DOI: 10.6977/IJoSI.202504_9(2).0007
S. Viharika, N.A. Balaji/Int. J. Systematic Innovation, 9(2), 74-84 (2025)

Optimizing cloud-based intrusion detection systems through hybrid data sampling and feature selection for enhanced anomaly detection

Sadargari Viharika*, N. Alangudi Balaji

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, India

*Corresponding author E-mail: reddyviharika266@gmail.com

(Received 09 October 2024; Final version received 03 February 2025; Accepted 13 February 2025)

Abstract

To enhance detection accuracy in network intrusion scenarios, this study proposes an optimized intrusion detection system (IDS) framework that integrates advanced data sampling, feature selection, and anomaly detection techniques. Leveraging random forest (RF) and genetic algorithm, the framework optimizes sampling ratios and identifies critical features. In contrast, the isolation forest algorithm detects and excludes outliers, refining dataset quality and classification performance. Evaluated on the UNSW-NB15 dataset, comprising over 2.5 million records and 42 diverse features, the proposed framework demonstrates significant improvements in anomaly detection, including reduced false alarm rates and enhanced identification of rare threats, such as shellcode, worms, and backdoors. Experimental results reveal that the RF-based model achieves an F1 score of 91.8% and an area under the curve (AUC) of 96%, outperforming traditional machine learning models and standalone RF classifiers. The integration of extreme gradient boosting (XGB) and its optimized variant, XGBGA, further enhances the framework, with XGBGA achieving the highest performance metrics, including an F1 score of 92.8% and an AUC of 97%. These findings underscore the importance of data optimization strategies in improving the accuracy and reliability of IDSs, particularly in handling imbalanced datasets and diverse network traffic. Future work will focus on real-time processing capabilities to handle streaming data and expanding the framework's applicability to domains such as fraud detection and cybersecurity, where precise anomaly detection is essential.

Keywords: Anomaly Detection, Data Optimization, Intrusion Detection System, Machine Learning

1. Introduction

The importance of network security has escalated with the expansion of the Internet and the corresponding rise in the volume and complexity of network traffic. As part of the defense against hostile activities on networks, intrusion detection systems (IDSs) have become essential tools. These systems analyze network traffic and identify anomalies that may indicate security breaches. IDSs are generally categorized into two major groups: signature-based systems and anomaly-based systems. Signature-based IDSs, such as Snort, operate by maintaining large databases of known attack signatures and comparing incoming traffic to these predefined patterns to detect intrusions (Ahmad et al., 2021; Heidari et al., 2023; Heidari et al., 2024). These systems excel at identifying known threats, but face significant challenges in detecting new or evolving threats due to their reliance on existing signatures.

On the other hand, anomaly-based IDSs construct models of normal network behavior and flag deviations from these models as potential threats. These systems are particularly valuable in detecting previously unknown threats as they do not depend on predefined signatures. However, anomaly-based IDSs often suffer from high false alarm rates and reduced detection accuracy, mainly due to the large volume of network data and the skewed distribution between normal and anomalous activities. This data imbalance can result in the model focusing too heavily on more common behaviors, which can reduce its ability to detect rare but important anomalies (Chkirbene et al., 2020; Junwon et al., 2022).

To address these issues, recent studies have explored integrating data optimization techniques with machine learning models to enhance IDS performance. The present study proposes a hybrid data optimization-based IDS, termed RFGA, which combines data sampling and feature selection techniques to improve the accuracy and efficiency of anomaly detection (Hassan et al., 2024). Data sampling techniques such as oversampling and undersampling are used to resolve the issue of imbalanced data distribution by either amplifying the representation of rare events or reducing the frequency of common events (Heidari et al., 2024). Feature selection further refines the model by retaining only the most relevant features that help distinguish between normal and anomalous behaviors when discarding redundant or irrelevant data. The proposed RFGA employs the random forest (RF) algorithm as its core classification technique, utilizing optimized data inputs to build a more effective and robust detection system.

The major contributions of this paper are as follows:

1) The paper introduces a novel RFGA framework that combines isolation forest (iForest), genetic algorithm (GA), and RF to optimize data sampling and feature selection, significantly improving intrusion detection accuracy.
2) The system demonstrates superior performance in detecting rare and severe network anomalies, such as backdoors, worms, and shellcodes, compared to traditional methods.
3) The RFGA framework is rigorously evaluated using the UNSW-NB15 dataset, showing significant reductions in false alarm rates and better handling of imbalanced datasets.

The structure of this paper is as follows: A thorough assessment of relevant work on the subject of IDSs is given in Section 2, with an emphasis on different machine learning strategies and data optimization tactics. The main techniques used in the development of RFGA are introduced in Section 3, including GA, RF, and iForest. The RFGA framework's architecture and execution are described in detail in Section 4. The effectiveness of the suggested system is illustrated by a discussion of the experimental findings and their implications. Section 5 wraps up the investigation and makes recommendations for further research avenues.

2. Literature Review

The identification and mitigation of network anomalies have been central to the development of effective IDSs. Given the growing complexity of network environments, traditional IDS approaches have increasingly been supplemented by advanced data mining and machine learning techniques. These approaches aim to improve the detection rate of IDSs when minimizing false alarms, a balance that has proven difficult to achieve due to the inherent challenges in handling large and imbalanced datasets.

2.1. Data Sampling in IDSs

One of the primary challenges in intrusion detection is the uneven distribution of network data, where normal activities vastly outnumber anomalous ones. This imbalance can skew the performance of IDSs, making it difficult to detect rare but potentially severe threats. Data sampling techniques, such as oversampling and undersampling, have been used to tackle this issue. Oversampling methods, like the synthetic minority oversampling technique, increase the representation of minority classes by generating synthetic samples, thereby balancing the dataset (Heidari et al., 2024). Conversely, undersampling techniques reduce the number of majority class instances, as demonstrated by methods like EasyEnsemble and BalanceCascade, which selectively downsample the dataset to achieve a more balanced distribution (Heidari et al., 2024).

The effectiveness of data sampling in enhancing IDS performance has been highlighted in several studies. Research has explored the use of data sampling to improve the accuracy and speed of intrusion detection, utilizing the least squares support vector machine (SVM) to identify suspicious network activities. Findings indicated that sampling techniques could effectively select representative data subsets, thereby improving the detection capabilities of IDSs (Molina-Coronado et al., 2020). Further advancements include the integration of modified K-means clustering with machine learning techniques. The modified K-means algorithm identified common patterns across datasets, enabling more effective data compression and reducing the computational burden on the IDS. By combining K-means with the C4.5 decision tree algorithm and further enhancing detection through SVM and extreme learning machine techniques, this method significantly improved the efficacy and accuracy of IDSs, particularly in identifying Denial-of-Service attacks (Bukhari et al., 2024; Heidari et al., 2023).

2.2. Feature Selection in IDSs

Feature selection is as vital as data sampling in enhancing IDS performance. By removing superfluous or irrelevant features and selecting the most pertinent ones, feature selection approaches aim to reduce the dimensionality of the data. The filter, wrapper, and embedding methods are the primary strategies for selecting features.

Filter techniques assess each feature using statistical metrics like divergence or correlation, selecting the features most likely to improve classification performance (Deebak & Hwang, 2024). In contrast, wrapper techniques choose or eliminate features based on how well they contribute to the accuracy of the model, using a specific learning algorithm to assess their significance. Embedding approaches, such as decision trees, perform feature selection simultaneously with model training, guided by the learned weights of the features (Ahmad et al., 2021).

In IDSs, feature selection plays a crucial role, as several studies have demonstrated. For example, some studies employed logarithmic marginal density ratios to adjust initial features in SVM-based detection systems, yielding higher-quality features and improved classification efficiency (Hassan et al., 2024). Another approach used a hybrid classification algorithm that significantly enhanced the model's training data by combining correlation-based feature selection with fuzzy C-means clustering. This demonstrated how advanced machine learning techniques combined with feature selection could enhance the detection of unusual network behaviors (Hnamte et al., 2023).

Further advancements in feature selection have been achieved through integrating GA with machine learning models. GAs, known for their global optimization capabilities, have been extensively applied in network security for feature selection and parameter tuning. For instance, one application used logistic regression (LR) with GA to select the most effective feature subset, showing improved detection performance of decision tree-based methods when optimized with GA (Heidari et al., 2023). In addition, combining GA with fuzzy logic has been explored, where fuzzy logic assesses whether network events are indicative of anomalies, while GA generates digital signatures for network segments under investigation (Heidari et al., 2024).

In the context of cloud-based IDSs, hybrid approaches that combine multiple machine-learning techniques have shown promise in enhancing detection accuracy. For example, an advanced spam detection system integrated GA with a random weight network, achieving significant improvements in accuracy, precision, and recall (Hassan et al., 2024). Similarly, an IDS for wireless mesh networks combined GA-based feature selection with multiple SVM classifiers, resulting in a highly accurate and efficient detection mechanism (Bukhari et al., 2024).

3. Key Methodologies

The foundational approaches for the suggested RFGA are presented in this section. In particular, it discusses the use of RF, GA, and iForest to improve the efficacy and precision of the IDS.

3.1. Isolation Forest

Liu et al. (2012) presented the tree-based approach known as the iForest. It is intended to provide high accuracy and minimal time complexity for locating outliers in large, highly dimensional datasets. Since anomalies are "few and different," it is easier to isolate them from the rest of the data, which is the fundamental tenet of iForest. By recursively splitting the data, the iForest creates a series of binary trees called isolation trees (iTrees). Every tree is constructed by selecting a feature at random, followed by the selection of a random split value between the feature's maximum and minimum values (Drewek-Ossowicka et al., 2021; Ferrag et al., 2019*)*. This random partitioning process continues until every data point is isolated in a separate leaf node. Because of their unique properties, anomalies are predicted to be separated at shorter travel lengths than typical sites. Points with shorter pathways are regarded as anomalies, and the average path length over all trees is calculated. Because of its linear time complexity and capacity to handle high-dimensional data without the need for labeled data, iForest is very useful for huge datasets. This versatility makes it extremely adaptable to a wide range of applications.

3.2. GA

The GA is a search heuristic paradigm that draws inspiration from the mechanism of natural selection. The algorithm produces solutions of superior quality for optimization and search problems by emulating the fundamental principles of biological evolution. The DO IDS framework uses genetic GA to optimize the sampling ratio and feature selection processes. The technique starts by encoding potential solutions into a genotype string structure, where different combinations of these strings represent different potential solutions or chromosomes. A randomly generated initial population of chromosomes is utilized, with each chromosome indicating a potential solution to the problem (Nguyen et al., 2020). To assess the quality of each solution, a fitness function is employed, which exhibits variability contingent upon the specific problem being addressed.

In the framework of RFGA, the fitness function is derived from the F1 score, which takes into account both precision and recall, rendering it a suitable metric for the analysis of classification tasks. The GA enhances the population by employing a selection process that prioritizes the fittest people, conducting crossover events to facilitate the interchange of genetic material between chromosomes, and introducing mutations

to uphold genetic variety. These mechanisms enable GAs to explore the solution space systematically and ultimately converge toward an optimal or nearly optimal solution through iterative generations (Fig. 1). The DO IDS framework utilizes GA to optimize the sample ratios and identify the most pertinent features, hence improving the overall accuracy of the system's detection.

3.3. RF and Extreme Gradient Boosting (XGB): Ensemble Learning Techniques

Ensemble learning methods combine the predictions of multiple individual models to improve overall performance. Both RF and XGB are robust ensemble algorithms that utilize decision trees, but they differ significantly in their approach. Random RF, developed by Leo Breiman, builds an ensemble of decision trees using a technique called bootstrap sampling, where each tree is trained on a different subset of the data sampled with replacement. During the construction of each tree, a random subset of features is selected at every split, ensuring that each tree is independent, thereby reducing inter-tree correlation. This randomness enhances the diversity of the trees and improves the overall model's predictive performance. RF is particularly known for its resistance to overfitting and ability to provide valuable insights into the importance of features. The algorithm is effective for both binary and multi-class classification tasks and has been widely adopted due to its simplicity, robustness, and ease of use.

On the other hand, XGB follows a different approach by using boosting instead of bagging. While RF constructs each tree independently, XGB builds trees sequentially, where each new tree is trained to correct the errors made by the previous one. This boosting mechanism allows XGB to focus more on the difficult-to-classify samples, iteratively improving model performance with each tree. XGB has gained significant popularity due to its scalability, efficiency, and ability to handle complex data with high accuracy. The algorithm incorporates regularization techniques like L1 (alpha) and L2 (lambda) to prevent overfitting, making it particularly suitable for high-dimensional datasets. In addition, XGB includes features like subsampling and colsample_bytree, which allow for further optimization of the model's performance by randomly sampling subsets of the training data and features at each boosting round.

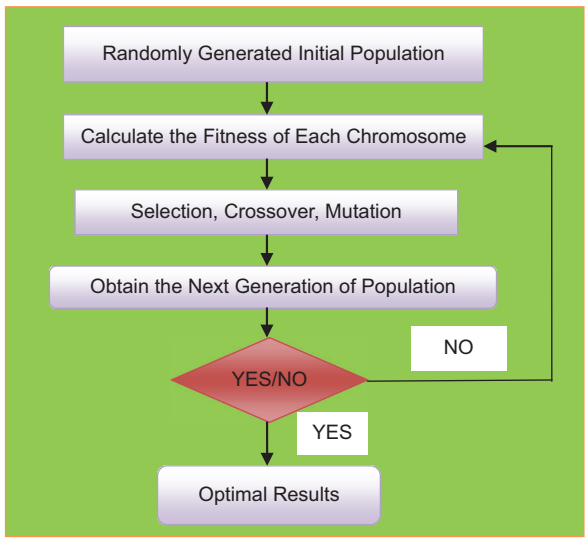

Fig. 1. Workflow diagram

Despite the differences in their approaches, both RF and XGB offer mechanisms to evaluate feature importance, which is crucial for identifying the most relevant features for a given classification task. This ability helps practitioners improve model interpretability and selection of key performance indicators. While RF tends to be more interpretable due to its simple randomization process, XGB generally provides superior predictive accuracy, especially for more complex tasks. In practical applications, both algorithms can be tuned for optimal performance, and they can be highly effective for tasks like anomaly detection in network security.

3.4. Proposed RFGA Framework

The proposed RFGA framework represents a novel approach to intrusion detection, integrating GA with RF to address the challenges of imbalanced data and improve detection accuracy. Unlike traditional models, the RFGA framework leverages the optimization capabilities of GA to refine both data sampling and feature selection processes, ensuring a robust classification system. In addition, experiments incorporate XGB and an enhanced version, XGBGA, along with traditional models such as LR, Naïve Bayes (NB), K-nearest neighbors (KNN), SVM, and DT. This comprehensive comparison ensures that the RFGA framework's performance is thoroughly evaluated across a diverse range of methodologies.

3.4.1. Data sampling and optimization using GA

In the RFGA framework, GA plays a pivotal role in optimizing data sampling ratios and feature selection, addressing the critical issue of imbalanced datasets. The fitness function used is the F1 score, a balanced metric that combines precision and recall to assess classification performance. GA initiates with a population of potential solutions, where each chromosome encodes a candidate sampling ratio or feature subset. Through iterative processes of selection, crossover, and mutation, GA evolves these

chromosomes to converge on optimal solutions. By refining sampling ratios, GA ensures that rare anomalies are adequately represented in the training data, significantly enhancing the system's ability to detect unusual patterns within network traffic. This systematic approach not only addresses class imbalance but also improves the overall accuracy and reliability of the IDS.

3.4.2. Feature selection using GA

Feature selection is a critical component of the RFGA framework, designed to reduce data dimensionality and enhance classification accuracy by focusing on the most relevant features. Initially, the dataset undergoes preprocessing using techniques like outlier detection to remove noise. Subsequently, GA optimizes the feature subset selection process, where each chromosome encodes a subset of features, and the fitness function evaluates the classifier's performance using these features. This iterative optimization identifies the most pertinent features, which are then used to train classifiers. The reduction in data dimensionality not only decreases computational complexity but also improves detection performance by eliminating irrelevant or redundant features. This targeted feature selection ensures that the RFGA framework achieves superior accuracy when maintaining computational efficiency.

3.4.3. RF classifier training

The RF algorithm, developed by Leo Breiman, is an ensemble learning method that combines multiple decision trees to improve classification accuracy and robustness. Known for its effectiveness in both binary and multi-class tasks, RF's design inherently mitigates overfitting. The construction of an RF model involves generating numerous decision trees, each trained on a unique bootstrap-sampled subset of the data, sampled with replacement. At each tree node, a random subset of features is chosen, and the best split is selected based on this subset, reducing inter-tree correlation and enhancing model diversity. Once trained, each tree votes on the predicted class label, and the final classification is determined by the majority vote across all trees. RF's structure offers several advantages, including high accuracy, resistance to overfitting, and the ability to process large datasets with numerous features. In addition, RF provides valuable insights into feature importance, which supports feature selection within the RFGA framework and refines the system's accuracy and interpretability.

Fig. 2 illustrates the role of IDS in securing network environments by analyzing network traffic to detect potential threats. IDS is typically classified into

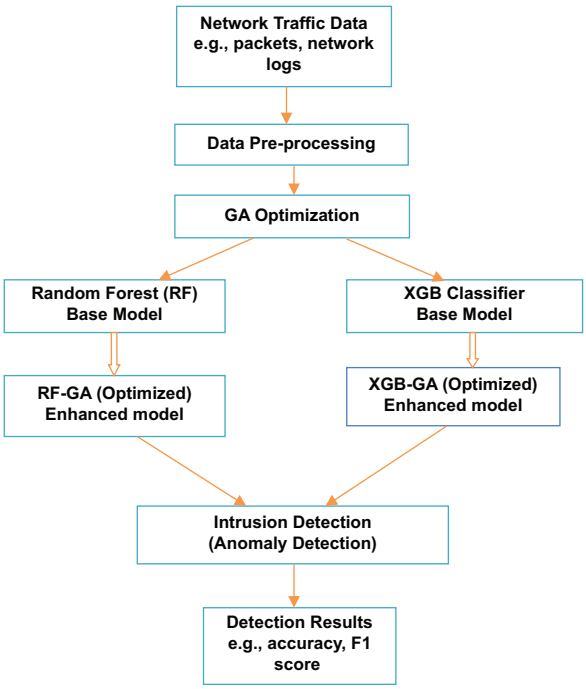

Fig. 2. Intrusion Detection System (IDS) Framework with GA Optimization for RF and XGB Classifiers

two types: signature-based IDS, which identifies threats by comparing data patterns against known signatures, and anomaly-based IDS, which flags deviations from normal network behavior that might indicate new or unknown threats. The RFGA framework employs a hybrid optimization approach, integrating data sampling and feature selection techniques to improve the performance of anomaly detection, particularly by reducing false alarm rates and enhancing detection accuracy.

Algorithm 1. Optimized Intrusion Detection System (XGBGA)

Input: Network traffic data [X1, X2,…, Xn]
Output: Optimized intrusion detection with high accuracy
Data Sampling:
• Apply Isolation Forest (iForest) to detect and remove outliers.
• Use Genetic Algorithm (GA) to optimize sampling ratios based on the F1 score.
Feature Selection:
• Encode feature subsets using GA.
• Optimize feature selection to maximize F1 score.
Classifier Training:
• Train Random Forest (RF) or Train extreme gradient boosting (XGB) with the optimized data.
• Use the trained RF or XGB model to classify and detect anomalies.
Result: Improved intrusion detection with enhanced accuracy and reduced false alarms.

Table 1. Literature review comparison

Study (Year), dataset	Methodology	F1 score (%)	Key contributions
Bukhari et al. (2024), WSN	SCNN-BiLSTM (Federated learning)	92.6	Secure IDS for wireless sensor networks
Hanafi et al. (2024), IoT	Binary Golden Jackal+LSTM	92.3	Improved IoT intrusion detection
Belouch & Hadaj (2017), NSL-KDD	Ensemble learning	88.4	Comparison of ensemble methods for IDS
Wu (2020), UNSW-NB15	Deep learning (CNN, RNN)	91.1	IDS using computational intelligence
Vashishtha et al. (2023), Cloud IDS	Hybrid (RF+CNN)	93.5	Hybrid IDS for cloud with feature selection
Hnamte et al. (2023), KDD Cup 99	LSTM-AE	91.7	Two-stage deep learning IDS
Talukder et al. (2023), CICIDS 2017	Hybrid machine learning	94.0	Reliable IDS hybrid model
Henry et al. (2023), UNSW-NB15	Hybrid deep learning+Feature optimization	92.6	IDS with feature optimization
Hnamte & Hussain (2023), KDD Cup 99	Deep CNN-BiLSTM	94.5	Hybrid CNN-BiLSTM IDS model.
Mohamed & Ismael (2023), IoT	Fog-to-cloud computing	90.2	Hybrid IoT IDS based on fog-to-cloud
Wang et al. (2023), NSL-KDD	RF+Autoencoder	92.0	Hybrid RF-Autoencoder IDS
Mehmood et al. (2022), UNSW-NB15	Hybrid (RF+SVM)	91.3	Hybrid RF-SVM IDS model
Zhang & Wei (2021), UNSW-NB15	XGBoost	93.1	Optimized XGBoost IDS model
Singh et al. (2020), KDD Cup 99	RF with SMOTE	89.9	SMOTE applied to RF for imbalanced data
Ahmed et al. (2018), NSL-KDD	SVM with feature selection	88.8	Feature selection with SVM for IDS
Sharma et al. (2024), IoT	BiLSTM with attention	94.2	BiLSTM with attention to IoT IDS

Abbreviations: BiLSTM: Bidirectional long short-term memory; CNN: Convoluted neural network; IDS: Intrusion detection system; IoT: Internet of Things; RF: Random forest; RNN: Recurrent neural network; SMOTE: Synthetic minority oversampling technique; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

4. Experimental Setup

4.1. Data Sampling

In this initial phase, the iForest method is employed to identify and eliminate outliers, aiming to reduce the impact of data imbalance on the classification process. GA is then used to optimize the sampling ratio for each class, with chromosomes representing possible sampling ratios and genes corresponding to the proportion of outliers in the sample. The F1 score is used as the fitness function to evaluate the performance of different sampling ratios, and GA iteratively refines these ratios to maximize the F1 score. The sampling ratios and their effectiveness are summarized in Table 2.

Table 2. Optimal sampling ratios for random forest and extreme gradient boosting

Class	Optimal sampling ratio (Random forest)	Optimal sampling ratio (Extreme gradient boosting)
Anomalous 1	0.85	0.88
Normal 1	0.92	0.94
Anomalous 2	0.78	0.80
Normal 2	0.89	0.91
Anomalous 3	0.88	0.90
Normal 3	0.91	0.93
Anomalous 4	0.80	0.83
Normal 4	0.94	0.96

4.2. Feature Selection

Feature selection is another critical step in the DO IDS framework, aimed at reducing the dimensionality of the data and eliminating irrelevant or redundant features. In this process, each chromosome represents a subset of features, with genes indicating whether a feature is included or excluded. The fitness function, based on the F1 score, evaluates the classifier's performance using the selected feature subset. GA searches for the optimal feature subset, which is then used to train the RF classifier. The optimal feature subsets for each class are detailed in Table 3.

4.3. Classifier Training

Finally, the RF classifier is trained using the refined dataset and feature subsets. Through this optimization, the RF classifier achieves enhanced accuracy in detecting a wide range of anomalies. The majority voting system in RF ensures reliable classification, where each decision tree votes on the predicted class label. Fig. 3 illustrates the training process and the integration of optimized components in the RFGA framework, highlighting the combined power of iForest, GA, and RF to achieve robust intrusion detection. The overall training process and the integration of the optimized components are illustrated in Fig. 3.

4.4. Dataset Description

The experimental evaluation of the RFGA framework was conducted on a system equipped with an Intel Core i5-4460 CPU at 3.6 GHz and 8GB of RAM, running Python applications within the PyCharm IDE. The UNSW-NB15 dataset, developed by the Australian Centre for Cyber Security, was selected for evaluation due to its comprehensive representation of modern network traffic patterns. This dataset comprises 2,540,044 records, divided into training and testing subsets, with 42 features detailing various network behaviors. A detailed description of the dataset, including sample size, feature attributes, and label distribution, is provided in Table 4. By combining this information, the RFGA framework ensures transparency and thorough evaluation of its effectiveness across diverse scenarios.

Table 3. Optimized feature sets for random forest and extreme gradient boosting

Class	Selected features (Random forest)	Selected features (Extreme gradient boosting)
Anomaly group 1	Features 1, 3, 7, and 9	Features 1, 3, 7, 9, and 10
Normal group 1	Features 2, 5, 6, and 8	Features 2, 4, 6, 8, and 11
Anomaly group 2	Features 1, 4, and 10	Features 1, 4, 10, and 12
Normal group 2	Features 2, 3, and 11	Features 3, 5, 7, and 11
Anomaly group 3	Features 3, 6, and 12	Features 3, 6, 8, and 9
Normal group 3	Features 4, 7, and 8	Features 4, 7, 9, and 11
Anomaly group 4	Features 5, 9, and 11	Features 5, 8, 9, and 12

The experimental results of the RFGA framework demonstrate its superior performance in intrusion detection compared to traditional models and state-of-the-art approaches. Specifically, the number of total, training, and testing samples used in the experiment are as follows: a total of 2,540,044 samples, with 1,750,000 used for training and 790,044 reserved for testing. These figures ensure a representative distribution of both normal and anomalous network traffic, which is critical for evaluating the performance of the detection models.

The impact of optimal sampling ratios on the results is significant. These ratios, which determine how the data are sampled for training the models, are optimized through the GA to address the issue of imbalanced datasets, where anomalous events are far less frequent than normal events. The optimization of sampling ratios ensures that anomalies are adequately represented in the training data, improving the model's ability to detect rare and novel threats. However, the reason the sampling ratios are not the same for RF and XGB is attributed to the different nature of these models. While RF builds trees independently from random subsets of data, XGB follows a boosting process, where each new tree corrects errors made by the previous one. As a result, the models interact differently with the data and benefit from distinct sampling strategies, leading to optimized ratios that are specific to each model's architecture and learning approach.

The results, as shown in the performance metrics, reflect the effectiveness of these optimized sampling ratios. Both RFGA and XGBGA significantly outperform traditional machine learning models such as LR, NB, KNN, and SVM. In particular, XGBGA achieves the highest overall performance, with the highest precision, recall, and F1 score among all models. These superior results highlight how both the optimal sampling ratios and feature selection processes, tailored to each classifier, contribute to better performance in detecting intrusions when minimizing false alarms.

4.5. Challenges and Future Directions

Despite its promising results, the RFGA framework has certain limitations. First, the

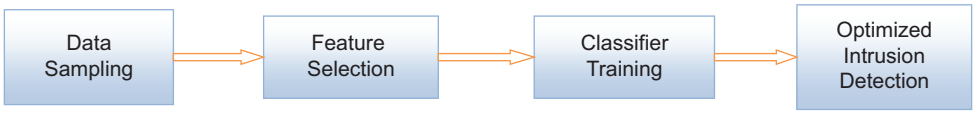

Fig. 3. Training process and integration of optimized components

Table 4. Detailed dataset description

Feature name	Description
Duration	Total duration of the connection (in seconds).
Protocol type	Type of protocol (e.g., TCP, UDP, ICMP).
Service	Network service on the destination (e.g., HTTP, FTP, SMTP).
Flag	Status flag for the connection.
Source bytes	Number of bytes transferred from source to destination.
Destination bytes	Number of bytes transferred from destination to source.
Land	1 if the source and destination IP/port are identical, otherwise 0.
Wrong fragment	Number of incorrect fragments in the connection.
Urgent	Number of urgent packets in the connection.
Hot	Number of "hot" indicators (e.g., logins, file accesses).
Failed logins	Number of failed login attempts.
Logged in	1 if successfully logged in, otherwise 0.
Root shell	1 if root shell access is obtained, otherwise 0.
File creation	Number of file creation operations.
Shell commands	Number of shell command operations.
Accessed files	Number of accessed files.
Outbound commands	Number of outbound commands in an FTP session.
Is host login	1 if the login belongs to the host, otherwise 0.
Is guest login	1 if the login is a guest login, otherwise 0.
Count	Number of connections to the same host as the current connection.
Same host rate	Percentage of connections to the same host.
Same service rate	Percentage of connections to the same service.
Diff service rate	Percentage of connections to different services.
Source failures	Number of failed source connections.
Destination errors	Number of error responses from destination.
Avg packet size	Average size of packets exchanged.
Total packets	Total packets exchanged in the connection.

computational demands of GA can hinder scalability in large-scale environments. Second, the reliance on

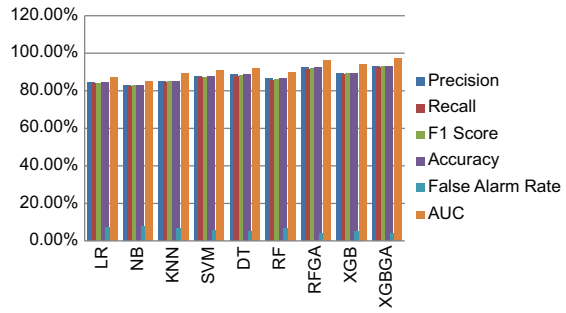

Fig. 4. Model accuracy and false alarm rate comparison

offline processing restricts its applicability to real-time anomaly detection. Finally, the evaluation is limited to the UNSW-NB15 dataset, which may not fully represent all network environments. Addressing these limitations is crucial for future development. Future work will focus on developing real-time processing capabilities to enable the detection of anomalies in streaming data. In addition, efforts will aim to optimize the computational efficiency of the RFGA framework, ensuring scalability for deployment in large-scale environments. Validation of the framework using diverse datasets and its application to other domains, such as fraud detection and cybersecurity, will further enhance its robustness and versatility. By addressing these areas, the RFGA framework can be extended to meet the demands of evolving network security challenges. In conclusion, the RFGA framework demonstrates significant advancements in intrusion detection by integrating GA with RF and XGB to optimize data sampling and feature selection. Experimental results confirm its superiority over traditional and state-of-the-art models, making it a reliable and robust solution for network security. Future enhancements will focus on scalability, real-time processing, and broader applicability to ensure its continued relevance and effectiveness.

Despite the significant advancements in IDS technology, challenges persist in achieving real-time anomaly detection, particularly in cloud environments, where the volume and diversity of network traffic continue to grow. Existing data stream management systems often struggle to process network streams quickly enough to detect anomalies in real time, a limitation exacerbated by the computational complexity of anomaly detection algorithms and the high false-positive rates associated with traditional detection methods (Heidari et al., 2024).

To address these challenges, hybrid data processing approaches that combine convolutional neural networks with optimization techniques, such as grey wolf optimization, have been proposed. These approaches, such as the TopoMAD stochastic seq2seq model, utilize advanced machine learning

Table 5. Performance metrics, including area under the curve

Model	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)	False alarm rate (%)	Area under the curve
Logistic regression	84.3	83.7	84.0	84.3	7.2	0.87
Naïve-bayes	82.9	82.3	82.6	82.9	7.8	0.85
K-nearest neighbor	85.1	84.5	84.8	85.1	6.8	0.89
Support vector machine	87.6	86.9	87.2	87.6	5.5	0.91
Decision trees	88.5	87.9	88.2	88.5	5.1	0.92
Random forest	86.7	85.4	85.8	86.7	6.9	0.90
RFGA	92.4	91.2	91.8	92.4	4.2	0.96
Extreme gradient boosting	89.5	88.7	89.1	89.5	5.0	0.94
XGBGA	93.1	92.5	92.8	93.1	3.8	0.97

techniques to capture the spatial and temporal dependencies of network data, enabling more robust and accurate anomaly detection (Vashishtha et al., 2023). Furthermore, the integration of autoencoders with traditional machine learning models has shown potential in enhancing the resilience of IDSs to corrupted data, improving their ability to detect anomalies in complex network environments (Wang et al., 2023).

In conclusion, the ongoing evolution of IDS technology is driven by the need for more accurate, efficient, and scalable detection systems. The integration of data sampling, feature selection, and advanced machine learning techniques offers a promising pathway toward achieving these goals. The proposed DO IDS framework, which combines these approaches, represents a significant step forward in the development of robust and reliable IDSs, particularly in the context of cloud computing and other large-scale network environments.

5. Conclusion

The RFGA framework, which integrates iForest, GA, and RF, has proven highly effective in enhancing the accuracy and robustness of IDSs by optimizing data sampling and feature selection, particularly for imbalanced and high-dimensional datasets. Experimental results confirm that RFGA outperforms traditional machine learning models, notably in detecting rare network anomalies like shellcode, worms, and backdoors. By combining GA to optimize sampling ratios and feature subsets with RF's strong classification capabilities, RFGA offers a reliable solution for network security. Moving forward, the primary goals are to develop real-time processing capabilities to detect anomalies in streaming data and to adapt the framework to other critical applications, such as fraud detection, where the accuracy and handling of imbalanced data are essential. In addition, future efforts will aim to reduce computational demands, thus enhancing the framework's scalability and making it practical for deployment across large-scale environments.

References

Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32, e4150.

Belouch, M., & Hadaj, S.E. (2017). Comparison of ensemble learning methods applied to network intrusion detection. In: Proceedings of the ACM Conference, pp. 1–4.

Bukhari, S.M.S., Zafar, M.H., Abou Houran, M., Moosavi, S.K.R., Mansoor, M., Muaaz, M., & Sanfilippo, F. (2024). Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-BiLSTM for enhanced reliability. *Ad Hoc Networks*, 155(103), 407. https://doi.org/10.1016/j.adhoc.2024.103407

Chkirbene, Z., Erbad, A., Hamila, R., Mohamed, A., Guizani, M., & Hamdi, M. (2020). TIDCS: A dynamic intrusion detection and classification system based feature selection. *IEEE Access*, 8, 95864–95877. https:///doi.org/10.1109/ACCESS.2020.2994931

Deebak, B.D., & Hwang, S.O. (2024). Healthcare applications using blockchain with a cloud-assisted decentralized privacy-preserving framework. *IEEE Transactions on Mobile Computing*, 23(5), 5897–5916. https://doi.org/10.1109/TMC.2023.3315510

Dey, A. (2020). Deep IDS: A Deep Learning Approach for Intrusion Detection Based on IDS 2018. In: 2020 *2nd International Conference on*

Sustainable Technologies for Industry 4.0 (STI). IEEE, p. 1–5.

Drewek-Ossowicka, A., Pietrołaj, M., & Rumiński, J. (2021). A survey of neural networks usage for intrusion detection systems. *Journal of Ambient Intelligence and Humanized Computing*, 12, 497–514. https://doi.org/10.1007/s12652-020-02014-x

Ferrag, M.A., Maglaras, L., Janicke, H., & Smith, R. (2019). Deep Learning Techniques for Cyber Security Intrusion Detection: A Detailed Analysis. In: *6th International Symposium for ICS SCADA Cyber Security Research* (ICS-CSR 2019), Athens, 10–12 September.

Halbouni, A., Gunawan, T.S., Habaebi, M.H., Halbouni, M., Kartiwi, M., & Ahmad, R. (2022). CNN-LSTM: Hybrid deep neural network for network intrusion detection system. *IEEE Access*, 10, 99837–99849.

Hanafi, A.V., Ghaffari, A., Rezaei, H., Valipour, A., & Arasteh, B. (2024). Intrusion detection in Internet of things using improved binary golden jackal optimization algorithm and LSTM. *Cluster Computing*, 27(3), 2673–2269. https://doi.org/10.1007/s10586-023-04102-x

Hassan, S.R., Rehman, A.U., Alsharabi, N., Arain, S., Quddus, A., & Hamam, H. (2024). Design of load-aware resource allocation for heterogeneous fog computing systems. *PeerJ Computer Science*, 10, e1986. https://doi.org/10.7717/peerj-cs.1986

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14, e1520. https://doi.org/10.1002/widm.1520

Heidari, A., Navimipour, N.J., & Unal, M. (2023). A secure intrusion detection platform using blockchain and radial basis function neural networks for the internet of drones. *IEEE Internet of Things Journal*, 10, 8445–8454. https://doi.org/10.1109/JIOT.2023.3237661

Henry, A., Gautam, S., Khanna, S., Rabie, K., Shongwe, T., Bhattacharya, P., Sharma, B., & Chowdhury, S. (2023). Composition of hybrid deep learning model and feature optimization for intrusion detection system. *Sensors*, 23(2), 890. https://doi.org/10.3390/s23020890

Hnamte, V., & Hussain, J. (2023). DCNNBiLSTM: An efficient hybrid deep learning-based intrusion detection system. *Telematics and Informatics Reports*, 10, 100053. https://doi.org/10.1016/j.teler.2023.100053

Hnamte, V., Nhung-Nguyen, H., Hussain, J., & Hwa-Kim, Y. (2023). A novel two-stage deep learning model for network intrusion detection: LSTM-AE. *IEEE Access*, 11, 37131–37148. https://doi.org/10.1109/ACCESS.2023.3266979

Liu, F., Ting, K.M., & Zhou, Z.H. Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), IEEE, 2012, p. 413–422.

Mehmood, M., Javed, T., Nebhen, J., Abbas, S., Abid, R., Bojja, G.R., & Rizwan, M. (2022). A hybrid approach for network intrusion detection. *Computers*, *Materials and Continua*, 70, 91–107. https://doi.org/10.32604/cmc.2022.019127

Mohamed, D., & Ismael, O. (2023). Enhancement of an IoT hybrid intrusion detection system based on fog-to-cloud computing. *Journal of Cloud Computing*, 12(1), 41. https://doi.org/10.1186/s13677-023-00420-y

Molina-Coronado, B., Mori, U., Mendiburu, A., & Miguel-Alonso, J. (2020). Survey of network intrusion detection methods from the perspective of the knowledge discovery in databases process. *IEEE Transactions on Network and Service Management*, 17(4), 2451–2479. https://doi.org/10.1109/TNSM.2020.3016246

Pingale, S.V., & Sutar, S.R. (2022). Analysis of Web Application Firewalls, Challenges, and Research Opportunities. In: *Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications (ICDSMLA 2020)*. Singapore: Springer, p. 239–248.

Pingale, S.V., & Sutar, S.R. (2022). Automated network intrusion detection using multimodal networks. *International Journal of Computational Science and Engineering*, 25(3), 339–352. https://doi.org/10.1504/IJCSE.2022.123123

Pingale, S.V., & Sutar, S.R. (2022). Remora whale optimization hybrid deep learning for network intrusion detection using CNN features. *Expert Systems with Applications*, 210, 118476. https://doi.org/10.1016/j.eswa.2022.118476

Pingale, S.V., & Sutar, S.R. (2023). Remora-based Deep Maxout Network model for network intrusion detection using convolutional neural network features. *Computers and Electrical Engineering*, 110, 108831. https://doi.org/10.1016/j.compeleceng.2023.108831

Ravikumar, C., Ravi Kumar, R., Sarada, M., Pabba, K., & Pasha, M.A. (2024). A comprehensive exploration of machine learning in early detection with a focus on lung and pancreatic cancer for revolutionizing cancer diagnostics. *International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications* (ICETCS 2024).

Ravikumar, C.H., Batra, I., & Malik, A. (2023). Block

chain based secure with improvised bloom filter over a decentralized access control network on a cloud platform. *Journal of Engineering Science and Technology Review*, 16(2), pp. 123–130. https://doi.org/10.25103/jestr.162.16

Ravikumar, C.H., Sridevi, M., Ramchander, M., Ramesh, V., & Kumar, V.P. (2024). Enhancing digital security using signa-deep for online signature verification and identity authentication. *International Journal of Systematic Innovation*, 8(2), pp. 58–69. https://doi.org/10.6977/IJoSI.202406_8(2).0005

Rekha, G., Malik, S., Tyagi, A.K., & Nair, M.M. (2020). Intrusion detection in cyber security: Role of machine learning and data mining in cyber security. *Advances in Science, Technology and Engineering Systems Journal*, 5(3), 72–81. https://doi.org/10.25046/aj050310

Talukder, M.A., Hasan, K.F., Islam, M.M., Uddin, M.A., Akhter, A., Yousuf, M.A., Alharbi, F., & Moni, M.A. (2023). A dependable hybrid machine learning model for network intrusion detection. *Journal of Information Security and Applications*, 72, 103405. https://doi.org/10.1016/j.jisa.2022.103405

Vashishtha, L.K., Singh, A.P., & Chatterjee, K. (2023). HIDM: A hybrid intrusion detection model for cloud-based systems. *Wireless Personal Communications*, 128, 2637–2666. https://doi.org/10.1007/s11277-022-10063-y

Wang, C., Sun, Y., Wang, W., Liu, H., & Wang, B. (2023). Hybrid intrusion detection system based on combination of random forest and autoencoder. *Symmetry*, 15(3), 568.

Wu, P. (2020). *Deep Learning for Network Intrusion Detection: Attack Recognition with Computational Intelligence* (PhD Thesis). UNSW Sydney.

Xu, Z., Zhang, W., Li, Y., & Li, W. (2024). Secure and efficient intrusion detection in IoT using deep reinforcement learning. *Journal of Computer Science and Technology*, 39(3), 552–570.

AUTHOR BIOGRAPHIES

Ms.Sadargari Viharika. She received her bachelor's degree from ECEW, JNTUH in 2014. She attained her M.Tech degree from MRIET, JNTUH, and Hyderabad in 2020. She is pursuing her Ph.D. in the Computer Science and Engineering Department at KL University from 2022. Presently, she is working as an Assistant Professor in the Department of Information Technology at MLR Institute of Technology. Her areas of interest include Cloud Computing, Machine Learning, Deep Learning, and Artificial Intelligence. She can be contacted at reddyviharika266@gmail.com

N. Alangudi Balaji. He is affiliated with the Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram. Dr. N. Alangudi Balaji is an academic and researcher affiliated with the Department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation in Vaddeswaram, Andhra Pradesh, India. With over two decades of experience in engineering education, his primary focus areas include cloud computing, machine learning, and information security. He has published extensively on deep learning models, convolutional neural networks, and Internet of Things (IoT) applications in various reputed journals indexed under Scopus and Web of Science. E-Mail: alangudibalaji@gmail.com