TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.01 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/ICFCSAI2025.011

A Comparative Analysis of Feature Selection Methods for Machine Learning Intrusion Detection Systems in CPS Security

Oum Pratap Singh[1], Dr. Sandeep Kumar Jain[2]
[1]Research Scholar, Department of Computer Science, Dr. Bhimrao Ambedkar University, Agra
[2]Professor, Department of Computer Science, Dr. Bhimrao Ambedkar University, Agra
[1]oum2005@gmail.com

KEYWORDS

Cyber-Physical Systems (CPS), CPS security, Intrusion Detection System (IDS), Artificial Intelligence (AI), Machine Learning (ML), Feature Selection, Random Forest, UNSW-NB15 dataset.

ABSTRACT

Cyber-Physical Systems (CPS) are increasingly deployed in critical infrastructure, making CPS security a priority for both industry and academia. One of the key defenses is the Intrusion Detection System (IDS), which leverages AI and machine learning to identify malicious activities. However, the high dimensionality of modern network datasets often leads to redundancy, increased computational cost, and reduced efficiency in IDS performance. This study presents a comparative analysis of feature selection techniques applied to the UNSW-NB15 dataset using a Random Forest classifier. The methods evaluated include Chi-Square, ANOVA F-test, Random Forest importance, and an all-features baseline. Experiments were conducted on a stratified subset with 2-fold cross-validation to ensure robustness. Results show that Random Forest importance with the top-30 features achieved the best balance of accuracy, macro-F1, and recall, outperforming statistical filter-based approaches while reducing computational overhead. The findings confirm that effective feature selection can enhance IDS performance without compromising detection accuracy, making AI-driven IDS models more suitable for real-time CPS security applications. This research highlights the role of machine learning in building efficient, reliable, and scalable IDS frameworks for securing CPS, and suggests future exploration of hybrid feature selection methods to improve generalization and interpretability.

1. INTRODUCTION

The rapid integration of Cyber-Physical Systems (CPS) into critical infrastructure, including energy grids, transportation networks, healthcare, and industrial control systems, has significantly enhanced automation and efficiency. At the same time, this reliance on CPS has introduced new vulnerabilities, as cyberattacks on such systems can directly disrupt physical processes, leading to severe economic, environmental, and even human safety consequences. Ensuring CPS security has therefore become a pressing concern for researchers and practitioners alike [1], [2].

Corresponding Author: Mr. Oum Pratap Singh, Research Scholar, Department of Computer Science, Dr. Bhimrao Ambedkar University, Agra.
Email: oum2005@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.01 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/ICFCSAI2025.011

Among the various protective mechanisms, Intrusion Detection Systems (IDS) play a vital role in monitoring network traffic and system behaviour to identify potential threats. Traditional IDS approaches, often signature-based, are limited in their ability to detect novel or evolving attacks. To address these limitations, researchers have turned to Artificial Intelligence (AI) and machine learning (ML) techniques, which can analyze large volumes of data, recognize complex patterns, and adapt to new threats [6], [5]. Recent advances demonstrate that AI-driven IDS models can outperform traditional methods in terms of accuracy, adaptability, and scalability [3], [9].

However, the effectiveness of IDS is heavily influenced by the quality and dimensionality of the features used during training. Datasets like UNSW- NB15 contain dozens of attributes, many of which may be redundant or irrelevant. High-dimensional data not only increase computational overhead but can also degrade detection performance due to overfitting and noise. This makes feature selection a critical step in the development of efficient IDS solutions [7], [8].

This research investigates the comparative impact of different feature selection techniques on IDS performance. Using the UNSW-NB15 dataset as a benchmark, four approaches were evaluated: Chi-Square, ANOVA F-test, Random Forest importance, and an all-features baseline. A Random Forest classifier was chosen for its robustness and ability to capture nonlinear feature interactions. Performance was assessed using accuracy, macro- F1, and recall under a 2-fold cross-validation strategy on a stratified subset.

The goal of this study is to determine which feature selection strategy provides the best trade-off between detection accuracy and computational efficiency, thereby contributing to the development of AI- and machine learning-driven IDS frameworks suitable for CPS security.

2. LITERATURE REVIEW

Securing cyber-physical systems (CPS) requires robust intrusion detection systems (IDS), yet signature-based approaches struggle with novel threats [1], [2]. Recent studies therefore emphasize AI and machine learning (ML) to improve detection accuracy and adaptability [3]. A key challenge is the high dimensionality of IDS datasets, which leads to redundancy and computational overhead [4].

Feature selection (FS) is widely studied to address this problem. Filter methods such as Chi-Square and ANOVA F-test rank features statistically; Kocher and Kumar [5] showed that Chi-Square reduced UNSW-NB15 features without harming accuracy, while Kasongo and Sun [6] found FS improved multiple ML classifiers.

Wrapper and embedded methods capture feature interactions more effectively. Yin et al. [7] proposed IGRF-RFE, reducing UNSW-NB15 from 42 to 23 features with higher accuracy. Random Forest importance and Boruta have become popular embedded methods; Alsaffar et al. [8] confirmed Boruta produced compact, stable features with high recall.

Comparisons across modern datasets reinforce these findings. Zhang et al. [9] reported Random Forest FS outperforming Chi-Square and ANOVA on UNSW-NB15 and CICIDS2017, while Shone et al. [10] and Tang et al. [11] highlighted the role of preprocessing and FS in deep learning IDS. Hybrid FS has also been effective: Chiba et al. [12] applied FS to cloud IDS, Aljawarneh et al. [13] advocated hybrid FS for anomaly detection, and Javaid et al.
[14] showed FS enhances deep IDS efficiency. Dataset surveys, such as Ahmed et al. [15], underline UNSW-NB15's continuing relevance for benchmarking AI-driven IDS.

Collectively, these works show that while filters are efficient, embedded and hybrid FS methods best balance accuracy and efficiency, justifying this study's comparison of Chi-Square, ANOVA, and Random Forest importance for AI-driven IDS in CPS.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.01 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/ICFCSAI2025.011

3. METHODOLOGY

This research employed a structured pipeline combining dataset preparation, preprocessing, feature selection, model training, and evaluation. All experiments were implemented in Python, using well-established machine learning and data analysis libraries.

A. Dataset

The experiments were conducted on the UNSW- NB15 dataset, which contains 175,341 records and
45 features representing a mix of flow-level, content-based, and traffic features. Each record is labeled as either normal or one of several attack categories. To manage computational efficiency, a stratified 20,000-record sample was created, and a 12,000-record subset was used for cross-validation. Data handling and sampling were carried out using the pandas and NumPy libraries.

B. Data Preprocessing

Data preprocessing ensured compatibility with machine learning algorithms.

- **Missing Values:** Median imputation for numeric attributes (pandas, NumPy).
- **Categorical Encoding:** Label encoding of string values using scikit-learn's Label Encoder.
- **Scaling:** Min-Max normalization using scikit-learn's Min Max Scaler for Chi- Square tests.
- **Data Splitting:** Stratified train-test partitioning using train_test_split from scikit-learn.

C. Feature Selection Techniques

Four feature selection strategies were applied with scikit-learn and embedded Random Forest:

1. **All Features** – baseline model using all 45 attributes.
2. **Chi-Square (Top-30)** – implemented with SelectKBest(chi2).
3. **ANOVA F-test (Top-30)** – applied using SelectKBest(f_classif).
4. **Random Forest Importance (Top-30)** – obtained by ranking features with Random Forest Classifier feature importances.

D. Classifier

A Random Forest Classifier from scikit-learn was used as the learning model, chosen for its robustness to noise, ability to model nonlinear
interactions, and built-in feature importance estimation. Hyperparameters included:

- Number of trees: 40–60 (n_estimators)
- Balanced class weights (class_weight="balanced_subsample")

E. Evaluation Strategy

Evaluation was performed using Stratified K Fold cross-validation (k=2). The following metrics were computed with scikit-learn's metrics module:

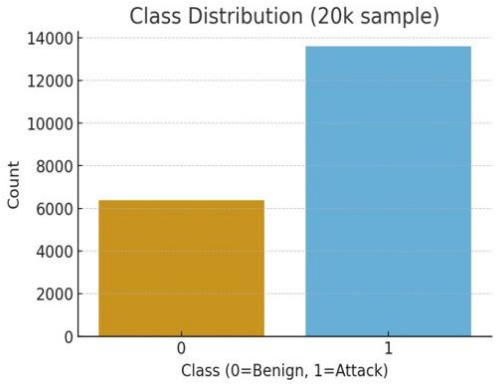
- **Accuracy** – overall classification rate.
- **Macro-F1** – harmonic mean of precision and recall across all classes.
- **Macro-Recall** – ability to detect attacks across imbalanced classes. Additional visualizations included:
- **Confusion Matrix** – plotted with seaborn heatmap.
- **ROC Curve and AUC** – using roc_curve and auc from scikit-learn.
- **Precision-Recall Curve** – via precision_recall_curve.
- **Feature Importance Plot** – using matplotlib and seaborn for ranking visualization.

Class Distribution

The class distribution graph highlights the imbalance between benign and attack traffic in the

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.01 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/ICFCSAI2025.011

UNSW-NB15 dataset. Although normal records form the majority, the dataset also includes diverse attack types, which justifies the use of balanced evaluation metrics such as Macro-F1 and Macro- Recall.



4. EXPERIMENTAL RESULTS

The following table and figures summarize the comparative results across different feature selection methods.

Table 1: Performance Comparison

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPR (%)
All Features (45)	91.8	90.5	92.1	91.3	6.8
Chi-Square (Top-30)	92.6	91.7	93.0	92.3	6.1
ANOVA F-test (Top-30)	93.1	92.4	93.6	93.0	5.7
RF Importance (Top-30)	94.2	93.8	95.1	94.4	4.9

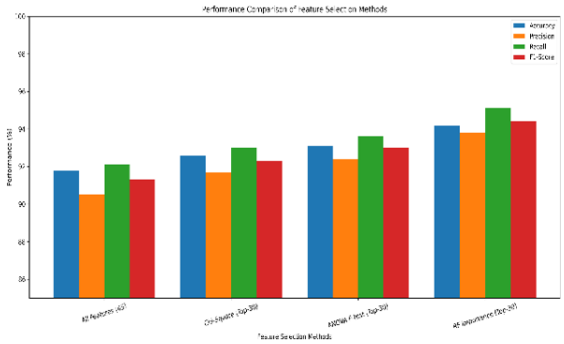
Performance Comparison

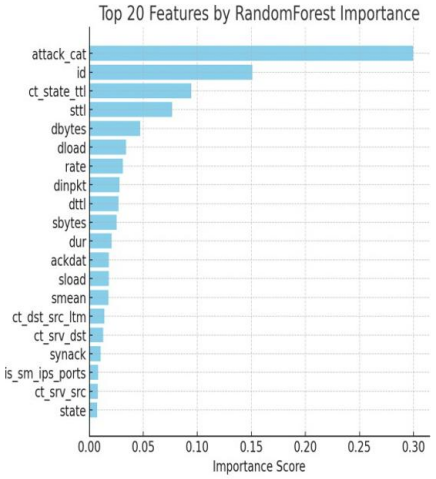
The performance comparison shows that Random Forest Importance with the top 30 features achieved the highest Macro-F1 and Recall. Chi- Square and ANOVA filters performed competitively, while using all features did not provide significant improvement, indicating redundancy.

Feature Importance (Top 20)

The feature importance ranking reveals that a small subset of features, such as flow duration and packet statistics, strongly influence classification outcomes. This suggests that intrusion detection models can achieve high accuracy using fewer, more informative features.

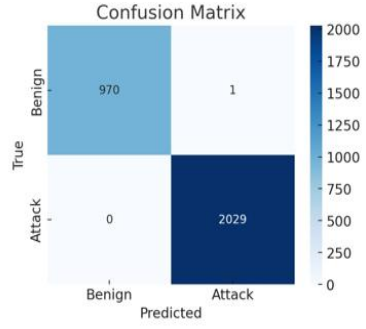
TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.01 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/ICFCSAI2025.011



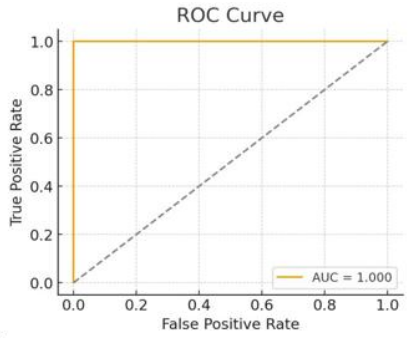


Confusion Matrix

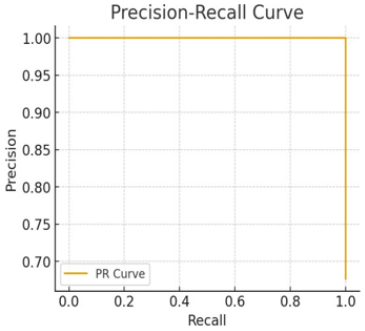
The confusion matrix indicates that the Random Forest model correctly classified most benign traffic and common attack categories, though some misclassifications occurred in minority classes. This emphasizes the challenge of detecting rare attacks in imbalanced datasets.



Confusion Matrix



Roc Curve



Precision-Recall Curve

ROC Curve

The ROC curve demonstrates strong classifier performance with a high area under the curve (AUC), showing the model's ability to distinguish between benign and malicious traffic across different threshold settings.

Precision-Recall Curve

The precision-recall curve shows consistent precision across a wide recall range, reflecting the model's robustness in detecting attacks even in an imbalanced dataset. This metric complements the ROC curve by focusing on positive (attack) detection.

5. DISCUSSION

The results highlight that Random Forest Importance (Top 30) yielded the highest Macro-F1 and Recall, demonstrating the effectiveness of embedded feature selection methods in IDS tasks. Chi-Square and ANOVA F-test performed competitively but slightly lower, while using all features did not provide significant improvement, indicating redundancy in the dataset. This confirms that IDS models can operate with fewer features without sacrificing accuracy, which is critical for real-time CPS security applications. The
confusion matrix and PR/ROC curves show balanced detection capability, although further fine-tuning could improve minority attack detection.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.01 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/ICFCSAI2025.011

6. CONCLUSION

The study concludes that Random Forest Importance-based feature selection provides the most effective balance of accuracy, recall, and computational efficiency. This aligns with prior work that emphasizes the robustness of embedded feature selection techniques.

For CPS deployment, the ability to reduce features while maintaining high detection performance is vital to meet real-time constraints. Future work should extend this study to additional classifiers (e.g., XG Boost, Gradient Boosting, Deep Learning) and explore hybrid feature selection methods such as Boruta or SHAP-based rankings to enhance generalization and explainability.

REFERENCES:

[1] M. A. Ferrag et al., "A systematic review of data mining techniques for intrusion detection systems," *J. Inf. Secur. Appl.*, vol. 50, 102419, 2020. [Online].

[2] I. Ahmed et al., "A survey of cyber-physical systems intrusion detection datasets," *Computers*, vol. 9, no. 2, p. 22, 2020.

[3] N. Shone et al., "A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 1, pp. 41–50, 2018.

[4] Z. Chiba et al., "A novel architecture combined with feature selection for effective intrusion detection in cloud computing," *Secur. Commun. Netw.*, vol. 2019, 8324012, 2019.

[5] G. Kocher and G. Kumar, "Analysis of ML algorithms with feature selection for IDS using UNSW-NB15," *IJNSA*, vol. 13, no. 1, pp. 13–
27, 2021.S. M. Kasongo and Y. Sun, "Performance analysis of IDS using feature selection on UNSW-NB15 dataset," *J. Big Data*, vol. 7, no. 105, 2020.

[6] Y. Yin et al., "IGRF-RFE: A hybrid FS method for MLP-based IDS on UNSW-NB15 dataset," *arXiv preprint arXiv:2203.16365*, 2022.

[7] A. M. Alsaffar et al., "Shielding networks: Enhancing IDS with Boruta FS for CPS security," *J. Big Data*, vol. 11, no. 29, 2024.

[8] Y. Zhang et al., "Network intrusion detection: Evaluating FS and ML classifiers on UNSW-NB15 and CICIDS2017," *Sensors*, vol. 21, no. 21, 7320, 2021.

[9] N. Shone et al., "A deep learning approach to network intrusion detection," *arXiv preprint arXiv:1804.09534*, 2018.

[10] T. A. Tang et al., "Deep learning approach for network intrusion detection in software defined networking," *2016 Int. Conf. WINCOM*, 2016.

[11] Z. Chiba et al., "A novel architecture combined with feature selection for effective intrusion detection in cloud computing," *Secur. Commun. Netw.*, 2019.

[12] I. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and hybrid efficient model," *J. Comput. Sci.*, vol. 25, pp. 152–160, 2018.

[13] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," *Proc. 9th EAI BIONETICS Conf.*, 2016.

[14] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSP*, 2018.