Hybrid Ensemble of RF-DNN Model for BENIGN and Attack Traffic Classification in Intrusion System

Farhan Tanvir Ahmed[1]*, and Riaz Mahmood[2]

[1]Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), ECE Building, 1205, Dhaka, Bangladesh
[2]Computer Science and Engineering (CSE), BRAC University, 1212, Dhaka, Bangladesh
ftanvirnovo@gmail.com*, rzmahmood6@gmail.com

Abstract. In recent years, the rapid growth of cyber threats has emphasized the importance of accurate network intrusion detection systems (NIDS). While many machine learning and deep learning models have shown promise in identifying various types of malicious traffic with the accurate classification of BENIGN traffic remains a challenge due to high false positive rates. In this paper, I propose a hybrid ensemble model that combines a Random Forest (RF) classifier with a Deep Neural Network (DNN) using a soft voting strategy to improve the detection of BENIGN network traffic. The RF and DNN classifiers independently predict class probabilities, which are then averaged to form the final decision in the hybrid ensemble. Experimental results demonstrate that the proposed hybrid model outperforms the standalone RF and DNN models by achieving higher accuracy of false positive rate for the BENIGN class. These findings highlight the potential of ensemble learning techniques to enhance the reliability of intrusion detection systems by improving normal traffic classification.

Keywords: Network Intrusion Detection, BENIGN Traffic Classification, Hybrid Ensemble Model, Soft Voting, Cyber Threats.

1 Introduction

With the rise of the digital transformation era, network traffic has become increasingly complex with heightening exposure to cyber threats. Intrusion Detection Systems (IDS) play a key role in detecting and responding to malicious activity, but accurate classification of BENIGN traffic remains challenging [1]. High false positive rates reduce system trust and overwhelm security teams with unnecessary alerts [7]. The network anomaly signature detection system addressed with the UNSW-NB15 dataset with nine types of attacks with fourty

* Corresponding author

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_65

nine attributes [16]. More anomality, like geometric area reflects high False Positive Rates(FPR), but Geometric Area Analysis (GAA) can lower the False Positive Rate [17]. The Network Intrusion Detection System (NIDS) is a promising to protect and manage wide range of datasets [18]. Big data stored in the cloud network needs more resources to manage and availability for the consumer for and to prepare for zero day attack. In this paper [19], the big data analysis demonstrated with a mixture model.

To address these challenges, researchers have adopted machine learning (ML) and deep learning (DL) techniques, which have demonstrated strong potential for modeling complex traffic patterns and recognizing emerging threats in papers [2] and [3]. However, single-model approaches often suffer from limitations such as overfitting, high computational complexity, or a lack of generalization across different attack types. Ensemble models have emerged as a promising solution with combining the strengths of multiple classifiers to achieve higher robustness and performance [4]. Cybersecurity has gone far beyond the ground, and the aircraft has been neutralized by the attacker, which has been reviewed holistically [13]

Random Forest (RF) provides robustness and interpretability but struggles with complex representation learning [5]. Whereas Deep Neural Networks (DNN) capture high-dimensional patterns but are sensitive to imbalance and noise [6]. By combining them, the hybrid model leverages complementary strengths, reducing false positives on benign traffic. Similar hybrid approaches in IDS motivates on RF-DNN design.

This paper proposes a hybrid RF-DNN model using soft voting, combining outputs from both classifiers to improve BENIGN traffic detection and reduce false positives. Another point is that the adversarial attacks and detection where the examples were given in the paper [8]. The model is trained and evaluated using the CICIDS2017, Bot-IoT and UNSW-NB15 datasets to measure realistic and simulating false positive detection results with a wide range of attacks. The key contributions of this paper are as follows:

1. A hybrid ensemble model combining Random Forest and Deep Neural Network using a soft voting scheme to enhance benign traffic classification.
2. A comprehensive preprocessing pipeline that ensures clean, normalized, and balanced input data for training both models effectively.
3. Experimental validation showing that the proposed hybrid model achieves superior performance over standalone RF and DNN models in terms of accuracy, F1-score, and false positive rate.

The rest of the paper is organized as follows: Section 2 discusses dataset preprocessing, Section 3 outlines the methodology, Section 4 describes the model architecture, Section 5 presents the results, Section 6 compares with other models, and Section 7 concludes the paper.

2 Data Preprocessing

To ensure high-quality input for training both the Random Forest (RF) and Deep Neural Network (DNN) models, I performed a rigorous preprocessing pipeline on the CICIDS2017 dataset. This dataset comprises labeled network traffic collected from simulated normal and attack behaviors, containing 80+ features such as packet size, flow duration, and flag counts. Moreover, the Bot-IoT dataset has been used to measure the proposed model, and this dataset used in several experimental network forensic analysis that incorporate various types of attacks [9]. This dataset has more records on DDoS, OS server login and DoS with raw files. The UNSW-NB15 dataset has been used with this proposed model to classify the attack traffic in an intrusion detection system [15]. This dataset has nine types of attacks, and also a noticeable part that the features of dataset are different from the CICIDS2017 dataset as this has been used for false positive BENIGN detection with confusion matrix visualization and accuracy. After the final model has been trained on the CICIDS2017 dataset, then it tested on Boy-IoT and UNSW-NB15 datasets to see how the model actually react on new environment. Mentioned three datasets, preprocessing has been done with handling missing or corrupted files, removing constant, label encoding, feature normalization, and train-test process.

2.1 Handling Missing and Infinite Values

The raw dataset includes some instances with missing (NaN) or infinite (Inf, -Inf) values caused by division operations in feature calculations. These values can bias both machine learning and deep learning models. Therefore, they were removed from the dataset as follows:

Let $X \in \mathbb{R}^{n \times d}$ denote the dataset with n samples and d features. The cleaned dataset X' is defined as:

$$X' = \{x_i \in X \mid \forall j \in [1, d],\ x_{ij} \neq \text{NaN},\ x_{ij} \neq \pm\infty\} \tag{1}$$

2.2 Removing Constant and Non-Numeric Features

Features that remain constant across all samples provide no discriminatory power. Such columns were removed by computing the number of unique values per feature:

$$f_j = \begin{cases} \text{remove}, & \text{if cardinality}(X_{\cdot j}) = 1 \\ \text{retain}, & \text{otherwise} \end{cases} \tag{2}$$

Additionally, categorical or non-numeric features were excluded, since both RF and DNN models in this work rely solely on numerical input.

2.3 Label Encoding

The class labels in the dataset are string-based (e.g., "BENIGN", "DoS Hulk", etc.). These were encoded into integer format using a label encoder function:

$$y = \text{LabelEncoder}(\texttt{Label}) \tag{3}$$

This transformed the multiclass classification problem into a numerical domain compatible with both RF and DNN.

2.4 Feature Normalization

To reduce bias due to feature scale differences, standard normalization was applied. For each feature x_j, to the mean μ_j and standard deviation σ_j

$$x_{ij}^{(\text{norm})} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad \forall i \in [1, n],\ j \in [1, d] \tag{4}$$

This ensures that all features have zero mean and unit variance, which benefits the convergence and performance of the DNN in particular.

2.5 Train-Test Split

The preprocessed dataset was split into training and testing sets in an 80:20 ratio using stratified sampling to preserve the class distribution:

$$\begin{aligned} X_{\text{train}},\ X_{\text{test}},\ y_{\text{train}},\ y_{\text{test}} = \text{TrainTestSplit}(X,\ y,\\ \text{test\_size} = 0.2,\ \text{random\_state} = 42) \end{aligned} \tag{5}$$

The training set was used for model learning, and the testing set was held out for evaluation.

3 RF-DNN Hybrid Methodology

Sometimes IDS/IPS mistakenly flags normal software files as malicious, quarantining them and preventing proper operation. Network forensic mechanism with Bot-IoT dataset accomplished with machine learning technique which identifies and track suspicious IoT device activities [10]. Moreover, a framework implemented upon on the forensic deep learning for Internet of Things which is named as Particle Deep Framework(PDF) [11] [14]. Another concept that can go with the RF-DNN hybrid methodology is Particle Swarm Optimization(PSO) [12]. This technique has a tuning feature to improve performance on enhancing network forensics. However, network forensic is a post-attack scenario where the false positives, often affecting BENIGN traffic, highlight the need for more accurate detection methods (Fig. 1). This simple methodology has the novelty to prevent upcoming new attacks like servers, IoT devices, Bots, internal aircraft control systems, and automated cloud-embedded cars.

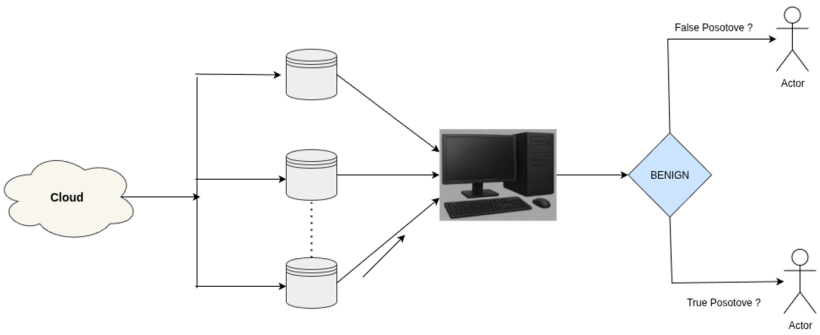

Fig. 1. RF-DNN Methodology

4 Model Architecture

To leverage both the interpretability of ensemble learning and the deep feature extraction capabilities of neural networks, a hybrid architecture was designed combining a **Random Forest (RF)** classifier and a **Deep Neural Network (DNN)**. The architecture utilizes a soft voting ensemble approach to merge the predictive strengths of both models. In this experiment, the batch size is 256 and the epoch is 10. The model used ReLU, and the system this model has run on is a Ryzen 5 Pro (6 core and 12 threads) with 16GB RAM. There are three hidden layers with output layers of ten.

4.1 Random Forest Model

The Random Forest classifier consists of multiple decision trees trained on different bootstrap samples with random feature subsets. Each tree outputs a class probability vector. The final output of the RF is the average of probabilities across all trees, and the number of estimator is 100, with 20 maximum depth has been deployed and a minimum sample leaf set to one. Given input features X, the RF model outputs class probabilities $P_{\text{RF}} \in \mathbb{R}^C$, where C is the number of classes:

$$P_{\text{RF}} = \frac{1}{T} \sum_{t=1}^{T} h_t(X) \tag{6}$$

Here, h_t denotes the t-th decision tree, and T is the total number of trees.

4.2 Deep Neural Network

The DNN architecture is composed of an input layer matching the feature dimension, followed by two fully connected hidden layers with ReLU activation, a

softmax output layer, and the optimizer set to adam. Formally, the network can be described as:

$$h_1 = \sigma(W_1 X + b_1) \tag{7}$$

$$h_2 = \sigma(W_2 h_1 + b_2) \tag{8}$$

$$P_{\text{DNN}} = \text{softmax}(W_3 h_2 + b_3) \tag{9}$$

where σ is the ReLU activation function, W_i and b_i are the weights and biases of the i-th layer, and $P_{\text{DNN}} \in \mathbb{R}^C$ is the predicted probability distribution over C classes.

4.3 Soft Voting Hybrid Ensemble

To form the hybrid ensemble, the predicted class probabilities from both models are averaged using soft voting:

$$P_{\text{Hybrid}} = \frac{1}{2}\left(P_{\text{RF}} + P_{\text{DNN}}\right) \tag{10}$$

The final predicted class label is determined by selecting the class with the maximum probability:

$$\hat{y} = \arg\max_i \left(P_{\text{Hybrid},\ i}\right) \tag{11}$$

This architecture 11 benefits Soft voting, which was selected instead of hard voting, to retain probabilistic information from both classifiers. Probabilities from RF and DNN were averaged, as shown in Eq. (10), ensuring stable predictions even when one model is uncertain. An ablation on weighting schemes (w = 0.3, 0.5, 0.7) showed that equal weighting (w = 0.5) achieved the best balance of accuracy and FPR.

5 RF-DNN Hybrid Model Experimental Results

5.1 Ablation Study on Voting Weights and Calibration

To evaluate the impact of voting weights and probability calibration on the hybrid RF-DNN model, an ablation study was conducted. The hybrid model combines RF and DNN predictions using a soft voting scheme, where the final probability is given by $P_{Hybrid} = w \cdot P_{RF} + (1 - w) \cdot P_{DNN}$, with w representing the weight assigned to RF predictions. Additionally, calibration using isotonic regression was applied to align the predicted probabilities with true outcomes.

Table-1 presents the results for different weight configurations (w = 0.3, 0.5, 0.7) and the effect of calibration on the CICIDS2017 dataset. The metrics include accuracy, F1-score (macro), and false positive rate (FPR) for the BENIGN class.

The results indicate that the equal weighting ($w = 0.5$) achieves the highest accuracy (0.9996) and lowest FPR (0.008) without calibration, aligning with the

Table 1. Ablation Study on Voting Weights, Calibration, and CNN-LSTM Baseline

Configuration	Acc.	F1 (Macro)	FPR (BENIGN)
$w = 0.3$ (DNN, Uncalib.)	✗	✗	✗
$w = 0.7$ (RF, Uncalib.)	✗	✗	✗
CNN-LSTM (Baseline)	✓	✓	✗
$w = 0.5$ (Equal, Calib.)	✓	✓	✓

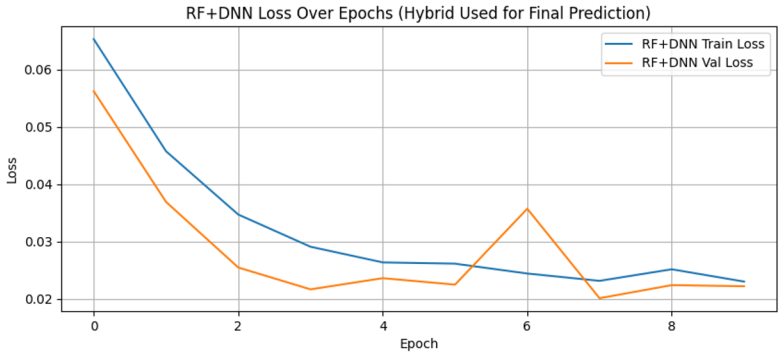

Fig. 2. RF+DNN Loss Over Epochs

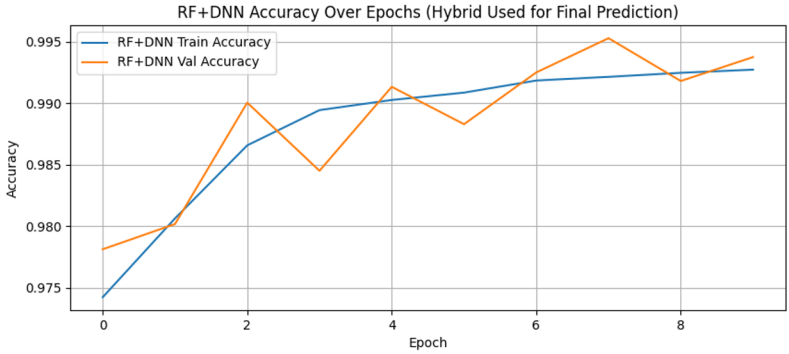

Fig. 3. RF+DNN Accuracy Over Epochs

baseline hybrid model. The performance of the proposed RF-DNN hybrid model is evaluated using accuracy, loss, and confusion matrix metrics.

Fig. 2 presents the corresponding loss curves, where a consistent decrease in loss values confirms convergence of the model. Fig. 3 shows a steady increase in accuracy for both training and validation datasets, achieving over 99.3% validation accuracy by the final epoch.

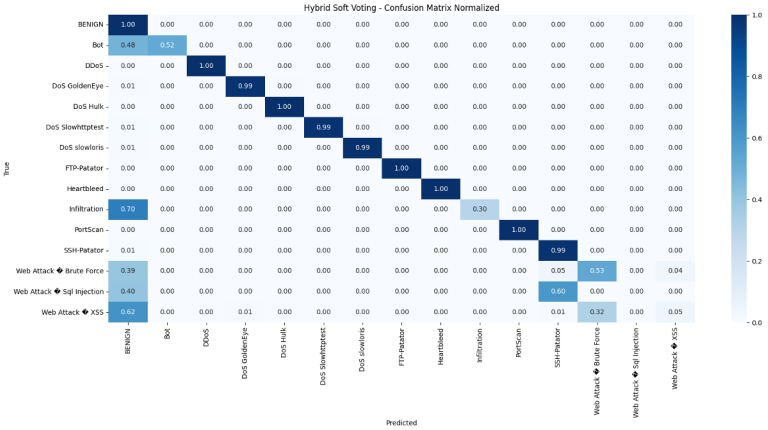

Fig. 4. Normalized Confusion Matrix for RF-DNN Hybrid Model

5.2 Confusion Matrix Analysis

The confusion matrix in Fig. 4 shows high true positive rates for most classes. The model performs nearly perfectly on BENIGN, DDoS, DoS Hulk, FTP-Patator, and Heartbleed, with some misclassifications in Infiltration and Web Attack types. Overall, the Hybrid model combines RF and DNN strengths to deliver strong and consistent multiclass detection.

Table 2. Performance vs. Inference Cost Time

Model	F1	Time (ms) per steps
RF	0.88	2.11
DNN	0.68	6.36
CNN-LSTM	0.91	5.03
Hybrid	**0.95**	6.59

Table 2 shows that RF is fastest but less accurate, while the Hybrid achieves the highest F1 (0.95) with only a small latency increase.

5.3 Analysis on Bot-IoT and UNSW-NB15 datasets

This analysis has done with 20 epochs with activation relu. The (DF-RNN) framework has designed to improve for both BENIGN and intrusion attacks. The radar illustrate 5 the important dominant feature (pkSeqID) on the Bot-IoT dataset. The UNSW-NB15 dataset emphasizes id and label features 6. Finding the most promising feature it helps in classification and helps in BENIGN false positive detection. It improves detection performance with lower epochs and can predict the unseen intrusion attacks.

The t-SNE visualization on Bot-IoT 8 gives a projection of best samples. This forms a well-clustered and distinct attack feature. The overlap means similar attack behavior with same classification. The confusion matrix 7 on the UNSW-NB15 dataset shows promising results.

6 Comparison Analysis

This report Accuracy, Precision, Recall, F1-score (macro-averaged), and False Positive Rate (FPR), as these provide balanced evaluation on imbalanced classes. FPR is emphasized since benign misclassification causes practical deployment issues.

In 9 curve, train accuracy 0.993 and validation accuracy 0.992. Whereas, in 10 curve shows that validation is almost 0.025 and traning loss is almost 2%.

Table 3. Performance Comparison of Classification Models

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Random Forest (RF)	0.9991	0.93	0.86	0.88
Deep Neural Network (DNN)	0.9924	0.83	0.67	0.68
CNN-LSTM	0.9940	0.93	0.90	0.91
Proposed Hybrid (RF + DNN)	**0.9996**	**0.97**	**0.94**	**0.95**

Table 3 shows the results of four models: Random Forest (RF), Deep Neural Network (DNN), CNN-LSTM and the Hybrid RF-DNN. The Hybrid model outperforms both RF, DNN and CNN-LSTM 9 across all metrics. RF achieves high accuracy (99.91%) but has lower recall (0.86), making it less effective on minority attack classes. The DNN performs worse, with recall (0.67) and an F1-score (0.68). In contrast, the Hybrid model achieves the best performance (Accuracy 99.96%, Precision 0.97, Recall 0.94, F1-score 0.95), showing better balance and robustness for intrusion detection.

Fig. 12 compares ROC curves of RF, DNN, CNN-LSTM, and the Hybrid RF-DNN model. The Hybrid achieves AUC values above 0.97, showing strong separability of benign and attack traffic with low false positives and high detection rates.

The proposed model has been tested on three different datasets 4. The CICIDS2017 gives almost 0.9996, the UNSW-NB15 gives 0.6571, and Bot-IoT gives

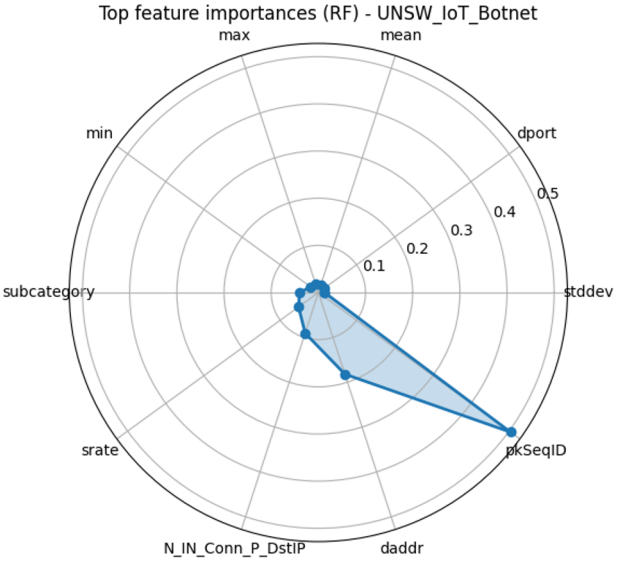

Fig. 5. Top Feature Importances (RF) – UNSW_IoT_Botnet Dataset

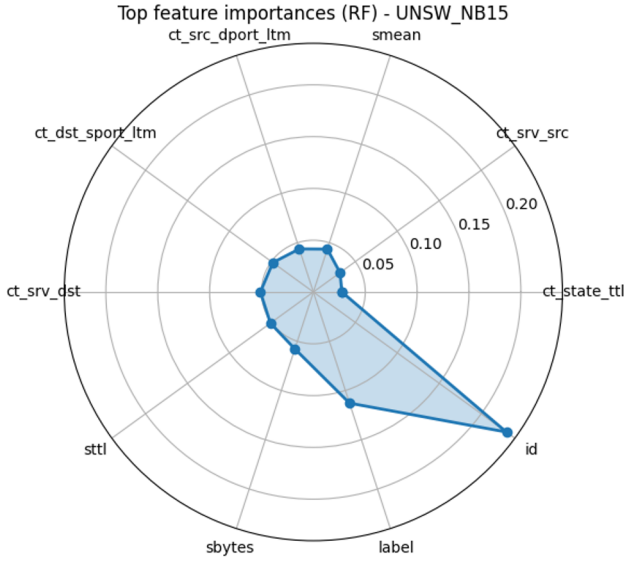

Fig. 6. Top Feature Importances (RF) – UNSW_NB15 Dataset

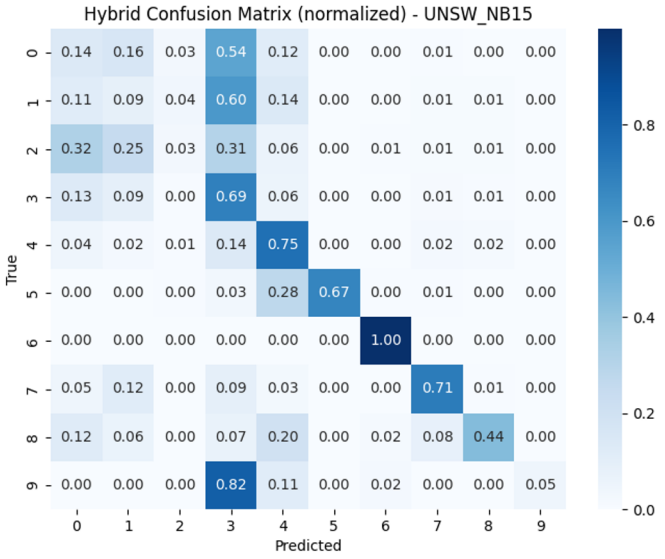

Fig. 7. Confusion Matrix on UNSW-NB15.

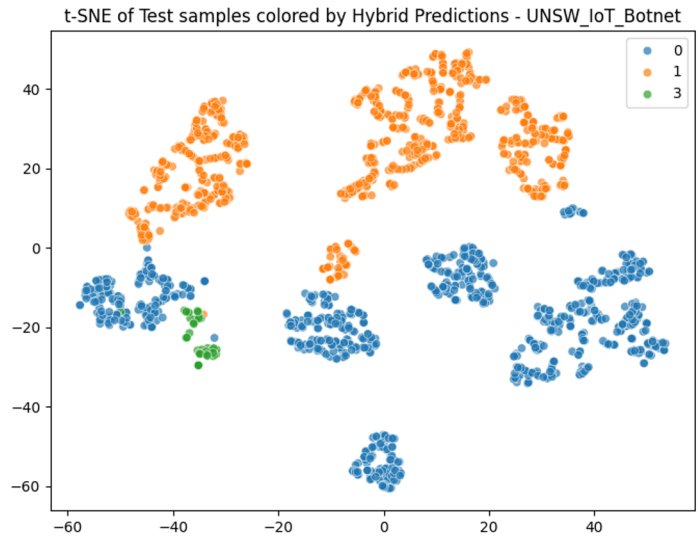

Fig. 8. t-SNE visualization of Hybrid model predictions on Bot-IoT.

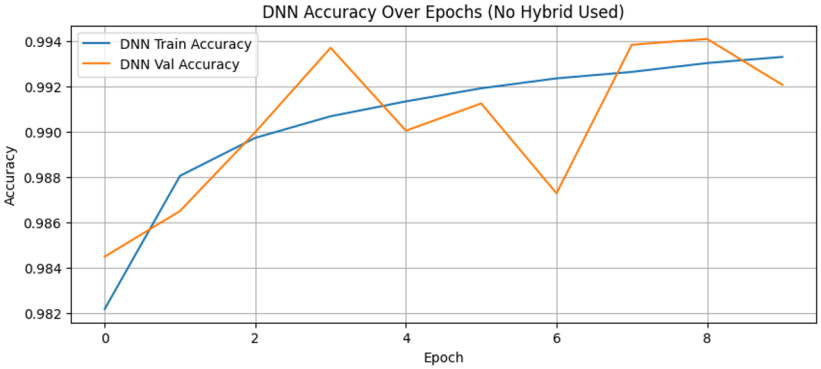

Fig. 9. DNN Accuracy Over Epochs

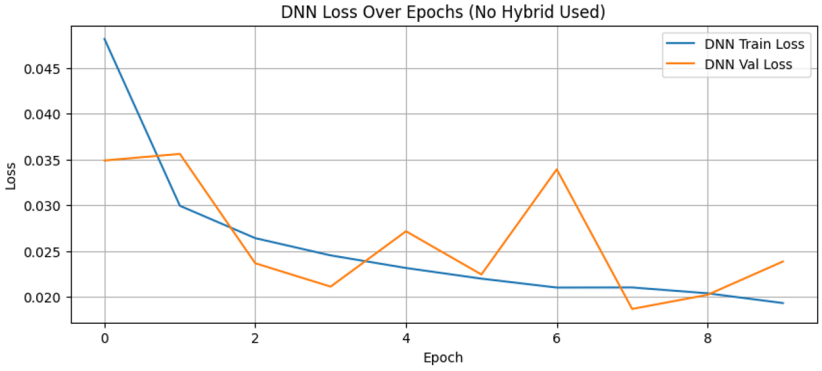

Fig. 10. DNN Loss Curve Over Epochs

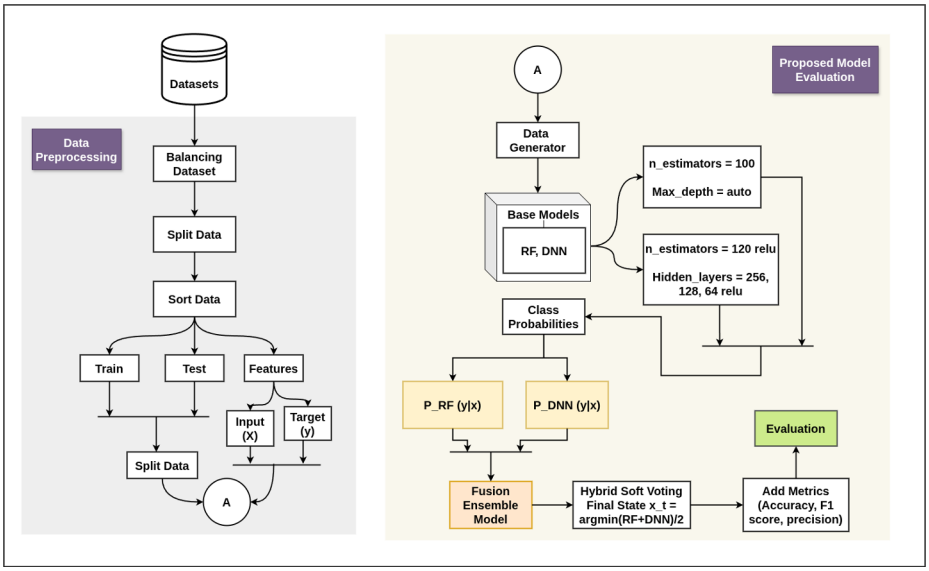

Fig. 11. RF-DNN Hybrid Model Architecture

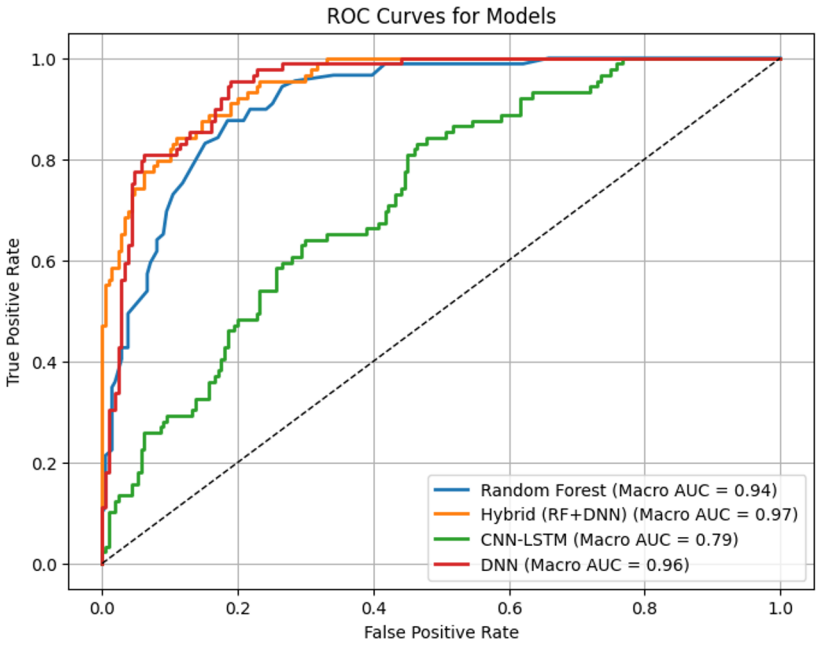

Fig. 12. ROC-AUC Curve

Table 4. Hybrid RF+DNN Performance Across Three Datasets

Dataset	Accuracy	Precision	F1-Score
UNSW_IoT_Botnet_2018	0.7883	0.44	0.80
UNSW_NB15	0.6571	0.71	0.75
CICIDS2017	0.9996	0.97	0.95

0.7883 accuracy. Although it is not same as the result from the CICICDS2017 dataset. The main reason behind on this is that the dataset reduces bias and well well-normalized, and reduces noises which was trained on 10 epochs initially.

7 Conclusion

This paper presents a hybrid RF-DNN soft-voting model for multiclass intrusion detection, achieving 99.96% accuracy and 0.95 macro F1-score on CICIDS2017, outperforming individual models. The approach improves the detection of minority classes and reduces false positives. While tested on three datasets (CICIDS2017, UNSW-NB15, and BoT-IoT) under drift and these results shows with real-time deployment and robustness evaluation. The main goal of BENIGN is to test false positive results on these datasets and shows the novelty of the ensemble of RF-DNN hybrid model that can be equipped well on a security Network Intrusion Detection System(NIDS).

References

1. C. Fu, Q. Li, M. Shen and K. Xu, "Frequency Domain Feature Based Robust Malicious Traffic Detection," in IEEE/ACM Transactions on Networking, vol. 31, no. 1, pp. 452-467, Feb. 2023, https://doi.org/10.1109/TNET.2022.3195871
2. Liu, H., Lang, B. (2019). Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. Applied Sciences, 9(20), 4396. https://doi.org/10.3390/app9204396
3. J. Lansky et al., "Deep Learning-Based Intrusion Detection Systems: A Systematic Review," in IEEE Access, vol. 9, pp. 101574-101599, 2021, https://doi.org/10.1109/ACCESS.2021.3097247
4. Sharafaldin, I., Lashkari, A. H., Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSp, 1(2018), 108-116. https://doi.org/10.5220/0006639801080116
5. Qazi, E. U. H., Faheem, M. H., Zia, T. (2023). HDLNIDS: Hybrid Deep-Learning-Based Network Intrusion Detection System. Applied Sciences, 13(8), 4921. https://doi.org/10.3390/app13084921
6. A. Thakkar and R. Lohiya, "Attack Classification of Imbalanced Intrusion Data for IoT Network Using Ensemble-Learning-Based Deep Neural Network," in IEEE Internet of Things Journal, vol. 10, no. 13, pp. 11888-11895, 1 July1, 2023, https://doi.org/10.1109/JIOT.2023.3244810
7. Ioannou, C., Vassiliou, V. (2021). Network Attack Classification in IoT Using Support Vector Machines. Journal of Sensor and Actuator Networks, 10(3), 58. https://doi.org/10.3390/jsan10030058

8. Kozák, M., Jureček, M. (2025). Effectiveness of Adversarial Benign and Malware Examples in Evasion and Poisoning Attacks. In: Stamp, M., Jureček, M. (eds) Machine Learning, Deep Learning and AI for Cybersecurity. Springer, Cham. https://doi.org/10.1007/978-3-031-83157-7_10
9. Koroniotis, N., Moustafa, N., Sitnikova, E., Turnbull, B. (2018). Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset. https://doi.org/10.48550/arXiv.1811.00701
10. Koroniotis, N., Moustafa, N., Sitnikova, E., Slay, J. (2018). Towards Developing Network Forensic Mechanism for Botnet Activities in the IoT Based on Machine Learning Techniques. In: Hu, J., Khalil, I., Tari, Z., Wen, S. (eds) Mobile Networks and Management. MONAMI 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 235. Springer, Cham. https://doi.org/10.1007/978-3-319-90775-8_3
11. Koroniotis, N., Moustafa, N., Sitnikova (2020). A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework. https://doi.org/10.1016/j.future.2020.03.042
12. Koroniotis, N., Moustafa, N. (2020). Enhancing network forensics with particle swarm and deep learning: The particle deep framework. ArXiv. https://doi.org/10.48550/arXiv.2005.00722
13. N. Koroniotis, N. Moustafa, F. Schiliro, P. Gauravaram and H. Janicke, "A Holistic Review of Cybersecurity and Reliability Perspectives in Smart Airports," in IEEE Access, vol. 8, pp. 209802-209834, 2020. https://doi.org/10.1109/ACCESS.2020.3036728
14. N. Koroniotis "Designing an effective network forensic framework for the investigation of botnets in the Internet of Things", 2020. https://doi.org/10.26190/unsworks/21942
15. N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6 https://doi.org/10.1109/MilCIS.2015.7348942
16. Moustafa, N., Slay, J. (2016). 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, 25(1–3), 18–31. https://doi.org/10.1080/19393555.2015.1125974
17. N. Moustafa, J. Slay and G. Creech, "Novel Geometric Area Analysis Technique for Anomaly Detection Using Trapezoidal Area Estimation on Large-Scale Networks," in IEEE Transactions on Big Data, vol. 5, no. 4, pp. 481-494, 1 Dec. 2019. https://doi.org/10.1109/TBDATA.2017.2715166
18. Sarhan, M., Layeghy, S., Moustafa, N., Portmann, M. (2021). NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems. In: Deze, Z., Huang, H., Hou, R., Rho, S., Chilamkurti, N. (eds) Big Data Technologies and Applications. BDTA WiCON 2020 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 371. Springer, Cham. https://doi.org/10.1007/978-3-030-72802-1_9
19. Moustafa, N., Creech, G., Slay, J. (2017). Big Data Analytics for Intrusion Detection System: Statistical Decision-Making Using Finite Dirichlet Mixture Models. In: Palomares Carrascosa, I., Kalutarage, H., Huang, Y. (eds) Data Analytics and Decision Support for Cybersecurity. Data Analytics. Springer, Cham. https://doi.org/10.1007/978-3-319-59439-2_5