Journal of Business and Digital Innovation (JBDI)
Volume 1, Issue 1, January 2026, PP 86-103
ISSN 2978-4522 (Print) & ISSN 2978-4530 (Online)
https://doi.org/10.19164/jbdi.v1i1.1818

Comparative Analysis of Traditional Machine Learning, Deep Learning, and Hybrid Ensemble Models for Anomaly Detection and Web Application Firewall Optimisation

Shiva Nezamzadeh & Dilek Celik
Northumbria University, London

Abstract
Anomaly detection is an important component of cybersecurity, particularly in safeguarding web application firewalls (WAFs) from malicious traffic. In this study, we perform a comparative analysis of three Machine Learning (ML) approaches: Random Forest (RF), Convolutional Neural Network (CNN), and a stacking ensemble combining RF and CNN with Logistic Regression (LR) as the meta-learner to explore the most effective approach for anomaly detection. To ensure a fair comparison, we trained all models under consistent preprocessing pipelines, including data class balancing using the SMOTE technique to address the common imbalance in attack data. The results of this study showed that the stacking ensemble outperformed the other models, achieving the highest accuracy (99.97%). The CNN model followed closely with comparable accuracy (99.94%), while also offering significant advantages in terms of computational efficiency and interpretability, particularly when supplemented with SHAP analysis. In contrast, the RF model achieved moderate accuracy (80.41%) but demonstrated strengths in interpretability and efficiency. These findings highlight that, with effective preprocessing, a standalone CNN can provide a practical and resource-efficient alternative to more complex ensemble models. The findings of this study highlight the importance of preprocessing in optimising model performance and propose CNN as a suitable solution for real-time cybersecurity applications. Future research should explore these models across diverse datasets, further investigate hybrid deep learning (DL) frameworks, and integrate advanced interpretability methods to enhance model transparency and trust in ML-based security systems.

Keywords: Machine Learning, Deep Learning, Hybrid Ensemble Model, Anomaly Detection, Web Application Firewall Optimisation
Wordcount: 232

1.0 Introduction

The role of anomaly detection using ML models in network traffic cannot be overstated, particularly in light of increasingly sophisticated cyberattacks that are testing conventional models of security. Indeed, it can be seen that there is an urgent requirement for intelligent security models that can adapt to changing security challenges. WAFs are also of critical utility in network security, as they continually filter HTTP traffic in search of malicious traffic that can compromise web applications. However, it must also be recognized that there are limitations in terms of security that are tied to the models of detection that are adopted. Indeed, there have been several suggestions that ML can be of significant utility in terms of making WAF more responsive, particularly in terms of unearthing sophisticated cyberattacks that are using zero-day exploits and hidden attack models that are not easily detectable using conventional models of security, as has been underscored in previous studies (Nassif et al., 2021; Nti et al., 2022; Alghanmi, Alotaibi, & Buhari, 2022); however, there are also certain limitations in terms of models that can yield more practical results, including ML, DL, and ensemble models.

This study examines the effectiveness of three ML model categories of traditional ML algorithms, DL models, and hybrid ensemble approaches to enhance anomaly detection within WAFs. Accuracy, precision, recall, F1-score, and other performance metrics including computational efficiency, feature significance, and interpretability are used to assess the models. By applying consistent preprocessing and evaluation criteria, the study seeks to reveal the optimal model for enhancing WAF performance. This work contributes to the field of study by providing a systematic, side-by-side comparison of various traditional ML, DL, and hybrid ensemble models under a unified preprocessing, feature-selection, and evaluation pipeline, facilitating fair and reproducible performance assessment.

The main aim of this research is to increase the security of web applications through ML. Specifically, it compares various model categories in their ability to distinguish between legitimate and malicious traffic, providing insights into their practicality for real-world implementation.

1.2. Research objectives:

- To distinguish the strengths and limitations of traditional ML, DL, and hybrid ensemble models for anomaly detection in WAF.
- To identify the most effective model to enhance WAF security.
- To assess how different anomaly detection techniques impact model performance and efficiency.

1.3. Research Questions:

- How do traditional ML, DL, and hybrid ensemble models differ in their capacity to detect anomalies in web application traffic?
- Which model offers the best balance of accuracy, computational efficiency, and interpretability for anomaly detection in WAF?

The remainder of this paper is organised as follows: Section 2 provides details about the literature review, outlining existing approaches, key challenges, and research gaps. Section 3 presents the methodology, including the dataset, preprocessing steps, model selection, and evaluation metrics. Section 4 provides the experimental setup and model performance results. Section 5 delivers a comparative analysis, discussing each model's strengths, limitations, and practical implications. Finally, Section 6 concludes this research and offers recommendations for future research.

2.0. Literature Review

2.1. Anomaly Detection

Anomaly detection has evolved from statistical techniques to adaptive ML methods that can identify previously unseen attacks (Kalariya, Jethva and Alginahi, 2024). By implementing these techniques, Betarte, Pardo and Martinez (2018) proposed a model that uses HTTP header and payload attributes, achieving 98.4% accuracy on the OWASP dataset, highlighting the importance of resilient detection protocols.

2.2 Traditional Machine Learning Models and Their Ensembles in WAFs

Traditional ML models—such as RF, LR, Support Vector Machines (SVM), and Gradient Boosting—have been widely used in WAFs detect threats due to their simplicity and interpretability. SVM performs very well with high-dimensional data but may become inefficient with larger datasets. RF is a powerful method for network traffic classification, and it is robust against overfitting and high accuracy; however, it requires considerable computational resources, which may impact real-time applicability (Alserhani and Aljared, 2023; Acito, 2023). Ensemble models such as RF and Gradient Boosting can further improve accuracy by reducing false positives (FPs), which makes them suitable for WAFs. Gradient Boosting improves performance with sequential optimisation, but it is prone to overfitting and requires longer training times (Athief, Kishore and Paranthaman, 2024; Acito, 2023). Despite such limitations, ensemble approaches often outperform individual models, which are more vulnerable to high FP rates (Alserhani and Aljared, 2023).

Bagging, Stacking, and AdaBoost are common ensemble methods which improve classification performance by emphasising misclassified instances, a property particularly suited to binary anomaly detection tasks. However, AdaBoost can be sensitive to noisy data and outliers, which may negatively affect performance in real-world network traffic (Jeffrey et al., 2024; Odeh and Taleb, 2024). Advanced ensemble approaches combining multiple models have demonstrated strong performance. For instance, Tama et al. (2020) proposed a stacked ensemble integrating RF, Gradient Boosting Machine, and XGBoost, which achieved improved accuracy and reduced false positives across two datasets using grid search and k-fold cross-validation. Although computationally intensive, such approaches can deliver substantial performance gains.

Overall, traditional ML models such as RF, SVM, and LR continue to play a key role in web security due to their interpretability and efficiency. While these models face challenges with high-dimensional or large-scale data and potential overfitting, ensemble methods help mitigate many of these limitations. Nonetheless, ensuring robustness and effective integration across diverse models remains a challenge in ensemble design. These characteristics of traditional ML models and their ensemble strategies are summarised in Figure 1.

Figure 1. Traditional ML (Conceptual diagram and annotations created by the authors; embedded image reproduced from CrowdforThink, 2019*)*

2.3 DL Approaches and Their Ensembles in WAFs

DL models have become more important in improving WAFs. Tian et al. (2020) suggested a distributed DL framework for detecting online attacks on resource-limited edge devices. They showed that DL can support scalable and real-time threat detection while maintaining high accuracy. Among widely implemented DL models, Long Short-Term Memory (LSTM) networks have demonstrated strong performance in capturing long-term sequential patterns. However, despite achieving high accuracy in DDoS detection (97.57%), LSTM models are prone to overfitting and require substantial computational power, which can limit their practicality in real-time WAF implementations (Dawadi, Adhikari and Srivastava, 2023; Ali et al., 2022).

CNN is another DL model that shows strong relevance to cybersecurity. Although CNNs were originally developed for image processing, they have been effectively used for malware and network traffic analysis by automatically learning discriminative characteristics. Nevertheless, their high computing requirements can limit suitability for real-time WAF deployment, especially when detecting infrequent attack patterns (Ali et al., 2022; Kimanzi et al., 2024).

Although DL ensembles have shown performance gains in WAF-based intrusion detection, their effectiveness closely depends on careful model selection and integration. For instance, PANACEA—an ensemble that combined CNNs, RNNs, and Autoencoders—achieved high detection accuracy but faced difficulties in identifying the dominant contributing model. Similarly, ResNet-18 has reported standalone classification accuracy of around 77%, which highlights the significance of selecting robust base models when designing ensemble architectures (AL-Essa et al., 2024; Tan, 2023; Li et al., 2023; Alanazi et al., 2022).

To manage the complexity of DL ensembles, various aggregation strategies have been proposed; model performance and computing limitations are usually used to guide selection (Waheed et al., 2023). Average voting, majority voting, and optimal weighting are common approaches. Average voting is frequently stated to deliver more stable performance by balancing model predictions, while majority voting offers simplicity at the cost of decreased sensitivity. By giving each model a certain level of priority, optimal weighting can increase accuracy even more, although it adds more computing overhead (Alanazi et al., 2022).

In summary, DL models such as LSTM and CNN have enhanced WAF capabilities by achieving high detection accuracy. However, their practical deployment in real-time environments can be constrained by overfitting risks and computational demands. While DL ensembles can further improve performance, they introduce additional complexity in model selection and system integration. These strengths and limitations of DL approaches are summarised in Figure 2.

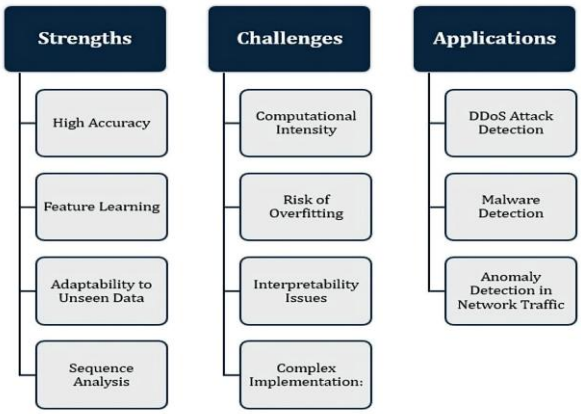

Figure 2. DL approaches (Conceptual diagram created by the authors)

2.4 Ensemble Techniques combining Traditional and DL models

Hybrid ensemble methodologies, which merge conventional ML and DL models, leverage the synergistic benefits inherent in both approaches, demonstrating considerable promise for anomaly detection within WAFs. Although DL models possess robust feature-learning capabilities, they frequently exhibit high computational demands. Traditional ML models, including RF, Gradient Boosting, and SVM, can augment DL models by providing enhanced efficiency and interpretability. Through feature-level or decision-level fusion, hybrid ensembles can attain elevated classification accuracy while addressing the constraints of individual models (Tan, 2023; Ovi, Rahman and Hossain, 2024).

A pertinent illustration is the hybrid ensemble introduced by Abdelmounaim and Madani (2024), which integrated traditional ML models (XGBoost and RF) with DL models (CNN and RNN) employing a stacking architecture. Their methodology exhibited superior classification accuracy compared to standalone models, thereby indicating improved robustness and adaptability within intrusion detection applications.

Notwithstanding these advantages, hybrid ensembles present increased complexity. The heightened model heterogeneity introduces difficulties in training, tuning, and deployment, especially within real-time or resource-limited settings (Xu et al., 2024). Moreover, ensemble strategies like stacking and blending could potentially amplify the vulnerabilities of individual models if the integration process is not meticulously planned (Li et al., 2023).

Overall, hybrid ML–DL ensemble models provide significant performance enhancements for cybersecurity applications; however, challenges concerning interpretability, scalability, and deployment intricacy could restrict their practical utility in specific WAF implementations.

2.5 Research Gap

This review of existing research critically examines important approaches to anomaly detection using ML, including traditional ML algorithms, DL architectures, and combined ensemble methods. While previous research has studied the strengths and weaknesses of these methods separately, there is a lack of research that systematically compares all three within a single experimental setup for WAF optimisation (Xu et al., 2024; Alanazi et al., 2022; Ali et al., 2022). This research gap makes it challenging to fully understand the trade-offs between prediction accuracy, computational efficiency, and practical use. To address this, this study gives a comparative evaluation of models from each category, using consistent preprocessing and evaluation methods. By clarifying the relative advantages of each method, this research aims to help choose scalable, reliable, and understandable ML solutions to improve WAF-based cybersecurity.

3.0. Research Methodology and Planning

3.1. Research Design

This study, which uses a quantitative and comparative approach, evaluates different ML models for finding anomalies, following the CRISP-DM framework. The research is based on positivism, assuming that how well models perform provides an objective view of reality. It also takes a realist approach, considering anomaly detection as something that can be observed and understood through data analysis. Additionally, the study is empirically driven, aiming to create knowledge through systematic experimentation and quantitative analysis. Methodologically, the research follows the CRISP-DM process, which includes five main stages: (1) Data Understanding, (2) Data Preparation, (3) Modelling, (4) Evaluation, and (5) Interpretation (Gill et al., 2023; Brodie, 2024). (Figure 3)

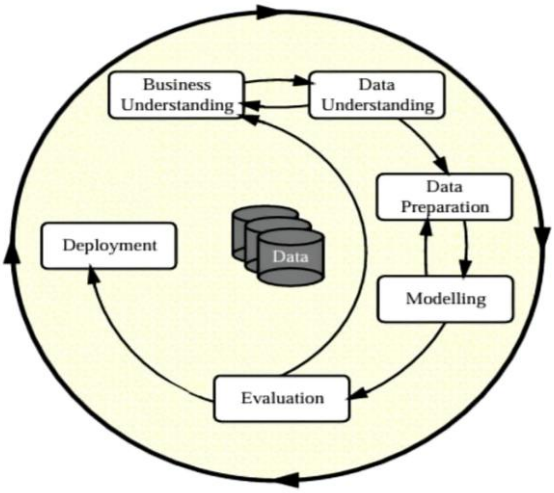

Figure 3. CRISP-DM process (Wirth and Hipp, n.d.)

3.2. Data Collection Methods and Data Understanding

This study employs the UNSW_NB15 dataset, a publicly accessible resource acquired from Kaggle, and extensively utilised in academic investigations pertaining to intrusion detection. The dataset offers distinct training and testing files, each encompassing in excess of 50,000 network traffic instances, thereby facilitating robust model training and evaluation for the purpose of anomaly detection. All records within the dataset are completely anonymised, ensuring the absence of any personally identifiable information. Given the absence of human subjects and the non-utilisation of sensitive personal data, prior ethical approval was deemed unnecessary for the present study.

3.3. Data Processing and Preparation

3.3.1 Dataset Overview

The dataset includes 45 features, including 41 numerical and 4 categorical attributes. The binary target variable (Label) denotes normal traffic (0) and attacks (1). No missing or duplicate values were found.

3.3.2 Data Engineering

One-hot encoding was used to represent categorical features, while numerical features were standardised using Z-score normalisation.

3.3.3 Outlier Handling and Correlation Analysis

Initial exploratory analysis included visual inspections (histograms, boxplots, scatter plots) and correlation assessment to understand feature behaviour. Since the dataset contains non-Gaussian distributions and a mixture of linear and non-linear relationships, model-based methods were prioritised for robust outlier handling and variable relevance assessment.

3.3.4 Model-based feature selection

To improve model performance and address dataset complexity, RF and XGBoost were employed for feature selection and outlier detection. Feature importance scores from these models were used to rank attributes, with attack_cat_Normal identified as having the strongest influence on the target variable (Label). Instances misclassified by either model were flagged as potential outliers. Crucially, no extreme values were removed, capped, or otherwise altered, as they were considered intrinsic to the dataset and representative of authentic network behaviour. Preserving these extreme values allows the models to learn from the complete spectrum of normal and malicious traffic, which is essential for realistic intrusion detection. Combined with model-driven feature selection, this approach reduced dataset dimensionality and optimised it for subsequent modelling. To address class imbalance, SMOTE and undersampling techniques were used. This study combined RF and XGBoost with both SMOTE and undersampling to create a balanced dataset optimised for anomaly detection (Vibhute et al., 2024; Pansari et al., 2024; Putra, 2024).

3.4. Data Analysis Methods

The main objective of this study is to identify the most suitable model from each category (traditional ML, DL, and hybrid ensembles). Only ensemble-based models were selected from the ML and DL categories to avoid bias from comparing single models to ensembles. Collectively, these models were chosen to represent complementary strengths across interpretability (RF), deep feature learning (CNN), and heterogeneous model integration

(stacking), thereby ensuring a systematic comparison across traditional ML, DL, and hybrid ensemble paradigms. RF, an ensemble CNN, and a stacking approach were chosen as representative models based on their characteristics, computational requirements, and suitability for the dataset. RF was used as the traditional machine learning ensemble baseline due to its robustness to noisy and high-dimensional data, strong resistance to overfitting, interpretability, and comparatively low computational complexity (Panasov and Nechitaylo, 2021), making it a well-established benchmark in intrusion detection research.

An ensemble CNN was chosen to represent DL-based methods because of its ability to automatically learn hierarchical and non-linear feature representations, while ensemble aggregation enhances generalisation performance. CNN ensembles have consistently achieved high detection accuracy and demonstrated strong scalability in network intrusion detection, although they are associated with higher computational intensity (Yang, Lv and Chen, 2022). Stacking was employed as a hybrid ensemble learning representative to explicitly evaluate whether combining heterogeneous models with fundamentally different architectures can enhance detection performance. Unlike voting or boosting techniques, stacking enables a meta-learner to learn optimal combinations of base model predictions. In this study, logistic regression (LR) was adopted as a lightweight meta-learner to minimise additional complexity while integrating RF and CNN outputs (Zhang and Wang, 2023).

Other candidate models, such as support vector machines, gradient boosting techniques, and recurrent neural networks, were excluded because they either conceptually overlapped with the selected approaches or relied on temporal assumptions outside the scope of this study. This selection strategy enables a focused yet representative comparison across traditional ML, DL, and hybrid ensemble paradigms for intrusion detection (Azam et al., 2023; Javaid et al., 2016).

3.5. Validation and Testing

Since the dataset provided distinct training and testing files, data splitting was inherently addressed, and preprocessing was performed separately on each subset. Stratified K-Fold Cross-Validation was used during model training to preserve class balance and reduce overfitting (Chen et al., 2023). Hyperparameter tuning was conducted using GridSearchCV, which also incorporated stratified folds. Although computationally intensive, Grid Search was selected for its reliability in cybersecurity contexts, where accuracy is critical. Given the computational demands of the CNN model, a reduced hyperparameter grid was applied. This ensured adequate optimisation without imposing excessive resource demands (Masum et al., 2021; Franceschi et al., 2024).

Model performance was evaluated using a comprehensive set of metrics to ensure balanced and reliable comparisons. Accuracy, Precision, Recall, and F1-Score were used to assess predictive performance. These metrics capture the trade-off between FPs and false negatives (FNs), which is crucial in anomaly detection (Yang et al., 2023; Li et al., 2023). ROC-AUC was applied to evaluate the models' ability to discriminate across thresholds. Balanced Accuracy was included to address class imbalance and ensure fair evaluation across both classes (Owusu-Adjei et al., 2023). To strengthen reliability, agreement metrics such as Matthews Correlation Coefficient (MCC) and Cohen's Kappa were incorporated (Chicco and Jurman, 2020). Additional error-based metrics including Specificity, Fall-out (FPR), Negative Predictive Value (NPV), False Discovery Rate (FDR), and False Omission Rate (FOR) were used to capture the impact of misclassification (Larner, 2024).

Finally, confusion matrices were generated to provide a clear breakdown of classification outcomes. To assess computational efficiency and interpretability, inference speed, memory usage, and SHAP values were included (Wang and Wang, 2022). RF was further assessed through model complexity, permutation importance, and feature importance (Ahsan et al., 2021), and model size was specifically measured for CNN (Saleem et al., 2022). Results were visualised to provide a clear comparison of overall model efficiency.

4.0. Results and Analysis

4.1. Explanatory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to identify key patterns in the dataset and guide subsequent modelling decisions. The distribution of the Label variable indicated a clear class imbalance, with malicious traffic dominating in both the training and test datasets (Figure 4).

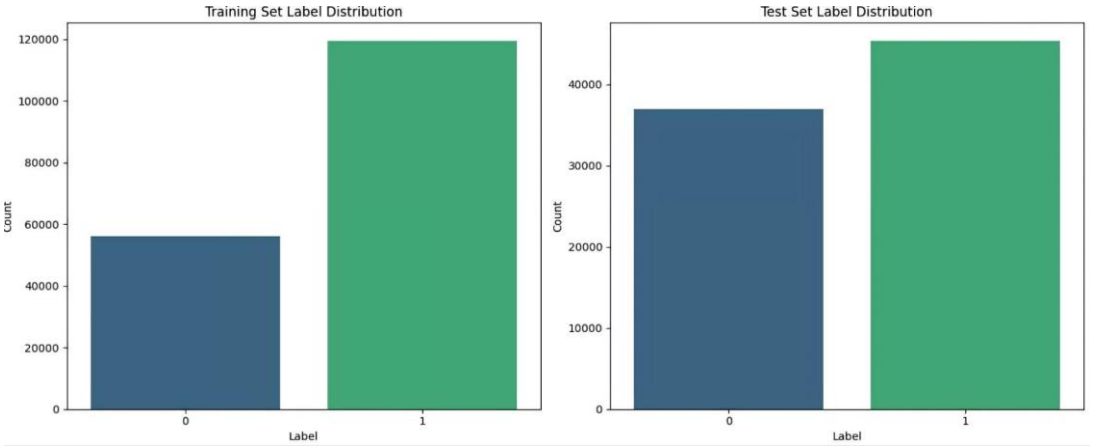

Figure 4. Label distribution of the training and test datasets

Correlation analysis showed that sttl (source time-to-live) and ct_dst_sport_ltm (count of connections to the destination over time) had the strongest positive associations with the Label variable. In contrast, swin (source window size) and dload (data transmission rate from destination to source) displayed the strongest negative correlations (Figure 5). Similar correlation trends were observed in the test dataset, indicating consistent feature behaviour across data splits.

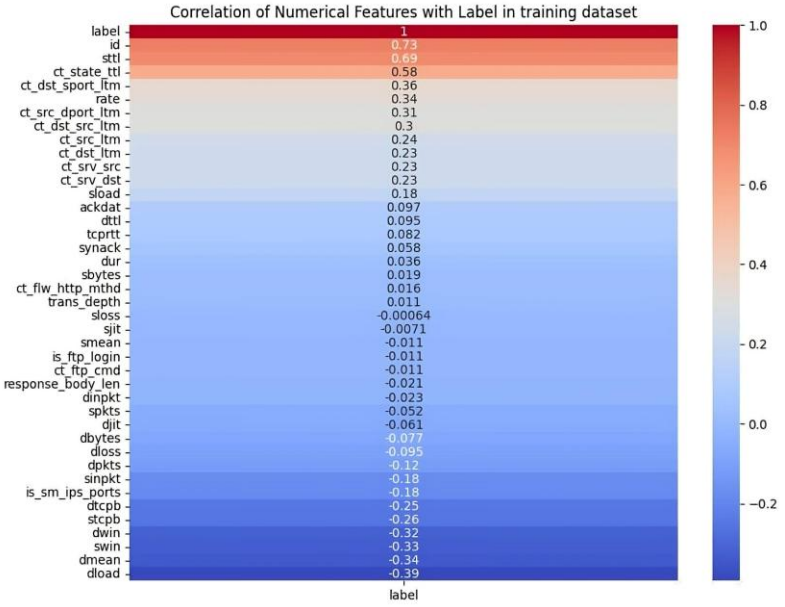

Figure 5. Correlation of 'Label' with other features in the training dataset.

Additional analyses, including service type distributions and attack category comparisons, were performed to further characterise the dataset. Overall, the EDA identified meaningful patterns and confirmed the dataset's complexity, underscoring the need for robust preprocessing prior to model training.

4.2 Outlier handling, Correlation Analysis, Model-based feature selection

4.2.1 Outlier Detection, Distribution, and Correlation Analysis

Exploratory analysis using histograms, boxplots, and scatter plots indicated that many numerical features deviated from normality, with substantial skewness and kurtosis. This observation motivated the use of non-parametric methods in subsequent analysis. Figure 6 presents a representative example using the spkts feature (the number of packets transmitted from the source during a network flow), which demonstrates a highly skewed distribution. Given these characteristics, Spearman's rank correlation was applied to examine relationships between numerical features and the target variable (Label), revealing varying degrees of association. For categorical variables, ANOVA tests were used to identify both significant and non-significant relationships. These findings informed the feature selection process and subsequent model design.

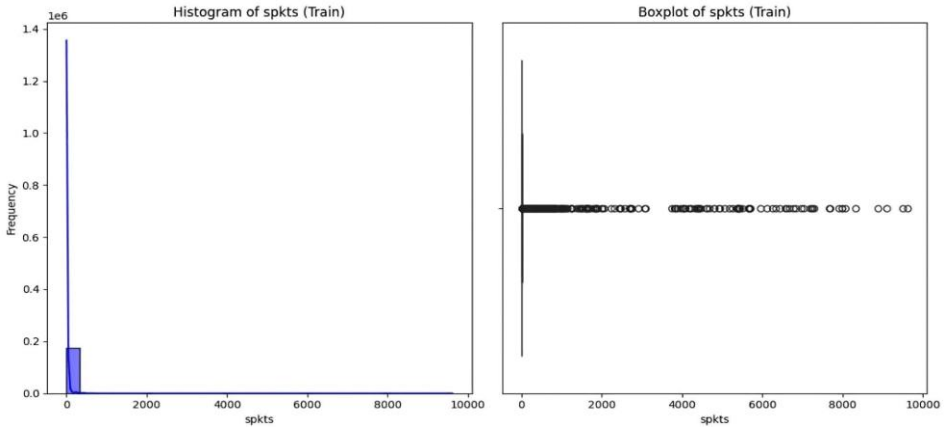

Figure 6. Histogram and Boxplot of 'spkts' in the training dataset

4.2.2 Model-based feature selection

RF and XGBoost were utilised for feature selection and outlier analysis to improve model performance and manage dataset complexity. Feature importance rankings from these models were used to prioritise the most influential attributes, with attack_cat_Normal identified as the most dominant predictor of the target variable (Label). Potential outliers were identified through the models' misclassification signals. Since neither model flagged extreme values for removal, all observations were retained, suggesting that the extreme values are intrinsic to the dataset. This approach reduced dataset dimensionality and prepared the data for subsequent modelling. Furthermore, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training dataset to address class imbalance.

4.3 Models Application

4.3.1 Random Forest

The RF model was optimised using GridSearchCV with Stratified K-Fold Cross-Validation, as outlined in Section 3. It achieved an overall accuracy of 80.4% and a balanced accuracy of 78.4%, indicating consistent performance despite class imbalance. In cybersecurity contexts, such results constitute a reasonable baseline, although recall and discrimination capability are more critical than accuracy alone. The RF model achieved a recall of 80.4% and a ROC–AUC score of 0.94, which demonstrated strong power to distinguish between normal and anomalous traffic (Table 1). Confusion matrix analysis (Table 2) confirms effective identification of attack instances with a relatively low number of false negatives, although some benign traffic was misclassified.

Table 1. RF model evaluation metrics

Metrics	Value
Accuracy	0.8041
Balanced Accuracy	0.7842
Precision	0.8422
Recall	0.8041
F1- Score	0.7939
MCC	0.6336
Cohen's Kappa	0.5895
Specificity	0.5873
ROC AUC Score	0.9420
Sensitivity (TPR)	0.8041
FPR	0.4127
NPV	0.9620
FDR	0.2556
FOR	0.0380

Table 2. RF model confusion matrix

Metrics	Value
TP	44,473
TN	21,731
FP	15,269
FN	859

From a computational perspective, the RF model achieved high inference speed and low memory consumption (Table 3), owing to its shallow tree structure. This highlights its suitability for real-time or resource-constrained deployment scenarios.

Table 3. RF model computational metrics

Metric	Value
Inference Speed	6,305,931.79 samples per second
Memory Usage	491.4296875
Number of Estimators	5
Maximum Depth	3
Maximum Features	sqrt

Model interpretability was examined by using built-in feature importance, permutation importance, and SHAP analysis. Across all approaches, sttl and attack_cat_Normal emerged as the most influential features, indicating transparent model behaviour. The SHAP interaction analysis further illustrated the combined influence of these features on prediction outcomes (Tables 4–5; Figure 7).

Table 4. RF model feature importance based on the built-in feature importance

Feature	Importance
Sttl	0.202839
Ct_state_ttl	0.184023
Sload	0.155977
Dload	0.127539
rate	0.106322

Table 5. RF model feature importance based on permutation

Feature	Permutation Importance
Attack_cat_Normal	0.112580
State_INT	0.032521
Smean	0.017683
Tcprtt	0.015555
demean	0.012592

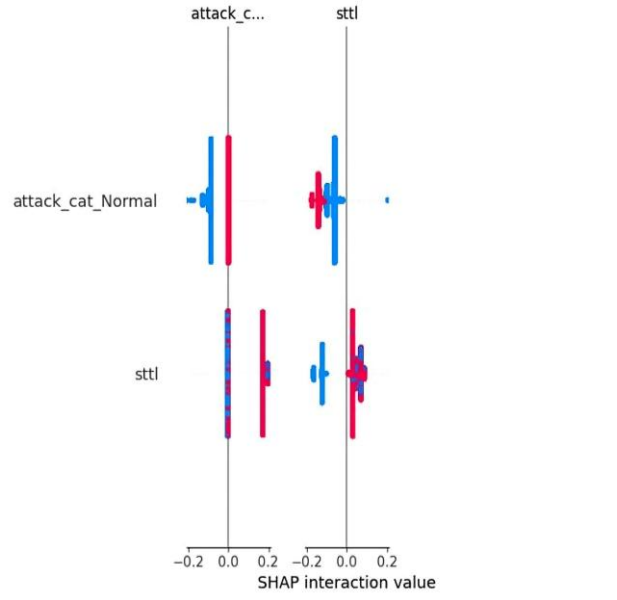

Figure 7. RF model SHAP interaction plot

4.3.2 CNN

To ensure reproducibility, random seeds were fixed for both NumPy and TensorFlow, and deterministic data splitting was applied throughout the experimental pipeline. The CNN model operated on a reshaped one-

dimensional feature vector using a single Conv1D layer with 32 filters and a kernel size of 3, followed by batch normalisation and max pooling with a pool size of 2. A global average pooling layer was then applied to summarise feature maps and reduce model complexity. The extracted features were passed to a fully connected dense layer with eight neurons and ReLU activation, followed by a dropout layer with a rate of 0.5 for regularisation. The final classification layer used a softmax activation function.

Hyperparameter tuning was performed using stratified three-fold cross-validation to evaluate combinations of learning rate, number of filters, and dropout rate. Learning rate 0.0005, 32 filters, and dropout rate 0.5, which were selected based on mean cross-validation accuracy, were the optimal configuration. The model was subsequently trained using an 80/20 training–validation split for 10 epochs. Analysis of the learning curves showed close alignment between training and validation accuracy and loss, indicating stable convergence and no evidence of overfitting. (Figure 8)

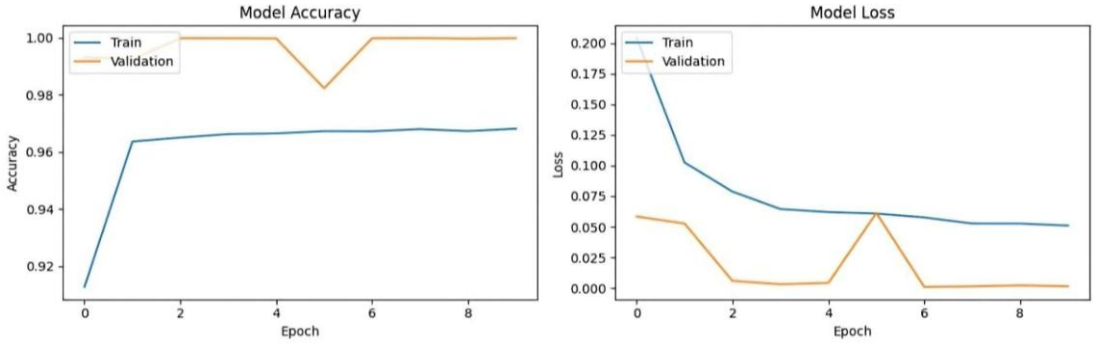

Figure 8. Accuracy and Loss of the train and validation set over 10 epochs

The CNN model demonstrated consistently high performance across all evaluation metrics, with accuracy, precision, recall, and F1-score all exceeding 99% (Table 6). The balanced accuracy of 99.81% and ROC–AUC of 0.9999 indicate excellent discrimination between normal and attack traffic. Confusion matrix results (Table 7) reveal minimal false positive and false negative rates, which highlight the model's robustness in detecting both attack and benign traffic.

Table 6. CNN model evaluation metrics

Metrics	Value
Accuracy	0.9994
Balanced Accuracy	0.9981
Precision	0.9994
Recall	0.9994
F1- Score	0.9994
MCC	0.9963
Cohen's Kappa	0.9963
Specificity	0.9993
ROC AUC Score	0.9999
TPR	0.9994
FPR	0.0007
NPV	0.9993
FDR	0.0006
FOR	0.0007

Table 7. CNN model confusion matrix

Metrics	Value
TP	45,306
TN	36,973
FP	27
FN	26

The CNN model demonstrated inference performance suitable for large-scale data processing. While its memory usage was higher than that of traditional ML models, the relatively small model size reflects an efficient architecture given its strong predictive performance (Table 8).

Table 8. CNN model computational metrics

Metric	Value
Inference Speed	19311.66
Average Memory Usage	5297.43
Model Size (MB)	0.05

SHAP analysis was used to examine feature contributions. The results identified *sttl* and *attack_cat_Normal* as the most influential features shaping the CNN model's predictions (Table 9; Figure 9).

Table 9. CNN model feature importance based on SHAP analysis

Feature	Importance Value
Sttl	0.174911
Attack_cat_Normal	0.173987
Ct_state_ttl	0.160816
Dload	0.160719
Attack_cat_Fuzzers	0.082288

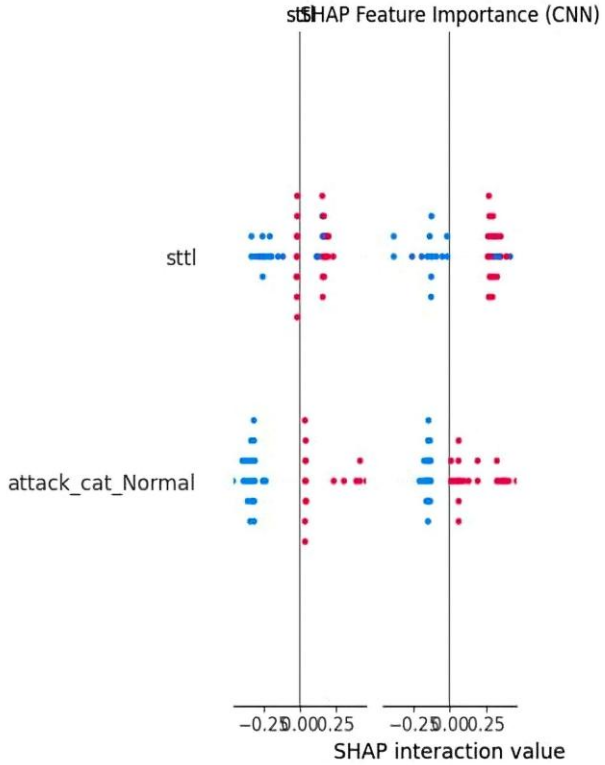

Figure 9. CNN model SHAP Interaction plot

4.3.3 Stacking hybrid ensemble

The stacking hybrid ensemble combined two trained base models: an RF, which generated class probabilities, and a CNN, which produced predictions. These outputs were combined into a new feature set used to train a lightweight LR meta-model. The ensemble delivered near-perfect classification performance, with accuracy, precision, recall, and F1-score all reaching approximately 99.97% (Table 10). Moreover, high MCC and Cohen's Kappa values indicate excellent agreement and robustness. As shown in the confusion matrix (Table 11), misclassification rates were minimal, with only 5 false positives and 21 false negatives, confirming the ensemble's detection capability.

Table 10. Hybrid Ensemble evaluation metrics

Metrics	Value
Accuracy	0.9997
Balanced Accuracy	0.9997
Precision	0.9997
Recall	0.9997
F1- Score	0.9997
MCC	0.9994
Cohen's Kappa	0.9994
Specificity	0.9999
ROC AUC Score	0.9997
TPR	0.9995
FPR	0.0001
NPV	0.9994
FDR	0.0001
FOR	0.0006

Table 11. Hybrid Ensemble confusion matrix

Metrics	Value
TP	45,311
TN	36,995
FP	5
FN	21

The computational characteristics of the stacking ensemble were analysed to complement the performance evaluation (Table 12). The reported inference speed, memory usage, and model size reflect the computational behaviour of the ensemble architecture itself and should be interpreted as indicative rather than exhaustive, as the overall computational intensity depends on the underlying base models.

Table 12. Hybrid Ensemble computational metrics

Metric	Value
Inference Speed	60657902.15
Average Memory Usage	5885.76
Model Size	19.05

Permutation importance analysis was used to examine feature contributions within the stacking ensemble (Figure 10). The results indicate that the CNN-derived features dominate the ensemble's decision-making. This confirms that the ensemble's predictions are primarily driven by the CNN component rather than an equal contribution from both base models.

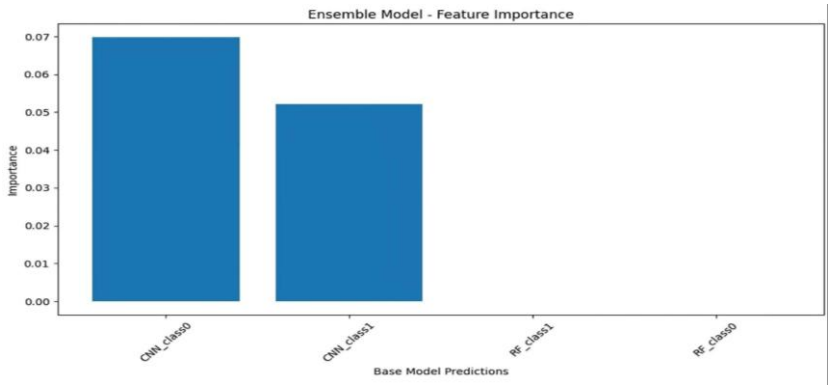

Figure 10. Hybrid Ensemble Feature Importance based on Permutation

5.0. Discussion and Implications

5.1 Model Performance

Figures 11–13 present a comparative analysis of the three evaluated models. Both the CNN model (accuracy: 99.94%) and the ensemble model (accuracy: 99.97%) illustrate similarly high levels of accuracy, substantially outperforming the RF model (accuracy: 80.41%). While the RF model achieved reasonably good performance, it was clearly less effective than the other two models.

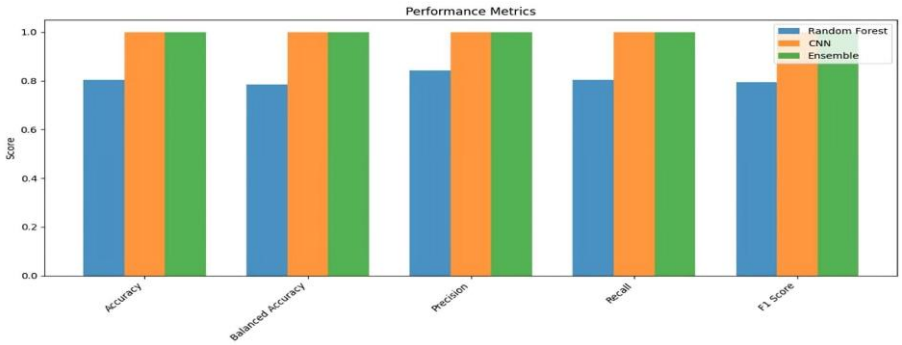

Figure 11. Comparison of models: Evaluation Metrics

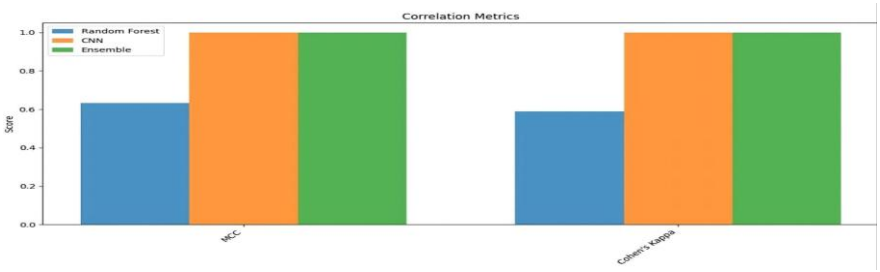

Figure 12. MCC and Cohen's Kappa comparison of models

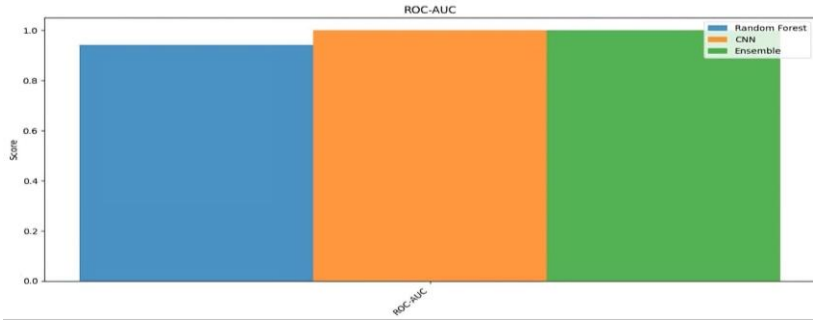

Figure 13. ROC-AUC comparison of models

Table 13 summarises the computational intensity metrics for each model. It is important to note that the reported inference speed of the ensemble model may not accurately reflect its true computational cost, as it is influenced by the combined performance of its constituent models. With this consideration, the ensemble model exhibits the highest computational intensity, likely due to its layered architecture. The CNN model shows moderate computational demand, while the RF model remains the most computationally efficient model.

Table 13 Comparison of computational metrics of models

Model	Inference Speed	Average Memory Usage	Model Size	Additional Details
RF	6,305,931.79	491.43	N/A (Tree-based)	5 estimators, max depth 3
CNN	19,311.66	5,297.43	0.05	N/A
Ensemble	60,657,902.15	5,885.76	19.05	These results should be interpreted alongside the computational characteristics of the RF and CNN base models

The overview of the feature importance and model interpretability is provided in Table 14. As shown, *'stl'* and

'attack_cat_Normal' emerged as the most influential features in both the RF and CNN models. The RF model is inherently interpretable, owing to its tree-based structure, allowing for clear insight into feature contributions. In contrast, the CNN model's deeper architecture introduces challenges in interpretability. The stacking ensemble model further increases the model complexity by integrating multiple learners. Nonetheless, analysis revealed that the CNN was the dominant model driving the ensemble's predictions.

Table 14. Comparison of feature importance and interpretability of models

Model	Key Contributing Features	Interpretability Assessment
RF	'sttl', 'attack_cat_Normal'	High, due to the intrinsic feature-based structure
CNN	'attack_cat_Normal', 'sttl'	Moderate; relies on complex internal representations
Ensemble	'CNN_class0', 'CNN_class1'	Moderate; interpretable via constituent models' contributions

A unified preprocessing and feature extraction was applied across all models, including RF, CNN, and a stacking ensemble with LR as the meta-learner. While both the CNN and the ensemble models demonstrated strong and comparable predictive performance, the RF model delivered comparatively lower results. Feature importance analysis confirmed the substantial contribution of the CNN model to the ensemble's performance. In terms of computational efficiency, the RF model required the lowest computational resources while offering the highest interpretability; by contrast, the CNN and ensemble models demanded greater computational costs, with the ensemble being the most resource-intensive. Despite the inherent complexity of the CNN model, its interpretability can be enhanced through post hoc explanation methods. Overall, the CNN model emerged as the most practical and effective model within the scope of this research.

5.2 Comparison with previous studies

This study employed a unified preprocessing and feature selection pipeline combining XGBoost, RF, and SMOTE across all models to ensure a fair comparison. Inconsistent preprocessing can inflate model-specific performance while weakening comparative validity, whereas systematic pipelines improve generalisation and interpretability (Farha and Ahmed, 2024; Ahsan et al., 2021). Consistent with these findings, Vibhute et al. (2024) demonstrated that RF-based feature selection significantly enhances CNN performance on the UNSW-NB15 dataset, achieving accuracy comparable to more complex ensemble models. Effective preprocessing allows simpler architectures to rival computationally intensive ensembles, a conclusion reinforced by the strong performance of the standalone CNN in this study. Hybrid ensemble models are often employed to improve detection accuracy; however, they frequently bring higher computational cost. For example, Farha and Ahmed (2024) reported resource limitations in a stacking ensemble incorporating an FNN. In contrast, the present study demonstrates that a simpler CNN-based design, when supported by effective preprocessing, can deliver competitive performance without excessive computational demands.

5.3 Contribution to the field

This study advances cybersecurity research by clarifying the relationship between preprocessing strategies, model architecture, and computational efficiency. By standardising preprocessing and feature selection, the research ensured methodological consistency and reduced bias in model comparisons. The integration of traditional ML and DL within the ensemble framework demonstrated that high predictive performance can be achieved without reliance on a complex meta-model. Importantly, the results show that, with well-structured preprocessing and feature extraction, a single CNN model can match the performance of an ensemble model, while requiring lower computational intensity. The application of SHAP and permutation importance enhanced the interpretability of the models. Overall, these findings contribute to a more balanced understanding of how predictive accuracy, interpretability, and computational efficiency can be jointly optimised in the design of practical intrusion detection systems.

5.4 Implications for Business Digital Innovation and Cybersecurity Strategy

Prior research highlights the strategic role of artificial intelligence in strengthening organisational decision-making and digital resilience by improving risk awareness, supporting data-driven management, and guiding cybersecurity investment planning. From a business perspective, AI-driven cybersecurity improves operational efficiency, reduces financial losses and service disruptions, and strengthens regulatory compliance through explainable security mechanisms. Together, these benefits underscore the role of AI-enabled security solutions in enhancing

digital service reliability, fostering organisational trust and long-term competitive advantage (Daram and Senthilkumar, 2025; Sissodia et al., 2025).

Building on this foundation, the findings of this study suggest that selecting an anomaly detection model should be viewed as a strategic business decision rather than a purely technical choice. Models that combine high detection accuracy with manageable computational requirements can reduce infrastructure and operational costs, making them suitable for long-term deployment. Concurrently, stronger detection performance reduces the likelihood of successful cyberattacks, thereby limiting financial losses, service downtime, and reputational damage. This research explicitly balances predictive accuracy, interpretability, and computational efficiency, offering organisations a practical framework for selecting intrusion detection solutions that align with budgetary constraints and risk tolerance. Such a balanced approach supports cost-effective cybersecurity investment and maximises the long-term return on investment in AI-enabled WAF solutions.

5.5. Limitations

Despite the strong performance reported in this study, several limitations should be acknowledged. First, the experimental evaluation was conducted using a single intrusion detection dataset. Although UNSW-NB15 is widely used, model performance may vary across datasets with different traffic characteristics and attack distributions, potentially limiting generalisability.

Second, the results are influenced by the available computational resources. Variations in hardware can affect training efficiency, model complexity, and, to a limited extent, predictive performance; therefore, the findings should be interpreted within the context of the specific experimental setup used in this study.

Third, while representative models from traditional ML, DL, and hybrid ensemble categories were deliberately selected to enable systematic comparison, this choice may introduce methodological bias. Alternative architectures or more complex model configurations may achieve better performance, but their inclusion was constrained by practical computational considerations.

Finally, the exceptionally high accuracy achieved by the CNN and stacking ensemble should be interpreted with caution. Feature importance and SHAP analyses consistently identified *attack_cat_Normal* as a highly influential feature, which may contribute to optimistic performance estimates. This effect was explicitly analysed to ensure transparency. Accordingly, the primary contribution of this study lies in the controlled and fair comparison of traditional ML, DL, and hybrid ensemble models under a unified preprocessing framework, highlighting practical trade-offs between accuracy, interpretability, and computational cost.

6.0. Conclusions and Recommendations

6.1 Conclusion

This study shows that the CNN model achieved the highest overall performance for anomaly detection, with the stacking ensemble delivering comparable accuracy but at a significantly higher computational cost. The RF model indicated acceptable performance but remained less accurate than the DL approaches. Feature importance analyses revealed 'sttl' and 'attack_cat_Normal' as key predictors in both the CNN and RF models, with the ensemble's decisions primarily influenced by the CNN component. Consistent preprocessing using XGBoost and RF for feature extraction and SMOTE for class balancing was critical to achieving fair comparisons and strong model performance. These preprocessing steps enabled the CNN model to perform comparably to more complex models, confirming that effective data preparation can reduce the reliance on computationally intensive designs. Overall, these findings support the selection of intrusion detection solutions that balance detection effectiveness with computational efficiency, enabling sustainable and cost-conscious deployment in real-world environments.

6.2 Recommendation for future findings

Based on the findings of this study, several directions for future research can be identified. First, validating the proposed models on multiple and more diverse intrusion detection datasets would help assess their robustness and improve generalisability across different real-world network environments.

Second, future work could explore more advanced deep learning architectures. Integration of multiple deep learning techniques within a unified framework, such as dual CNN structures or CNN-RNN combinations, may enhance detection performance while maintaining acceptable computational efficiency.

Finally, improving model interpretability represents an important avenue for future study. Investigating advanced explainability methods, including saliency maps, Layer-wise Relevance Propagation (LRP), and attention-based visualisation techniques, could enhance model trustworthiness and support practical deployment (Ahmed and Jalal, 2024; Achtibat et al., 2024; Wang, Ouyang and Zeng, 2024).

References

Abdelmounaim, K. & Madani, M.A. (2024) 'A hybrid ensemble approach integrating machine learning and deep learning with sentence embeddings for webpage content classification', in *Modern Artificial Intelligence and Data Science 2024: Tools, Techniques and Systems*. Cham: Springer Nature Switzerland, pp. 85–97.

Achtibat, R., Hatefi, S.M.V., Dreyer, M., Jain, A., Wiegand, T., Lapuschkin, S. & Samek, W. (2024) *Attnlrp: Attention-aware layer-wise relevance propagation for transformers*. arXiv:2402.05602.

Acito, F. (2023) 'Ensemble models', in *Predictive Analytics with KNIME: Analytics for Citizen Data Scientists*. Cham: Springer Nature Switzerland, pp. 255–265.

Ahmed, M.W. & Jalal, A. (2024) 'Robust object recognition with genetic algorithm and composite saliency map', in *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*. IEEE, pp. 1–7.

Ahsan, M., Gomes, R., Chowdhury, M.M. & Nygard, K.E. (2021) 'Enhancing machine learning prediction in cybersecurity using dynamic feature selector', *Journal of Cybersecurity and Privacy*, 1(1), pp. 199–218.

Alanazi, F., Jambi, K., Eassa, F., Khemakhem, M., Basuhail, A. & Alsubhi, K. (2022) 'Ensemble deep learning models for mitigating DDoS attacks in software-defined networks', *Intelligent Automation & Soft Computing*, 33(2).

Al-Essa, M., Andresini, G., Appice, A. & Malerba, D. (2024) 'PANACEA: A neural model ensemble for cyber-threat detection', *Machine Learning*, pp. 1–44.

Alghanmi, N., Alotaibi, R. & Buhari, S.M. (2022) 'Machine learning approaches for anomaly detection in IoT: An overview and future research directions', *Wireless Personal Communications*, 122(3), pp. 2309–2324.

Ali, R., Ali, A., Iqbal, F., Hussain, M. & Ullah, F. (2022) 'Deep learning methods for malware and intrusion detection: A systematic literature review', *Security and Communication Networks*, 2022(1), p. 2959222.

Alserhani, F. & Aljared, A. (2023) 'Evaluating ensemble learning mechanisms for predicting advanced cyber attacks', *Applied Sciences*, 13(24), p. 13310.

Athief, R., Kishore, N. & Paranthaman, R.N. (2024) 'Web application firewall using machine learning', in *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, pp. 1–7.

Azam, Z., Islam, M.M. & Huda, M.N. (2023) 'Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree', *IEEE Access*, 11, pp. 80348–80391.

Betarte, G., Pardo, Á. & Martínez, R. (2018) 'Web application attacks detection using machine learning techniques', in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 1065–1072.
Brodie, M.L. (2024) *A framework for understanding data science*. arXiv:2403.00776.

Chen, C. et al. (2023) 'Application of GA-WELM model based on stratified cross-validation in intrusion detection', *Symmetry*, 15(9), p. 1719.

Chicco, D. & Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21(1), p. 6.

Choudhury, M.P. & Choudhury, J.P. (2022) 'Machine learning-based model to find out firewall decisions towards improving cyber defence', in *International Conference on Internet of Things and Connected Technologies*. Singapore: Springer Nature Singapore, pp. 179–195.

CrowdforThink (2019) *Bringing AI and machine learning accessible to enterprises: Credit to cloud*. Available at: https://crowdforthink.com/blogs/bringing-ai-and-machine-learning-accessible-to-enterprises-credit-to-cloud

Daram, K. & Senthilkumar, P. (2025) 'Optimizing Cloudflare security and performance with AI-based web application firewall and anomaly detection', *International Journal on Smart Sensing and Intelligent Systems*, 18(1). https://doi.org/10.2478/ijssis-2025-0040

Dawadi, B.R., Adhikari, B. & Srivastava, D.K. (2023) 'Deep learning technique-enabled web application firewall for the detection of web attacks', *Sensors*, 23(4), p. 2073.

Farha, F. & Ahmed, M.U. (2024) 'Heterogeneous ensemble approach in intrusion detection using stacking technique', in *2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS)*. IEEE, pp. 1–6.

Franceschi, L. et al. (2024) *Hyperparameter optimisation in machine learning*. arXiv:2410.22854.

Gandomi, A.H., Chen, F. & Abualigah, L. (2022) 'Machine learning technologies for big data analytics', *Electronics*, 11(3), p. 421.

Gill, M.S. et al. (2023) 'Integration of domain expert-centric ontology design into the CRISP-DM for cyber-physical production systems', in *2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, pp. 1–8.

Javaid, A., Niyaz, Q., Sun, W. & Alam, M. (2016) 'A deep learning approach for network intrusion detection system', *EAI Endorsed Transactions on Security and Safety*, 3(9), p. 21.

Jeffrey, N., Tan, Q. & Villar, J.R. (2024) 'Using ensemble learning for anomaly detection in cyber–physical systems', *Electronics*, 13(7), p. 1391.

Kalariya, P., Jethva, M. & Alginahi, Y. (2024) 'ML assisted web application firewall', in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, pp. 1–6.

Kimanzi, R., Kimanga, P., Cherori, D. & Gikunda, P.K. (2024) *Deep learning algorithms used in intrusion detection systems—A review*. arXiv:2402.17020.

Kook, L. et al. (2022) *Deep interpretable ensembles*. arXiv:2205.12729.

Larner, A.J. (2024) 'Paired measures', in *The 2×2 Matrix: Contingency, Confusion and the Metrics of Binary Classification*. Cham: Springer International Publishing, pp. 17–53.

Li, M., Gao, Q. & Yu, T. (2023) 'Kappa statistic considerations in evaluating inter-rater reliability between two raters', *BMC Cancer*, 23(1), p. 799.

Li, W. et al. (2023) *Deep model fusion: A survey*. arXiv:2309.15698.

Li, Z. et al. (2023) 'Towards inference efficient deep ensemble learning', in *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7), pp. 8711–8719.

Masum, M. et al. (2021) 'Bayesian hyperparameter optimization for deep neural network-based network intrusion detection', in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 5413–5419.

Mi, X., Zou, B., Zou, F. & Hu, J. (2021) 'Permutation-based identification of important biomarkers for complex diseases via machine learning models', *Nature Communications*, 12(1), p. 3008.

Mungoli, N. (2023) *Adaptive ensemble learning: Boosting model performance through intelligent feature fusion in deep neural networks*. arXiv:2304.02653.

Nassif, A.B., Talib, M.A., Nasir, Q. & Dakalbab, F.M. (2021) 'Machine learning for anomaly detection: A systematic review', *IEEE Access*, 9, pp. 78658–78700.

Nti, I.K., Quarcoo, J.A., Aning, J. & Fosu, G.K. (2022) 'A mini-review of machine learning in big data analytics: Applications, challenges, and prospects', *Big Data Mining and Analytics*, 5(2), pp. 81–97.

Odeh, A. & Taleb, A.A. (2024) 'Ensemble learning techniques against structured query language injection attacks', *Indonesian Journal of Electrical Engineering and Computer Science*, 35(2), pp. 1004–1012.

Ovi, M.S.I., Rahman, M.H. & Hossain, M.A. (2024) *PhishGuard: A multi-layered ensemble model for optimal phishing website detection*. arXiv:2409.19825.

Owusu-Adjei, M. et al. (2023) 'Imbalanced class distribution and performance evaluation metrics', *PLOS Digital Health*, 2(11), p. e0000290.

Panasov, V.L. & Nechitaylo, N.M. (2021) 'Decision trees-based anomaly detection in computer assessment results', *Journal of Physics: Conference Series*, 2001(1), p. 012033.

Pansari, N. et al. (2024) 'Attack classification using machine learning on UNSW-NB15 dataset', in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. IEEE, pp. 1–9.

Putra, Z.P. (2024) 'Evaluating the performance of classification algorithms on the UNSW-NB15 dataset for network intrusion detection', *Jurnal Ilmiah FIFO*, 16(1), p. 84.

Saleem, M.A. et al. (2022) 'Comparative analysis of recent architecture of convolutional neural network', *Mathematical Problems in Engineering*, 2022(1), p. 7313612.

Sambandam, R.K. et al. (2023) 'Comparison of machine learning-based intrusion detection systems using UNSW-NB15 dataset', in *International Conference on Artificial Intelligence on Textile and Apparel*. Singapore: Springer Nature Singapore, pp. 311–324.

Sissodia, R. et al. (2025) 'Artificial intelligence (AI) in cybersecurity', in *Advances in Computational Intelligence and Robotics*. Hershey, PA: IGI Global, pp. 121–152.

Stormit.cloud (2022) *StormIT achieves AWS service delivery for AWS WAF*. Available at: https://www.stormit.cloud/blog/aws-waf-service-delivery/

Tama, B.A., Nkenyereye, L., Islam, S.R. & Kwak, K.S. (2020) 'An enhanced anomaly detection in web traffic using a stack of classifier ensemble', *IEEE Access*, 8, pp. 24120–24134.

Tan, P. (2023) *Ensemble-based hybrid optimization of Bayesian neural networks and traditional machine learning algorithms*. arXiv:2310.05456.

Tian, Z. et al. (2019) 'A distributed deep learning system for web attack detection on edge devices', *IEEE Transactions on Industrial Informatics*, 16(3), pp. 1963–1971.

Vibhute, A.D. et al. (2024) 'Network anomaly detection and performance evaluation of convolutional neural networks on UNSW-NB15 dataset', *Procedia Computer Science*, 235, pp. 2227–2236.

Waheed, M. et al. (2023) *An evaluation and ranking of different voting schemes for improved visual place recognition*. arXiv:2305.05705.

Wang, Y. & Wang, X. (2022) *A unified study of machine learning explanation evaluation metrics*. arXiv:2203.14265.

Wang, Z., Ouyang, Y. & Zeng, H. (2024) 'ARFN: An attention-based recurrent fuzzy network for EEG mental workload assessment', *IEEE Transactions on Instrumentation and Measurement*.

Wirth, R. & Hipp, J. (2000) 'CRISP-DM: Towards a standard process model for data mining', in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 1, pp. 29–39.

Xu, M. et al. (2024) *A survey of resource-efficient LLM and multimodal foundation models*. arXiv:2401.08092.

Yang, F. et al. (2023) 'Assessing inter-annotator agreement for medical image segmentation', *IEEE Access*, 11, pp. 21300–21312.

Yang, Y., Lv, H. & Chen, N. (2023) 'A survey on ensemble learning under the era of deep learning', *Artificial Intelligence Review*, 56(6), pp. 5545–5589.

Zhang, X.Y. & Wang, M.M. (2023) 'An efficient combination strategy for hybrid quantum ensemble classifier', *International Journal of Quantum Information*, 21(06), p. 2350027.