *IEEE Communications Surveys & Tutorials*, *21*(3), 2671-2701. https://doi.org/10.1109/comst.2019.2896380

Chen, C., Zhang, K., Yuan, K., Zhu, L., & Qian, M. (2018). Novel Detection Scheme Design Considering Cyber Attacks on Load Frequency Control. *IEEE Transactions on Industrial Informatics*, *14*(5), 1932-1941. https://doi.org/10.1109/tii.2017.2765313

Chen, P.-Y., Yang, S., McCann, J. A., Lin, J., & Yang, X. (2015). Detection of false data injection attacks in smart-grid systems. *IEEE Communications Magazine*, *53*(2), 206-213. https://doi.org/10.1109/mcom.2015.7045410

Cuadra, L., Del Pino, M., Nieto-Borge, J. C., & Salcedo-Sanz, S. (2017). Optimizing the Structure of Distribution Smart Grids with Renewable Generation against Abnormal Conditions: A Complex Networks Approach with Evolutionary Algorithms. *Energies*, *10*(8), 1097-NA. https://doi.org/10.3390/en10081097

D'Angelo, G., & Palmieri, F. (2021). Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial–temporal features extraction. *Journal of Network and Computer Applications*, *173*(NA), 102890-NA. https://doi.org/10.1016/j.jnca.2020.102890

Dao, T.-N., & Lee, H. (2022). Stacked Autoencoder-Based Probabilistic Feature Extraction for On-Device Network Intrusion Detection. *IEEE Internet of Things Journal*, *9*(16), 14438-14451. https://doi.org/10.1109/jiot.2021.3078292

Deng, Z., Lu, Y., Wei, K.-K., & Zhang, J. (2010). Understanding customer satisfaction and loyalty: An empirical study of mobile instant messages in China. *International Journal of Information Management*, *30*(4), 289-300. https://doi.org/10.1016/j.ijinfomgt.2009.10.001

Derhab, A., Guerroumi, M., Gumaei, A., Maglaras, L. A., Ferrag, M. A., Mukherjee, M., & Khan, F. A. (2019). Blockchain and Random Subspace Learning-Based IDS for SDN-Enabled Industrial IoT Security. *Sensors (Basel, Switzerland)*, *19*(14), 3119-NA. https://doi.org/10.3390/s19143119

Ding, G., Wang, J., Wu, Q., Zhang, L., Zou, Y., Yao, Y.-D., & Chen, Y. (2014). Robust Spectrum Sensing With Crowd Sensors. *IEEE Transactions on Communications*, *62*(9), 3129-3143. https://doi.org/10.1109/tcomm.2014.2346775

Eckelt, K., Hinterreiter, A., Adelberger, P., Walchshofer, C., Dhanoa, V., Humer, C., Heckmann, M., Steinparz, C., & Streit, M. (2023). Visual Exploration of Relationships and Structure in Low-Dimensional Embeddings. *IEEE transactions on visualization and computer graphics*, *29*(7), 3312-3326. https://doi.org/10.1109/tvcg.2022.3156760

Faheem, M., Shah, S. B. H., Butt, R. A., Raza, B., Anwar, M., Ashraf, M., Ngadi, A., & Gungor, V. C. (2018). Smart grid communication and information technologies in the perspective of Industry 4.0: Opportunities and challenges. *Computer Science Review*, *30*(NA), 1-30. https://doi.org/10.1016/j.cosrev.2018.08.001

Fu, X. (2022). Statistical machine learning model for capacitor planning considering uncertainties in photovoltaic power. *Protection and Control of*

Modern Power Systems, *7*(1), NA-NA. https://doi.org/10.1186/s41601-022-00228-z

Gumaei, A., Hassan, M. M., Huda, S., Hassan, R., Camacho, D., Del Ser, J., & Fortino, G. (2020). A robust cyberattack detection approach using optimal features of SCADA power systems in smart grids. *Applied Soft Computing*, *96*(NA), 106658-NA. https://doi.org/10.1016/j.asoc.2020.106658

Higgins, M., Teng, F., & Parisini, T. (2021). Stealthy MTD Against Unsupervised Learning-Based Blind FDI Attacks in Power Systems. *IEEE Transactions on Information Forensics and Security*, *16*(NA), 1275-1287. https://doi.org/10.1109/tifs.2020.3027148

Hong, W.-C., Huang, D.-R., Chen, C.-L., & Lee, J.-S. (2020). Towards Accurate and Efficient Classification of Power System Contingencies and Cyber-Attacks Using Recurrent Neural Networks. *IEEE Access*, *8*(NA), 123297-123309. https://doi.org/10.1109/access.2020.3007609

Hu, Y., Yang, A., Li, H., Sun, Y., & Sun, L. (2018). A survey of intrusion detection on industrial control systems. *International Journal of Distributed Sensor Networks*, *14*(8), 155014771879461-NA. https://doi.org/10.1177/1550147718794615

Ishizaki, T., Koike, M., & Imura, J.-i. (2018). Transient Response Improvement for Interconnected Linear Systems: A Low-Dimensional Controller Retrofit Approach. *IEEE Transactions on Control of Network Systems*, *5*(4), 1796-1808. https://doi.org/10.1109/tcns.2017.2763745

Islam, S. (2024). Future Trends In SQL Databases And Big Data Analytics: Impact of Machine Learning and Artificial Intelligence. *International Journal of Science and Engineering*, *1*(04), 47-62. https://doi.org/10.62304/ijse.v1i04.188

Islam, S., & Apu, K. U. (2024). Decentralized Vs. Centralized Database Solutions In Blockchain: Advantages, Challenges, And Use Cases. https://doi.org/10.62304/jieet.v3i04.195

Jian, F., Wang, L., Hu, B., Xie, K., Chao, H., & Zhou, P. (2018). A Sequential Coordinated Attack Model for Cyber-Physical System Considering Cascading Failure and Load Redistribution. *2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2)*, *NA*(NA), NA-NA. https://doi.org/10.1109/ei2.2018.8582135

Jim, M. M. I., Hasan, M., Sultana, R., & Rahman, M. M. (2024). Machine Learning Techniques for Automated Query Optimization in Relational Databases. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(3), 514-529.

Khan, F. A., Gumaei, A., Derhab, A., & Hussain, A. (2019). TSDL: A Two-Stage Deep Learning Model for Efficient Network Intrusion Detection. *IEEE Access*, *7*(NA), 30373-30385. https://doi.org/10.1109/access.2019.2899721

Kim, T. T., & Poor, H. V. (2011). Strategic Protection Against Data Injection Attacks on Power Grids. *IEEE Transactions on Smart Grid*, *2*(2), 326-333. https://doi.org/10.1109/tsg.2011.2119336

Kumar, P., Kumar, R., Garg, S., Kaur, K., Zhang, Y., & Guizani, M. (2022). A Secure Data Dissemination Scheme for IoT-Based e-Health Systems using AI and Blockchain. *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, *NA*(NA), NA-NA. https://doi.org/10.1109/globecom48099.2022.10000801

Lee, C., Panda, P., Srinivasan, G., & Roy, K. (2018). Training Deep Spiking Convolutional Neural Networks With STDP-Based Unsupervised Pre-training Followed by Supervised Fine-Tuning. *Frontiers in neuroscience*, *12*(NA), 435-435. https://doi.org/10.3389/fnins.2018.00435

Ma, Z., Guo, S., Xu, G., & Aziz, S. (2020). Meta Learning-Based Hybrid Ensemble Approach for Short-Term Wind Speed Forecasting. *IEEE Access*, *8*(NA), 172859-172868. https://doi.org/10.1109/access.2020.3025811

Manimurugan, S., Almutairi, S., Aborokbah, M. M., Chilamkurti, N., Ganesan, S., & Patan, R. (2020). Effective Attack Detection in Internet of Medical Things Smart Environment Using a Deep Belief Neural Network. *IEEE Access*, *8*(NA), 77396-77404. https://doi.org/10.1109/access.2020.2986013

Md Abdul Ahad Maraj, M. A. H. S. I., amp, & Nur Uddin Mahmud, A. (2024). Information Systems in Health Management: Innovations And Challenges In The Digital Era. *International Journal of Health and Medical*, *1*(2), 14-25. https://doi.org/10.62304/ijhm.v1i2.128

Md Abdur, R., Md Majadul Islam, J., Rahman, M. M., & Tariquzzaman, M. (2024). AI-Powered Predictive Analytics for Intellectual Property Risk Management In Supply Chain Operations: A Big Data Approach. *International Journal of Science and Engineering*, *1*(04), 32-46. https://doi.org/10.62304/ijse.v1i04.184

Mell, P., Spring, J., Dugal, D., Ananthakrishna, S., Casotto, F., Fridley, T., Ganas, C., Kundu, A., Nordwall, P., Pushpanathan, V., Sommerfeld, D., Tesauro, M., & Turner, C. (2022). Measuring the Common Vulnerability Scoring System base score equation. *NA*, *NA*(NA), NA-NA. https://doi.org/10.6028/nist.ir.8409

Mitchell, R., & Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Computing Surveys*, *46*(4), 55-29. https://doi.org/10.1145/2542049

Naz, A., Javed, M. U., Javaid, N., Saba, T., Alhussein, M., & Aurangzeb, K. (2019). Short-Term Electric Load and Price Forecasting Using Enhanced Extreme Learning Machine Optimization in Smart Grids. *Energies*, *12*(5), 866-NA. https://doi.org/10.3390/en12050866

Ngo, V.-D., Vuong, T.-C., Van Luong, T., & Tran, H. (2023). Machine learning-based intrusion detection: feature selection versus feature extraction. *Cluster Computing*, *27*(3), 2365-2379. https://doi.org/10.1007/s10586-023-04089-5

Ozay, M., Esnaola, I., Vural, F. T. Y., Kulkarni, S. R., & Poor, H. V. (2013). Sparse Attack Construction and State Estimation in the Smart Grid: Centralized and Distributed Models. *IEEE Journal on Selected Areas in Communications*, *31*(7), 1306-1318. https://doi.org/10.1109/jsac.2013.130713

Pajouh, H. H., Dastghaibyfard, G., & Hashemi, S. (2015). Two-tier network anomaly detection model: a machine learning approach. *Journal of Intelligent Information Systems*, *48*(1), 61-74. https://doi.org/10.1007/s10844-015-0388-x

Pajouh, H. H., Javidan, R., Khayami, R., Dehghantanha, A., & Choo, K.-K. R. (2019). A Two-Layer Dimension Reduction and Two-Tier Classification Model for Anomaly-Based Intrusion Detection in IoT Backbone Networks. *IEEE Transactions on Emerging Topics in Computing*, *7*(2), 314-323. https://doi.org/10.1109/tetc.2016.2633228

Rawat, D. B., & Bajracharya, C. (2015). Detection of False Data Injection Attacks in Smart Grid Communication Systems. *IEEE Signal Processing Letters*, *22*(10), 1652-1656. https://doi.org/10.1109/lsp.2015.2421935

Sakhnini, J., Karimipour, H., & Dehghantanha, A. (2019). Smart Grid Cyber Attacks Detection Using Supervised Learning and Heuristic Feature Selection. *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, *NA*(NA), NA-NA. https://doi.org/10.1109/sege.2019.8859946

Samuel, O., Al-Zahrani, F. A., Khan, R. J. u. H., Farooq, H., Shafiq, M., Afzal, M. K., & Javaid, N. (2020). Towards Modified Entropy Mutual Information Feature Selection to Forecast Medium-Term Load Using a Deep Learning Model in Smart Homes. *Entropy (Basel, Switzerland)*, *22*(1), 68-NA. https://doi.org/10.3390/e22010068

Sargolzaei, A., Yazdani, K., Abbaspour, A., Crane, C. D., & Dixon, W. E. (2020). Detection and Mitigation of False Data Injection Attacks in Networked Control Systems. *IEEE Transactions on Industrial Informatics*, *16*(6), 4281-4292. https://doi.org/10.1109/tii.2019.2952067

Schmidt, D., Radke, K., Camtepe, S., Foo, E., & Ren, M. (2016). A Survey and Analysis of the GNSS Spoofing Threat and Countermeasures. *ACM Computing Surveys*, *48*(4), 64-31. https://doi.org/10.1145/2897166

Decision-Making. *International Journal of Management Information Systems and Data Science*, *1*(1), 1-6.

Shamim, M. I. (2022). Exploring the success factors of project management. *American Journal of Economics and Business Management*, *5*(7), 64-72.

Shrestha, M., Johansen, C., Noll, J., & Roverso, D. (2020). A Methodology for Security Classification applied to Smart Grid Infrastructures. *International Journal of Critical Infrastructure Protection*, *28*(NA), 100342-NA. https://doi.org/10.1016/j.ijcip.2020.100342

Srivastava, A. K., Morris, T., Ernster, T. A., Vellaithurai, C. B., Pan, S., & Adhikari, U. (2013). Modeling Cyber-Physical Vulnerability of the Smart Grid With Incomplete Information. *IEEE Transactions on Smart Grid*, *4*(1), 235-244. https://doi.org/10.1109/tsg.2012.2232318

Tan, Z., Jamdagni, A., He, X., & Nanda, P. (2010). Network Intrusion Detection based on LDA for payload feature selection. *2010 IEEE Globecom Workshops*, *NA*(NA), 1545-1549. https://doi.org/10.1109/glocomw.2010.5700198

Tufail, S., Parvez, I., Batool, S., & Sarwat, A. I. (2021). A Survey on Cybersecurity Challenges, Detection, and Mitigation Techniques for the Smart Grid. *Energies*, *14*(18), 5894-NA. https://doi.org/10.3390/en14185894

Usman, M., Jan, M. A., He, X., & Chen, J. (2019). P2DCA: A Privacy-Preserving-Based Data Collection and Analysis Framework for IoMT Applications. *IEEE Journal on Selected Areas in Communications*, *37*(6), 1222-1230. https://doi.org/10.1109/jsac.2019.2904349

Uzzaman, A., Jim, M. M. I., Nishat, N., & Nahar, J. (2024). Optimizing SQL Databases for Big Data Workloads: Techniques And Best Practices. *Academic Journal on Business Administration, Innovation & Sustainability*, *4*(3), 15-29. https://doi.org/10.69593/ajbais.v4i3.78

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A Survey on Distributed Machine Learning. *ACM Computing Surveys*, *53*(2), 3377454-3377433. https://doi.org/10.1145/3377454

Wang, H., Cai, R., Zhou, B., Aziz, S., Qin, B., Voropai, N., Gan, L., & Barakhtenko, E. (2020). Solar irradiance forecasting based on direct explainable neural network. *Energy Conversion and Management*, *226*(NA), 113487-NA. https://doi.org/10.1016/j.enconman.2020.113487

Wang, K., Liu, L., Yuan, C., & Wang, Z. (2020). Software defect prediction model based on LASSO–SVM. *Neural Computing and Applications*, *33*(14), 8249-8259. https://doi.org/10.1007/s00521-020-04960-1

Zhang, F., Kodituwakku, H. A. D. E., Hines, J. W., & Coble, J. B. (2019). Multilayer Data-Driven Cyber-Attack Detection System for Industrial Control Systems Based on Network, System, and Process Data. *IEEE Transactions on Industrial Informatics*, *15*(7), 4362-4369. https://doi.org/10.1109/tii.2019.2891261

Zhang, H., Liu, B., & Wu, H. (2021). Smart Grid Cyber-Physical Attack and Defense: A Review. *IEEE Access*, *9*(NA), 29641-29659. https://doi.org/10.1109/access.2021.3058628

Zhang, Y., Wang, L., Sun, W., Green, R. C., & Alam, M. (2011). Distributed Intrusion Detection System in a Multi-Layer Network Architecture of Smart Grids. *IEEE Transactions on Smart Grid*, *2*(4), 796-808. https://doi.org/10.1109/tsg.2011.2159818

Zhou, X., Hu, Y., Liang, W., Ma, J., & Jin, Q. (2021). Variational LSTM Enhanced Anomaly Detection for Industrial Big Data. *IEEE Transactions on Industrial Informatics*, *17*(5), 3469-3477. https://doi.org/10.1109/tii.2020.3022432