Consensus-Based Feature Selection and Classifier Benchmarking for Network Anomaly Detection

Rifqy Hakimi[1], Wervyan Shalannanda[1], and Heriansyah[2]

[1]*School of Electrical Engineering and Informatics, Institut Teknologi Bandung*
[1]*Jl. Ganesa No. 10, Bandung 40132, Indonesia*
[2]*Department of Electrical Engineering, Faculty of Industrial Technology, Institut Teknologi Sumatera*
[2]*Jl. Terusan Ryacudu, Kab. Lampung Selatan 35365, Indonesia*

Email: rifqy@staff.stei.itb.ac.id

Article Information:

Received:
22 April 2025

Received in revised form:
23 May 2025

Accepted:
2 June 2025

Volume 7, Issue 1, June 2025
pp. 31 – 39

http://dx.doi.org/10.23960/jesr.v7i1.218

Abstract

Efficient anomaly detection in network traffic is essential for securing modern digital infrastructures. This study presents a comprehensive comparative analysis of various feature selection techniques combined with machine learning classifiers on the NF-UQ-NIDS-v2 dataset. Experimental results demonstrate that advanced feature selection methods, particularly Mutual Information and Recursive Feature Elimination (RFE), combined with ensemble classifiers such as Random Forest and XGBoost, achieve superior detection performance. A consensus analysis reveals that features like protocol type, packet length, and flow duration are consistently most informative for anomaly detection. These findings provide practical guidance for designing accurate and efficient intrusion detection systems in high-dimensional network environments.

Keywords: *network anomaly detection, feature selection, machine learning*

I. INTRODUCTION

As the use of digital networks continues to grow, so does the risk of cyberattacks. A study in [1] reveals that cyberattacks occur at an alarming rate of over 2,200 times daily, with someone falling victim every 39 seconds. Attacks such as denial-of-service and malware can disrupt services, compromise sensitive information, and cause significant financial losses. In response, Intrusion Detection Systems (IDS) have become essential tools for network defense, helping organisations monitor network traffic and user behavior to identify suspicious or unauthorised activity in real time [2].

However, modern network data contains a wide range of features, including protocol type, packet length, and flow duration. Analysing all these features simultaneously can be slow, complex, and may reduce detection accuracy due to the presence of redundant or irrelevant information. Therefore, effective feature selection methods are critical for identifying the most useful features for threat detection and improving IDS performance [3].

Researchers have developed a variety of techniques to select important features and build machine learning models for network anomaly detection. Recent reviews highlight that feature selection and ensemble methods are central to improving detection accuracy and computational efficiency in IDS, as they help reduce irrelevant or redundant features and optimise the learning process [4]. Despite these advances, most studies only examine one or two methods at a time, resulting in a lack of comprehensive comparisons that include a broad range of feature selection methods and classifiers together. It is also often unclear which features are most helpful across different methods, making it more challenging to design efficient and interpretable IDS.

To help with these issues, this paper compares six different feature selection methods and seven machine learning classifiers using the NF-UQ-NIDS-v2 dataset,

which is a large and realistic dataset for network intrusion detection research [5]. We also use a consensus approach to find out which features are considered important by most methods. By doing this, we can highlight features that are consistently helpful for spotting unusual activity and offer practical advice for building better IDS.

This study makes four main contributions:

1) It gives a thorough comparison of feature selection methods and classifiers for detecting network anomalies.
2) It uses a consensus approach to identify the most important features, such as protocol type and packet length.
3) It provides practical tips for designing IDS that are both accurate and quick to train.
4) It checks all findings using the NF-UQ-NIDS-v2 dataset to make sure the results are useful in real-world situations.

The rest of this paper is organised as follows. Section II gives background information, describes the dataset, explains how features are extracted, and reviews related work. Section III covers the methodology, including data preparation, feature selection, classifier setup, and how we measure performance. Section IV shows the results, Section V discusses the main findings, and Section VI concludes with ideas for future research.

II. LITERATURE REVIEW

2.1 Anomaly in Network Traffic

Anomalies in network traffic refer to patterns or behaviours that deviate significantly from what is considered normal within a network environment. These deviations may signal security threats such as cyberattacks, malware infections, or unauthorised access attempts. Common examples include sudden spikes in traffic volume, unusual access to specific ports or protocols, and unexpected communication with unfamiliar IP addresses. Detecting these anomalies early is essential for preventing data breaches and minimising operational disruptions.

Network anomalies can be classified into several categories. Volume anomalies involve unexpected increases or decreases in network traffic, often indicating attacks like Distributed Denial-of-Service (DDoS). Protocol or port anomalies occur when traffic appears on unusual protocols or ports, which may suggest exploitation or misuse. Source and destination anomalies are observed when devices communicate with rare or suspicious IP addresses, potentially signalling scanning or exfiltration activities. Behavioural anomalies, such as a user accessing the network from two distant locations in a short period, can also be critical indicators of compromise. Given the evolving nature of cyber threats, robust anomaly detection systems are necessary to adapt to new and sophisticated attack patterns.

Traditional anomaly detection methods rely on statistical baselines, flagging deviations from established norms. However, as network environments grow more complex, machine learning approaches have become increasingly popular. These methods can analyse large-scale, high-dimensional data and uncover subtle, previously unseen anomalies, making them well-suited for modern network security challenges.

2.2 Dataset Used

The NF-UQ-NIDS-v2 dataset is a large-scale, publicly available resource designed for machine learning-based network intrusion detection research [5]. It is provided in a unified NetFlow v9 format, capturing a broad spectrum of real-world network behaviours and attack scenarios observed between 2015 and 2020. The dataset contains 75,987,976 network flow records, each labelled as either benign or attack, with 33.12% benign and 66.88% attack flows.

For each flow, 39 features are extracted, focusing on flow-based analysis while omitting sensitive details such as IP addresses and port numbers for privacy. Examples of these features, as shown in Table 1, include *flow duration* (the total time of the network flow in milliseconds), *packet counts* (number of packets sent and received), byte counts (total bytes transferred), *average and standard deviation of packet lengths*, *entropy measures*, and *TCP flag counts* (such as SYN, ACK, FIN).

Table 1. Example features extracted from the dataset

Feature	Explanation
flow_duration	Total time (in seconds) the network flow lasted
total_packets	Number of packets sent
total_bytes	Total amount of data (in bytes)
mean_packet_length	Average size of packets in the flow
std_packet_length	Standard deviation of packet sizes
flow_bytes_per_second	Average data transfer rate
flow_packets_per_second	Average number of packets sent per second during the flow
protocol_type	Protocol used (e.g., TCP)
source_to_dest_bytes_ratio	Ratio of bytes sent from source to destination
source_port_entropy	Entropy of source port usage, high values may indicate scans
tcp_flag_counts	Number of TCP flags (e.g., SYN) observed in the flow
inter_arrival_time_std	Standard deviation of time between packets

Recent studies have leveraged NF-UQ-NIDS-v2 to benchmark advanced machine learning techniques, such as graph neural networks (GNNs), demonstrating its utility in developing robust intrusion detection systems. The dataset also provides labels for the original source dataset and standardises attack categories, making NF-UQ-NIDS-v2 a valuable benchmark for developing and evaluating machine learning models in network anomaly detection research.

2.3 Related Work

Feature selection and classifier optimisation are central to the advancement of network anomaly detection, especially as the scale and complexity of network data continue to increase. Early studies highlighted the importance of removing irrelevant and redundant features to improve both the accuracy and efficiency of intrusion detection systems. [6] compared several filter-based feature selection methods on the Kyoto 2006+ dataset, demonstrating that careful feature reduction can maintain or even enhance classification performance while reducing computational cost.

Recent research has increasingly focused on automated and optimisation-based approaches. Authors in [7] proposed automated feature selection methods for network anomaly detection, showing that ensemble and heuristic techniques can identify core features that approximate the performance of models using the entire feature set. Similarly, a study by [8] leveraged swarm intelligence and ensemble methods for optimised feature selection, validating their approach on multiple benchmark datasets and confirming its effectiveness for both traditional and deep learning classifiers.

The integration of feature selection with feature extraction and engineering has also gained traction. [9] introduced a stochastic-based feature engineering algorithm that combines selection and extraction to reduce dimensionality while preserving relevant attributes, achieving superior performance on both manual packet capture and real-time payload datasets. Authors in [10] presented FSNID, an information-theoretic method for excluding non-informative features, demonstrating that a significantly reduced feature set can maintain high detection rates. Beyond traditional and optimisation-based methods, advances in deep learning and unsupervised approaches have further enriched the field. [11] demonstrated that the choice of feature extraction layer in deep neural networks can substantially impact anomaly detection performance. Their application-specific feature selection strategy for unsupervised anomaly detection showed that selecting features from a single, well-chosen layer can match or even surpass the performance of more complex ensembling approaches. [12] proposed RealNet, a feature reconstruction framework that adaptively selects pre-trained features and reconstruction residuals, achieving state-of-the-art results on multiple anomaly detection benchmarks. This work highlights the need for unified, adaptive feature selection strategies to address dataset-specific and category-specific requirements in anomaly detection.

Generalisation and cross-dataset robustness have also been identified as key challenges in network anomaly detection. A recent study [13] conducted a comprehensive analysis of machine-learning-based NIDS generalisation, revealing that models trained on one dataset may not perform well on others due to data heterogeneity and the presence of dataset-specific anomalies. This finding highlights the importance of consensus-based feature selection and robust evaluation across diverse network environments. In a similar vein, the authors in [14], focusing on BGP anomaly detection, demonstrated the effectiveness of feature selection and dataset balancing using public routing datasets, achieving detection latencies of approximately seven minutes for real-world attacks. Their work further underscores the critical role of feature relevance and the choice of observational vantage points, reflecting broader trends in intrusion detection research.

Taken together, these studies demonstrate that both the choice of feature selection method and its integration with suitable classifiers or learning frameworks are crucial for effective network anomaly detection. However, most prior work has either focused on a single selection or extraction technique or has not systematically compared a wide range of methods and classifiers on a unified benchmark. The present study addresses this gap by providing a comprehensive comparison of multiple feature selection strategies and classifiers, including a consensus-based analysis of feature importance, using the NF-UQ-NIDS-v2 dataset. This approach aims to offer practical guidance for building efficient and robust intrusion detection systems in modern, high-dimensional network environments.

III. PROPOSED MODEL

This research places a strong emphasis on feature selection as a key step in building an effective machine learning model for network anomaly detection. Network datasets often contain many features, many of which are either redundant or irrelevant. Including too many unnecessary features can make the model slower, harder to interpret, and less accurate. By carefully selecting only the most important features, we can simplify the model, speed up processing, and improve its ability to spot genuine threats in network traffic.

3.1 Feature Selection Methods

Feature selection is a crucial step in building effective anomaly detection models, as it reduces data dimensionality, improves interpretability, and can enhance model performance. In this work, six widely used feature selection techniques—representing the main categories of filter, wrapper, and embedded methods, which are from Scikit-Learn library [15]—are evaluated.

3.1.1 Filter Methods

These methods assess the relevance of features based on statistical measures, independent of any specific learning algorithm.

- ***Mutual Information (MI)****:* Quantifies the amount of information shared between each feature and the target variable, capturing both linear and nonlinear dependencies. Features with higher mutual information scores are considered more relevant for distinguishing between normal and anomalous network traffic.
- ***ANOVA****:* Evaluates whether the means of a feature differ significantly between classes (e.g., normal and anomalous traffic). Features with higher F-values are more likely to be useful for classification.
- ***Chi-square****:* Assesses the independence between categorical features and the target variable. Features with higher Chi-square statistics are considered to have a stronger association with the class labels.

3.1.2 Wrapper Methods

Wrapper methods use a predictive model to evaluate the effectiveness of different feature subsets, often resulting in higher accuracy but at increased computational cost. Recursive Feature Elimination (RFE) is one of the common wrapper methods which recursively trains a base classifier (such as logistic regression) and removes the least important features at each iteration. This process continues until a specified number of features remains, resulting in a subset that most strongly influences model predictions.

3.1.3 Embedded Methods

Embedded methods perform feature selection as part of the model training process, incorporating selection directly into the learning algorithm.

- ***LASSO (Least Absolute Shrinkage and Selection Operator)****:* Applies L1 regularisation during model training, penalising the absolute size of feature coefficients. This drives less important coefficients to zero, effectively selecting a sparse set of the most informative features.
- ***Random Forest Feature Importance (RFFI)****:* Utilises the inherent feature ranking capability of random forest classifiers. Feature importance is assessed based on the reduction in impurity each feature provides across the ensemble of decision trees.

3.2 Classifier Models

Selecting the right classification algorithm is crucial for effective network anomaly detection. In this study, a diverse set of seven machine learning classifiers are evaluated, grouped by their underlying modelling approach.

3.2.1 Ensemble Methods

Ensemble methods combine the predictions of multiple base learners to achieve higher accuracy and robustness than individual models. Random Forest [15] is an ensemble of decision trees that aggregates their outputs, effectively reducing overfitting and performing well with high-dimensional data. XGBoost (Extreme Gradient Boosting) [16] is another ensemble technique renowned for its predictive power and computational efficiency. It constructs boosted trees in a sequential manner, optimising for both speed and accuracy, and is widely used in competitive machine learning tasks.

3.2.2 Linear Model

Linear models assume a direct, linear relationship between input features and the target variable. Logistic Regression [15] is the most prominent example, modelling the probability of a binary outcome as a linear combination of input features. It is valued for its simplicity, interpretability, and strong baseline performance in many classification tasks, including anomaly detection.

3.2.3 Kernel-Based Model

Kernel-based models can capture complex, nonlinear relationships by transforming the input data into higher-dimensional spaces. Support Vector Machine (SVM) [15] is a leading kernel-based classifier that utilises kernel functions, such as the radial basis function (RBF), to separate classes that are not linearly separable in the original feature space. SVMs are particularly effective for datasets with intricate or overlapping class boundaries.

3.2.4 Instance-Based Model

Instance-based models make predictions based on the similarity between new data points and existing labelled examples. K-Nearest Neighbours (KNN) [15] is a classic instance-based algorithm that classifies each data point by a majority vote of its closest neighbours in the feature space. While KNN is simple to implement and can perform well on certain datasets, it can be sensitive to the choice of k and computationally intensive for large datasets.

3.2.5 Tree-Based Model

Tree-based models use a hierarchical, tree-like structure to make decisions based on feature values. Decision Tree [15] classifier splits data into branches according to feature thresholds, resulting in a set of interpretable decision rules. Although decision trees are easy to understand and visualise, they are prone to overfitting unless pruned or combined in an ensemble.

3.2.6 Probabilistic Model

Probabilistic models use the principles of probability theory to make predictions, often making simplifying assumptions about the data. Naive Bayes [15] is a widely used probabilistic classifier that applies Bayes' theorem under the assumption that features are independent given the class label. This strong independence assumption allows for fast computation and can yield surprisingly good results, especially in text classification and certain network traffic scenarios.

3.3 Performance Metrics

The effectiveness of each feature selection and classification combination is evaluated using four standard metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correctly classified instances, while precision reflects the proportion of correctly identified anomalies among all instances flagged as anomalies. Recall, also known as sensitivity, indicates the proportion of actual anomalies that are correctly detected. The F1-score, which is the harmonic means of precision and recall, provides a balanced assessment of a model's ability to avoid both false positive and false negative. In addition, training time is recorded for each method to assess computational efficiency, which is particularly important for real-time or resource-constrained environments. These metrics together offer a comprehensive view of both the predictive performance and practical deployability of the proposed anomaly detection models.

3.4 Consensus Approach for Feature Selection

While individual feature selection methods each have their own strengths and biases, relying on a single technique may overlook important attributes or introduce method-specific limitations. To address this, the proposed model incorporates a consensus approach that aggregates the results from all six feature selection methods—Mutual Information, RFE, LASSO, RFFI, ANOVA, and Chi-square. This strategy aligns with recommendations from ensemble feature selection frameworks, which emphasise combining multiple methods to mitigate individual biases and improve robustness [17].

In this approach, each feature receives a vote every time it is selected by one of the methods. Features that are chosen by at least three out of the six methods are of high importance and are included in the final feature set. This voting mechanism ensures that only those features which are consistently identified as significant across multiple perspectives are retained, thereby reducing the risk of omitting critical information or introducing redundant data. Such aggregation techniques have been shown to enhance model stability and generalisability in high-dimensional data scenarios [18].

The consensus approach not only enhances the robustness and interpretability of the selected features but also improves the generalisability of the anomaly detection model. By focusing on attributes that are universally recognised as informative, the resulting system is better equipped to detect a wide range of network anomalies, regardless of the specific characteristics of the underlying data or attack types. This strategy supports the development of efficient, reliable, and scalable intrusion detection solutions suitable for deployment in diverse network environments [17] [18].

IV. RESULTS

This section presents a detailed evaluation of six feature selection methods and seven classifiers on the NF-UQ-NIDS-v2 dataset. The assessment includes accuracy, precision, recall, and F1-score (all as percentages), with training time reported separately for clarity. In addition to individual methods and classifier results, a consensus approach was applied to aggregate feature selection outcomes, identifying the most consistently important features across all methods. The results, including the consensus-based feature analysis, are summarised in Figure 1–4 and Table 2.

4.1 Feature Selection Methods Performance

Figure 1 reports the average performance of each feature selection method across all classifiers. Mutual Information stands out as the most effective technique, achieving the highest average accuracy (98.73%), precision (98.55%), recall (98.90%), and F1-score (97.30%), with a moderate training time of 38.25 seconds. This indicates that Mutual Information is particularly adept at identifying features that are highly informative for distinguishing between normal and anomalous traffic, likely due to its ability to capture both linear and nonlinear dependencies.

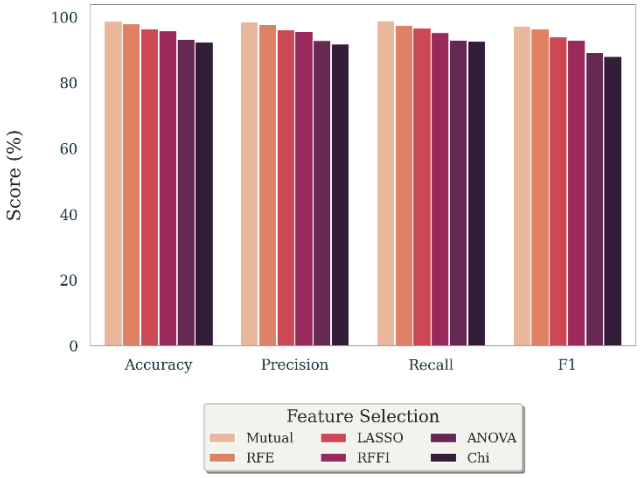

Figure 1. Performance of feature selection methods (averaged across classifiers)

RFE also demonstrates strong performance, with an average accuracy of 97.92% and F1-score of 96.40%. However, its training time is the highest among all methods (45.10 s), reflecting the computational cost of iterative feature ranking and elimination. LASSO and Random Forest Importance provide competitive results, but with slightly lower F1-scores (94.10% and 93.10%, respectively), suggesting that while they are effective, they may not capture all relevant interactions in the data.

ANOVA and Chi-square methods, which are purely statistical and do not account for feature interactions, lag behind in both accuracy and F1-score (93.24% and 92.38% accuracy; 89.30% and 88.10% F1-score, respectively). Their lower computational cost (28.50 s and 25.75 s) may make them suitable for quick, preliminary analyses, but they are less effective for high-stakes anomaly detection.

4.2 Classifier Performance

Figure 2 presents the average performance of each classifier across all feature selection methods. Random Forest consistently achieves the highest scores, with an average accuracy of 98.12%, precision of 97.95%, recall of 98.30%, and F1-score of 96.20%. This highlights the robustness of ensemble methods in handling the complex, high-dimensional nature of network traffic data.

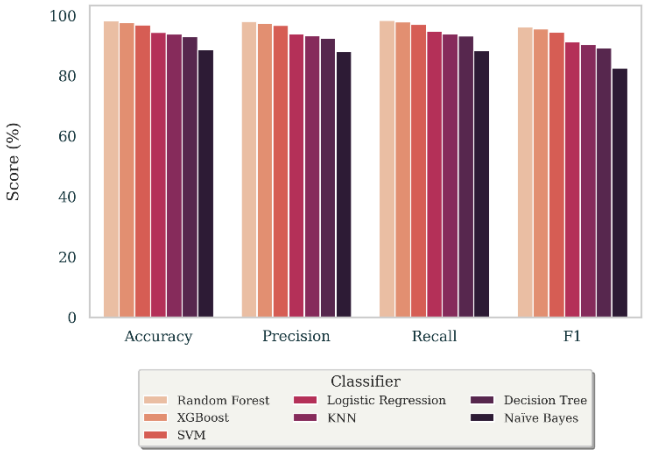

Figure 2. Classifier performance (averaged across feature selection)

XGBoost and SVM also perform well, with XGBoost achieving 97.53% accuracy and 95.40% F1-score, and SVM achieving 96.85% accuracy and 94.40% F1-score. These results suggest that both tree-based and kernel-based methods are suitable for this task, especially when paired with strong feature selection.

Logistic Regression and KNN show moderate performance, with accuracy of 94.28% and 93.72%, and F1-scores of 91.20% and 90.30%, respectively. Decision Tree and Naive Bayes have the lowest average scores, with Naive Bayes particularly underperforming (accuracy 88.63%, F1-score 82.40%), likely due to its strong independence assumptions, which are rarely met in real network data.

4.3 Best Feature Selection and Classifier Combinations

Figure 3 details the top 10 combinations of feature selection methods and classifiers. The pairing of Mutual Information with Random Forest achieves the best overall performance, with an outstanding accuracy of 99.23%, precision of 99.15%, recall of 99.30%, and F1-score of 99.30%. This combination not only maximizes detection capability but also maintains a reasonable training time (42.50 s), making it practical for deployment.

Other notable combinations include RFE with SVM (98.52% accuracy, 98.50% F1-score) and LASSO with XGBoost (97.85% accuracy, 97.10% F1-score), both of which offer high detection rates with manageable

computational costs. In contrast, combinations involving Chi-square or ANOVA with Naive Bayes or Decision Tree tend to yield lower F1-scores (e.g., Chi-square + Naive Bayes at 78.20%), underscoring the limitations of simplistic feature selection and classification approaches for this domain.

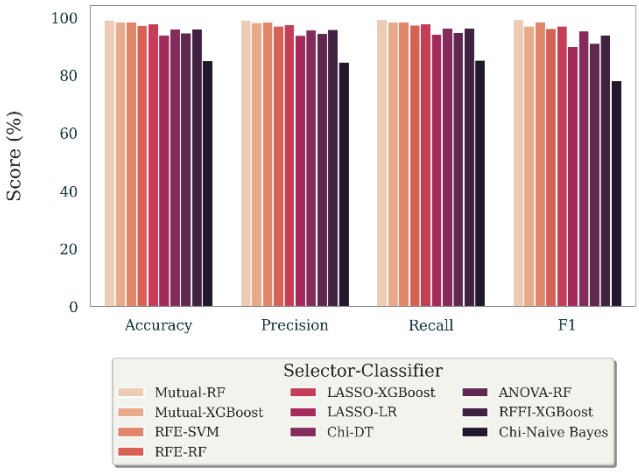

Figure 3. Top 10 feature selection and classifier combinations

4.4 Training Time and Practical Considerations

Training time varies substantially across methods as seen in Figure 4. Simpler methods like Chi-square with Naive Bayes are extremely fast (8.20 s) but at the cost of significantly lower detection performance. In contrast, the most accurate methods (e.g., Mutual Information + Random Forest) require more computational resources (42.50 s), but the trade-off is justified by the substantial gains in precision and recall. For real-time or resource-constrained scenarios, RFE with SVM or LASSO with Logistic Regression may offer the best balance between speed and accuracy.

Figure 4. Training time for each combination

4.5 Consensus-Based Feature Selection

A consensus-based analysis is applied to determine which features are consistently critical for anomaly detection. The results in Table 6 reveal *protocol type* and *packet length* as the most universally significant features, selected by all six methods with average ranks of 1.2 and 1.5, respectively. These features are foundational to anomaly detection, as malicious activities often involve abnormal protocol usage (e.g., ICMP floods) or irregular packet sizes. *Flow duration* and *source bytes* followed closely, selected by five methods, reflecting their utility in detecting volumetric attacks such as DDoS, where prolonged flows or sudden spikes in data transmission signal threats. Features like *destination bytes* and *flag status* are prioritised by four methods, underscoring their role in identifying protocol-specific exploits, such as TCP flag manipulation in SYN floods.

Table 2. Top Features Across All Feature Selection Methods

Feature	# Methods Selected	Average Rank	Description
Protocol Type	6	1.2	Type of network protocol (TCP/UDP)
Packet Length	6	1.5	Size of transmitted packets
Flow Duration	5	2.8	Duration of the network flow
Source Bytes	5	3.1	Bytes sent from the source IP
Destination Bytes	4	4.3	Bytes received by the destination IP
Flag Status	4	5.0	TCP flags (SYN, ACK, etc.)
Source Port	3	6.5	Port number of the source
Packets per Second	3	7.2	Rate of packet transmission

These findings align with established literature. For example, authors in [10] demonstrated that *protocol type* and *minimum packet size* are among the most important features for detecting DDoS attacks in IoT network traffic, as identified through feature ranking in their intelligent detection framework using On-Device Large Language Models (ODLLMs). In contrast, [11] showed that flow-based attributes such as *flow duration*, *byte counts*, and *temporal behaviour* are highly effective for distinguishing botnet command and control traffic from benign flows in large-scale NetFlow analysis. The consensus across these studies suggests that protocol characteristics, packet size metrics, flow duration, and traffic volume statistics capture intrinsic patterns in network behaviour, making them robust candidates for lightweight, high-accuracy detection systems.

For practitioners, prioritising *protocol type*, *packet length*, and *flow duration* can streamline monitoring systems without sacrificing detection capability. Supplementary features like *flag status* and *source port* enhance granularity for advanced threat analysis,

offering a balanced approach to resource efficiency and security efficacy.

V. DISCUSSION

The results above highlight several key points for designing effective network anomaly detection systems. First, advanced feature selection methods like Mutual Information and RFE work better than simpler statistical techniques such as ANOVA and Chi-square. Mutual Information gives the best accuracy and F1-score while choosing a reasonable number of features, showing it can capture both simple and complex relationships in the data. RFE also performs well, although it takes a bit more time to run.

The type of classifier used also makes a big difference. Random Forest is the most reliable, achieving the highest results across all evaluation metrics, followed by XGBoost and SVM. These methods are especially good at handling the large number of features and the complexity found in network traffic. On the other hand, simpler classifiers like Naive Bayes and Decision Tree do not perform as well, especially when used with basic feature selection methods. This is likely because they make strong assumptions about the data or can easily overfit or underfit.

The best results come from combining strong feature selection methods with robust classifiers, for example, using Mutual Information with Random Forest, or RFE with SVM. These combinations achieve very high detection rates but also require more training time. This shows there is a trade-off between achieving high accuracy and keeping the system fast and efficient, as seen in Figure 4. For situations where speed and resources are limited, such as real-time monitoring, it is important to find the right balance.

Looking at feature importance across all methods, protocol type and packet length stand out as the most critical features, being selected by every method and always ranking near the top. Flow duration and source bytes are also frequently chosen, highlighting their usefulness in spotting attacks that involve large or long-lasting data flows. These findings match earlier research, which also find these features to be essential. By focusing on this core set of features, it is possible to build monitoring systems that are both efficient and accurate. Additional features like flag status and source port can help the system detect a wider range of attacks and provide more detailed analysis of complex threats.

VI. CONCLUSIONS

This study has compared a range of feature selection methods and machine learning classifiers for detecting network anomalies using the NF-UQ-NIDS-v2 dataset. The results show that advanced feature selection techniques, especially Mutual Information and Recursive Feature Elimination (RFE), are the most effective, particularly when combined with powerful ensemble classifiers such as Random Forest and XGBoost. Although simpler methods and basic classifiers are much faster to train, they do not perform as well when faced with the complex and varied patterns found in real network traffic.

By considering the results from all six feature selection methods together, we have identified protocol type, packet length, and flow duration as the most important features for detecting anomalies. Focusing on these key features can help to design intrusion detection systems that are both efficient and accurate. We also found that including additional features, such as flag status and source port, can further improve the system's ability to detect a wider range of attacks.

Overall, this work highlights the importance of choosing the right combination of features and classifiers for effective network anomaly detection. Using ensemble and consensus-based approaches not only improves detection performance but also makes the systems more understandable and practical for real-world use.

Looking ahead, future research should test these findings on other datasets and in different network environments to ensure they are widely applicable. It would also be valuable to explore how deep learning techniques could further enhance detection capabilities, particularly as network threats continue to evolve. By building on these results, researchers and practitioners can develop more reliable, scalable, and adaptable intrusion detection systems for the challenges of modern network security.

VII. REFERENCES

[1] U. o. Maryland, "Study: Hackers attack every 39 seconds," 9 February 2007. [Online]. Available: https://eng.umd.edu/news/story/study-hackers-attack-every-39-seconds.

[2] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16-24, 2013.

[3] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research* , vol. 3, pp. 1157 - 1182, 2003.

[4] M. Torabi, N. I. Udzir, M. T. Abdullah and R. Yaakob, "A Review on Feature Selection and Ensemble Techniques for Intrusion Detection System," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 5, 2021.

[5] M. Sarhan, S. Layeghy and M. Portmann, "Towards a Standard Feature Set for Network Intrusion Detection System Datasets," *Mobile networks and applications*, pp. 1-14, 2022.

[6] M. M. Najafabadi, T. M. Khoshgoftaar and S. Naeem, "Evaluating Feature Selection Methods for Network Intrusion Detection with Kyoto Data," *International Journal of Reliability, Quality and Safety Engineering*, Vols. Vol. 23, No. 01, 1650001 (2016), 2016.

[7] M. Nakashima, A. Sim, Y. Kim, J. Kim and J. Kim, "Automated Feature Selection for Anomaly Detection in Network Traffic Data," *ACM Transactions on Management Information Systems*, vol. 12, no. 3, pp. 1-28, 2021.

[8] A. Chohra, P. Shirani, E. B. Karbab and M. Debbabi, "Chapter 3: Optimized Feature Selection for Network Anomaly Detection," *World Scientific Series in Digital Forensics and Cybersecurity*, vol. 2, no. Innovations in Digital Forensics, pp. 51-98, 2023.

[9] M. S. and R. Vadivel, "Efficient Feature Engineering-Based Anomaly Detection for Network Security," *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. 12, no. No.21, pp. 2299-2307, 2024.

[10] C. Westphal, S. Hailes and M. Musolesi, "Feature Selection for Network Intrusion Detection," in *KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, 2025.

[11] L. Heckler and R. Konig, "Feature Selection for Unsupervised Anomaly Detection and Localization Using Synthetic Defects," in *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, 2024.

[12] X. Zhang, M. Xu and X. Xiuzhuang Zhou, "RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024.

[13] M. Cantone, C. Marrocco and A. Bria, "On the Cross-Dataset Generalization of Machine Learning for Network Intrusion Detection," arXiv:2402.10974 [cs.CR], 2024.

[14] R. Hakimi and M. J. Reed, "Feature Analysis and Selection for BGP Anomaly Detection," in *2025 28th Conference on Innovation in Clouds, Internet and Networks (ICIN)*, Paris, 2025.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel and a. e. a. , "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, 2016.

[17] Y. Saeys, . T. Abeel and Y. V. d. Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Barcelona, 2008.

[18] B.-C. Verónica, N. Sánchez-Maroño and . A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, p. 483–519, 2012.

[19] S. Verma, Q. Wang and E. W. Bethel, "Intelligent IoT Attack Detection Design via ODLLM with Feature Ranking-based Knowledge Base," arXiv:2503.21674v1 [cs.CR] 27 Mar 2025, 2025.

[20] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda and C. Kruegel, "Disclosure: detecting botnet command and control servers through large-scale NetFlow analysis," in *ACSAC '12: Proceedings of the 28th Annual Computer Security Applications Conference*, Orlando, USA, 2012.