Cybersecurity Threat Detection

Hepple Xi
Department of Computer Science
Western University
London, Canada
hxi26@uwo.ca

Jinqi Liu
Department of Computer Science
Western University
London, Canada
jliu3422@uwo.ca

Cheng Guo
Department of Computer Science
Western University
London, Canada
cguo266@uwo.ca

***Abstract*—Cybersecurity threats are rapidly evolving, presenting significant challenges to conventional detection methods, especially regarding the management of imbalanced network traffic data. This research critically examines the effectiveness of various machine learning approaches—including Logistic Regression, Support Vector Machines (SVM), Random Forest, XGBoost, Neural Networks, and a proposed Ensemble method—on the UNSW-NB15 dataset. Through rigorous feature selection and systematic performance evaluation, this study identifies XGBoost as the top-performing model, achieving notable accuracy and balanced detection metrics. The introduced Ensemble approach demonstrates robust and stable performance by leveraging complementary strengths across individual classifiers. This work provides valuable insights and establishes benchmarks for future cybersecurity solutions, emphasizing advanced feature engineering, balanced classification strategies, and ensemble modeling. Further research will explore refined hyperparameter optimization and advanced ensemble frameworks to enhance real-world cybersecurity effectiveness.**

***Index Terms*—Cybersecurity, Intrusion Detection, Machine Learning, Ensemble Model, UNSW-NB15 Dataset**

I. Introduction

The increasing prevalence of cybersecurity threats poses a severe challenge to digital infrastructure, particularly as attack techniques become more sophisticated and widespread. Effective intrusion detection systems (IDS) have become indispensable in safeguarding networks, especially in the context of growing attack volumes and the proliferation of zero-day exploits [1]. Traditional IDS methods often fail to adapt to dynamic traffic patterns, struggle with imbalanced datasets, and exhibit limitations in generalizing to unseen threats.

Machine learning (ML) models have emerged as promising alternatives that enable data-driven approaches to anomaly and threat detection. However, despite their potential, current ML-based solutions often suffer from inconsistent performance across different attack types, underutilization of ensemble strategies, and limited evaluation in comprehensive datasets. Furthermore, deep learning models are frequently deployed without adequate comparative benchmarking, leaving a gap in understanding their practical efficacy versus traditional methods.

This study addresses these challenges by systematically evaluating six widely used ML models: Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, Neural Networks, and Ensemble Model, using the publicly available UNSW-NB15 dataset. The UNSW-NB15 dataset was selected for this study due to its realism, diversity, and suitability for modern intrusion detection research. Unlike older datasets such as KDDCup99 [2] or NSL-KDD, UNSW-NB15 captures realistic network traffic generated by IXIA PerfectStorm and labeled using Bro IDS, offering a more accurate reflection of contemporary Internet behavior. It includes a wide range of attack categories, such as DoS, Exploit, and Reconnaissance, and provides 49 well-structured features without missing values. Its balanced structure, high data quality, and public availability make it an ideal benchmark for evaluating traditional and deep learning-based detection models under practical conditions. We compare models across several key performance metrics: Accuracy, Precision, Recall, F1-score, AUROC, and AUPRC. The primary objectives are to assess the relative performance of these models, mitigate the effects of class imbalance, and explore the viability of ensemble learning for robust attack detection.

The experimental results show that XGBoost achieves the best overall performance with an accuracy of 89.62% and an F1-score of 91.09%. Random Forest and Logistic Regression also deliver strong, balanced results. Neural Networks yield high recall (96.18%) but suffer from lower precision, reflecting a trade-off in misclassification costs. A majority-vote ensemble strategy is also introduced, wherein a traffic instance is only classified as malicious when four of five models agree, aiming to enhance reliability.

The novelty of this work lies in its thorough, cross-model benchmarking on a real-world, imbalanced dataset, coupled with the integration of an ensemble voting strategy that prioritizes conservative decision-making. This research contributes to both theoretical and applied domains: it provides actionable insights for selecting and tuning detection models, and proposes a framework for combining diverse classifiers to boost trustworthiness in critical applications.

The rest of this report is organized as follows: Section II presents the background and related works. Section III outlines the dataset, preprocessing techniques, and experimental setup. Section IV details the performance evaluation and comparative result analysis of the models. Section V discusses the conclusion, future works and lessons learned.

II. Background&Related Work

Cybersecurity plays a pivotal role in protecting digital infrastructure across governments, enterprises, and academic

institutions. With the increasing sophistication of attack techniques and the rising frequency of breaches, the need for effective threat detection has become more urgent than ever. A recent study by Sommer and Paxson [3] reviewed anomaly-based intrusion detection systems (IDS) and highlighted their potential to uncover novel threats, but also emphasized limitations such as high false-positive rates and poor adaptability to evolving traffic patterns.

Numerous examples underscore the real-world consequences of cyberattacks. In 2021, the Colonial Pipeline ransomware attack disrupted fuel supply chains across the U.S. East Coast [4]. In 2023, MOVEit Transfer software was exploited, exposing sensitive data from government agencies and corporations [5]. If an academic platform like Brightspace were compromised, it could lead to leaked grades, identity theft, or manipulated coursework, posing major risks to educational integrity and privacy.

Motivated by these concerns, this project aims to develop a machine learning-based detection system capable of accurately classifying network traffic and quickly identifying malicious activities. Such a tool could be instrumental in preventing escalation and responding rapidly to cyber threats, thereby enhancing real-time defense mechanisms.

Previous research has applied various models to intrusion detection tasks. For instance, Mukkamala and others [6] evaluated neural networks and SVMs on KDDCup99, showing promising results, though limited by synthetic and outdated data. Similarly, Moustafa and Slay [7], the creators of the UNSW-NB15 dataset, demonstrated that tree-based models could perform well but lacked ensemble or deep learning exploration. Other works (e.g., [8], [9]) have explored hybrid approaches but often lacked systematic benchmarking across models.

This research addresses the gap by providing a comparative evaluation of five machine learning methods, including traditional classifiers, tree ensembles, and deep learning on a real-world, imbalanced dataset. We further extend prior work by integrating a voting-based ensemble model that enhances decision reliability through multi-model agreement.

III. Methods

This section outlines the methodological framework adopted in this study for cybersecurity threat detection using machine learning. The workflow includes defining specific research objectives, selecting and describing the dataset, applying robust data preprocessing and feature selection techniques, training six representative classification models, and finally evaluating their performance using a combination of statistical metrics and visual tools.

Our methodology is designed to support both academic exploration and practical applicability: from selecting statistically meaningful features to ensuring model robustness under imbalanced conditions. Each component of the pipeline is described in detail in the following subsections.

A. Research Objectives

a) Objective 1: Evaluate the Effectiveness of Classical and Advanced Machine Learning Models in Cyber Threat Detection: The first objective is to conduct a comparative study of traditional classifiers (e.g., Logistic Regression, SVM) and advanced models (e.g., XGBoost, Neural Network, Ensemble methods) to determine their effectiveness in detecting network intrusions. This has both theoretical and practical importance: understanding which models are best suited for high-dimensional, imbalanced cybersecurity datasets contributes to machine learning literature, while also guiding practical deployments in real-world intrusion detection systems (IDS).

b) Objective 2: Identify the Most Informative Features Using Statistical Techniques for Cybersecurity Applications: This objective aims to apply statistical methods such as ANOVA and Pearson correlation analysis to select a minimal but highly discriminative set of features from the UNSW-NB15 dataset. Reducing feature space enhances model interpretability, improves training efficiency, and prevents overfitting. This supports academic efforts in explainable AI and aids practitioners in deploying lightweight and interpretable IDS models.

c) Objective 3: Develop an Ensemble-Based Threat Detection Framework with Enhanced Precision and Robustness: The third objective is to construct an ensemble model that combines multiple classifiers to achieve higher accuracy, precision, and robustness against different types of network attacks. This aligns with current trends in robust AI systems and provides practical value by reducing false alarms and improving detection consistency in diverse network environments.

B. Dataset

The dataset used in this study is the UNSW-NB15, a modern and comprehensive benchmark dataset for intrusion detection research, developed by the Australian Centre for Cyber Security in 2015. Unlike older datasets such as KDDCup99 and NSL-KDD, UNSW-NB15 was designed to closely mimic real-world network environments by generating realistic traffic using the IXIA PerfectStorm tool and capturing it through the Bro (Zeek) Intrusion Detection System.

The dataset includes a diverse set of 49 features(Table I simply lists 6 of them) extracted from both packet headers and flow metadata, covering aspects such as connection duration, source/destination byte counts, time-to-live values, flag combinations, and statistical behaviors across recent connection windows. Importantly, each record is labeled as either normal (label = 0) or attack (label = 1).

TABLE I: Descriptions of Some Example Features

Feature Name	Description
srcip	Source IP address
dstip	Destination IP address
dur	Record Total duration
proto	Transaction protocol
service	http, ssh, dns, and (-) if not much used service
label	0 for normal and 1 for attack records

A total of 257,673 records are provided, split into: Training set: 175,341 records, Testing set: 82,332 records. The data distribution of Training Set and Testing Set is shown in the following Fig 1.

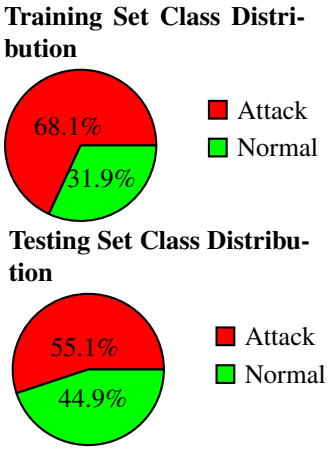


Fig. 1: Class distribution in training and testing subsets of the UNSW-NB15 dataset.

Figure 2 illustrates the Pearson correlation coefficients between all numeric features in the UNSW-NB15 dataset. The heatmap visualizes the strength and direction of linear relationships among features, with red indicating strong positive correlation, blue indicating strong negative correlation, and white representing near-zero or no correlation. Diagonal entries are all 1.0, as each feature is perfectly correlated with itself.

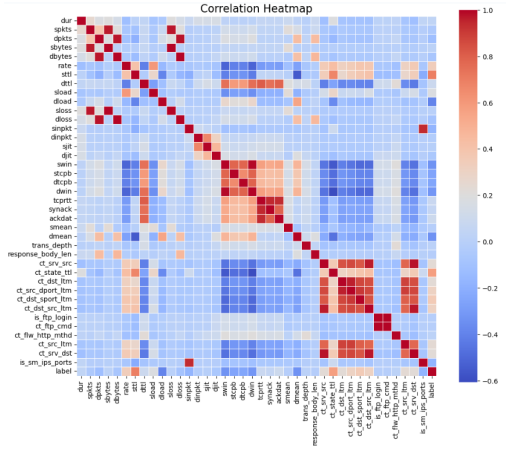


Fig. 2: Correlation heatmap of numeric features in the UNSW-NB15 dataset.

The attacks are classified into 9 types, including DoS, Exploits, Reconnaissance, Fuzzers, Generic, Backdoor, Shellcode, Worms, and Analysis, representing a wide spectrum of contemporary threats, they all labeled as 1 in the dataset.

C. Data Preprocessing

From the heatmap, it is apparent that several traffic-related statistical features such as `dpkts`, `spkts`, `sbytes`, `dbytes`, and `rate` exhibit high positive correlation with one another, with coefficients often exceeding 0.8. These highly correlated features introduce redundancy, which may lead to multicollinearity and negatively affect model stability and generalization. To mitigate this, only one representative feature from such correlated clusters was retained during feature selection.

Moreover, features associated with network path or connection state—such as `ct_state_ttl`, `sttl`, and `dttl`—show moderate to strong positive correlation with the target variable `label`, with coefficients typically ranging between 0.3 and 0.6. This suggests a meaningful statistical relationship with the label, indicating their relevance for classification.

In addition, features such as `sbytes`, `rate`, and `ct_dst_ltm` also show noticeable correlation with the label, reinforcing their importance in attack detection and justifying their inclusion in the final candidate feature set.

On the other hand, protocol flag-related features such as `is_ftp_cmd` and `ct_flw_http_mthd` exhibit near-zero correlation with most features and the label, indicating limited statistical significance and suggesting that they may be excluded from modeling.

In summary, the Pearson correlation heatmap provides a valuable quantitative and visual reference for identifying redundant or weakly informative features, guiding effective feature engineering and ensuring the robustness and efficiency of the subsequent learning models.

Before model training, comprehensive data preprocessing was conducted to ensure data quality and enhance model performance. The following steps were systematically applied:

First, non-numeric features such as IP addresses, timestamps, and protocol identifiers were removed, retaining only numerical attributes that could be directly utilized by machine learning models. This step eliminates categorical noise and simplifies downstream processing.

Second, a missing value analysis was performed across all features. It was confirmed that the UNSW-NB15 dataset contains no missing entries, ensuring the integrity of the training and testing splits without the need for imputation.

Third, to improve model efficiency and reduce overfitting, a two-stage feature selection strategy was employed.

- **Correlation Analysis:** Pearson correlation coefficients between each feature and the binary label were calculated. Features with absolute correlation values close to zero were considered weakly informative and were candidates for exclusion. Highly correlated feature pairs were also reviewed to avoid multicollinearity.
- **ANOVA Score Selection:** An ANOVA F-test (`f_classif`) was used to evaluate the statistical significance of each feature concerning the class labels. Features with an F-score greater than 2000 were retained,

indicating strong discrimination power between attack and normal traffic.

Finally, features were selected based on a combined criterion:

- ANOVA F-score > 2000
- Absolute Correlation > 0.1

This dual filtering ensures that only features that are both statistically significant and meaningfully correlated with the classification task are used for model training. A table(Table II) summarizing the selected features and their correlation and ANOVA scores was compiled to guide the modeling phase.

Through this preprocessing pipeline, the dimensionality of the data was reduced without sacrificing too much important information, leading to improved model training efficiency and potentially better generalization on unseen traffic samples.

TABLE II: Selected Features Based on ANOVA and Correlation

Feature Name	ANOVA Score	Correlation
dpkts	2501.116244	-0.118591
rate	22611.817828	0.337979
sttl	161780.526952	0.692741
sload	6066.435816	0.182870
dload	32170.312851	-0.393739
sinpkt	5612.172675	-0.176110
swin	21961.796620	-0.333633
stcpb	12194.973688	-0.255006
dtcpb	11723.168274	-0.250340
dwin	19905.909669	-0.319626
dmean	23195.039749	-0.341808
ct_srv_src	9707.760648	0.229044
ct_state_ttl	87830.779493	0.577704
ct_dst_ltm	9783.377995	0.229887
ct_src_dport_ltm	18059.210546	0.305579
ct_dst_sport_ltm	25645.955324	0.357213
ct_dst_src_ltm	17835.399757	0.303855
ct_src_ltm	10549.386809	0.238225
ct_srv_dst	9618.746606	0.228046
is_sm_ips_ports	6191.368613	-0.184679

D. Models

To systematically investigate the applicability and effectiveness of different classification paradigms in cybersecurity threat detection, this study selects six widely used and theoretically grounded machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, a Multi-layer Perceptron Neural Network (implemented in TensorFlow), and a majority-vote based Ensemble Model. These models span linear classifiers, kernel-based methods, ensemble learning techniques, and deep learning architectures, enabling comprehensive horizontal comparison from both interpretability and performance perspectives.

Logistic Regression is included as a baseline due to its simplicity and high interpretability. While its linear decision boundary limits performance in non-linear scenarios, it remains a valuable benchmark and a tool for feature impact analysis in intrusion detection tasks.

SVM, based on margin maximization theory and structural risk minimization, is suitable for high-dimensional and sparse domains. Its kernel trick allows non-linear classification without explicit transformation of the feature space, making it a classic and strong baseline in security-related classification.

Random Forest and **XGBoost** represent two dominant forms of ensemble learning. Random Forest employs bootstrap aggregation to reduce overfitting and improve generalization, while XGBoost leverages gradient boosting and regularization to optimize residual errors iteratively. XGBoost has been widely adopted in security competitions and industry settings for its strong handling of class imbalance and high-dimensional data.

Neural Network uses a fully connected feedforward architecture to capture complex, non-linear patterns. Although performance may degrade with limited feature richness, it serves as a reference point for deep learning capacity in tabular datasets.

The Ensemble Model aggregates predictions from the five individual classifiers using a majority-vote scheme. A sample is labeled as an attack only if the prediction meets a specified vote threshold. This fusion strategy aims to reduce individual model variance and enhance prediction robustness for deployment in real-world systems.

E. Training Procedure and Strategy

All models were trained using feature subsets filtered by a hybrid method combining Pearson correlation analysis and ANOVA F-test. This dual strategy ensures that selected features are statistically significant and relevant for classification.

For models sensitive to feature scaling (e.g., Logistic Regression, SVM), standardization was applied using `StandardScaler`. Tree-based models such as Random Forest and XGBoost were trained using default parameters due to their scale invariance and robustness.

The neural network consisted of three dense layers with 128, 64, and 32 neurons, respectively, each followed by ReLU activation. Dropout and Batch Normalization were applied to prevent overfitting and stabilize training. The model was optimized using the Adam optimizer with binary cross-entropy loss. To handle class imbalance, class weights were applied, and early stopping was used on a 20% validation split to monitor convergence.

The training process was visualized in terms of both loss and accuracy over 200 epochs, as shown in Figure 3. The training loss decreased steadily and plateaued after approximately 100 epochs, while the validation loss showed fluctuations yet maintained a downward trend overall, indicating good generalization with minimal overfitting. The accuracy curves further confirm effective learning: validation accuracy rapidly increased during early epochs and stabilized above 95%, consistently outperforming training accuracy. These curves demonstrate that the neural network successfully converged and maintained strong generalization capability.

The ensemble model loaded the five trained base classifiers and aggregated their predictions via majority voting. The decision threshold was adjustable (e.g., predicting ”attack”

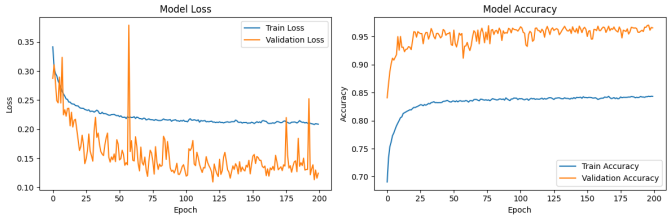


Fig. 3: Training and validation loss and accuracy curves of the neural network over 200 epochs.

only if at least four classifiers agree), allowing control over precision-recall balance in deployment.

IV. Results

To comprehensively assess model performance on the intrusion detection task, we adopted six commonly used classification metrics: Accuracy, Recall, Precision, F1-Score, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC). These metrics capture both general and class-specific predictive performance, particularly in the context of imbalanced datasets such as UNSW-NB15.

- **Accuracy**: Accuracy measures the proportion of total correct predictions over all samples. While widely used, accuracy alone may be misleading in imbalanced settings where the majority class dominates, masking poor performance on the minority (attack) class:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: the proportion of correctly predicted positive instances among all predicted positives, which indicates the proportion of predicted attack labels that are truly attacks. A model with high recall but low precision will produce many false positives, potentially overwhelming security analysts. Hence, precision helps evaluate the trustworthiness of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** (also called Sensitivity or true positive rate): the proportion of correctly predicted positive instances among all actual positives, which quantifies the proportion of actual attack samples correctly identified by the model. High recall is critical in cybersecurity, where failing to detect an attack (false negative) is typically more harmful than raising a false alarm:

$$\text{Recall} = \frac{TP}{TP + FN}$$

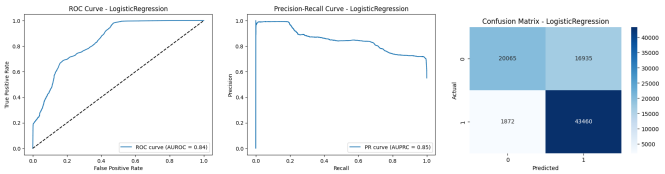
- **F1-Score**: the harmonic mean of Precision and Recall, providing a balanced measure that penalizes extreme disparities between them. It is especially useful when seeking a balance between missing threats and generating false alerts:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **AUROC** (Area Under the ROC Curve): the area under the Receiver Operating Characteristic (ROC) curve, which plots True Positive Rate (Recall) vs. False Positive Rate (FPR). It reflects the model's ability to distinguish between classes across different thresholds.
- **AUPRC** (Area Under the Precision-Recall Curve): the area under the curve that plots Precision vs. Recall. This metric is particularly informative in imbalanced settings, as it focuses on the performance of positive class prediction.

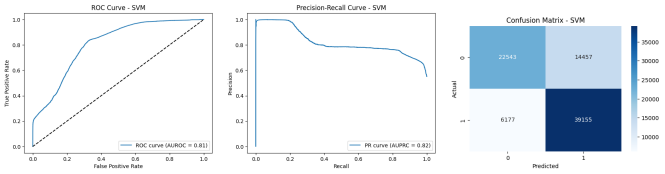
The selection of these metrics was motivated by the practical demands of real-world intrusion detection: minimizing false negatives (missed attacks), controlling false positives (alarm fatigue), and ensuring robust performance regardless of class distribution. By jointly analyzing these six metrics, we obtain a nuanced and application-relevant understanding of each model's strengths and weaknesses.

To comprehensively evaluate the performance of all classification models, we present the Confusion Matrices, ROC curves, and PRC (Precision-Recall) curves for each model. These visualizations offer insight into detection quality, class separation capabilities, and false alarm behavior. The results are illustrated in Figures 4 to 9.



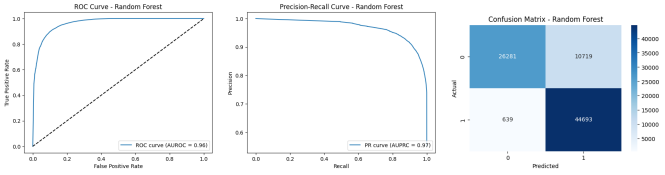
(a) ROC Curve (b) PRC Curve (c) Confusion Matrix

Fig. 4: Evaluation Results of Logistic Regression



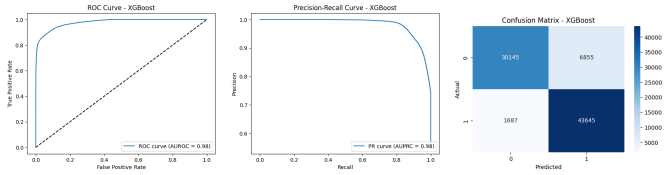
(a) ROC Curve (b) PRC Curve (c) Confusion Matrix

Fig. 5: Evaluation Results of Support Vector Machine (SVM)



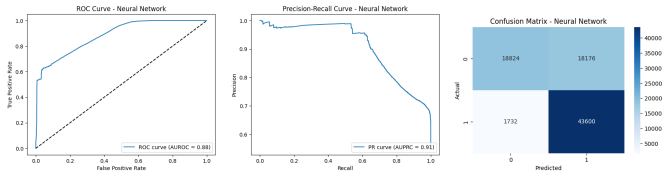
(a) ROC Curve (b) PRC Curve (c) Confusion Matrix

Fig. 6: Evaluation Results of Random Forest



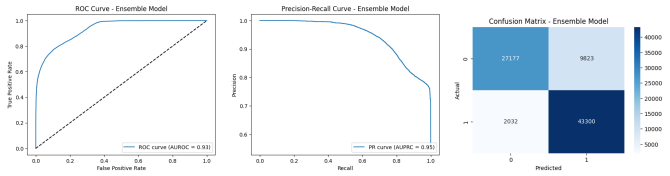
(a) ROC Curve (b) PRC Curve (c) Confusion Matrix

Fig. 7: Evaluation Results of XGBoost



(a) ROC Curve (b) PRC Curve (c) Confusion Matrix

Fig. 8: Evaluation Results of Neural Network



(a) ROC Curve (b) PRC Curve (c) Confusion Matrix

Fig. 9: Evaluation Results of Ensemble Model

Experimental results demonstrate that **XGBoost** achieves the highest overall performance across key metrics, confirming its effectiveness in capturing non-linear decision boundaries under class imbalance. **Random Forest** performed consistently and was easier to tune, making it advantageous in engineering applications. **Logistic Regression** offered reasonable accuracy and excellent interpretability, suggesting suitability for lightweight or explainable security systems.

SVM performance was acceptable but highly dependent on kernel choice and hyperparameters, with training time as a limiting factor. **Neural Network** achieved the third highest recall but at the cost of low precision, indicating a tendency to over-detect attacks, which is acceptable in high-sensitivity scenarios. The **Ensemble Model** did not achieve the best score in any single metric but delivered stable and balanced results across all metrics, validating the benefit of combining diverse decision boundaries in security-critical contexts.

The performance of all six models was evaluated using the selected metrics discussed previously. Table III presents the results across Accuracy, Recall, Precision, F1-Score, AUROC, and AUPRC for each classifier.

From the table, **XGBoost** emerges as the top-performing individual model, achieving the highest accuracy (89.62%), precision (86.43%), F1-score (91.09%), and tied highest AUROC and AUPRC (0.98). This confirms its strength in handling imbalanced, tabular data and learning complex decision boundaries.

TABLE III: Performance Comparison of All Models on UNSW-NB15 Dataset

Model	Accuracy	Recall	Precision	F1 Score	AUROC	AUPRC
SVM	0.7494	0.8637	0.7303	0.7915	0.81	0.82
Random Forest	0.8620	**0.9859**	0.8066	0.8873	0.96	0.97
XGBoost	**0.8962**	0.9628	**0.8643**	**0.9109**	**0.98**	**0.98**
Neural Network	0.7582	0.9618	0.7058	0.8141	0.88	0.91
Logistic Regression	0.7716	0.9587	0.7196	0.8221	0.84	0.85
Ensemble Model	0.8560	0.9552	0.8151	0.8796	0.93	0.95

Random Forest also performs strongly, particularly in recall (98.59%), demonstrating its capability in identifying attack samples with high sensitivity, while maintaining good precision and overall balance. The **Ensemble Model** achieved a well-rounded performance, especially in AUPRC (0.95), indicating reliable detection capability across all thresholds. Although it does not surpass XGBoost in individual metrics, it delivers consistent and stable results, reinforcing the benefits of model aggregation.

Neural Network attained the third highest recall (96.18%) but at the cost of lower precision (70.58%), highlighting a tendency to over-detect attacks and raise false alarms. **Logistic Regression** maintained moderate performance with decent recall and simplicity, while **SVM** achieved solid recall but lower precision and accuracy, possibly due to kernel sensitivity and hyperparameter tuning limitations.

These results suggest that tree-based ensemble methods (XGBoost and Random Forest) are highly effective for this intrusion detection task, offering both high accuracy and strong recall. Neural network approaches can enhance recall but require further refinement for precision control. The ensemble strategy proves valuable in producing robust and balanced outcomes, suitable for real-world applications where both detection coverage and decision confidence are critical.

V. Conclusions&Future Work

This study conducted a comprehensive evaluation of machine learning models for cybersecurity threat detection on the UNSW-NB15 dataset, emphasizing robustness, interpretability, and performance under data imbalance. Through rigorous feature selection and systematic evaluation, tree-based ensemble methods—particularly XGBoost—demonstrated superior effectiveness, achieving the highest accuracy (89.62%) and F1-score (91.09%). Random Forest and Logistic Regression also provided reliable performance with strong interpretability, while Neural Networks, despite achieving high recall, suffered from lower precision, highlighting trade-offs inherent in deep learning models with limited features.

The results underscore three key insights: (1) feature engineering and model selection critically influence detection efficacy under class imbalance; (2) accuracy alone is insufficient for model evaluation, necessitating complementary metrics such as AUROC and AUPRC; and (3) ensemble strategies can enhance system stability, even if individual models outperform them in isolated metrics.

Future research will focus on two primary directions. First, model-specific hyperparameter tuning through grid search and cross-validation will be pursued to further optimize performance. Second, the ensemble strategy will be refined by exploring soft-voting, stacking, and adaptive aggregation techniques. Additionally, extending the system to handle real-time streaming data and applying online learning paradigms will enhance its applicability in dynamic threat environments.

Overall, this work provides a reproducible framework for evaluating cybersecurity detection models and offers actionable guidance for building resilient and trustworthy intrusion detection systems in real-world settings.

REFERENCES

[1] M. Soltani, B. Ousat, M. J. Siavoshani, and A. H. Jahangir, "An adaptable deep learning-based intrusion detection system to zero-day attacks," *arXiv preprint arXiv:2108.09199*, 2021.

[2] Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set." Information Security Journal: A Global Perspective (2016): 1-14.

[3] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *IEEE Symposium on Security and Privacy*, 2010.

[4] Cybersecurity and Infrastructure Security Agency (CISA), "Colonial Pipeline Ransomware Attack Overview", 2021.

[5] TechCrunch, "MOVEit breach hits US agencies, corporations", 2023.

[6] S. Mukkamala, G. Janoski, and A. H. Sung, "Intrusion detection using neural networks and support vector machines," in *IEEE Int. Joint Conf. Neural Networks*, 2002.

[7] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)."Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

[8] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications*, 2014.

[9] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Expert Systems with Applications*, 2009.