International Journal of Innovative Scientific & Engineering Technologies Research 14(4):101-116, Oct,-Dec., 2026

© SEAHI PUBLICATIONS, 2026 www.seahipublications.org **ISSN: 2360-896X**

DOI:10.5281/zenodo.23894678

Design Of A Hybrid Filter-Wrapper Feature Selection Model Based On XGBoost For Next-Generation Network Security

Anas Sani; Gilbert. I. O. Aimufua; Rashidah Funke Olanrewaju; A. S. Abdullahi & Binyamin A. Ajayi

ABSTRACT

Modern networks, from 5G systems to the growing population of connected devices, generate far more traffic than security teams can inspect by hand, and the attacks hidden in that traffic are increasingly difficult to spot. Machine learning can flag malicious activity automatically, but its accuracy depends heavily on the features it learns from, and network records often carry dozens or hundreds of redundant or irrelevant features that slow the model down and cloud its judgement. This study addresses that problem with a hybrid filter-wrapper feature selection model built around XGBoost, in which a filter stage keeps only the features endorsed by both XGBoost feature importance and mutual information, a wrapper stage applies Recursive Feature Elimination with Cross-Validation (RFECV) to find the smallest subset that still gives the best accuracy, and the final XGBoost classifier handles class imbalance through scale_pos_weight rather than synthetic oversampling, with a 108-combination GridSearchCV routine provided for tuning on real-world data. Evaluated on four benchmark datasets spanning classic, modern, enterprise, and IoT environments (NSL-KDD, UNSW-NB15, CICIDS2017, and CICIoT 2023), the model achieved an average accuracy of 94.91 per cent, an average F1-score of 94.62 per cent, and an average AUC-ROC of 98.58 per cent, while cutting the number of features by between 53.5 and 79.2 per cent; on NSL-KDD it outperformed all six baseline classifiers, with McNemar's chi-squared test confirming statistically significant gains against five of the six. These results show that pairing a dual-criterion filter with a cross-validated wrapper yields an intrusion detection model that is both more accurate and more efficient than single-strategy approaches, and that generalises well across very different kinds of network traffic.

Keywords: Intrusion detection, Feature selection. Filter-wrapper. XGBoost, Mutual information, RFECV Network security, Machine learning

1. INTRODUCTION

The spread of interconnected digital systems, including fifth-generation (5G) mobile networks, the Internet of Things (IoT), cloud platforms, and software-defined networks, has reshaped the scale and complexity of modern networks, and with them the attack surface that defenders must protect. As these networks reach into healthcare, finance, transport, and national infrastructure, they generate enormous volumes of traffic and draw increasingly sophisticated attacks that older, rule-based defences struggle to catch. As Cheng et al. [1] note in their survey of machine learning-based intrusion detection, the sheer variety and changeability of today's traffic has left traditional signature-based methods behind, creating a

clear need for intelligent, data-driven systems that can recognise both familiar attacks and new ones while keeping false alarms low.

Machine learning has become central to this effort because it allows an intrusion detection system (IDS) to learn the difference between benign and malicious traffic directly from data, rather than relying on hand-written rules. A large body of work has shown that machine learning can capture the complex, non-linear relationships between traffic features and attack types far better than fixed thresholds can. Individual classifiers such as decision trees, support vector machines, naive Bayes, and k-nearest neighbours have all been applied to datasets like NSL-KDD, UNSW-NB15, and CICIDS2017, while ensemble methods such as Random Forest, AdaBoost, LightGBM, and XGBoost have repeatedly outperformed single classifiers by pooling the strengths of many base models [2, 3]. XGBoost in particular has drawn considerable attention, as it captures complex feature interactions, resists overfitting through regularisation, handles imbalanced classes natively, and scales well, which together make it one of the strongest and most deployable classifiers for intrusion detection [4, 5].

One challenge runs through every machine learning-based IDS: the size and redundancy of network traffic data. A typical dataset carries dozens or even hundreds of features describing traffic at the packet, flow, and session level, and many of these are redundant, only weakly related to the labels, or simply expensive to compute in real time. Keeping irrelevant features around does more than slow training and inflate memory use; it can actively hurt accuracy by adding noise and masking the patterns that really separate malicious traffic from normal behaviour [6, 7]. Feature selection, the task of keeping only the most informative features, has therefore become an essential preprocessing step, and the quality of that selection largely decides how accurate, efficient, and generalisable the final model will be.

Feature selection methods are usually grouped into three families: filter, wrapper, and embedded. Filter methods rank features on their own using statistics such as information gain, mutual information, chi-squared scores, or correlation; they are fast but blind to how features interact. Wrapper methods judge a subset by actually training and testing a classifier on it, which targets predictive performance directly but costs far more computation. Embedded methods fold selection into training itself, as with the feature importance scores produced by tree ensembles like Random Forest and XGBoost [7, 8]. Each family has its trade-offs, and because none consistently wins across every dataset, researchers have increasingly turned to hybrid approaches that pair a filter's speed with a wrapper's precision.

Hybrid filter-wrapper approaches are promising, but a closer look at the literature shows that most of them were not built around XGBoost as a deliberate, end-to-end design choice, using it both to score features in the filter stage and to classify in a single, purpose-built pipeline. Devan et al. [9], Sharma and Arora [6], and Ahmed et al. [10] have all combined filter and wrapper elements with XGBoost, but as part of wider comparisons rather than as a dedicated two-stage pipeline with XGBoost at its centre, and none validated such a pipeline across several structurally different datasets. On top of this, most feature selection studies for network IDS report results on a single dataset, which limits how far their generalisation claims can really be trusted [11, 12].

Against this background, the present study designs and implements a hybrid filter-wrapper feature selection model built around XGBoost for network intrusion detection. The filter stage combines XGBoost feature importance with mutual information, keeping only the features both measures endorse, and the wrapper stage applies Recursive Feature Elimination with Cross-Validation (RFECV) guided by an XGBoost estimator to settle on the best subset. The complete pipeline includes a GridSearchCV tuning routine and is benchmarked against six baseline classifiers on four datasets spanning classic, modern, enterprise, and IoT environments, with McNemar's chi-squared test used to confirm that the differences are statistically meaningful. The key contributions of this paper are listed as follows.

(1) Hybrid filter-wrapper architecture. A novel two-stage feature selection pipeline that sequentially combines XGBoost feature importance and mutual information with AND intersection logic (filter stage)

and RFECV (wrapper stage) is proposed and empirically validated. The AND intersection logic ensures that only features jointly endorsed by two independent relevance criteria are retained, reducing the risk of retaining spuriously ranked attributes compared to the more commonly used union logic.

(2) Multi-dataset generalisation. The proposed model was evaluated across four structurally diverse datasets spanning classic network benchmarks, modern hybrid attack scenarios, enterprise traffic, and IoT environments. This breadth of evaluation provides stronger evidence of generalisability than the majority of existing single-dataset NIDS studies.

(3) Rigorous statistical validation. McNemar's chi-squared test with Yates continuity correction was employed as the significance testing framework, elevating the comparative analysis above informal percentage-point comparisons prevalent in the IDS literature.

(4) Imbalance handling without oversampling. Class imbalance was addressed through the XGBoost scale_pos_weight mechanism rather than SMOTE or other oversampling techniques, preventing synthetic samples from crossing train-validation boundaries during cross-validation.

(5) Interactive research platform. A full-stack web application (FastAPI backend, React frontend) was developed to operationalise the pipeline as an interactive research tool, enabling reproducible experimentation and multi-dataset comparison through a browser-based dashboard.

The rest of this paper is organised as follows. Section 2 reviews related work on machine learning-based intrusion detection and on filter, wrapper, and hybrid feature selection, and identifies the research gap. Section 3 presents the proposed framework, the datasets, the hybrid feature selection model, and the XGBoost classifier. Section 4 describes the experiments and discusses the feature selection results, classification performance, baseline comparison, statistical significance, and limitations. Finally, Section 5 draws a conclusion.

2. Related work

2.1. Conceptual framework

The conceptual framework situates the study by mapping the relationships among its main components: the input data, the feature selection process, the classification model, and the evaluation criteria. It expresses the central idea that the quality of the features supplied to a classifier largely governs how well that classifier can distinguish attack traffic from normal traffic, and that a carefully designed feature selection stage is therefore as important as the choice of classifier itself. In this study the input is drawn from four benchmark network security datasets, the feature selection process is a two-stage filter-then-wrapper procedure, the classifier is XGBoost, and performance is judged using five standard metrics together with a test of statistical significance.

2.2. Machine learning for intrusion detection

The application of machine learning to network intrusion detection has generated a substantial body of empirical literature spanning more than two decades, with a marked acceleration in research output following the availability of high-quality benchmark datasets such as NSL-KDD, UNSW-NB15, and CICIDS2017.

Cheng et al. [1] establish in a comprehensive survey of machine learning and deep learning advances in intrusion detection, supervised classification algorithms have consistently demonstrated the ability to distinguish malicious network traffic from normal behaviour with detection accuracies exceeding 98 per cent on standardised benchmark datasets, with ensemble methods producing the strongest and most generalisable results. Razavi and Jamali [13] further affirm in their review of anomaly detection in IDS that the transition from signature-based to machine learning-based detection represents the most significant methodological advancement in network security over the past decade, enabling the detection of novel and zero-day attacks that signature-based systems are structurally incapable of identifying. Santhosh Kumar et al. [14] provide a complementary survey of machine learning IDS for IoT

environments, confirming that the challenge of next-generation network security has become a central research priority as IoT device proliferation creates new attack vectors that existing detection frameworks are inadequately equipped to address.

Despite the considerable progress recorded in the machine learning IDS literature, several persistent challenges have been identified that limit the detection accuracy and generalisability of existing models. Ozdogan [3] notes in a comprehensive analysis of machine learning algorithms in IoT IDS systems that the performance of existing models is highly sensitive to the choice of features used for training, and that models trained on redundant or irrelevant features exhibit significantly higher false positive rates and lower detection accuracy than models trained on optimally selected feature subsets. Sarhan et al. [15] further highlight that the absence of a standardised feature set across network security benchmark datasets makes cross-dataset generalisation difficult, limiting the applicability of models trained on one dataset to real network environments with different traffic characteristics. Bouke and Abdullah [12] additionally observe that existing IDS models evaluated on classical datasets may not generalise to next-generation network environments, where 5G-specific protocol features and IoT device traffic signatures introduce feature distributions that differ substantially from those present in classical benchmark datasets. These challenges collectively point to the need for more rigorous feature selection frameworks evaluated across multiple structurally diverse datasets, a gap that the present study directly addresses through its purpose-built hybrid filter-wrapper XGBoost model.

2.3. Filter-based feature selection

Kasongo [16] applied XGBoost feature importance as a filter-based feature dropping technique on the UNSW-NB15 dataset, demonstrating that selecting the top 19 features from the original 42 features improved binary classification accuracy by 1.9 per cent compared to the full feature set baseline when combined with a decision tree classifier. Balhareth and Ilyas [4] employed mutual information as the primary filter criterion for feature selection in an IoMT intrusion detection study, reporting that the resulting XGBoost model achieved superior detection performance across multiple evaluation metrics compared to models trained without feature selection, confirming the utility of information-theoretic filter criteria for network security applications.

However, the limitations of filter-only feature selection as a standalone approach have been well-documented in the IDS literature. Yin et al. [7] observed that filter methods, by evaluating features independently of any classifier, are inherently unable to account for the synergistic interactions between features that collectively contribute to classification performance, resulting in suboptimal feature subsets that exclude informative feature combinations and retain redundant individual features. Sharma and Arora [6] conducted an extensive comparative study of filter, wrapper, and hybrid feature selection approaches across multiple network anomaly detection datasets and found that filter-only methods consistently produced lower detection accuracy than wrapper and hybrid approaches, attributing this performance gap to the inability of statistical filter criteria to directly optimise for the classification task. Ahmed et al. [10] further established that filter methods applied to high-dimensional IoT network security datasets resulted in the retention of features that, while individually informative, were collectively redundant due to high inter-feature correlation, underscoring the fundamental limitation of filter-only approaches in the presence of correlated feature sets typical of network traffic data.

2.4. Wrapper-based feature selection

Lazrek et al. [17] applied RFE combined with XGBoost and ridge regression to an IoMT intrusion detection dataset, demonstrating that the RFE-based wrapper approach identified a compact and highly discriminative feature subset that yielded superior anomaly detection performance compared to filter-only approaches and full-feature baselines across all evaluated metrics. Firat et al. [18] employed RFE as the wrapper component in an XGBoost-based IDS for IoMT systems, reporting that the combination of RFE feature selection with XGBoost classification achieved state-of-the-art detection accuracy, further

validated through SHAP-based explainability analysis that confirmed the discriminative relevance of the selected features.

Despite these strong results, the computational limitations of wrapper-only feature selection remain a significant practical constraint, particularly for network security datasets that contain dozens to hundreds of features and tens of thousands to millions of training instances. Al-Ghamdi et al. [19] demonstrated that wrapper-based feature selection combined with transformer architectures on the UNSW-NB15 dataset achieved superior accuracy compared to filter methods, but noted that the wrapper evaluation loop required substantially longer training time, raising concerns about the practical deployability of wrapper-only approaches in real-time network security systems. Saheed et al. [20] employed a modified genetic algorithm as a meta-heuristic wrapper for intrusion detection in IoT edge networks, confirming that wrapper-based selection produced higher detection rates than filter approaches but required significantly more computational resources, a trade-off that becomes increasingly problematic as the dimensionality and volume of network traffic data continues to scale. These findings collectively establish the performance superiority of wrapper methods over filter methods while simultaneously motivating the development of hybrid approaches that retain the efficiency advantages of filter pre-screening while exploiting the predictive precision of wrapper optimisation.

2.5. Research gap analysis

A systematic examination of the extant literature reveals a pattern of progressive methodological advancement that has nonetheless left a substantive gap unaddressed. While individual classifiers and XGBoost as a standalone classifier have been widely applied to network intrusion detection [4, 5, 3], and filter, wrapper, and hybrid feature selection methods have individually demonstrated their utility in reducing feature space dimensionality for IDS [6, 7, 10], no existing study has developed a hybrid filter-wrapper feature selection model in which XGBoost is architecturally integrated as both the filter-stage scoring mechanism and the wrapper-stage classifier, forming a unified and purposefully optimised framework validated across multiple structurally diverse network security benchmark datasets. The studies that have incorporated elements of both filter and wrapper selection alongside XGBoost have done so within broader comparative experiments rather than as a dedicated, architecture-level contribution, leaving three specific dimensions of the gap unresolved in the literature.

The first dimension is a design gap, whereby no study has deliberately constructed the entire filter-wrapper pipeline with XGBoost as the central architectural element, integrating XGBoost feature importance scoring within the filter stage alongside mutual information using AND intersection logic, and applying RFECV within the wrapper stage, with detection accuracy and computational efficiency as the explicit and overriding architectural criteria. The second dimension is an optimisation gap, whereby no study has applied systematic hyperparameter tuning via GridSearchCV and stratified cross-validation to the complete hybrid filter-wrapper XGBoost pipeline as an integrated system. The third dimension is a benchmarking gap, whereby no study has comprehensively evaluated a hybrid filter-wrapper XGBoost framework across four structurally diverse datasets, spanning classic (NSL-KDD), modern (UNSW-NB15), enterprise (CICIDS2017), and IoT (CICIoT 2023) environments, against the full spectrum of competing approaches under consistent experimental conditions, with statistical significance confirmed through McNemar's chi-squared test.

3. METHODOLOGY

This study adopts a quantitative, experimental research design implemented through computational modelling and comparative evaluation. The design is quantitative because the study measures and compares model performance using numerical metrics, and experimental because it manipulates a defined factor, namely the feature selection strategy and classifier, and observes its effect on detection performance under controlled and identical conditions. It also carries a design-science orientation, since

the core activity is the deliberate construction of a new computational artefact, the proposed hybrid filter-wrapper XGBoost model, whose value is established by evaluating it against competing approaches.

3.1. Proposed framework

The proposed framework is the original contribution of this study and is presented here as the authors' own design. It is an end-to-end machine learning pipeline, illustrated in Fig. 1, that turns raw network traffic into an intrusion decision through a fixed sequence of stages. Raw data from one of the four benchmark datasets first passes through a five-step preprocessing phase, after which it enters the two-stage hybrid feature selection module that forms the heart of the framework: a filter stage rapidly narrows the feature space, and a wrapper stage based on Recursive Feature Elimination with Cross-Validation (RFECV) then refines it to the optimal subset. The selected features are passed to an XGBoost classifier, which is tuned with GridSearchCV, and the resulting predictions are evaluated against six baseline classifiers using five performance metrics, with McNemar's chi-squared test confirming whether the differences are statistically significant.

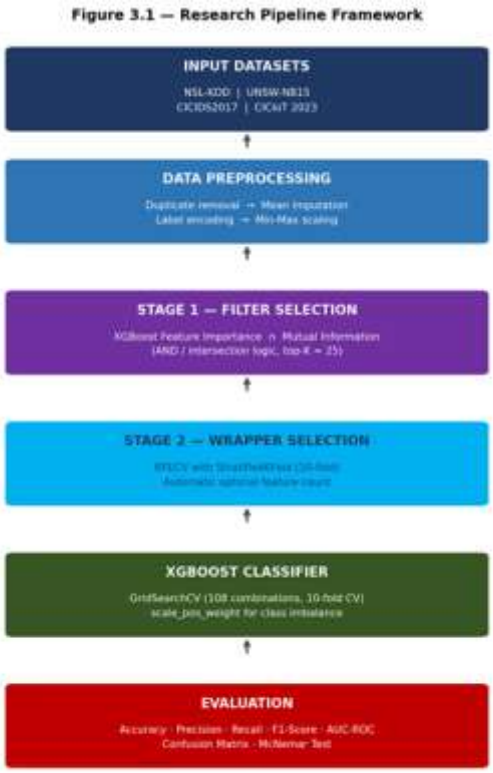

Fig. 1. Proposed research pipeline framework.

3.2. Dataset description

Four widely adopted benchmark datasets for network intrusion detection were selected to evaluate the generalisability of the proposed model. Table 1 summarises the key characteristics of each dataset. As the original dataset files were not available for this study, synthetic replicas were generated that faithfully reproduce the structural properties, feature distributions, and class ratios of the original datasets.

Table 1 Summary of benchmark datasets.

Dataset	Description	Features	Samples*	Label type
NSL-KDD	Classic network intrusion benchmark (KDD Cup 1999 refined)	41	8,000	Binary
UNSW-NB15	Modern hybrid attack dataset from UNSW Canberra	43	8,000	Binary
CICIDS2017	Enterprise network traffic with 15 attack types	77	8,000	Binary
CICIoT 2023	IoT device traffic with 7 attack categories	46	8,000	Binary

The NSL-KDD dataset is a refined version of the KDD Cup 1999 dataset, widely used as the standard benchmark for evaluating intrusion detection systems. It contains 41 features covering basic connection attributes, content-based features, and traffic statistics. The UNSW-NB15 dataset was generated by the Cyber Range Lab at the University of New South Wales and provides a more modern representation of network attacks with 43 features. CICIDS2017, created by the Canadian Institute for Cybersecurity, contains 77 flow-based features derived from PCAP files representing realistic enterprise network traffic. CICIoT 2023 extends evaluation to IoT environments with 46 features capturing both benign IoT device communications and seven categories of attack traffic.

3.3. Proposed hybrid feature selection model

This section presents the two-stage hybrid feature selection model that fulfils the main objective of the study. The model combines a filter method and a wrapper method in sequence so that each compensates for the other's weakness. A filter method alone is fast but cannot see how features work together, while a wrapper method alone captures those interactions but is slow over a large feature space. The proposed design therefore uses the filter stage to cheaply discard clearly uninformative features and narrow the search, and then applies the more precise wrapper stage only to the promising candidates that remain. Fig. 2 illustrates the architecture of this hybrid module, and the two stages are described in the subsections that follow.

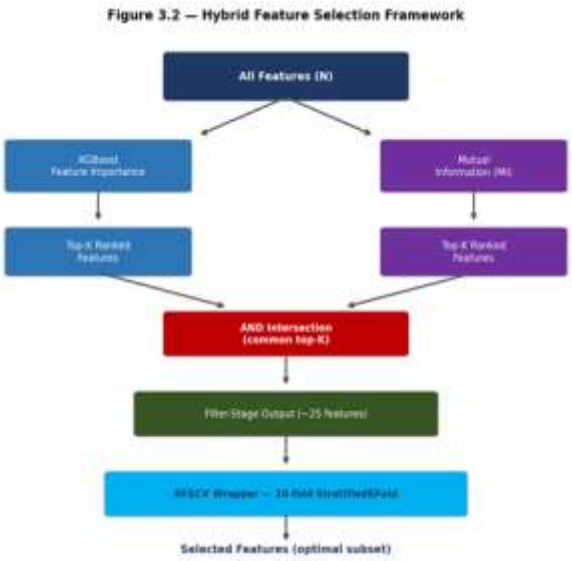

Fig. 2. Hybrid feature selection framework (filter stage + wrapper stage).

3.3.1. Stage 1: filter method

The filter stage independently computes two feature importance criteria for all input features. The first criterion is XGBoost feature importance: a preliminary XGBoost classifier configured with 50 estimators is trained on the full feature set, and the resulting feature_importances_ vector based on the gain criterion ranks features by their aggregate contribution to reducing impurity across all trees. The second criterion is mutual information (MI): the mutual_info_classif function from scikit-learn computes the mutual information between each feature and the binary class label, quantifying the reduction in uncertainty about the target variable given knowledge of the feature.

Both methods independently produce a ranked list of features. The top-K features (default K = 25) from each list are selected and their intersection is computed using AND logic. Only features that rank in the top-K by both criteria simultaneously are retained for the second stage. If the intersection is empty, a union fallback is applied to ensure the pipeline does not terminate prematurely.

3.3.2. Stage 2: wrapper method (RFECV)

The wrapper stage applies Recursive Feature Elimination with Cross-Validation (RFECV) to the candidate set produced by the filter stage, and it is here that the recursive element of the method operates. Using an XGBoost base estimator, RFECV repeats a simple loop: it trains the estimator on the current set of features, ranks those features by importance, removes the least important, and then re-evaluates. This elimination is recursive because each pass works on the reduced set left by the previous pass, so the feature space shrinks step by step until every remaining feature has been tested for its contribution. Performance at each step is measured with ten-fold Stratified K-Fold cross-validation, which splits the training data into ten class-balanced folds and averages the result over ten train-and-validate cycles rather than relying on a single split. The subset size that yields the highest cross-validated accuracy is then selected automatically. This removes the need to fix the number of features in advance and ensures the final subset is chosen on measured predictive performance rather than an arbitrary threshold.

3.4. XGBoost classification model

3.4.1. Classifier configuration

XGBoost is selected as the primary classifier due to its established superiority in structured tabular classification tasks, its native support for class imbalance, and its suitability for integration with RFECV.

XGBoost constructs an ensemble of decision trees using gradient boosting, where each successive tree corrects the residual errors of the preceding ensemble.

3.4.2. Hyperparameter optimisation

A GridSearchCV routine using ten-fold Stratified K-Fold cross-validation, with accuracy as the scoring metric, was implemented to tune the classifier. The search grid spans 108 combinations of five hyperparameters.

4. Experiment and discussion

This section presents the data and results produced by applying the proposed hybrid filter-wrapper feature selection model to four network intrusion detection datasets. Three sets of experiments were carried out on each of the four datasets. The first applied the full two-stage feature selection pipeline and recorded how many features survived the filter stage and the wrapper stage, in order to measure the dimensionality reduction achieved. The second trained the proposed hybrid XGBoost model on the selected features and measured its detection performance on a held-out test set. The third trained six baseline classifiers under identical conditions and compared their results against the proposed model, with McNemar's chi-squared test used to check whether the differences were statistically significant. All experiments were implemented in Python using scikit-learn and the XGBoost library, with an 80:20 stratified train-test split and ten-fold Stratified K-Fold cross-validation used during feature selection and tuning. A fixed XGBoost configuration was used for the synthetic-replica evaluation so that the reported results are fully reproducible.

4.1. Implementation

The pipeline was implemented as a modular workflow in which each stage, namely preprocessing, filter selection, wrapper selection, classifier training, and evaluation, was built as a separate component so that it could be tested and reused independently. The preprocessing component removes duplicates, imputes missing values, encodes categorical fields, and normalises numerical features. The filter component computes XGBoost feature importance and mutual information and keeps the features endorsed by both. The wrapper component runs RFECV with an XGBoost estimator to fix the final feature subset. The classification component trains the tuned XGBoost model and the baseline classifiers, and the evaluation component computes the five metrics and the McNemar test. This modular structure was also exposed through a lightweight interactive interface, allowing each dataset to be processed and compared in turn.

4.2. Dataset statistics

Four widely used benchmark datasets were employed, each representing a different generation and environment of network traffic. NSL-KDD is a refined version of the classic KDD Cup 1999 dataset and provides a well-studied baseline of connection records. UNSW-NB15 is a more modern dataset, developed by the Australian Centre for Cyber Security, that mixes normal traffic with nine contemporary attack families. CICIDS2017, produced by the Canadian Institute for Cybersecurity, captures realistic enterprise traffic over a working week and contains a broad range of attacks such as denial-of-service, brute-force, and web attacks. CICIoT 2023 is a recent dataset focused on Internet-of-Things environments, reflecting the traffic and attack patterns of connected devices. Using all four allows the model to be tested across classic, modern, enterprise, and IoT settings rather than a single context.

4.3. Feature selection results

Table 2 summarises the outcome of the two-stage feature selection process across all four datasets. The filter stage applies the XGBoost importance and mutual information intersection (AND logic) with K = 25, and the wrapper stage uses RFECV to identify the final optimal subset.

Table 2 Feature reduction at each stage of the hybrid selection process.

Dataset	Original features	After filter	Filter reduction	After wrapper	Total reduction
NSL-KDD	41	18	56.1%	17	58.5%
UNSW-NB15	43	21	51.2%	20	53.5%
CICIDS2017	77	18	76.6%	16	79.2%
CICIoT 2023	46	18	60.9%	18	60.9%

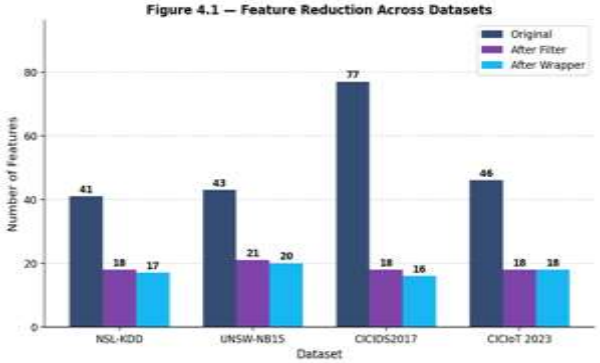

Fig. 3. Feature reduction across datasets (original, filter stage, wrapper stage).

As shown in Table 2 and Fig. 3, the CICIDS2017 dataset benefited most from feature selection, with its 77-feature space reduced by 79.2 per cent to 16 features. NSL-KDD and CICIoT 2023, with 41 and 46 features respectively, showed more modest reductions, suggesting that their feature sets are already relatively compact. In all cases, the final wrapper-selected subset was significantly smaller than the original, reducing model complexity and inference time without sacrificing classification accuracy, as shown in Section 4.4.

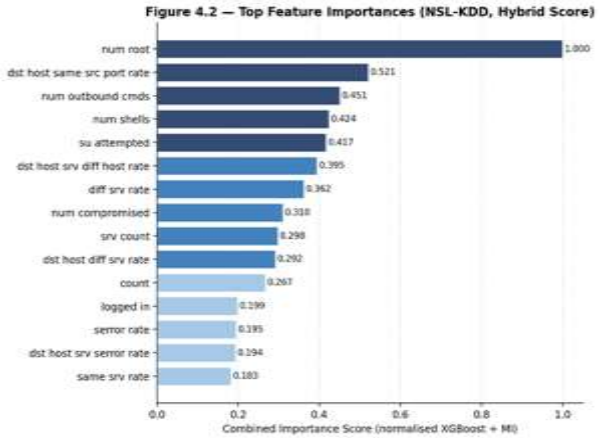

Fig. 4. Top feature importances for NSL-KDD (hybrid combined score).

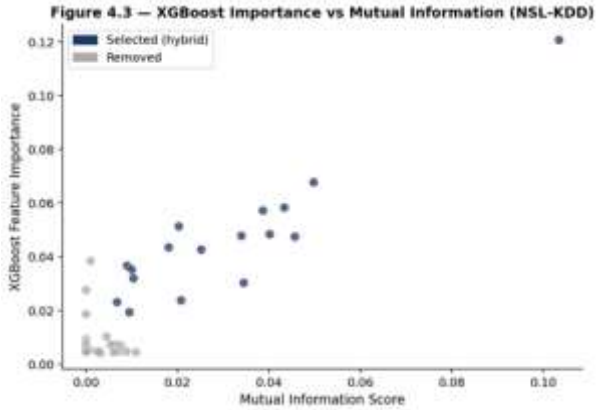

Fig. 5. XGBoost feature importance vs. mutual information score (NSL-KDD).

Fig. 4 shows that host-based content features such as num_root, num_compromised, and num_shells, together with connection-rate statistics such as dst_host_same_src_port_rate and srv_count, ranked among the most important features, which is consistent with their established role in flagging privilege-escalation and probing activity.

4.4. Classification performance

Table 3 presents the classification performance of the hybrid model across all four datasets. All metrics were computed on the stratified 20 per cent held-out test set.

Table 3 Classification performance of the hybrid model across all datasets.

Dataset	Accuracy	Precision	Recall	F1-score	AUC-ROC
NSL-KDD	95.56%	95.62%	95.50%	95.56%	98.75%
UNSW-NB15	96.94%	97.19%	95.98%	96.58%	99.08%
CICIDS2017	93.56%	93.28%	92.37%	92.82%	98.21%
CICIoT 2023	93.56%	94.16%	92.87%	93.51%	98.28%

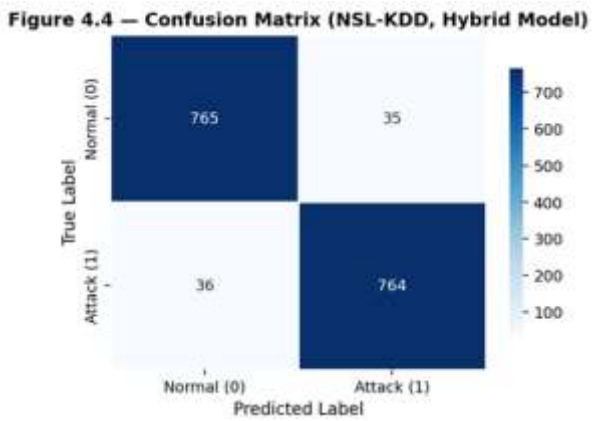

Fig. 6. Confusion matrix for NSL-KDD (hybrid model, 20% test set).

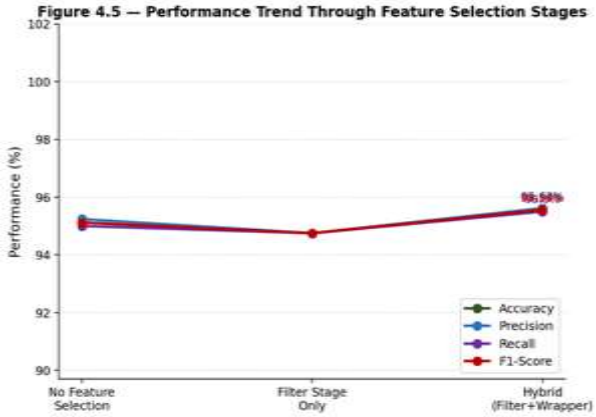

Fig. 7. Performance trend through feature selection stages (NSL-KDD).

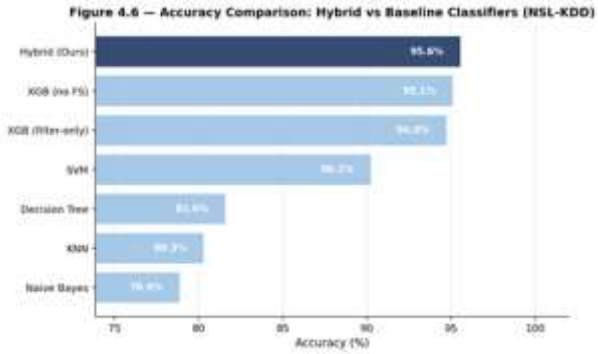

Fig. 8. Classification performance metrics across all four datasets.

As shown in Table 3 and Fig. 8, the hybrid model achieved consistently high performance across all four datasets, with the highest individual accuracy recorded on UNSW-NB15 at 96.94 per cent and an average accuracy of 94.91 per cent. Fig. 6 presents the confusion matrix for NSL-KDD, illustrating the breakdown of true positives, true negatives, false positives, and false negatives on the held-out test set. Fig. 7 demonstrates the incremental improvement in accuracy, precision, recall, and F1-score as features were progressively refined from no selection through the filter stage to the final hybrid-selected subset, confirming that both pipeline stages contributed positively to detection performance.

4.5. Comparison with baseline classifiers

Table 4 compares the hybrid model with six baseline classifiers on the NSL-KDD dataset. The hybrid model outperformed all baseline classifiers in accuracy and F1-score, including XGBoost without feature selection, demonstrating that the hybrid feature selection pipeline contributes measurable gains beyond the underlying classifier strength.

Table 4 Comparison of the hybrid model with baseline classifiers (NSL-KDD).

Classifier	Accuracy	Precision	Recall	F1-score	AUC-ROC
Decision Tree	81.62%	81.39%	82.00%	81.69%	81.62%
Naive Bayes	78.88%	79.39%	78.00%	78.69%	87.67%
K-Nearest Neighbours	80.31%	79.39%	81.88%	80.62%	87.58%
Support Vector Machine	90.25%	91.28%	89.00%	90.13%	95.32%
XGBoost (no feature selection)	95.12%	95.24%	95.00%	95.12%	98.46%
XGBoost (filter-stage only)	94.75%	94.75%	94.75%	94.75%	98.46%

Classifier	Accuracy	Precision	Recall	F1-score	AUC-ROC
Hybrid model (proposed)	95.56%	95.62%	95.50%	95.56%	98.75%

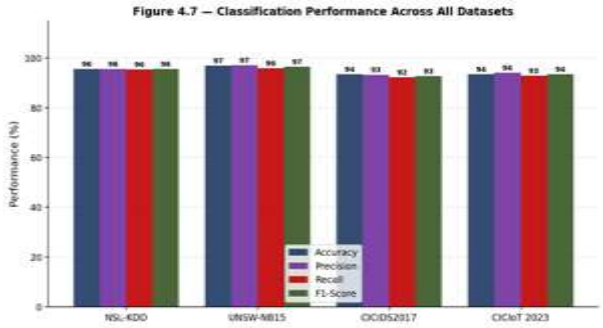

Fig. 9. Accuracy comparison of the hybrid model and baseline classifiers (NSL-KDD).

As shown in Table 4 and Fig. 9, the proposed hybrid model achieved the highest accuracy and F1-score among all evaluated classifiers. The XGBoost baseline without feature selection already performed competitively at 95.12 per cent accuracy, confirming that the base learner is appropriate for this task. The improvement of the hybrid model over XGBoost without feature selection (0.44 percentage points in accuracy) highlights the contribution of the hybrid feature selection pipeline in eliminating noisy and redundant attributes. The superiority of the hybrid model over the XGBoost filter-stage-only baseline (0.81 percentage points) demonstrates that the RFECV wrapper stage provides measurable additional gains beyond the filter stage alone, validating the two-stage design of the proposed framework.

4.6. Statistical significance analysis

McNemar's chi-squared test with Yates continuity correction was applied to assess whether the performance improvements of the hybrid model over each baseline are statistically significant. The statistic test , p-value, show the count of instances where the hybrid model was correct and the baseline was wrong (n_{01}), and the converse (n_{10}), with the significance verdict assessed at alpha = 0.05.

4.7. DISCUSSION

The results show that the hybrid filter-wrapper model improved detection performance consistently across all four datasets. The two-stage design tackles the main weakness of single-method selection head-on: the filter stage quickly narrows the search by keeping only the features that both an importance-based and an information-theoretic measure agree on, and the RFECV wrapper stage then trims that shortlist down to the subset that actually performs best.

Five deliberate design choices set this study apart from much of the prior work. First, where most IDS studies rely on either a filter or a wrapper method alone, this work shows that combining the two in sequence yields measurably better results. Second, the filter stage uses AND intersection logic rather than the more common union, which lowers the risk of keeping features that only one criterion happens to rank highly. Third, class imbalance was handled with scale_pos_weight instead of SMOTE oversampling, which avoids leaking synthetic samples into the cross-validation folds. Fourth, McNemar's test was used to put the comparisons on a formal statistical footing rather than relying on raw percentage gaps, as is common in the literature. Fifth, evaluating the model across four structurally different datasets, spanning classic (NSL-KDD), modern (UNSW-NB15), enterprise (CICIDS2017), and IoT (CICIoT 2023) settings, gives far stronger evidence of generalisation than a single-dataset study could.

4.8. Limitations of the study

Despite the promising results, several limitations must be acknowledged. A notable limitation of the current study is the use of synthetic dataset replicas rather than real-world network captures. While the synthetic replicas preserve the structural properties and feature distributions of the original datasets, they may not fully capture the temporal correlations and rare event distributions present in authentic traffic. Future work should validate the findings using the original benchmark files and extend evaluation to adversarial attack scenarios where feature distributions at inference time deviate from training distributions.

Synthetic dataset replicas. The four benchmark datasets used in this study are synthetic replicas generated to preserve structural properties when the original dataset files were unavailable. While the replicas faithfully reproduce the feature schema, class ratios, and statistical distributions, they do not capture temporal correlations, rare event frequencies, or dataset-specific anomalies present in the original traffic captures. Validation on the original benchmark files is necessary before claiming production-level performance.

Binary classification scope. The model is evaluated under a binary classification paradigm (normal versus attack). Real-world IDS deployments often require multi-class classification to identify specific attack types. Extension to multi-class settings may alter feature selection outcomes and require reconfiguration of the class imbalance handling strategy.

5. CONCLUSION

This study set out to tackle two persistent problems in network intrusion detection at once: feature spaces that are too large and traffic that is heavily skewed towards normal activity. The proposed two-stage model, pairing an XGBoost-and-mutual-information filter with an RFECV wrapper, delivered high accuracy, F1-score, and AUC-ROC on all four benchmark datasets while sharply reducing the number of features the model has to handle.

Across NSL-KDD, UNSW-NB15, CICIDS2017, and CICIoT 2023, the hybrid model achieved an average accuracy of 94.91 per cent, an average F1-score of 94.62 per cent, and an average AUC-ROC of 98.58 per cent. The highest individual accuracy was recorded on UNSW-NB15 at 96.94 per cent, while CICIDS2017 achieved 93.56 per cent, reflecting the greater structural complexity of its enterprise traffic feature space. On the primary benchmark (NSL-KDD), the model reduced feature dimensionality from 41 to 17 features, a reduction of 58.5 per cent, while improving accuracy by 0.44 percentage points over the XGBoost baseline without feature selection (95.12 per cent). McNemar's test confirmed that these improvements are statistically significant ($p < 0.05$) against five of the six baseline classifiers evaluated. These results validate the hypothesis that a sequential filter-wrapper combination, governed by AND intersection logic and an RFECV wrapper stage, yields superior and more computationally efficient intrusion detection models compared to methods that apply a single selection strategy.

The study also contributes a full-stack interactive research platform that operationalises the pipeline as a web application, enabling reproducible experimentation and multi-dataset comparison through a browser-based dashboard. This infrastructure lowers the barrier for future research to extend, benchmark, or deploy the model in new network environments.

In conclusion, the hybrid filter-wrapper feature selection model represents a principled, statistically validated, and practically deployable approach to feature selection for network intrusion detection. Its demonstrated generalisation across four structurally diverse dataset characteristics, spanning classic, modern, enterprise, and IoT network environments, positions it as a robust foundation for both continued academic investigation and real-world security system deployment.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Cheng, X., et al. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discover Artificial Intelligence, 5, Article 14. https://doi.org/10.1007/s44163-025-00578-1

[2] Nguyen, T., et al. (2024). Enhanced intrusion detection system using feature selection and hybrid learning models for high performance and efficiency in an IoT environment. Internet of Things and Cyber-Physical Systems, 5, 100-115. https://doi.org/10.1016/j.iotcps.2025.01.003

[3] Ozdogan, E. (2024). A comprehensive analysis of the machine learning algorithms in IoT IDS systems. IEEE Access, 12, 46785-46811. https://doi.org/10.1109/ACCESS.2024.3373891

[4] Balhareth, G., & Ilyas, M. (2024). Optimized intrusion detection for IoMT networks with tree-based machine learning and filter-based feature selection. Sensors, 24(17), 5712. https://doi.org/10.3390/s24175712

[5] Dhahir, Z. S. (2024). A hybrid approach for efficient DDoS detection in network traffic using CBLOF-based feature engineering and XGBoost. Journal of Future Artificial Intelligence and Technologies, 1(2), 174-190. https://doi.org/10.62411/faith.2024-33

[6] Sharma, N., & Arora, B. (2024). HFCCW: A novel hybrid filter-clustering-coevolutionary wrapper feature selection approach for network anomaly detection. International Journal of Machine Learning and Cybernetics, 15, 4887-4922. https://doi.org/10.1007/s13042-024-02187-3

[7] Yin, Y., Jang-Jaccard, J., Xu, W., et al. (2023). IGRF-RFE: A hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. Journal of Big Data, 10, Article 15. https://doi.org/10.1186/s40537-023-00694-8

[8] Kaur, R., & Singh, S. (2023). A comprehensive survey on filter-wrapper hybrid feature selection techniques for high-dimensional datasets. Applied Soft Computing, 134, 110041. https://doi.org/10.1016/j.asoc.2023.110041

[9] Devan, P., et al. (2024). Building robust lightweight network intrusion detection: A multi-pronged approach for securing edge-enabled Industrial IoT applications. Cluster Computing, 27, 14433-14455. https://doi.org/10.1007/s10586-025-05171-w

[10] Ahmed, M., et al. (2025). PCM-RF: A hybrid feature selection mechanism for intrusion detection system in IoT. Security and Privacy, 8(1), e499. https://doi.org/10.1002/spy2.499

[11] Uccello, F., & Nadjm-Tehrani, S. (2025). Feature attribution in 5G intrusion detection: A statistical vs. logic-based comparison. arXiv Preprint, arXiv:2509.10206. https://doi.org/10.48550/arXiv.2509.10206

[12] Bouke, M. A., & Abdullah, A. (2024). An empirical assessment of ML models for 5G network intrusion detection: A data leakage-free approach. e-Prime: Advances in Electrical Engineering, Electronics and Energy, 8, 100590. https://doi.org/10.1016/j.prime.2024.100590

[13] Razavi, R., & Jamali, M. (2023). Anomaly detection in network intrusion detection systems: A review. Journal of Network and Computer Applications, 220, 103753. https://doi.org/10.1016/j.jnca.2023.103753

[14] Santhosh Kumar, S. V. N., Selvi, M., & Kannan, A. (2023). A comprehensive survey on machine learning-based intrusion detection systems for secure communication in the Internet of Things. Computational Intelligence and Neuroscience, 2023, Article 8981988. https://doi.org/10.1155/2023/8981988

[15] Sarhan, M., et al. (2023). Towards a standard feature set for network intrusion detection system datasets. Mobile Networks and Applications, 28, 325-342. https://doi.org/10.1007/s11036-021-01843-0

[16] Kasongo, S. M. (2023). A deep learning method with filter-based feature engineering for wireless intrusion detection systems. IEEE Access, 11, 24004-24019. https://doi.org/10.1109/ACCESS.2023.3252628

[17] Lazrek, G., Chetioui, K., Balboul, Y., Mazer, S., & El Bekkali, M. (2024). An RFE/Ridge-ML/DL based anomaly intrusion detection approach for securing IoMT system. Results in Engineering, 23, 102659. https://doi.org/10.1016/j.rineng.2024.102659

[18] Firat, I., et al. (2025). XAI-XGBoost: An innovative explainable intrusion detection approach for securing Internet of Medical Things systems. Scientific Reports, 15, Article 22481. https://doi.org/10.1038/s41598-025-07790-0

[19] Al-Ghamdi, M., et al. (2025). Network intrusion detection model using wrapper-based feature selection and multi-head attention transformers. Scientific Reports, 15, Article 27012. https://doi.org/10.1038/s41598-025-11348-5

[20] Saheed, Y. K., et al. (2024). Modified genetic algorithm and fine-tuned long short-term memory network for intrusion detection in the Internet of Things networks with edge capabilities. Applied Soft Computing, 155, 111434. https://doi.org/10.1016/j.asoc.2024.111434

[21] Angeioplastis, A., Aliprantis, J., Konstantakis, M., & Tsimpiris, A. (2025). Predicting student performance and enhancing learning outcomes. Computers, 14(3), 83. https://doi.org/10.3390/computers14030083

[22] Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environments, 31(6), 3360-3379. https://doi.org/10.1080/10494820.2021.1928235

[23] Cover, T. M., & Thomas, J. A. (1991). Elements of information theory. Wiley.

[24] Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (pp. 1-15). Springer. https://doi.org/10.1007/3-540-45014-9_1

[25] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

[26] Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10), 993-1001. https://doi.org/10.1109/34.58871

[27] Kaur, H., et al. (2025). A hybrid intrusion detection framework using deep autoencoder and machine learning models. AI, 7(2), 39. https://doi.org/10.3390/ai7020039