CHAPTER 1

THE PROBLEM

1.1 Rationale

With the rapid advancement in information and communication technology, computer networks have become the core infrastructure for organizations, enterprises, and even individuals. However, with the increasing complexity of networks and the ever-growing presence of security threats, the protection of networks has become increasingly important. Cyberattacks such as denial-of-service (DoS), malware attacks, and intrusions into networks are a constant lurking threat. One of the most promising approaches to addressing these threats is by using a Network Intrusion Detection System (NIDS). NIDS aim to detect cyberattacks by analyzing network traffic and identifying suspicious or unusual behavior. However, with the increasing complexity of attacks and large volumes of data, traditional approaches in NIDS have shown their limitations. Machine learning (ML) has emerged as a promising solution in improving the detection capabilities of NIDS. By utilizing ML techniques such as classification, clustering, and data grouping, NIDS can learn from existing patterns in network traffic and automatically detect attacks that have never been detected before. Due to the urgent necessity to safeguard systems from unauthorized, uninvited, and unexpected intrusions, cybersecurity has become a global need in the modern day [1]. These incursions can take many different forms, such as threats to the functioning and integrity of the system or data breaches and information theft. Maintaining user confidence, safeguarding sensitive data, and facilitating seamless system operations all depend on threat protection [2]. Traditionally, intrusion detection systems (IDS) have been the mainstay of perimeter security [3, 4]. In cybersecurity, machine learning (ML) is a formidable technology that enhances systems' comprehension of various patterns and predicts possible data risks [12]. It streamlines training and processing processes to create models that can effectively protect systems against shady and malware activity [6, 13]. It is a game-changing technique that enables computers to make intelligent judgments without explicit programming by learning from data and adapting accordingly [14]. ML algorithms use historical and real-time data in the context of IDS to find patterns of typical activity and abnormalities that can point to security issues. These algorithms grow proficient at identifying novel and emerging attack vectors by training on a variety of datasets. By decreasing false positives,

improving threat detection speed and accuracy, and adjusting to changing threats, machine learning (ML) improves intrusion detection systems [15]. It gives security systems the ability to effectively protect networks and data from hostile activity and unwanted access [16]. In the current environment, building models that can effectively protect systems against shady and malware activity requires optimizing processing and training processes [12]. It's important to keep in mind, nevertheless, that a lot of modern ML-IDS solutions are frequently constrained by their dependence on tiny, antiquated, and well-balanced datasets for model building [17, 19]. The emphasis on smaller, frequently out-of-date datasets and data distribution imbalances, while enabling preprocessing and training with a variety of machine learning algorithms, raises concerns about the models' applicability in real-world situations, particularly when handling large amounts of data. The complexities of dataset preparation and appropriate algorithm selection frequently determine the achievable accuracy of these models, which adds another level of complexity to their effectiveness [20, 21].

Using the UNSW-NB15 data as a baseline, Moualla et al. [25] developed a ground-breaking network intrusion detection system (IDS) model that is essential to network security and counters current cyberattacks on networks. It was a multiclass ML-based network with multiple phases that was dynamically scalable. The imbalance was addressed using the SMOTE technique. Next, the ET Classifier was used, based on the Gini Impurity criterion. Finally, a trained Extreme Learning Machine (ELM) was used to categorize each assault using binary. A fully connected layer's outputs from the ELM classifier were fed into a logistic regression layer, which generated soft judgments for every class with an accuracy of 98.43%.

A filter-based feature-dropping method using the XGB algorithm was presented by Kasongo and Sun [26] on the UNSW-NB15 IDS dataset. The performance of the method was evaluated using a variety of predictive algorithms, such as Decision Tree (DT), Artificial Neural Network (ANN), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). They showed that by using their method, binary accuracy increased significantly, from 88.13% to 90.85%. The overall binary accuracy rates for DT, ANN, LR, KNN, and SVM were 90.85%, 84.4%, 77.64%, and 60.89%, respectively. Using the 19 best selected features, the accuracy rates for DT, ANN, LR, KNN, and SVM in the multiclass setting were 67.57%, 77.51%, 65.29%, and 53.95%, respectively.

Using Information Grain (IG) and Grain Ratio (GR), Nimbalkar and Kshirsagar [27] proposed a feature selection approach for intrusion detection systems (IDS) in which they picked 50% of the most important features to construct their model for detecting Denial of Service (DoS) and Distributed Denial of Service (DDoS) assaults. The investigations were conducted with popular datasets as BOT-IOT and KDDCUP'99. For the BOT-IOT and KDDCUP'99 datasets, they chose 16 and 19 features, respectively, and trained the model with the Jrip classifier to achieve the required performance for each. For the BOTIOT and KDDCUP'99 datasets, they obtained accuracy rates of 99.99% and 99.57%, respectively.

Tiawan et al. [36] provided a method for examining fundamental and significant elements of large-scale network data, boosting the speed and accuracy of traffic anomaly identification. They trained the dataset using a variety of ML classifiers after using the CIC-IDS2017 dataset, selecting significant and crucial features using IG and sorting and grouping features based on their minimal weight values. The quantity of pertinent and significant attributes that are produced by IG significantly influences execution time and correctness. The model was trained using a variety of machine learning (ML) techniques, including Random Tree (RT), RF, NB, Bayes Net (BN), and J48, although RF yielded the best accuracy. The J48 classifer algorithm, which employed 52 relevant selected features but required more time to run, had the highest accuracy of 99.87%, while the RF classifer, which utilized 22 relevant selected features, had the best accuracy of 99.86%.

The goal of uezzaz et al. [35] was to apply the Decision Tree (DT) algorithm to increase the dependability of Network Intrusion Detection (NID). The two main components of their method were feature selection based on entropy judgment and data quality enhancement. The DT classifier was then used to build a trustworthy NID system. Using two well-known datasets, NSLKDD and CIC-IDS2017, this proposed model was evaluated, and the results were impressive. In particular, the model demonstrated exceptional accuracy on the NSL-KDD dataset (99.42%) and the CICIDS2017 dataset (98.80%).

A model for identifying cyberattacks was published by Hammad et al. [34]. t-SNERF was utilized for feature correlations, data reduction, and RF training of the model. We used CIC-IDS2017 and Phishing to assess the UNSW-NB15 model. The novel methodology Te offered performed better than the ones used now. For UNSW-NB15, the accuracy rate was 100%; for Phishing, it was 99.70%; and for CIC-

IDS2017, it was 99.78%.

An IDS model that minimizes prediction latency by employing a hybrid feature selection (HFS) technique to minimize model sophistication was developed by Seth et al. [33]. The model was constructed using a fast gradient boosting method known as Light Gradient Boosting Machine (LightGBM). Using the CIC-IDS2018 dataset, this method reduced model construction time by 52.68% to 17.94% and prediction latency by 44.52% to 2.25%. When attribute selection and LightGBM are used, it can get exceptional accuracy. The created model yielded a remarkably low prediction latency along with accuracy, sensitivity, and precision rate of 97.73%, 96%, and 99.3%, respectively.

Applying this novel method to the KDDCUP'99 dataset, Talita et al. [32] combined the Naive Bayes (NB) classification algorithm with Particle Swarm Optimization (PSO) for feature selection. This dataset had over 400 thousand records with over 40 different features. PSO was used to narrow down the original list of more than 40 attributes to the 38 most pertinent ones in order to maximize computer resources and memory utilization. This approach produced a 99.12% accuracy rate, which is an amazing result compared to previous feature selection strategies. It also showed greater efficiency in terms of calculation time and classification accuracy.

Evolutionary approach-based feature selection (EFS) was used by [31] to reduce the dataset's volume, and a concurrent MapReduce technique was used to divide the incoming data into the most important attributes. Following that, the RF classifier was used to categorize their model as normal or attack. They assessed performance and distinguished between normal and abnormal behavior using the well-known KDDCUPP'99 dataset. They evaluated the performance of their model using 15 features, and found that it was nearly 93.9% accurate.

A filter-based feature selection method using the IG Ratio (IGR), Correlation Ratio (CR), and ReliefF (ReF) was presented by Kshirsagar and Kumar [30]. This method combined with a Subset Combination Strategy (SCS) to generate a feature subset based on the average weight of each classifier. The number of features was reduced from 77 to 24 for the CICIDS 2017 dataset and from 41 to 12 for the KDDCUP'99 dataset. Using PART, it completed the CIC-IDS2017 dataset with a 99.95% accuracy rate in 133.66 seconds, and the KDDCUP'99 dataset with a 99.32% accuracy rate in 11.22 seconds.

A novel machine learning technique based on feature clustering was presented by Ahmad et al. [29]. Flow, Message Queuing Telemetry Transport (MQTT), and Transmission Control Protocol (TCP) each have their own unique clusters established. This clustering technique effectively tackled the overfitting issues brought on by imbalances in the dataset and excessive dimensionality. They gave these clusters a variety of supervised machine learning techniques, such as SVM and RF. Their findings demonstrated that RF exhibited remarkable accuracy, attaining 98.67% in binary classification and 97.37% in multi-class classification using the UNSW-NB15 dataset for model training and validation.

A feature clustering-based machine learning model was presented by Ahmad et al. [29], in which clusters for Transmission Control Protocol (TCP), Message Queuing Telemetry Transport (MQTT), and Flow were used. Through the process of clustering, the overfitting caused by dimensionality and data-set inequality was removed. The clusters were subjected to a variety of supervised machine learning techniques, such as SVM and RF. They tested and trained the model using the UNSW-NB15 dataset, and they discovered that RF produced 97.37% accuracy in multiclass and 98.67% accuracy in binary.

Ahmad dkk. [29] presents a machine learning approach based on feature selection, wherein a cluster is applied to Transmission Control Protocol (TCP), Message Queuing Telemetry Transport (MQTT), and Flow. The process of fitting curves reduces overfitting, which is a result of data set dimensi and ketidaksetaraan. Every available machine learning method is applied to the cluster, including RF and SVM. They use the UNSW-NB15 dataset to train and test the model and find that the RF yields an accuracy of 98.67% in binary and 97.37% in multiclass.

To defend networks against contemporary dangers such DoS assaults or exploits, probes, generics, and so forth, [28] introduced an intrusion detection system (IDS) that identifies intrusions based on abuse. The UNSW-NB15 dataset was used to calculate the false alarm rate (FAR) and intrusion detection rate (IDR). When the IG and C5 classifiers were applied, the IG was able to identify 13 features out of 47 for feature selection, while the C5 had an accuracy rating of 99.37%.

Among the proposed methods, there are still some problems to improve and optimize network intrusion detection using machine learning quickly and accurately to cope with large and unbalanced datasets. The model used in this study includes ten different algorithms, including KNN, logistic regression, decision tree, random

forest, GBM, XGBM, adaboost, light gbm, and cat boost. We present a new network intrusion detection model that uses Random Oversampling (RO) to overcome data imbalance and a feature selection technique using random forest to select the most optimal features. In addition, we apply Principal Component Analysis (PCA) technique to reduce the dimensionality of the dataset.

This framework serves as a guide for structuring research, ensuring the system's architecture aligns with the principles of modern cybersecurity and machine learning applications.

1.2 Theorical Framework

There are various theoretical frameworks and concepts that must be referenced and utilized in the course of development and implementation of the Machine Learning based Intrusion Detection System (ML-IDS). Below is a structured theoretical framework that conceptualizes the research:

1. Machine Learning (ML) has emerged as a transformative approach in cybersecurity, particularly in the realm of Intrusion Detection Systems (IDS). Unlike traditional signature-based systems, ML-IDS leverage algorithms to analyze patterns in network traffic and identify anomalies indicative of potential threats. This includes supervised learning and unsupervised learning, model evaluation, and feature extraction. The theory behind these can facilitate a more in-depth understanding of how ML techniques can be successfully applied in intrusion detection
    - Data Normalization. This process ensures that the input features are on a similar scale, which is critical for many ML algorithms that are sensitive to the scale of input data. Theoretical concepts related to statistical normalization methods (e.g., Min-Max scaling, Z-score normalization) guide the implementation of this step.
    - Feature Resampling Oversampling. Addressing class imbalance through techniques like Synthetic Minority Over-sampling Technique (SMOTE) is crucial for improving the detection rates of minority classes. Theoretical frameworks surrounding resampling methods and their impact on model performance are integral to this aspect.
    - Techniques such as Principal Component Analysis (PCA) and feature selection are employed to reduce the feature space while retaining

essential information. Theoretical models of variance preservation and feature importance help in understanding how to select and transform features effectively.

2. Evaluation Metrics for Model Performance. Evaluating the effectiveness of the ML-IDS is paramount to ensure its reliability and accuracy. Metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are essential for assessing model performance. Theoretical frameworks surrounding statistical evaluation methods and their implications for cybersecurity applications guide the selection of appropriate metrics
3. Dynamic Dataset Creation. The development of comprehensive datasets that encompass a wide range of attack scenarios is crucial for training and validating ML models. Theoretical concepts related to dataset generation, including user profiling and traffic characterization, inform the creation of datasets that reflect real-world network behaviors. The framework emphasizes the importance of diverse attack scenarios, such as Denial of Service (DoS), Brute-force attacks, and Web attacks, to ensure robust model training.

This theoretical framework establishes the foundation for designing an ML-IDS that leverages advanced machine learning techniques for threat detection and incorporates robust data preprocessing, evaluation metrics, and visualization tools. Key concepts such as anomaly detection, data normalization, feature resampling, and model evaluation are integral to the system's conceptualization and implementation. This framework serves as a guide for structuring research, ensuring the system's architecture aligns with the principles of modern cybersecurity and machine learning applications.

1.3 Conceptual Framework/Paradigm

The conceptual framework for a machine learning-based network intrusion detection system (NIDS) designed to handle big and imbalanced datasets revolves around several critical variables that interact to enhance the detection of cyber threat. Therefore, we need to develop and validate ML-based intrusion detection for large, imbalanced datasets where all potential attack scenarios are encompassed. To address this gap, Our research places a signifcant emphasis on constructing a robust and well-structured framework that accommodates the detection of network intrusion in a more

efficient manner on substantial datasets. We employ data preprocessing techniques, including data normalization, feature resampling with oversampling, Feature selection using random forest PCA and K-vold validation.

1. The key techniques of our approach are as follows:
    - Oversampling: Oversampling techniques are applied to handle class imbalance in network intrusion datasets. Methods like SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic samples for underrepresented attack types, ensuring the model learns from a balanced dataset and improves classification performance.
    - Feature Selection using Random Forest: Random Forest assigns feature importance scores, selecting the most significant features where p-value > 0.5. This step reduces dimensionality while retaining key attributes that contribute most to attack detection accuracy.
    - Principal Component Analysis (PCA): PCA is used for feature extraction, reducing dataset complexity by transforming features into a smaller set of principal components while preserving variance. This step further optimizes model performance by eliminating redundancy.
    - K-fold Cross-Validation: K-fold validation ensures robust model evaluation by splitting the dataset into multiple subsets (e.g., K=10). Each fold acts as a test set once while the remaining are used for training. This iterative process enhances generalization and prevents overfitting in NIDS models.
2. Learning Dependent model variables in (observed distinguishing outcomes):
    - Forms Detection of accuracy. Attacks the with capability the of help the of machine features and data processing methods discussed above.
    - Model performance metrics. Performance measures like the precision, recall and F1 measure that assess how well an intrusion detection system is able to identify intrusions.
    - Processing time. The time it takes the system to through the process of analyzing the incoming network traffic, detecting an and attack logging the same. This is crucial for real time threat mitigation.
    - System robustness: The capacity types of of attacks the balanced and and system imbalanced different in data. data terms distribution, of which its is ability assessed to based sustain on high the detection system’s rates performance under on different both conditions.

This comprehensive approach seeks to bridge the gap in intrusion detection, accommodating the intricacies of large, imbalanced datasets, and improving the robustness of security measures in the face of evolving threats. Our proposed model's performance is rigorously evaluated across a spectrum of ML classifers, including Decision Tree (DT), Random Forest (RF), Extra Tree (ET), and Extreme Gradient Boosting (XGB). Tese classifers are trained using a reduced feature set. We assess our model using a comprehensive set of performance indicators, encompassing precision, recall, f1-score, confusion matrix, accuracy and the ROC curve. The ML algorithms integrated into our framework demonstrate an exceptional ability to detect attacks, consistently achieving accuracy rates exceeding 99.9%. This thorough performance evaluation ensures the robustness and reliability of our intrusion detection system, highlighting its efectiveness in identifying and countering potential threats. In summary, this paper's contribution can be encapsulated as follows:

This framework serves as a guide for structuring research, ensuring the system's architecture aligns with the principles of modern cybersecurity and machine learning applications.

1.4 Statement of the Problem

The problem formulations of this thesis are as follows:

1. How can we effectively manage class imbalance in network intrusion datasets? Class imbalance is a significant challenge in network intrusion detection, where the number of benign instances far exceeds the number of attack instances. This imbalance can lead to biased models that favor the majority class, resulting in poor detection rates for minority classes (i.e., various types of attacks). Traditional machine learning algorithms may struggle to learn the characteristics of minority classes, leading to high false negative rates.
2. What role does feature selection using random oversampling play in enhancing detection accuracy? Existing intrusion detection methodologies frequently exhibit performance variations due to their dependence on specific data distributions and feature sets. These variations can lead to inadequate detection scores, which diminish the effectiveness of the systems in real-world scenarios. Inconsistent true positive and false positive rates further complicate the evaluation of model performance, posing significant risks in practical applications where timely and accurate threat detection is critical.

3. Which feature extraction methods are most effective for dimensionality reduction? The highlighted challenges necessitate the development of innovative techniques and methodologies that can enhance the accuracy, robustness, and efficiency of intrusion detection systems. By focusing on creating adaptive models that can handle data imbalance and incorporate diverse datasets, researchers can improve network security measures and better mitigate the impact of evolving cyber threats, ensuring that detection systems remain effective in dynamic environments.

1.5 Hypothesis

This research aims to enhance the accuracy and robustness of intrusion detection systems (IDS) by employing various machine learning algorithms on large and imbalanced datasets, addressing the critical issue of data imbalance that can lead to variations in model performance, particularly in false positive and negative rates. It seeks to develop comprehensive benchmark datasets that reflect a wide range of network events and behaviors, facilitating systematic evaluation of IDS in real-world scenarios where data imbalance and computational efficiency are paramount, the research hypothesesare formulated as follows:

1. Hypothesis 1 (Validity Testing). Applying Random Oversampling (RO) for handling data imbalance, combined with feature selection using Random Forest and feature extraction using Principal Component Analysis (PCA), will significantly improve the accuracy and performance of machine learning models in Network Intrusion Detection Systems (NIDS).
2. Hypothesis 2 (Reliability Testing). Optimizing machine learning models such as KNN, logistic regression, decision tree, random forest, XGB, adaboost, light GBM, cat boost, and Extra Tree will enhance detection accuracy, potentially reaching up to 99% in identifying network intrusions.

1.6 Assumption

Some basic assumptions relevant to this study need to be clarified. These assumptions are derived from general and specific issues related to the design of security systems and are based on the belief that the network environment, software, and technological integration function smoothly as expected. Based on these considerations, the assumptions of this research are as follows:

1. Effectiveness of Machine Learning is assumed that machine learning (ML) algorithms can significantly enhance the accuracy and robustness of intrusion detection systems (IDS) when applied to large and imbalanced datasets.
2. Benchmark datasets is assumed that the CIC-IDS2017, CIC-IDS2018, and UNSW-NB15 datasets are representative of real-world network traffic and provide a comprehensive basis for evaluating the performance of various ML algorithms in detecting intrusions.
3. Performance Metrics is assumed that the selected performance metrics (accuracy, precision, recall, F1-score, ROC curve, and confusion matrix) are adequate for evaluating the effectiveness of the proposed ML models in the context of network intrusion detection

1.7 Scope and Delimitation

This study focuses on the development and evaluation of machine learning techniques to enhance network intrusion detection systems (IDS). It emphasizes the application of various algorithms, such as KNN, logistic regression, decision tree, random forest, XGB, adaboost, light GBM, cat boost, and Extra Tree etc, to improve detection accuracy and efficiency against cyber threats. The research also explores methods for feature selection and dimensionality reduction to optimize model performance while addressing challenges like overfitting

1. This research specifically focuses on the application of various machine learning algorithms, including KNN, logistic regression, decision tree, random forest, XGB, adaboost, light GBM, cat boost, and Extra Tree, to improve and optimize network intrusion detection.
2. This research specifically focuses on data imbalance using Random Oversampling (RO) and employs feature selection via random forest and extraction, also using dimensionality reduction using Principal Component Analysis (PCA) to enhance model performance and accuracy.
3. This research specifically focuses on creating a model that is efficient and accurate in handling large and unbalanced datasets in the context of network intrusion detection.
4. This research specifically focuses on modifications to the feature extraction and maximum optimization of each model and improves the accuracy of each model to as high as 99%

Table 1. 1 Research Plan and Action Point

No	Name of Activity	Month										
		Apr	May	Jun	Jul	Agt	Sep	Okt	Nov	Des	Jan	Feb
1	Requirements Analysis and Finding Dataset											
2	Designing the Network Intrusion Detection System											
3	Model Selection and Configuration											
4	Implementation of Improvements											
5	Performance Evaluation and Result Analysis											
6	Thesis Document and Evaluation											

1.8 Importance of the Study

This study significantly advances the field of network security by exploring the integration of Machine Learning (ML) techniques with Intrusion Detection Systems (IDS) to enhance threat detection capabilities. The findings provide valuable insights into how ML algorithms can be effectively utilized to identify and respond to a wide range of cyber threats, including both known and emerging attack vectors. By leveraging advanced data processing and feature selection methods, this research aims to improve the accuracy and efficiency of IDS, making them more resilient against sophisticated attacks.

The implications of this study are far-reaching, offering practical applications and strategies for enhancing network security through the adoption of ML-driven IDS solutions. The specific contributions are as follows:

1. Developing a Robust ML-Based IDS Framework: This study presents a comprehensive framework for designing an ML-based IDS that can adapt to the dynamic nature of network environments. It addresses the challenges of integrating various ML algorithms and optimizing them for real-time threat detection. The research provides actionable recommendations for building a flexible and scalable IDS that can evolve alongside emerging technologies and threat landscapes, ensuring sustained protection against cyber threats.
2. Enhancing Detection of Anomalous Behavior: The study investigates how ML algorithms can be employed to detect complex and evolving attack patterns, such as Distributed Denial of Service (DDoS) attacks, phishing attempts, and insider threats. By evaluating the performance of different ML models on diverse datasets, the findings will assist network security professionals in

enhancing their IDS capabilities to recognize both established and novel attack signatures, thereby improving overall threat detection efficacy.

3. Facilitating Timely and Effective Incident Response: This research explores the role of ML in enabling timely and effective responses to detected threats. It emphasizes the importance of developing intelligent response mechanisms that can automatically adjust based on the severity of the threat while minimizing disruptions to legitimate network activities. The study highlights strategies for implementing adaptive response protocols that enhance the resilience of network operations during security incidents.
4. Optimizing Data Processing and Feature Selection Techniques: The study delves into the technical aspects of data preprocessing and feature selection, demonstrating how these processes can significantly impact the performance of ML-based IDS. By providing practical examples of effective feature engineering and data augmentation techniques, the research aims to equip organizations with the tools necessary to improve their data analysis capabilities. This optimization will lead to more accurate threat detection and a deeper understanding of attack patterns, ultimately enhancing the overall security posture of network environments.

Through these contributions, this study aims to provide a solid foundation for the integration of ML techniques in IDS, offering valuable insights and practical guidance for organizations seeking to bolster their network security measures in an increasingly complex threat landscape.