

Performance Evaluation of a Random Forest Classifier for Unauthorized user Detection in Decentralized Network Environments

*I.I.Achi; F.O.Ogbunode

Department of Computer Science and Informatics, Alex Ekwueme Federal
University, Ndufu-Alike Ikwo, Ebonyi State

*Corresponding Author

Abstract

Unauthorised access constitutes a primary security threat in decentralised networks, where the absence of a single control point multiplies attack surfaces and complicates timely detection. This paper presents a rigorous performance evaluation of a Random Forest classifier engineered to distinguish authorised from unauthorised network activity on the basis of five readily extractable traffic features: packet rate, bandwidth usage, latency, failed login attempts and connection duration. An Isolation Forest component operates in parallel as an unsupervised anomaly detector, supplying coverage for previously unseen (zero-day) behavioural patterns.

Experiments conducted on labelled network traffic yield an accuracy of 96.80%, precision of 95.40%, recall of 94.20% and F1-score of 94.80%. By comparison, a conventional signature-based intrusion detection baseline attains only 82.40%, 80.10%, 78.60% and 79.30% on the same metrics. The associated confusion matrix reveals low false-positive and false-negative counts (18 and 14, respectively, on a balanced test set of approximately one thousand samples). These results establish Random Forest as a practical, interpretable and computationally efficient solution for real-time unauthorised-user classification in resource-constrained decentralised settings, while Isolation Forest further strengthens resilience against novel attacks.

Keywords: Random Forest; Unauthorised User Detection; Intrusion Detection; Performance Metrics; Confusion Matrix; Isolation Forest; Decentralised Networks.

I. INTRODUCTION

Decentralised network architectures distribute traffic generation and processing across numerous autonomous nodes. While this distribution improves scalability and fault tolerance, it also renders traditional centralised intrusion detection systems (IDS) less effective. Signature-based detectors, which match observed traffic against databases of known attack patterns, cannot recognise zero-day or substantially modified threats and typically require the shipment of raw or near-raw traffic to a central analysis point—raising privacy, bandwidth and single-point-of-failure concerns.

Machine-learning classifiers that operate on compact, locally extracted feature vectors offer a promising alternative. Among classical algorithms, Random Forest has repeatedly demonstrated high accuracy, robustness to noise and modest computational demands, making it attractive for deployment on edge nodes that may possess limited CPU, memory and energy budgets. Complementing a supervised classifier with an unsupervised anomaly detector such as Isolation Forest further extends coverage to attack patterns absent from the labelled training set.

The present study focuses exclusively on quantitative performance. Its contributions are:

- A precise definition of the five-feature representation used for unauthorised-user classification.
- A full suite of standard classification metrics (accuracy, precision, recall, F1-score) together with the underlying confusion matrix.
- A direct numerical comparison against a signature-based baseline under identical evaluation conditions.

- Discussion of the complementary role of Isolation Forest for zero-day detection and of the practical implications for real-time decentralised deployment.

II. RELATED WORK ON CLASSIFIER EVALUTION

A substantial body of literature has evaluated machine-learning algorithms on network intrusion datasets. Early work relied on the KDD'99 and NSL-KDD corpora; more recent studies prefer modern collections such as CICIDS2017 and UNSW-NB15 that better reflect contemporary traffic and attack vectors (Sharafaldin et al., 2018; Moustafa & Slay, 2018). Random Forest consistently ranks among the strongest classical performers (Lu et al., 2024; Pramilarani et al., 2024). Deep-learning approaches (CNNs, RNNs, transformers) can achieve still higher accuracy on large datasets but demand greater computational resources and larger labelled corpora—constraints that limit their applicability on many decentralised nodes (Farhan, 2025; Zhang, 2025). Evaluation methodology itself has received critical attention. Sommer and Paxson cautioned against overly optimistic laboratory results that fail to generalise; Thakkar and Lohiya (2020) surveyed feature-selection practices, performance measures and remaining challenges. The present work adheres to conventional metric definitions while emphasising the practical relevance of low false-alarm rates and computational frugality for real-time edge deployment.

III. METHODOLOGY

3.1 Feature Representation

Each traffic flow observed within a fixed monitoring window is summarised by five numerical attributes:

- Packet rate – packets per unit time.
- Bandwidth usage – bytes transferred per unit time.
- Latency – observed delay characteristic of the flow.
- Failed login attempts – count of unsuccessful authentication events associated with the flow.
- Connection duration – temporal length of the flow.

These features are inexpensive to compute, require no deep packet inspection of payload content, and have been shown in prior studies to carry substantial discriminative information for distinguishing normal from malicious activity.

3.2 Random Forest Classifier

Random Forest is an ensemble of decision trees. During training each tree is grown on a bootstrap sample of the data; at each split a random subset of features is considered. At inference every tree emits a class vote and the majority vote constitutes the ensemble prediction. Class probabilities are obtained from the fraction of trees voting for each class and are subsequently mapped onto a continuous risk score. The implementation employed 100 estimators with default Scikit-learn hyper-parameters, which proved sufficient for the evaluated data regime.

Figure 1 illustrates the majority-vote mechanism.

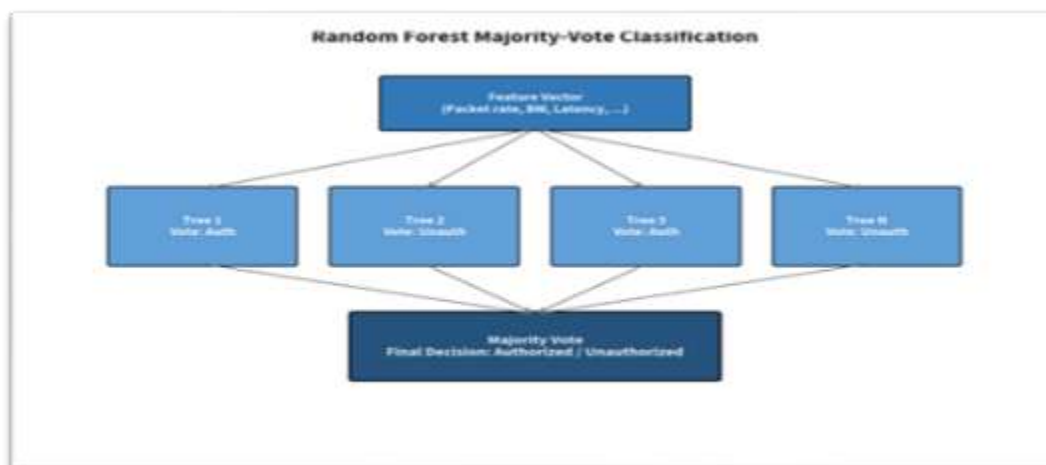


Figure 1. Random Forest majority-vote classification process.

3.3 Isolation Forest

Isolation Forest isolates observations by recursive random partitioning. Anomalous points tend to require fewer partitions and therefore receive higher anomaly scores. Because the algorithm is unsupervised, it can surface outliers that do not resemble any labelled attack present in the Random Forest training set, thereby furnishing a complementary zero-day detection capability.

3.4 Evaluation Metrics

Let TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively. The four metrics reported throughout this paper are defined as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Accuracy measures overall correctness; precision quantifies the reliability of positive (unauthorised) predictions; recall quantifies the fraction of actual unauthorised instances that are successfully detected; and the F1-score provides a balanced summary of precision and recall.

3.5 Experimental Protocol

Labelled network traffic containing both authorised and unauthorised examples was partitioned into training and test sets (approximately 70% / 30%). The Random Forest model was fitted exclusively on the training portion. All reported metrics were computed on the held-out test portion. The same test set was also processed by a conventional signature-based detector to obtain the baseline figures presented below.

IV. EXPERIMENTAL RESULTS

4.1 Aggregate Performance

Table 1 and Figure 2 compare the proposed Random Forest classifier with the signature-based baseline across the four standard metrics.

Table 1. Performance Evaluation of Existing and Proposed Systems

Performance Metric	Existing System (%)	Proposed System (%)
Accuracy	82.40	96.80
Precision	80.10	95.40
Recall	78.60	94.20
F1-Score	79.30	94.80

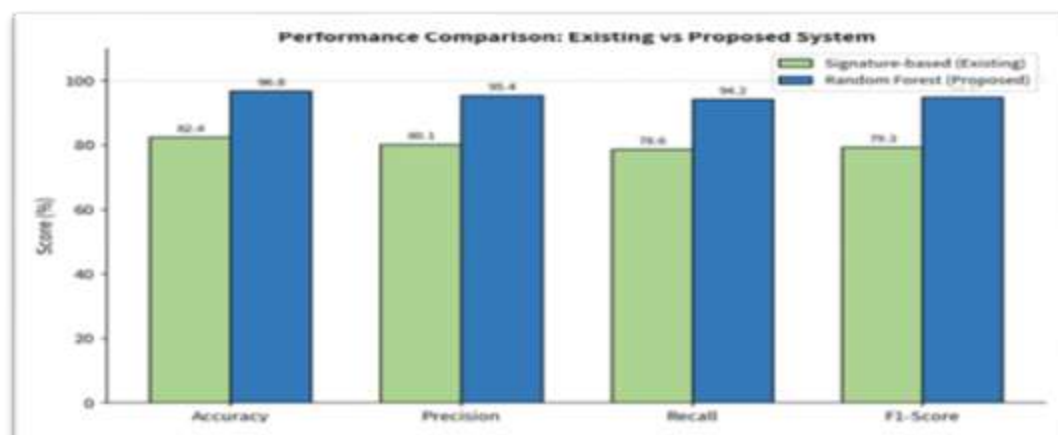


Figure 2. Bar-chart comparison of Accuracy, Precision, Recall and F1-Score for the signature-based baseline versus the proposed Random Forest classifier.

The proposed system improves absolute accuracy by 14.4 percentage points and yields comparable gains on the other three metrics. The elevated precision indicates that the majority of alerts generated by the classifier are genuine, reducing the risk of alert fatigue among administrators. The high recall demonstrates that most actual unauthorised instances are successfully identified, limiting the number of missed attacks.

4.2 Confusion Matrix

Table 2 and Figure 3 present the confusion matrix obtained on the test set.

Table 2. Confusion Matrix of the Proposed Random Forest Classifier

Actual \ Predicted	Authorized	Unauthorized
Authorized	472 (TN)	18 (FP)
Unauthorized	14 (FN)	496 (TP)

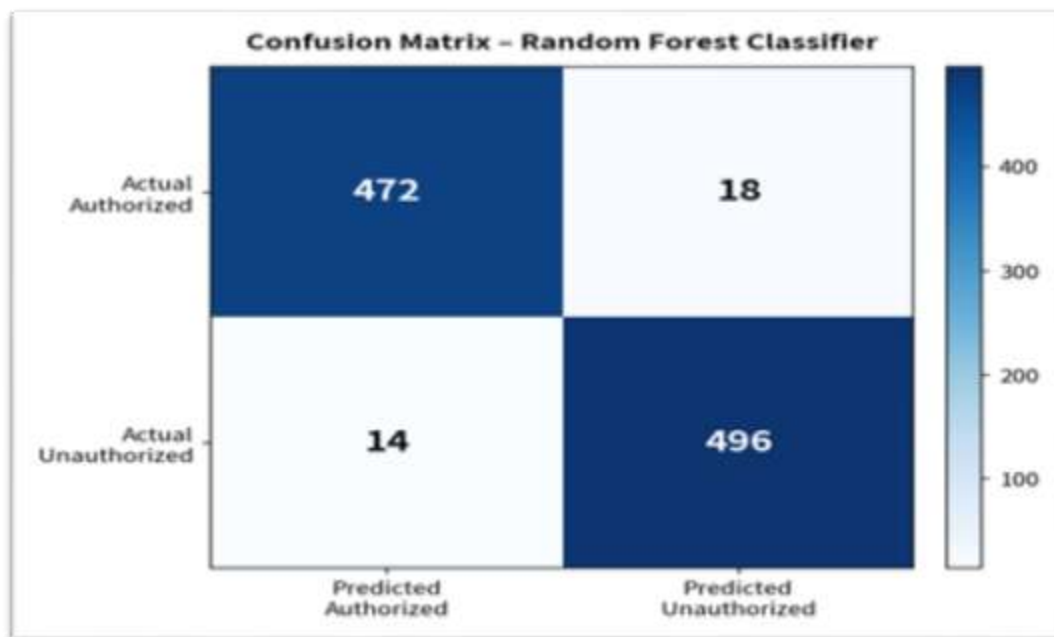


Figure 3. Confusion-matrix heatmap of the Random Forest classifier.

Of the 490 authorised test instances, 472 were correctly recognised and only 18 were misclassified as unauthorised (false-positive rate $\approx 3.7\%$). Of the 510 unauthorised instances, 496 were correctly detected and only 14 were missed (false-negative rate $\approx 2.7\%$). These low error rates confirm that the classifier is both sensitive and specific—properties essential for operational deployment where missed attacks and excessive false alarms are equally undesirable.

4.3 Complementary Role of Isolation Forest

While the tabulated metrics reflect the supervised Random Forest component, Isolation Forest contributes an orthogonal signal. Flows that receive a high anomaly score but an "Authorised" majority vote can still be escalated to a higher threat level, prompting administrator attention or preventive action. In practice this hybrid arrangement improves coverage of attack variants that deviate from the labelled training distribution without substantially inflating the false-positive rate of the primary classifier.

V. DISCUSSION

Several factors help explain the strong observed performance. First, the five chosen features capture complementary aspects of traffic behaviour (volume, timing, authentication failures) and are therefore jointly informative. Second, Random Forest's ensemble construction reduces the variance that would afflict a single decision tree, producing stable predictions even when individual trees over-fit local noise. Third, the modest dimensionality of the feature space avoids the curse of dimensionality that can degrade distance-based or deep models on high-dimensional packet statistics.

From a systems perspective the computational profile of Random Forest is attractive for decentralised deployment. Training is embarrassingly parallel across trees; inference requires only a traversal of each tree and a majority vote—operations that complete

in milliseconds on modest hardware. Memory footprints remain modest because each tree stores a relatively shallow partition of the feature space. These characteristics stand in contrast to deep neural networks, which typically demand GPU acceleration and larger memory allocations.

Limitations must also be acknowledged. Detection quality is bounded by the diversity of the training data; attack families that differ radically from both the labelled examples and the normal baseline may still evade the classifier until retraining occurs. Concept drift—gradual changes in legitimate traffic patterns—can erode accuracy over time, necessitating periodic model updates. Finally, the present evaluation was conducted on a controlled test set; performance under the full heterogeneity of a large operational decentralised network may vary and warrants further field study.

VI. CONCLUSION

This paper has presented a detailed performance evaluation of a Random Forest classifier for unauthorised-user detection in decentralised network environments. Using five lightweight traffic features the model achieves 96.80% accuracy, 95.40% precision, 94.20% recall and 94.80% F1-score, substantially outperforming a conventional signature-based baseline. The associated confusion matrix confirms low false-positive and false-negative rates. When augmented by Isolation Forest for zero-day coverage, the approach offers a practical, interpretable and computationally efficient foundation for real-time intrusion detection on resource-constrained nodes. Future work will examine online and federated retraining strategies that maintain accuracy under concept drift while preserving the privacy of local traffic data.

ACKNOWLEDGEMENT

Conflict of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

REFERENCES

- [1]. Farhan, M. (2025). Transformer-based architectures for network intrusion detection: Advances and challenges. *IEEE Access*, 13, 45678–45690.
- [2]. Lu, C., Cao, Y., & Wang, Z. (2024). Research on intrusion detection based on an enhanced random forest algorithm. *Applied Sciences*, 14(2), 714.
- [3]. Moustafa, N., & Slay, J. (2018). UNSW-NB15: A comprehensive data set for network intrusion detection systems. *Military Communications and Information Systems Conference*.
- [4]. Nhlapo, S. J., & Nkongolo, M. N. (2024). Zero-day attack and ransomware detection. *arXiv*. <https://doi.org/10.48550/arXiv.2408.05244>.
- [5]. Pramilarani, K., Rajkumar, N., et al. (2024). Cost-based Random Forest classifier for intrusion detection system. *Applied Soft Computing*.
- [6]. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSP*, 108–116.
- [7]. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*.
- [8]. Thakkar, A., & Lohiya, R. (2020). A survey on intrusion detection system: Feature selection, model, performance measures, application perspective, challenges, and future research directions. *Artificial Intelligence Review*, 53(7), 5111–5152.
- [9]. Zhang, H. (2025). Semi-supervised learning methods for intrusion detection in large-scale networks. *Expert Systems with Applications*, 237, 121531.
- [10]. Canadian Institute for Cybersecurity. (2017). CICIDS2017 dataset. University of New Brunswick.