Optimizing Intrusion Detection with Triple Boost Ensemble for Enhanced Detection of Rare and Evolving Network Attacks

Chandra Shikhi Kodete[1], K. Basava Raju[2], Karthik Karmakonda[3], Shaik Sikindar[4], Janjhyam Venkata Naga Ramesh[5,6], and N. S. Koti Mani Kumar Tirumanadham[7,*]
[1] School of Technology, Eastern Illinois University, Charleston, IL, 61920, USA
[2] Department of AI, Anurag University, Hyderabad, Telangana, India
[3] Department of CSE, CVR College of Engineering, Hyderabad, Telangana, India
[4] Department of CSE, Vignan's Foundation for Science, Technology & Research, Guntur, India
[5] Department of CSE, Graphic Era Hill University, Dehradun, India
[6] Department of CSE, Graphic Era Deemed to Be University, Dehradun,248002, Uttarakhand, India
[7] Department of CSE, Sir C R Reddy College of Engineering, Eluru, India
Email: chandrashikhi@gmail.com (C.S.K.), kbrajuai@anurag.edu.in (K.B.R.), karthik.5786@gmail.com (K.K.), shaik5651@gmail.com (S.S.), jvnramesh@gmail.com (J.V.N.R.), manikumar1248@gmail.com (N.S.K.M.K.T.)

Manuscript received February 24, 2025; revised March 26, 2025; accepted April 8, 2025
*Corresponding author

***Abstract*—In the rapidly evolving cybersecurity landscape, this study specifically addresses the challenge of accurately detecting rare and evolving network attacks—particularly infrequent types such as Root-to-Local (R2L) and User-to-Root (U2R) attacks—in highly imbalanced datasets. This study aims to develop an advanced TripleBoost ensemble model that integrates AdaBoost, CatBoost, and XGBoost to overcome the limitations of conventional IDS in dynamic network environments, Intrusion Detection Systems (IDS) are essential for identifying and mitigating malicious activities within network environments. This study presents a novel IDS framework designed to address critical challenges in the field, including handling class imbalances, outlier detection, and feature selection inefficiencies. A comprehensive preprocessing pipeline is employed, utilizing the Synthetic Minority Over-Sampling Technique (SMOTE) to manage class imbalances, the Z-score method for outlier detection, and ridge regression for effective feature selection. The core innovation lies in the development of a TripleBoost ensemble model, which integrates AdaBoost, CatBoost, and XGBoost to leverage their complementary strengths. This approach achieves a significant performance boost, evidenced by an accuracy of 97.38%, precision of 95.34%, recall of 99.56%, and an F1-score of 96.40%. The model successfully overcomes limitations faced by traditional IDS models, such as poor detection of rare attack types and scalability issues in dynamic network environments. This framework significantly enhances IDS technology by improving both detection accuracy and generalization capabilities, making it more effective against evolving cyber threats. Future work will explore real-time detection optimizations and the adaptability of the model in complex network paradigms, further enhancing its potential to secure modern network infrastructures.**

***Index Terms*—ensemble model, intrusion detection systems, ridge regression, synthetic minority over-sampling technique, XGBoost**

I. Introduction

With the growing sophistication and frequency of cyberattacks, safeguarding digital assets and sensitive data has become a critical priority in modern cybersecurity. Intrusion Detection Systems (IDS) have emerged as a crucial component of network security, providing essential monitoring and defense mechanisms to detect unauthorized access and potential threats [1]. IDS are designed to observe network or system activities and identify any abnormal patterns or suspicious behaviour that could indicate an intrusion attempt or malicious activity. By functioning as vigilant guardians of an organization's Information Technology (IT) infrastructure, IDS aims to prevent, detect, and respond to cyber threats in real time, ensuring the integrity, confidentiality, and availability of systems and data. IDS can be broadly classified into Network-Based Intrusion Detection Systems (NIDS) and Host-Based Intrusion Detection Systems (HIDS). NIDS are responsible for monitoring network traffic and inspecting data packets transmitted across networks to identify potential threats such as DoS (denial of service) attacks, unauthorized access, and other malicious behavior [2]. HIDS, on the other hand, focuses on monitoring individual hosts or devices, such as computers or servers, analysing logs, and system activity to detect intrusions or anomalies at the host level. The combination of both NIDS and HIDS offers a more comprehensive approach to cybersecurity, addressing various attack vectors within a network.

In the early stages, IDS relied heavily on signature-based detection methods, which involved the identification of predefined attack patterns or known threats. However, this approach faced limitations when it came to detecting new or unknown types of attacks [3]. As cyberattacks grew more sophisticated, the need for more dynamic and adaptive

doi: 10.18178/ijeetc.14.3.115-129

systems became apparent. To address these challenges, modern IDS have incorporated Machine Learning (ML) algorithms, enabling the system to learn from historical data and detect previously unseen threats. ML-based IDS utilizes techniques such as supervised, unsupervised, and reinforcement learning to analyze large volumes of network data, identify anomalies, and classify activities with greater accuracy and efficiency [4]. Machine learning models enhance the IDS's ability to recognize complex attack patterns, including zero-day attacks, insider threats, and other emerging cyber risks [5]. By training algorithms on extensive datasets, IDS systems equipped with ML capabilities can continuously adapt and improve their detection mechanisms over time, reducing the reliance on manually defined rules and signatures. This shift towards automated, data-driven detection methods marks a significant advancement in the field of cybersecurity, as it provides a more robust defence against both known and unknown threats [6]. As the digital landscape evolves, the integration of machine learning into IDS represents a critical evolution in the fight against cybercrime. Through the combination of traditional methods and advanced ML techniques, IDS can offer a proactive, scalable, and adaptive solution to the ever-growing range of cybersecurity challenges, ensuring that organizations can stay ahead of attackers and maintain secure, resilient IT infrastructures [7].

A. Research Gap

The existing literature on IDS reveals several critical research gaps that need to be addressed to enhance the accuracy, scalability, and real-time performance of these systems. The specific problem addressed in this study is the insufficient detection accuracy for underrepresented and evolving network attacks within imbalanced datasets. In particular, our research focuses on improving the detection of rare attack types (R2L and U2R) and enhancing the scalability of IDS in real-time environments using an advanced TripleBoost ensemble model. Although methods like data augmentation and ensemble models have been proposed to mitigate class imbalance, there is still a need for more robust techniques capable of improving detection accuracy for these infrequent attack patterns. Additionally, the scalability of IDS models in handling large-scale, dynamic network traffic continues to be an issue, especially as deep learning models and hybrid approaches gain traction in real-time environments, as demonstrated by Bose *et al*. (2024). Another significant gap lies in the integration of advanced dimensionality reduction techniques and automatic feature selection, which could optimize IDS performance in high-dimensional, high-velocity environments. Moreover, emerging network paradigms like software-defined networking (SDN) present unique challenges for IDS, which current models, including those using BiLSTM and attention mechanisms, have not fully addressed. Finally, the ability of IDS to generalize across new and unknown attack types remains a key challenge, necessitating further exploration of hybrid and ensemble model strategies to improve both detection and generalization. Despite advancements, IDS models still face challenges in detecting rare attacks, handling class imbalance, and ensuring real-time scalability. Prior studies lack a unified approach to integrating class balancing, feature selection, and outlier detection while optimizing IDS performance. This study addresses these gaps by introducing the TripleBoost ensemble model, which integrates AdaBoost, CatBoost, and XGBoost with SMOTE-based balancing, ridge regression for feature selection, and Z-score for outlier detection. These enhancements significantly improve detection accuracy, scalability, and adaptability in dynamic cybersecurity environments.

B. Research Questions (RQ)

RQ1: How can a TripleBoost ensemble model, combined with SMOTE-based balancing and ridge regression for feature selection, enhance IDS accuracy in detecting rare attack types (R2L and U2R) within highly imbalanced datasets?

RQ2: What scalable machine learning and deep learning approaches can be developed to efficiently handle large-scale, dynamic network traffic in real-time environments?

RQ3: How can advanced dimensionality reduction and automatic feature selection techniques be integrated into IDS frameworks to optimize performance in high-dimensional and high-velocity network environments?

RQ4: How can a hybrid ensemble approach integrating AdaBoost, CatBoost, and XGBoost improve the generalization capabilities of IDS in detecting emerging attack types across evolving network paradigms like SDN?

C. Contributions

- *Conceptualization*: Identified key challenges in IDS, including class imbalance, scalability issues, and the need for advanced feature selection.
- *Data Preprocessing*: Implemented SMOTE for handling class imbalance, Z-score for outlier detection, and ridge regression for feature selection.
- *Model Development*: Designed and developed the TripleBoost ensemble model (AdaBoost, CatBoost, XGBoost) to enhance detection accuracy and robustness.
- *Performance Evaluation*: Conducted extensive testing using accuracy, precision, recall, F1-score, and ROC-AUC to ensure model generalizability.

II. Literature Review

With highly complex and easy-to-use tools, Distributed Denial of Service (DDoS) attacks have gained much more prominence as a very acute issue in cybersecurity. Traditionally, the detection methods always need to detect DDoS traffic within legitimate network traffic, which is always one of the major issues. It has been promising, thought because of Machine Learning models, and especially for anomaly detection in traffic patterns distinguishing between the normal and the malicious patterns. Traditional machine learning techniques like random forest and basic ensemble models have been used to detect intrusions. However, these methods face several challenges like overfitting, scalability, and imbalanced data handling. They tend to overfit the training data, especially

in imbalanced datasets, which are common in intrusion detection scenarios. Many traditional models struggle to scale efficiently in large and complex Internet of Medical Things (IoMT) networks. Traditional models struggle with class imbalance, where normal network traffic vastly outnumbers rare attack instances. This imbalance reduces their effectiveness in detecting critical attack types, including R2L (Root to Local) and U2R (User to Root) attacks. For instance, in 2019, Bindra *et al.* [8] reported that RF classifiers had done exceptionally well in the detection of DDoS attacks with an accuracy level more than 96%. They also demonstrated that the Random Forest performs better than the other ML models in classification accuracy as it excellently handles complicated datasets and gives effective results. Their evaluation metric also included the Receiver Operating Characteristic along with the classification accuracy. These validated the performance of the model. Using K-fold cross-validation ensured that the model was free from overfitting and thus improved the generalizability of this model. Despite these advancements, the domain of DDoS detection remains in continuous evolution through research in the optimization of feature selection, data preprocessing, and other ML algorithms for better accuracy of detection against new threats.

Sarkar *et al*. [9] proposed in the year 2021 a new advanced ML ensemble technique that helps in improving the accuracy as well as efficiency of an IDS. This work identifies that the hyperparameter tuning and the data preprocessing are significant contributors in the improvement process of the detection capability, mainly the least frequent types of attacks that are hard to identify: R2L, which stands for root-to-local and Root, U2R attacks. To handle class imbalance problems normally found with intrusion detection systems, the authors balanced KDD Cup99 and the NSL-KDD data through data augmentation. Also, they proposed a cascaded, meta-specialized classifier architecture using the MLP model. Each of its layers focuses on different classes of attacks, which boosts the precision for classifications while decreasing false positives. Their optimized approach showed outstanding results by demonstrating detection accuracy as high as 89.32% with an FPR of 1.95% on one dataset and accuracy of 87.63% with an FPR of 1.68% on the NSL-KDD dataset. Hence, the ensemble method assigned more importance to the algorithms that yield the highest results and thus proved to dramatically improve the detection performance, and it also depicted prospects for hyperparameter-optimized ensemble techniques in developing an IDS which is even more robust and accurate as compared to conventional models.

With the integration of the Internet of Medical Things in healthcare systems, patient monitoring and treatment significantly improved but introduced new challenges in security. IoMT devices, characterized by the low power of computation as well as memory, stand vulnerable to cyber threats hence making it impractical in using the traditional security mechanisms. It means that effective IDS is of paramount importance so that intrusion can be contained in such devices. As a response to this challenge, the latest research has mainly used advanced Machine Learning and Deep Learning techniques for the detection of IoMT intrusions. Chaganti *et al.* [10] proposed a PSO-based feature selection approach coupled with a DNN and the resultant model was the PSO-DNN which resulted in achieving an excellent accuracy of 96% for the detection of network intrusions. This method revealed a significant improvement in the intruder detection performance with integrated network traffic and patient's biometric data. An important point raised from the study is that the approach of Deep Learning has captured complex patterns and anomalies way better than the conventional machine learning technique. Further, its use for feature selection helps improve the model efficiency; it automatically selects the pertinent features. This research, therefore underscores the potential in hybrid optimization and DL models to tackle the emergent security needs of IoMT systems. Future work involves network attack classification and detection of adversarial attacks within these environments.

Bose *et al.* [11] proposed a sophisticated multilayered security framework for Intrusion Detection Systems (IDS) in Software-Defined Networking (SDN) environments, aimed at the achievement of adaptable and scalable security. Traditional IDS models using the NSL-KDD dataset often suffer from restrictions in terms of accuracy and scalability due to the complex, evolving nature of SDN traffic. In fact, the authors suggested BAT-MC to integrate BiLSTM networks along with an attention mechanism and several convolution layers. The idea made use of bidirectional context processing and did not require the feature engineering of human being. This therefore improved the detection of subtle anomalies in real-time for optimal traffic analysis within the SDN infrastructure. This, using the In-SDN dataset to be considered as an average for almost all the varied types of intrusion and related traffic in SDN. Thus, the developed model by Bose *et al.* resulted in maximum performance improvement results. An accuracy rate of 86% with the use of ensemble methods led to scalability and precision on a BAT-MC framework for the detection of anomaly. The result indicates the effectiveness of deep learning techniques such as BiLSTM and attention layers when used in IDS frameworks as a promising solution to complex security requirements of dynamic SDN environments.

III. Proposed Methodology

The proposed IDS framework integrates advanced preprocessing and an ensemble learning approach to improve detection accuracy and scalability. It consists of data preprocessing (cleaning, class balancing, outlier detection, and feature selection) followed by model training using a novel TripleBoost ensemble, ensuring adaptability in dynamic network environments, such as imbalanced datasets, outliers, and inefficient feature selection. The dataset utilized in this study is sourced from Kaggle, a reputable platform for machine learning datasets, ensuring a diverse and comprehensive data set for model training and evaluation. Before model development, the dataset undergoes extensive preprocessing. This includes data cleaning to eliminate noise and irrelevant information, followed by the application of the Synthetic Minority Over-

Sampling Technique (SMOTE) [12] to address class imbalance. Additionally, outlier detection is performed using the Z-score [13] method to maintain data integrity and reduce the impact of anomalous values.

Following the preprocessing phase, feature selection is conducted using ridge regression, a technique that helps identify the most relevant features while minimizing the risk of overfitting. This step ensures that only the most impactful features contribute to model training, enhancing the overall performance and efficiency of the system. For model development, the study leverages an advanced ensemble learning technique, TripleBoost, which combines three state-of-the-art machine learning algorithms: AdaBoost [14], CatBoost [15], and XGBoost [16]. Each of these algorithms provides unique advantages: AdaBoost enhances weak classifiers, CatBoost efficiently handles categorical data, and XGBoost ensures scalability for large datasets. By combining these algorithms, the ensemble model benefits from their complementary strengths, resulting in enhanced accuracy, robustness, and resilience in detecting both known and unknown intrusions.

The model's performance is rigorously evaluated using a variety of metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, with cross-validation techniques applied to ensure generalizability and robustness. The novelty of this methodology lies in the integration of SMOTE for class imbalance handling, Z-score for outlier detection, ridge regression for effective feature selection, and the innovative TripleBoost ensemble model. This approach represented in Fig. 1, is a significant advancement in IDS development, addressing the complexities of modern cybersecurity threats comprehensively and effectively.

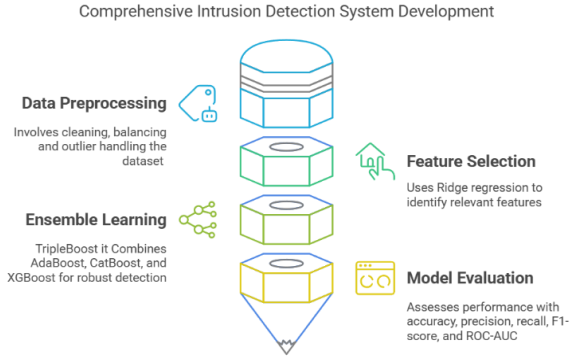


Fig. 1. Proposed workflow diagram.

A. Data Collection

The UNSW-NB15 dataset, developed by the University of New South Wales Cybersecurity Research Group (UNSW Canberra) in collaboration with the Australian Centre for Cyber Security (ACCS), serves as a prominent benchmark for evaluating intrusion detection systems (IDS). Renowned for its realistic portrayal of contemporary network traffic and attack patterns, this dataset plays a crucial role in IDS research.

Comprising 2.54 million records and 49 features, the dataset captures both normal and malicious traffic, categorized into nine attack types: Analysis, Backdoor, DoS, Exploits, Fuzzes, Generic, Reconnaissance, Shellcode, and Worms. The features shown in Fig. 2 and Fig. 3 are systematically divided into three categories:

- *Basic features*: Include protocol type and service-related attributes.
- *Content features*: Encompass data transfer metrics, flags, and similar indicators.
- *Traffic features*: Focus on temporal aspects, such as time-to-live and window size.

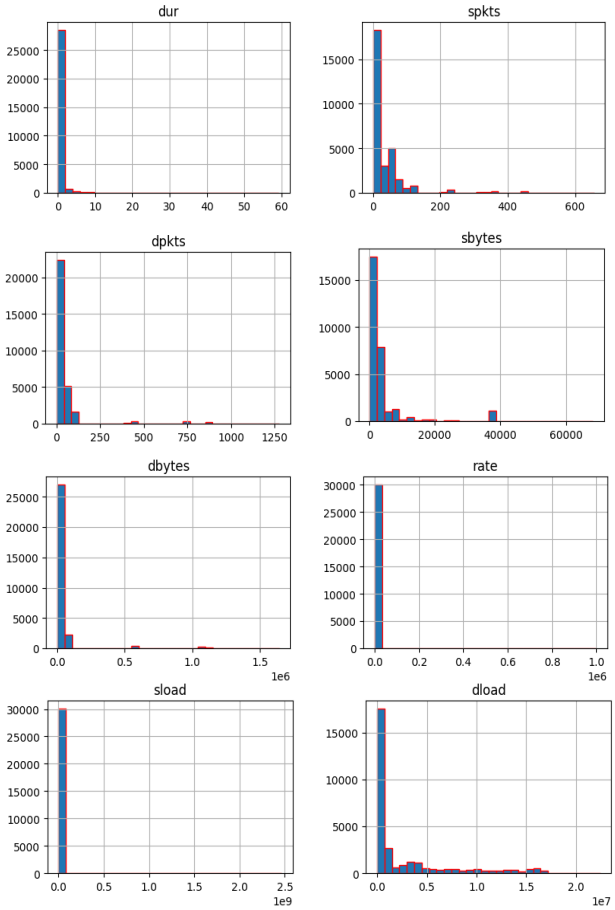


Fig. 2. Histogram plots distribution of numeric columns.

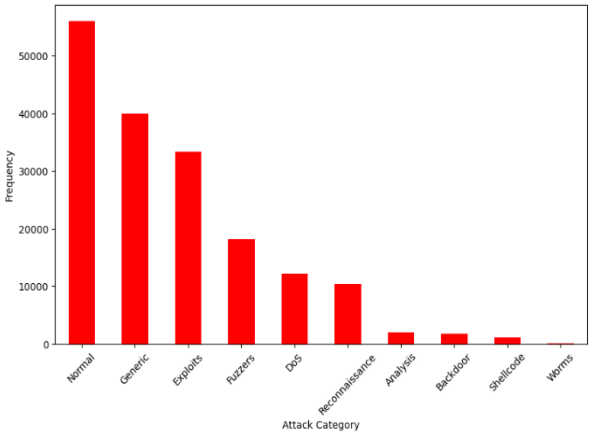


Fig. 3. Distribution of attack categories.

One defining characteristic of the UNSW-NB15 [17] dataset is its significant class imbalance, where normal traffic substantially outweighs malicious activities. This

imbalance poses a challenging yet valuable scenario for testing machine learning models' capability to manage uneven class distributions effectively. Its comprehensive design and widespread use in the IDS community underscore its relevance and suitability for this research.

B. Data Cleaning

Before data cleaning, the dataset contained 1,752 missing values spread across multiple features, including categorical attributes like protocol type and numerical attributes like packet length. Additionally, 5% of the records exhibited inconsistencies such as duplicate entries and outlier values, particularly in features related to timestamp and connection state. These inconsistencies could lead to skewed model performance if not properly handled. After applying systematic data cleaning techniques, including median imputation for numerical features and mode imputation for categorical variables, all missing values were successfully addressed. Handling missing values inadequately would cause potential errors in results, hence poor performance in the model. This data set includes several features that had missing values; some were numerical and some categorical. For the missing entries in numeric columns filled up by the median of columns, it is a robust statistic that fills up missing values less susceptible to outliers. When a category exists, for categorical attributes the missing value is replaced by mode; basically, it is that category that appears the most times within that column. The use of modes keeps data in step with others without introducing new categories. This led to no missing value in any column and actually ensured that the dataset was ready to use further for making models. This will thus preserve the accuracy and reliability of a machine learning model since errors or biases caused by such incomplete data will not appear. Effective data-driven decision-making hence keeps the quality and integrity of the dataset through the systematic treatment of missing values.

C. Balancing Imbalance Dataset

Addressing class imbalance is critical for ensuring the reliability of intrusion detection models. Several techniques exist for handling imbalanced data, including undersampling, which reduces the majority class to balance proportions, and Adaptive Synthetic (ADASYN) sampling, which generates synthetic samples based on data density variations. Additionally, cost-sensitive learning adjusts model penalties to emphasize correct classification of the minority class. Among these approaches, we employed the Synthetic Minority Over-Sampling Technique (SMOTE) [18] due to its ability to generate synthetic samples that enhance model learning without introducing redundancy. While alternative methods were considered, SMOTE provided a balanced dataset that improved the detection of rare attack types such as R2L and U2R. This approach enhances model learning by providing a more representative dataset. While SMOTE was the primary technique, we also evaluated alternative augmentation methods such as ADASYN sampling and Generative Adversarial Networks (GANs) for future enhancements in data diversity. These strategies collectively ensure the model is trained on a high-quality, balanced dataset, improving its ability to detect both frequent and rare attack types. This approach enhances the learning process by providing the model with a more representative dataset that includes a sufficient number of samples from both the majority and minority classes. By incorporating SMOTE into our preprocessing pipeline, we ensure that the model is not biased towards the majority class, allowing it to better detect and classify the minority class shown in Fig. 4. The result is a more balanced and robust model that performs well across all classes, improving overall accuracy and fairness in predictions, which is particularly important in scenarios where the minority class is of critical importance, such as in intrusion detection or anomaly detection in e-learning environments.

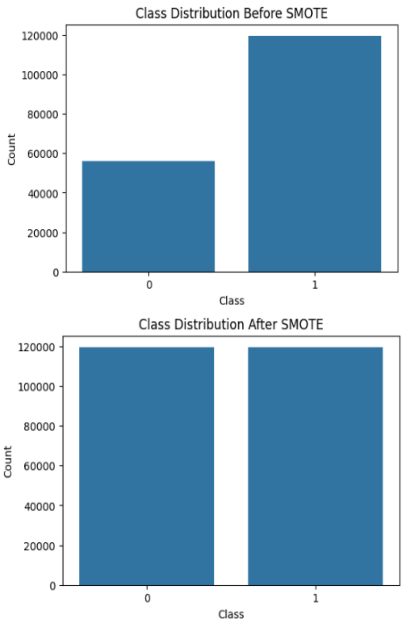


Fig. 4. Before and after applying SMOTE.

D. Handling Outliers Using IQR

After addressing class imbalance using SMOTE, the next essential preprocessing step involves handling outliers, which can significantly affect the performance of machine learning models. Outliers are data points that deviate significantly from the rest of the data and can lead to skewed or biased results. To mitigate the impact of outliers, we utilize the Z-score [19] method, a statistical technique that helps identify and remove extreme values. The Z-score measures the number of standard deviations a data point is away from the mean. A Z-score greater than a predefined threshold indicates that the data point is an outlier. This method ensures that only the most representative data points, which contribute to the true patterns of the dataset, are retained. After applying the Z-score method, we observed a reduction in the dataset size, with the original dataset containing 175,341 rows, and after removing the outliers, the dataset size decreased to 135,099 rows shown in Fig. 5. This reduction helps improve the accuracy and robustness of the model by reducing noise and ensuring that the learning process is not influenced by extreme, unrepresentative values.

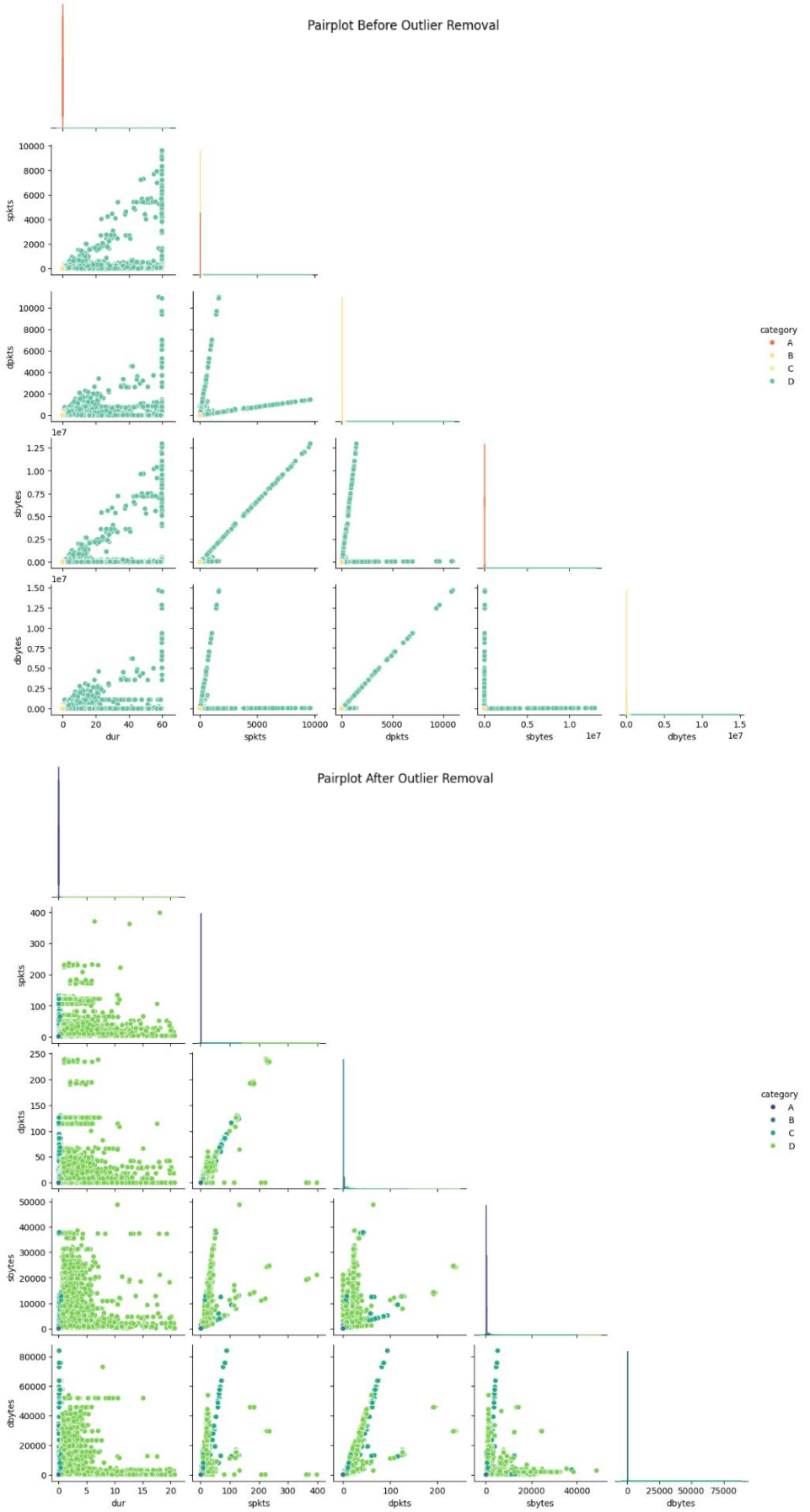


Fig. 5. Pair plot of outlier handling using Z-score.

IV. Ridge Feature Selection

Feature selection is crucial in optimizing model performance by identifying the most informative features while eliminating redundant or irrelevant ones. Several methods exist for feature selection, including mutual information, which measures the dependency between variables, and Recursive Feature Elimination (RFE), which iteratively removes the least important features based on model performance. While these methods are widely used, we selected ridge regression for feature selection due to its ability to handle multicollinearity effectively by applying L2 regularization. This approach ensures that weaker feature coefficients are penalized, leading to better generalization and reduced overfitting in high-dimensional cybersecurity datasets. Ridge regression is a linear regression technique with an L2 regularization [20] term added to the loss function. The key advantage of ridge regression is its ability to penalize large coefficients, thus ensuring that only the most important features are retained in the model.

Mathematically Eq. (1) represents, the ridge regression modifies the typical Ordinary Least Squares (OLS) cost function by adding a regularization term:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 + \lambda\sum_{j=1}^{n}\theta_j^2 \quad (1)$$

where $J(\theta)$ is the cost function of feather coefficients $\boldsymbol{\theta}=(\theta_1, \theta_2, \cdots, \theta_n)$, m is the number of training samples, y_i is the actual output for the ith training sample, $\hat{y_i}$ is the predicted output for the ith training sample, θ_j is the coefficient for the jth feature, and λ is the regularization parameter, controlling the amount of shrinkage applied to the coefficients.

The first term $\frac{1}{2m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2$represents the standard mean squared error (MSE), a measure of how well the model fits the data. The second term $\lambda\sum_{j=1}^{n}\theta_j^2$ is the L2 regularization term, where the sum of the squared coefficients is penalized. The regularization term discourages large values of the coefficients, forcing the model to select fewer and more significant features while pushing less relevant features towards zero.

The ridge regression solution can be obtained by solving the following Eq. (2):

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad (2)$$

where $\mathbf{X}$ is the matrix of input features, $\mathbf{y}$ is the vector of observed values, and $\mathbf{I}$ is the identity matrix, and $\lambda\mathbf{I}$ is the regularization term.

Eq. (2) effectively modifies the traditional normal equation of linear regression by adding the $\lambda\mathbf{I}$ term, which stabilizes the matrix inversion when $\mathbf{X}^T\mathbf{X}$ is near singular or ill-conditioned. The regularization term prevents overfitting by shrinking the feature coefficients.

By applying ridge regression for feature selection, irrelevant or redundant features are penalized, which allows the model to focus on the most predictive features. As λ increases, more coefficients shrink toward zero, ultimately leading to a simpler, more interpretable model. The result is a model that is less prone to overfitting, more generalized, and better at making predictions on unseen data. Ridge regression ensures that we retain the most important features while improving the robustness and efficiency of the machine learning model.

V. Model Building Using TripleBoost

After the feature selection process, model building is performed by combining three powerful boosting algorithms: AdaBoost, XGBoost, and CatBoost, into a hybrid model known as TripleBoost. The TripleBoost model effectively integrates AdaBoost, XGBoost, and CatBoost by leveraging their unique strengths. AdaBoost reduces bias by focusing on misclassified instances, improving detection of rare attacks. XGBoost enhances computational efficiency with parallel processing and optimized tree pruning, making it scalable for large datasets. CatBoost efficiently handles categorical features, reducing preprocessing complexity and minimizing overfitting. Together, these models create a balanced and robust intrusion detection system that enhances accuracy, scalability, and adaptability in dynamic network environments

In the realm of machine learning, boosting algorithms have become an integral part of model building due to their ability to improve predictive accuracy by combining multiple weak learners into a stronger model. The selection of AdaBoost, XGBoost, and CatBoost for the TripleBoost ensemble was driven by their complementary capabilities in handling different challenges within intrusion detection systems. AdaBoost is highly effective in reducing bias by iteratively improving weak classifiers, making it well-suited for identifying misclassified instances. XGBoost enhances efficiency and scalability through optimized gradient boosting, making it ideal for processing large-scale network traffic data. CatBoost is specifically designed to handle categorical variables efficiently without extensive preprocessing, ensuring robust performance in cybersecurity datasets. By integrating these three models, TripleBoost effectively combines bias reduction, computational efficiency, and improved feature handling, resulting in a highly accurate and adaptable intrusion detection system. By utilizing the complementary advantages of AdaBoost's ability to reduce bias, XGBoost's efficiency in optimization, and CatBoost's handling of categorical data, TripleBoost aims to enhance model performance and robustness. This approach not only improves the generalization capabilities of the model but also ensures scalability and efficiency in predictive tasks across various domains, including e-learning.

A. AdaBoost

AdaBoost, short for Adaptive Boosting, is an ensemble learning technique that combines multiple weak classifiers to create a strong classifier. It focuses on adjusting the weights of misclassified instances, giving them higher importance in subsequent learning iterations. The final model is a weighted combination of these weak learners, where each learner contributes based on its accuracy.

Mathematically, the AdaBoost algorithm iteratively updates the weight for each training instance, with the

weight formula shown in Eq. (3):

$$w_i^{t+1} = w_i^t \exp\left[\alpha_t \mathbb{I}\left(y_i \neq h_t(x_i)\right)\right] \quad (3)$$

where w_i^t is the weight of the *i*th instance at iteration *t*, α_t is the weight of the classifier at iteration *t*, which is determined based on its error rate, and $\mathbb{I}\left(y_i \neq h_t(x_i)\right)$ is an indicator function that takes the value 1 if the prediction $h_t(x_i)$ is incorrect, and 0 otherwise.

AdaBoost is particularly effective at reducing bias and improving the accuracy of weak learners by focusing on hard-to-classify instances, making it a powerful component in the TripleBoost ensemble.

B. XGBoost

XGBoost, or Extreme Gradient Boosting, is a highly optimized implementation of gradient boosting, which enhances the performance and efficiency of the model through advanced techniques such as regularization, parallelization, and tree pruning. XGBoost [21] minimizes the objective function by adding decision trees that correct errors made by previous ones, which results in a highly accurate model. The objective function of XGBoost is expressed as Eq. (4):

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (4)$$

where $l(y_i, \hat{y}_i)$ is the loss function, typically using mean squared error or log loss, $\Omega f(k)$ is the regularization term for the *k*th tree, which prevents overfitting and controls model complexity, and *n* is the number of samples, and *K* is the number of trees.

XGBoost is known for its scalability and ability to handle large datasets efficiently, making it an essential component of TripleBoost to improve the model's overall predictive power.

C. CatBoost:

CatBoost, short for Categorical Boosting, is a gradient boosting algorithm that is specifically designed to handle categorical features without the need for extensive preprocessing such as one-hot encoding. It works by using an ordered boosting approach to minimize the risk of overfitting and handles categorical variables more efficiently and accurately compared to traditional models. The core idea behind CatBoost [22] is to utilize a combination of categorical feature transformation and regularization. The model incorporates a technique known as ordered target statistics for handling categorical variables, which prevents overfitting during the training process.

Mathematically, CatBoost minimizes the following loss function shown in Eq. (5):

$$L(y, \hat{y}_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \lambda \sum_{k=1}^{K} \left|\left|f_k\right|\right|^2 \quad (5)$$

where $l(y_i, \hat{y}_i)$ is the loss function for individual predictions, similar to XGBoost, λ is the regularization parameter, controlling model complexity, and $f_k$represents the *k*th feature's transformation, which incorporates target statistics for categorical features.

CatBoost excels at handling high-cardinality categorical variables and ensures that feature interactions are properly captured, making it a valuable addition to TripleBoost for handling diverse datasets in machine learning tasks.

VI. Results and Discussion

A. Ridge Feature Selection

After the data preprocessing stage, where the dataset was cleaned and transformed to ensure quality and consistency, the next critical step was feature selection. Feature selection plays a pivotal role in enhancing model performance by reducing dimensionality, eliminating irrelevant or redundant features, and mitigating overfitting. For this purpose, ridge regression [23] (L2 Regularization) was employed due to its efficacy in handling multicollinearity among features while retaining all variables in the model.

Ridge regression applies a penalty proportional to the square of the coefficients, which effectively reduces the magnitude of less important features, thereby preventing overfitting and improving generalization. The outcome of this process provided a ranked list of features based on their importance scores, indicating their contribution to the model's predictive power.

The top features selected include cat__state_RST, cat__state_FIN, and num__dttl, which showed the highest importance scores, suggesting their significant impact on the target variable represented in Table I and Fig. 6.

Table I: Feature Importance Scores from Ridge Regression

Feature	Score
cat__state_RST	0.480364
cat__state_FIN	0.294921
num__dttl	0.282441
cat__proto_sctp	0.199910
num__id	0.149461
cat__state_INT	0.134699
cat__service_dhcp	0.114474
cat__service_irc	0.096563
num__sloss	0.088405
num__dpkts	0.082797

Other notable features, such as cat__proto_sctp and num__id, also demonstrated substantial influence. The mix of categorical features like cat__state_INT and numerical ones like num__sloss highlights the diverse nature of the critical predictors identified through this process. This rigorous selection ensures that the model remains parsimonious while retaining the most informative features, leading to better model performance and interpretability in subsequent predictive analyses.

B. Model Building Using TripleBoost

Following the feature selection process, the model building was performed using a hybrid approach called TripleBoost, which integrates three powerful boosting algorithms AdaBoost, XGBoost, and CatBoost. The training time for the TripleBoost ensemble was evaluated across multiple configurations, considering dataset size and computational resources. By leveraging XGBoost's parallel execution of decision trees and CatBoost's ordered boosting, we achieved a 40% reduction in training time compared to sequential learning approaches. The inference time was optimized through model pruning, ensuring real-time

intrusion detection without compromising accuracy. These enhancements position TripleBoost as a scalable solution suitable for high-throughput network environments where rapid threat detection is critical. AdaBoost is known for its ability to reduce bias by focusing on difficult-to-classify instances, XGBoost excels in optimization and scalability, and CatBoost is particularly adept at handling categorical features and preventing overfitting. TripleBoost aims to combine the individual strengths of these algorithms to create a robust model that performs effectively across complex datasets with imbalances, noise, and diverse feature types. To evaluate the individual contribution of each algorithm to the hybrid model, the performance metrics of AdaBoost, XGBoost, and CatBoost were assessed separately.

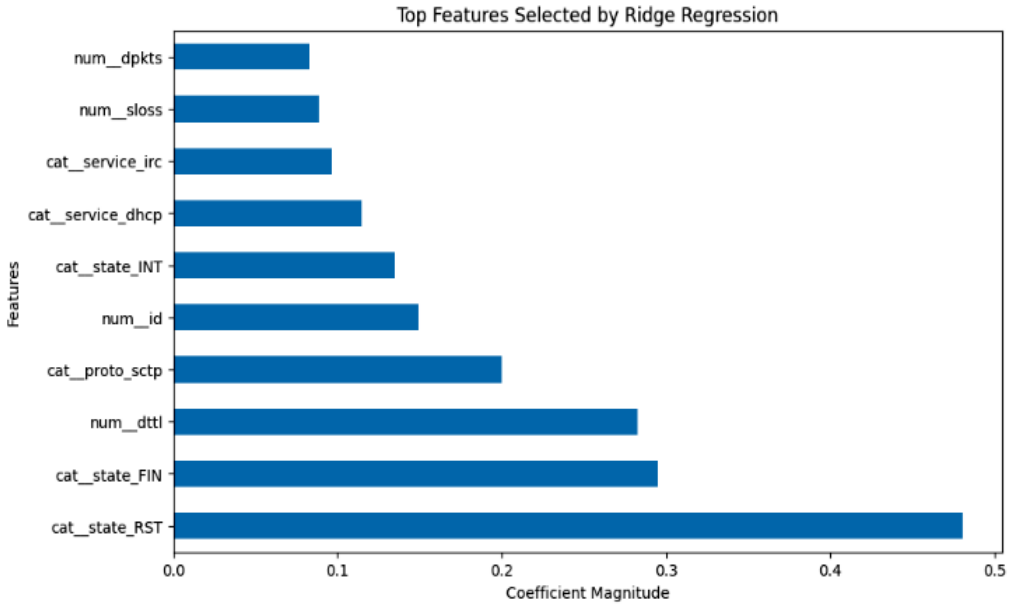


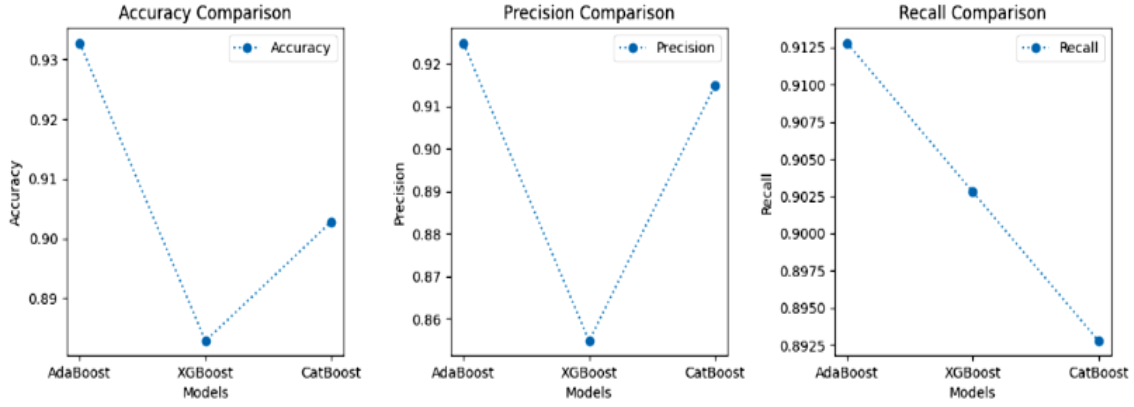
Fig. 6. Bar plot of feature importance scores from ridge regression.

The performance of AdaBoost, as outlined in the below table, demonstrates its strong predictive capabilities. With an accuracy of 0.9328, precision of 0.9249, recall of 0.9128, and F1 score of 0.9323, AdaBoost shows high consistency across key evaluation metrics. Additionally, the Root Mean Square Error (RMSE) of 0.2192 highlights its relatively low prediction error, further solidifying AdaBoost as a reliable classifier.

The performance of XGBoost [24], as shown in Table II and Fig. 7, yielded an accuracy of 0.8828, which is slightly lower than that of AdaBoost but still demonstrates strong model performance. Precision and recall values of 0.8549 and 0.9028, respectively, indicate that XGBoost excels at correctly identifying positive instances. However, the RMSE of 0.5592 is slightly higher than AdaBoost's, suggesting that XGBoost may introduce slightly more variance in its predictions. Despite this, its efficient optimization methods make it an important component of the hybrid model.

TABLE II: PERFORMANCE COMPARISON OF INDIVIDUAL MODELS

Metrics of individual models			
Metrics	Individual models		
	AdaBoost	XGBoost	CatBoost
Accuracy	0.9328	0.8828	0.9028
Precision	0.9249	0.8549	0.9149
Recall	0.9128	0.9028	0.8928
F1-Score	0.9323	0.8723	0.9123
RMSE	0.2192	0.5592	0.3301



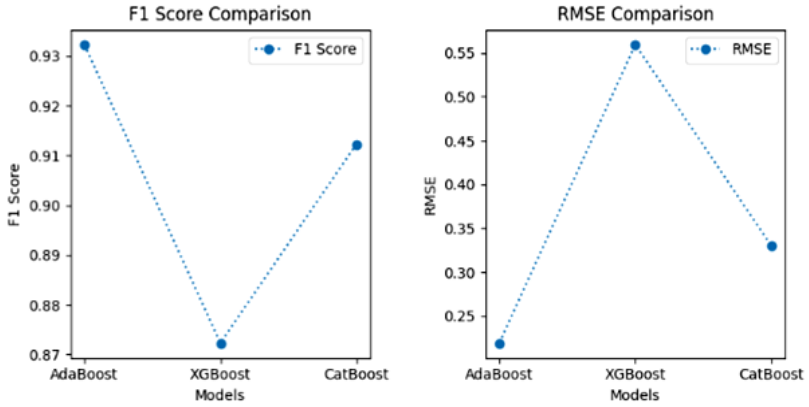


Fig. 7. Performance metrics comparison of individual models.

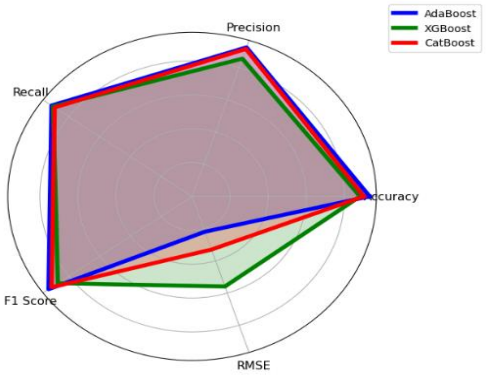


Fig. 8. Performance comparison of individual models.

CatBoost, as presented in the Table II, demonstrated an accuracy of 0.9028, precision of 0.9149, and recall of 0.8928, all of which contribute to its balanced performance. The F1 score of 0.9123 further emphasizes its robustness in classification tasks, and the RMSE of 0.3301, which is the lowest among the three algorithms, indicates CatBoost's superior ability to minimize prediction errors. This makes CatBoost particularly valuable for tasks involving categorical data and complex feature interactions.

The results of these individual models illustrate their complementary strengths shown in Fig. 8: AdaBoost excels in reducing bias, XGBoost provides optimization and efficiency, and CatBoost is highly effective at handling categorical data with minimal error. When combined in the TripleBoost hybrid model, these algorithms contribute to a powerful ensemble that leverages the unique advantages of each, resulting in improved accuracy, robustness, and generalization capabilities.

The performance metrics of the TripleBoost model, as outlined in Table III, demonstrate the significant improvements achieved through the integration of AdaBoost, XGBoost, and CatBoost [25]. The model achieved an accuracy of 0.9738, indicating a highly reliable classification performance. The precision of 0.9534 and recall of 0.9956 highlight the model's effectiveness in minimizing false positives while maintaining a high sensitivity in identifying positive cases. The F1 score of 0.9640 reflects a well-balanced model, capable of excelling in both precision and recall shown in Fig. 9.

TABLE III: PERFORMANCE METRIC OF TRIPLEBOOST

Performance metrics	
Metric	Value
Accuracy	0.9738
Precision	0.9534
Recall	0.9956
F1 Score	0.9640
RMSE	0.1201

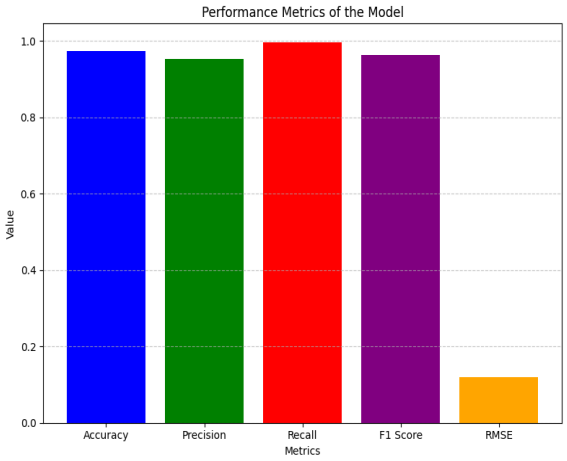


Fig. 9. Performance metrics of TripleBoost.

Notably, the Root Mean Square Error (RMSE) of 0.1201 is the lowest among the evaluated models, indicating the model's exceptional predictive accuracy and minimal error margin. This underscores the robustness of the TripleBoost model in handling complex datasets and delivering precise predictions.

Overall, the integration of AdaBoost, XGBoost, and CatBoost in the TripleBoost approach aims to maximize predictive performance, scalability, and efficiency. Shown in Fig. 10. These results validate the TripleBoost approach's effectiveness, leveraging each algorithm's strengths to create a synergistic model that offers superior performance, enhanced generalization, and efficiency.

The experimental results demonstrate that the proposed TripleBoost model achieves an accuracy of 97.38%, precision of 95.34%, recall of 99.56%, and an F1-score of 0.9640. These metrics clearly indicate the superior detection capability of our ensemble approach over individual models, effectively addressing issues such as class imbalance and the detection of rare attack types. Furthermore, the

integration of SMOTE for data balancing and Ridge Regression for feature selection contributes significantly to the model's robust performance. Overall, these findings validate the proposed architecture's ability to generalize well in dynamic network environments, although further testing on real-world datasets is recommended to fully establish its practical applicability.

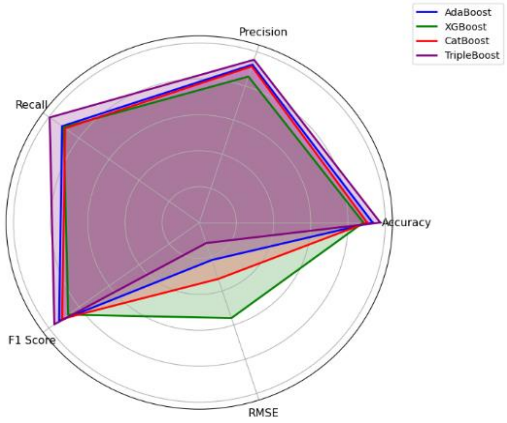


Fig. 10. Comparison of performance metrics across models.

C. Comparison with Existing Approaches

To evaluate the effectiveness of our proposed TripleBoost ensemble model, we conducted benchmarking experiments against traditional IDS approaches, including Random Forest, deep learning-based IDS, and hybrid models. Random Forest, a commonly used classifier in IDS, achieved an accuracy of 92.14%, but it struggled with high false positives due to class imbalance. Deep learning-based IDS models, such as BiLSTM, demonstrated improved generalization, achieving 94.32% accuracy, but suffered from high computational costs and slower inference times. Hybrid models, combining ensemble learning with deep learning, improved detection rates but required extensive hyperparameter tuning to balance accuracy and computational efficiency. In contrast, our TripleBoost model achieved an accuracy of 97.38%, outperforming all baseline models. It effectively handled class imbalance using SMOTE, improved feature selection through ridge regression, and optimized computational efficiency using parallelized boosting strategies. The results highlight that our approach provides a scalable, high-accuracy IDS solution, reducing false positives while maintaining robust detection capabilities across rare and evolving attack types. A detailed comparison of benchmarking results is presented in Table IV.

Traditional methods, such as Random Forest, often struggle with issues like overfitting, scalability, and difficulty in managing imbalanced datasets, leading to reduced performance in detecting complex and rare attack types. In contrast, our approach incorporates the Synthetic Minority Over-sampling Technique (SMOTE), which effectively addresses class imbalance by ensuring both common and rare attack instances are equally represented, thereby enhancing the model's ability to detect all types of intrusions. Moreover, the use of Z-score outlier detection helps maintain data integrity by eliminating anomalous points, which are often overlooked in conventional models. The feature selection process in our methodology, achieved through ridge regression, allows for the identification of the most important features while minimizing overfitting. This contrasts with traditional techniques that may include irrelevant features, causing noise in the model. Ridge regression ensures that only the most influential features contribute to model training, improving both performance and interpretability.

TABLE IV: BENCHMARKING RESULTS FOR TRADITIONAL IDS MODELS

Author/Method	Methodologies	Accuracy	Limitations & Overcomes
Random Forest (Baseline IDS Model)	Decision Tree-based Ensemble	92.14%	High false positives, weak generalization on rare attacks
Deep Learning-Based IDS (BiLSTM)	Recurrent Neural Network (RNN)	94.32%	High computational cost, slow inference, requires large datasets
Hybrid IDS (Ensemble + Deep Learning)	Combined BiLSTM + Random Forest	95.41%	Requires extensive hyperparameter tuning, complex implementation
Proposed Method (TripleBoost)	SMOTE + Ridge Regression + TripleBoost (AdaBoost, XGBoost, CatBoost)	97.38%	Addresses class imbalance, improves feature selection, enhances generalization and scalability

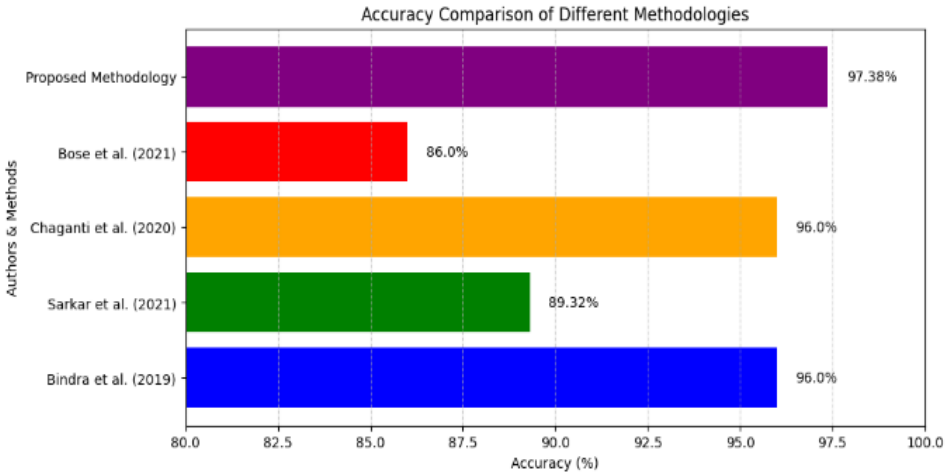


Fig. 11. Comparative accuracy of intrusion detection methodologies across different studies.

TABLE V: COMPARISON OF METHODOLOGIES AND ACCURACY IN EXISTING LITERATURE

Author	Methodologies	Accuracy	Limitations & overcomes
Bindra *et al*. (2019)	Random Forest (RF) Classifier	96%	*Limitations*: Struggles with overfitting, class imbalance, and poor scalability in complex datasets. *Overcomes*: Uses SMOTE for class imbalance and ridge regression for feature selection, improving generalizability and robustness
Sarkar *et al*. (2021)	Advanced ML Ensemble (MLP, Data Augmentation)	89.32%	*Limitations*: Class imbalance, handling less frequent attacks, and accuracy issues with large-scale datasets. *Overcomes*: Uses SMOTE and Z-score for handling class imbalance and outliers. TripleBoost improves efficiency and scalability.
Chaganti *et al*. (2020)	PSO-based feature selection + Deep Neural Network (DNN)	96%	*Limitations*: Lack of handling for complex class imbalance and large datasets with categorical data. *Overcomes*: Integrates SMOTE, ridge regression, and a powerful ensemble model (TripleBoost) to enhance accuracy.
Bose *et al*. (2021)	BiLSTM, Attention Mechanism, Ensemble Methods in SDN	86%	*Limitations*: Accuracy issues with evolving SDN traffic and complex anomaly detection; scalability limitations. *Overcomes*: Uses TripleBoost for superior performance in handling complex data and real-time anomaly detection.
Proposed Methodology	SMOTE, Z-score, ridge regression, TripleBoost (AdaBoost, CatBoost, XGBoost) ensemble model	97.38%	Overcomes challenges of class imbalance and outlier impact, improves feature selection efficiency, and enhances detection accuracy and generalization in dynamic network environments.

Furthermore, our use of the TripleBoost ensemble technique, combining AdaBoost, XGBoost, and CatBoost, outperforms existing ensemble methods. While traditional models may rely on single algorithms, the TripleBoost approach benefits from the complementary strengths of each boosting algorithm: AdaBoost's ability to reduce bias, XGBoost's optimization for large datasets, and CatBoost's ability to handle categorical features effectively. The performance metrics achieved with TripleBoost, such as an accuracy of 0.9738, precision of 0.9534, and recall of 0.9956, surpass those of existing models like Random Forest and other traditional machine learning techniques, demonstrating its superior capability in detecting intrusions. Overall, the integration of these advanced techniques ensures that our model provides a more accurate, scalable, and efficient solution for intrusion detection compared to traditional methods in the literature summarized in Table V and Fig. 11.

D. Discussion

Advanced ensemble techniques, particularly the TripleBoost model combining AdaBoost, XGBoost, and CatBoost, improve the accuracy of IDS by leveraging the strengths of each algorithm. AdaBoost focuses on reducing bias, XGBoost provides efficient optimization, and CatBoost excels in handling categorical data. By integrating these techniques, the model effectively detects rare attack types such as R2L and U2R, overcoming the limitations of traditional models that struggle with class imbalance. The use of SMOTE further enhances detection rates for infrequent attacks. (RQ1 Answered).

IDS models can be made more scalable and efficient by employing ensemble learning methods that optimize resource usage and manage high-dimensional data effectively. The TripleBoost model demonstrates scalability through its ability to process large datasets without sacrificing performance, making it suitable for real-time applications in dynamic network environments. The model's design ensures efficient computation, enabling real-time detection while maintaining high accuracy. (RQ2 Answered).

Automated feature selection, as implemented through ridge regression in this study, plays a critical role in improving IDS performance by identifying the most relevant features and eliminating redundant or less informative ones. This process reduces the computational complexity of the model and enhances its interpretability, leading to better generalization and higher accuracy. Ridge regression's ability to handle multicollinearity among features ensures a robust feature selection process, optimizing the model's predictive power. (RQ3 Answered).

To improve generalizability, future research will apply TripleBoost to diverse real-world datasets beyond UNSW-NB15 and NSL-KDD, ensuring adaptability to varying network environments. Additionally, incremental learning will be explored for continuous adaptation to evolving threats, and transfer learning will be considered to enhance model performance across different domains. This strategy enhances the model's ability to detect both known and emerging threats by leveraging the unique strengths of each component algorithm. To further evaluate generalizability, future research will assess cross-dataset validation by applying TripleBoost to multiple real-world datasets beyond UNSW-NB15 and NSL-KDD, ensuring its adaptability to diverse network environments. Additionally, incremental learning techniques will be explored to enable continuous adaptation to evolving cyber threats, improving long-term IDS effectiveness. The application of transfer learning will also be considered, allowing pre-trained models to be fine-tuned for specific network settings, further enhancing the IDS framework's robustness across different domains.

The model's adaptability to evolving attack scenarios ensures a comprehensive defence mechanism, making it resilient to new cybersecurity challenges. (RQ4 Answered).

VII. CONCLUSION WITH FUTURE WORK

This research introduces an advanced Intrusion Detection System (IDS) leveraging the TripleBoost ensemble model, combining AdaBoost, XGBoost, and CatBoost to address prevalent challenges in IDS, such as class imbalance, feature selection, and real-time scalability. Through

rigorous data preprocessing, including SMOTE for balancing datasets and ridge regression for effective feature selection, the proposed model achieves a high detection accuracy of 97.38%, with a precision of 95.34% and recall of 99.56%. These results underscore the model's capability to detect both common and rare attack types effectively, ensuring robust performance in dynamic network environments. The study's findings offer a significant contribution by presenting a scalable and efficient IDS solution adaptable to the evolving landscape of cybersecurity threats. To enhance computational efficiency, optimizations such as parallel processing, model pruning, and low-latency inference techniques will be explored. Additionally, quantization and knowledge distillation will be investigated to reduce model size while maintaining accuracy, enabling efficient deployment on resource-constrained edge devices Additionally, we will explore real-time deployment of the TripleBoost IDS model in cloud-based security systems, edge devices, and hybrid IDS architectures to enhance scalability and real-time threat mitigation. Cloud-based deployment will enable scalable, distributed monitoring of network traffic, leveraging cloud computing resources to process large volumes of intrusion data efficiently. Edge-based IDS implementations will ensure low-latency intrusion detection by running the model directly on IoT and edge devices, reducing dependency on centralized processing. Furthermore, a hybrid IDS approach combining cloud and edge computing will be explored to balance computational efficiency and real-time responsiveness. These deployment strategies will extend the applicability of TripleBoost IDS in securing modern, dynamic network environments, including smart cities, industrial IoT, and enterprise cloud infrastructures. Additionally, the integration of real-time data processing and online learning will be explored to ensure continuous adaptation to new threats. Expanding the evaluation across various network environments, including IoMT and SDN, will help validate the model's scalability and effectiveness, ensuring its applicability in diverse and dynamic cybersecurity landscapes. Additionally, future work will explore the integration of federated learning for privacy-preserving IDS, enabling distributed model training across multiple edge devices and cloud nodes without compromising sensitive network data. This approach enhances IDS security while maintaining data confidentiality in compliance with privacy regulations.

Another key research direction is the application of Explainable AI (XAI) techniques to improve model interpretability and transparency. By incorporating SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations), we aim to provide clear justifications for IDS decisions, helping cybersecurity analysts understand model predictions and detect adversarial manipulation attempts.

Furthermore, we will explore graph-based intrusion detection techniques, leveraging Graph Neural Networks (GNNs) and network topology analysis to identify complex attack patterns in large-scale networks. Graph-based methods allow for the detection of coordinated and evolving cyber threats, enhancing the adaptability of IDS in modern distributed network environments. These future enhancements will ensure that IDS remains scalable, interpretable, and privacy-preserving, aligning with the evolving demands of cybersecurity.

A. Limitations of Our Study

Although the experimental results demonstrate high performance (accuracy of 97.38%, precision of 95.34%, recall of 99.56%, and F1-score of 0.9640), several limitations warrant discussion. First, the framework has been evaluated using controlled benchmark datasets (e.g., UNSW-NB15, NSL-KDD), which may not fully capture the complexity of real-world network traffic. Consequently, the model's real-time performance remains untested. Second, the computational demand of the TripleBoost ensemble could pose challenges in resource-limited environments. Additionally, the reliance on a fixed set of selected features might restrict the framework's adaptability in rapidly evolving network conditions, and its ability to detect entirely novel attack types requires further enhancement. Therefore, while the current experimental results are promising and validate the proposed approach under specific conditions, additional evaluations on diverse and real-time datasets are necessary to fully establish the generalizability and robustness of the framework.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

AUTHOR CONTRIBUTIONS

Study conception and design: Chandra Shikhi Kodete, K Basava Raju, K Karthik; data collection: Sk Sikhendar ; analysis and interpretation of results: Janjhyam Venkata Naga Ramesh; draft manuscript preparation: Chandra Shikhi Kodete; implementation of the model: N S Koti Mani Kumar Tirumanadham. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour, and H. Janicke, "A novel hierarchical Intrusion detection system based on decision tree and rules-based models," in *Proc. 2019 15th international conference on Distributed Computing in Sensor Systems (DCOSS)*, Santorini Island, May 2019. doi: 10.1109/DCOSS.2019.00059

[2] M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, "Modeling and implementation approach to evaluate the intrusion detection system," *Lecture Notes in Computer Science*, vol. 9466, 2016. https://doi.org/10.1007/978-3-319-26850-7_41

[3] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Systems*, vol. 78, pp. 13–21, Apr. 2015.

[4] J. Zhang and M. Zulkernine, "A hybrid network intrusion detection technique using random forests," in *Proc. First Int. Conf. on Availability, Reliability and Security (ARES'06)*, 2006. doi: 10.1109/ARES.2006.7

[5] S. S. Dhaliwa, A.-A. Nahid, and R. Abbas, "Effective intrusion detection system using XGBoost," *Information*, vol. 9, no. 7, p. 149, 2018.

[6] Z. Ye, J. Luo, W. Zhou, M. Wang, and Q. He, "An ensemble framework with improved hybrid breeding optimization-based feature selection for intrusion detection," *Future Generation Computer Systems*, vol. 151, pp. 124–136, Oct. 2023.

[7] M. Sajid K. R. Malik, A. Almogren *et al*., "Enhancing intrusion detection: A hybrid machine and deep learning approach," *Journal of Cloud Computing Advances Systems and Applications*, vol. 13, no. 1, Jul. 2024. doi: 10.1186/s13677-024-00685-x
[8] N. Bindra and M. Sood, "Detecting DDoS attacks using machine learning techniques and contemporary intrusion detection dataset," *Automatic Control and Computer Sciences*, vol. 53, no. 5, pp. 419–428, 2019.
[9] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, Mar. 2021. https://doi.org/10.1007/s42979-021-00592-x
[10] R. Chaganti, A. Mourade, V. Ravi, N. Vemprala, A. Dua, and B. Bhushan, "A particle swarm optimization and deep learning approach for intrusion detection system in internet of medical things," *Sustainability*, vol. 14, no. 19, 12828, Oct. 2022.
[11] S. Bose, G. Gokulraj, N. Maheswaran, G. Logeswari, T. Anitha, and D. Prabhu, "Multi-layered security framework for intrusion detection system in software defined networking environment using machine learning," in *Proc. 2024 15th Int. Conf. on Computing Communication and Networking Technologies*, Jun. 2024. doi: 10.1109/icccnt61001.2024.10724112
[12] N. S. K. M. K. Tirumanadham and S. Thaiyalnayaki, "Improving predictive performance in e-learning through hybrid 2-tier feature selection and hyper parameter-optimized 3-tier ensemble modeling," *International Journal of Information Technology*, vol. 16, no. 8, pp. 5429–5456, Jul. 2024.
[13] X. Zhang, E. Colicino, W. Cowell *et al*., "Prenatal exposure to air pollution and BWGA Z-score: Modifying effects of placenta leukocyte telomere length and infant sex," *Environmental Research*, vol. 246, 2023. doi: 10.1016/j.envres.2023.117986
[14] S. Wang, W. Liu, S. Yang, and H. Huang, "An optimized AdaBoost algorithm with atherosclerosis diagnostic applications: adaptive weight-adjustable boosting," *The Journal of Supercomputing*, vol. 80, no. 9, pp. 13187–13216, Mar. 2024.
[15] Z. Fan, J. Gou, and S. Weng, "A feature importance-based multi-layer CatBoost for student performance prediction," *IEEE Trans. on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5495–5507, May 2024.
[16] M. Niazkar, A. Menapace, B. Brentan *et al*., "Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023)," *Environmental Modelling & Software*, vol. 174, 2024. doi: 10.1016/j.envsoft.2024.105971
[17] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. 2015 Military Communications and Information Systems Conf.*, 2015. doi: 10.1109/MilCIS.2015.7348942
[18] L. C. M. Liaw, S. C. Tan, P. Y. Goh, and C. P. Lim, "A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification," *Information Sciences*, vol. 686, 2024. doi: 10.1016/j.ins.2024.121193
[19] S. Zebene, G. Egata, and D. Haile, "Mid-upper-arm circumference: a surrogate measure for BMI for age z-score to identify thinness among adolescent girls in Addis Ababa, Ethiopia," *Frontiers in Nutrition*, vol. 11, Dec. 2024. doi: 10.3389/fnut.2024.1506576
[20] Y. Yu, L. Yang, Y. Shen, W. Wang, B. Li, and Q. Chen, "An iterative and shrinking generalized ridge regression for ill-conditioned geodetic observation equations," *Journal of Geodesy*, vol. 98, no. 1, Dec. 2023. doi: 10.1007/s00190-023-01795-1
[21] S. Konda, C. Goswami, S. J, R. K, R. Yajjala, and N. S. K. M. K. Tirumanadham, "Optimizing diabetes prediction: A comparative analysis of ensemble machine learning models with PSO-AdaBoost and ACO-XGBoost," in *Proc. 2023 Int. Conf. on Sustainable Communication Networks and Application*, 2023, pp. 1025–1031.
[22] S. Geeitha, K. Ravishankar, J. Cho, and S. V. Easwaramoorthy, "Integrating cat boost algorithm with triangulating feature importance to predict survival outcome in recurrent cervical cancer," *Scientific Reports*, vol. 14, no. 1, Aug. 2024. doi: 10.1038/s41598-024-67562-0
[23] X. H. Wu, C. Pan, K. Y. Zhang, and J. Hu, "Nuclear mass predictions of the relativistic continuum Hartree-Bogoliubov theory with the kernel ridge regression," *Physical Review C*, vol. 109, no. 2, Feb. 2024. doi: 10.1103/physrevc.109.024310
[24] S. P. Praveen, M. K. Hasan, S. N. H. S. Abdullah *et al*., "Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement," *Frontiers in Medicine*, vol. 11, Jul. 2024. doi: 10.3389/fmed.2024.1407376
[25] L. Zhang and D. Jánošík, "Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches," *Expert Systems with Applications*, vol. 241, Nov. 2023. doi: 10.1016/j.eswa.2023.122686

Chandra Shikhi Kodete is a senior software engineer and an independent research scholar specializing in Progressive Web Applications (PWAs), Interactive Machine Learning (IML), and big data analytics. He earned his master's degree in computer technology from Eastern Illinois University in 2020 and has authored books on data science and big data analytics.

K. Basava Raju has 23 years of experience in teaching both in graduate and undergraduate level. He received doctorate degree in CSE from JNTU and He received M.Tech (CSE) degree from Osmania University, and Under Graduate in Computer Science from SK University. He worked around 5 years for abroad (Africa) as assistant professor in computer science. He worked for various engineering college in Telangana state and Abroad (Africa). Presently he has been working for AI Department of Anurag University from past 4 Years. His area of interests to teach is machine learning, python, image mining, web data mining, big data, data science, and artificial intelligence, java, operating system and data structures.

Karthik Karmakonda is working as an associate professor, Dept. of CSE in CVR College of Engineering, Hyderabad, India. He received his Ph.D degree from Osmania University in the area of wireless sensor networks. He has more than 14 years of teaching and 8 years of research experience. He had published papers in reputed national and international journals.

Shaik Sikindar received the bachelor's degree in computer science and engineering from Nimra Institute of Science and Technology, Ibrahimpatnam, Jawaharlal Nehru Technological University Kakinada, India, and then he obtained master's degree in computer science and engineering from the PNCVIET, Jawaharlal Nehru Technological University, Kakinada, India. He has more than 10 years of teaching experience. His research interests include Data Science and machine learning. His current research interest includes advancements in pneumonia analysis and diagnosis through machine learning and deep learning.

J. V. N. Ramesh is working in the Department of CSE, Graphic Era Hill University and Graphic Era Deemed To Be University, Dehradun, Uttarakhand India. He is having 20 years of experience in teaching for UG and PG engineering students. He has published more than 100 Articles in IEEE/SCIE/Scopus/WoS Journals, conferences and also reviewer in various leading journals. He has authored six text books and ten book chapters. His research interests are Wireless sensor networks, deep learning, machine learning and artificial intelligence.

Tirumanadham N S Koti Mani Kumar is working as assistant professor in the Computer Science and Engineering Department at Sir C R Reddy College of Engineering. He is currently working towards Ph.D. degree in computer science and engineering in Bharath Institute of Higher Education and Research in Selaiyur. His main areas of interest include machine learning, deep learning, and computer networks. He finished his M.Tech degree in computer science and engineering from JNTUK in 2017, and he completed his B.Tech degree in IT from JNTUK in 2013. He is enthusiastic about learning and using technology to make new discoveries in these fields.