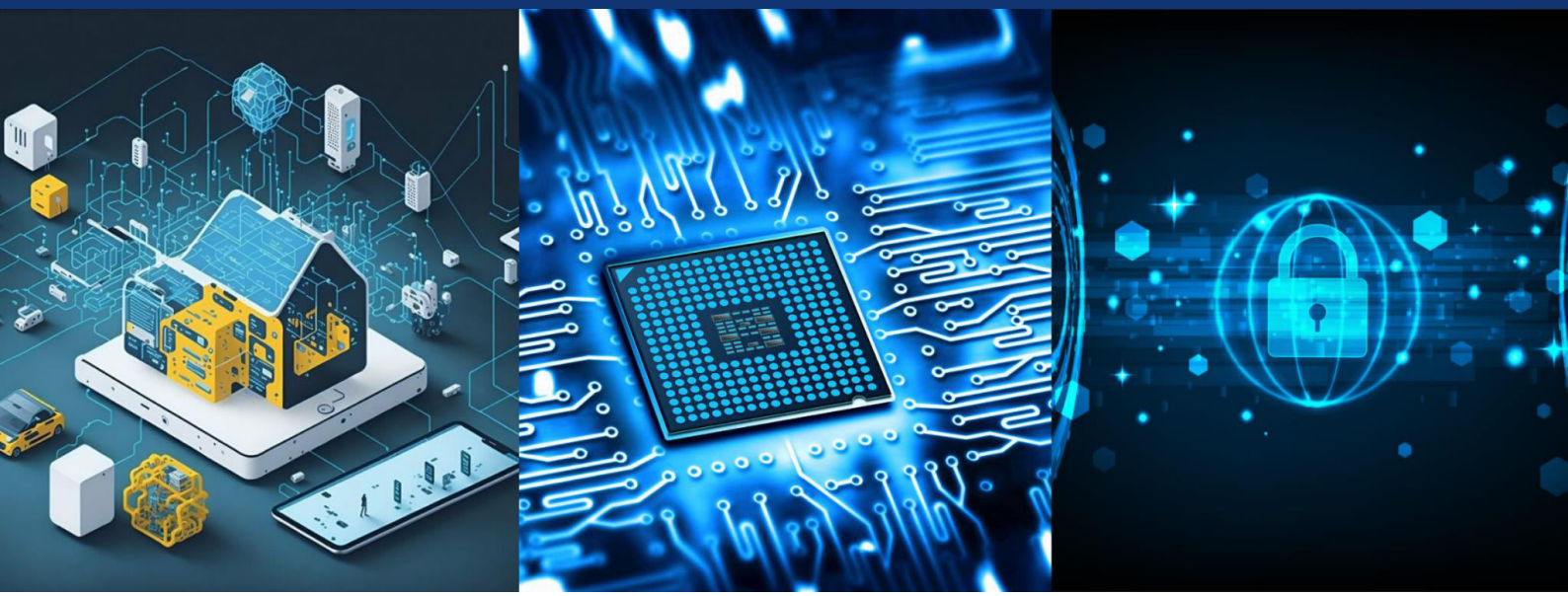
IJIRCCE

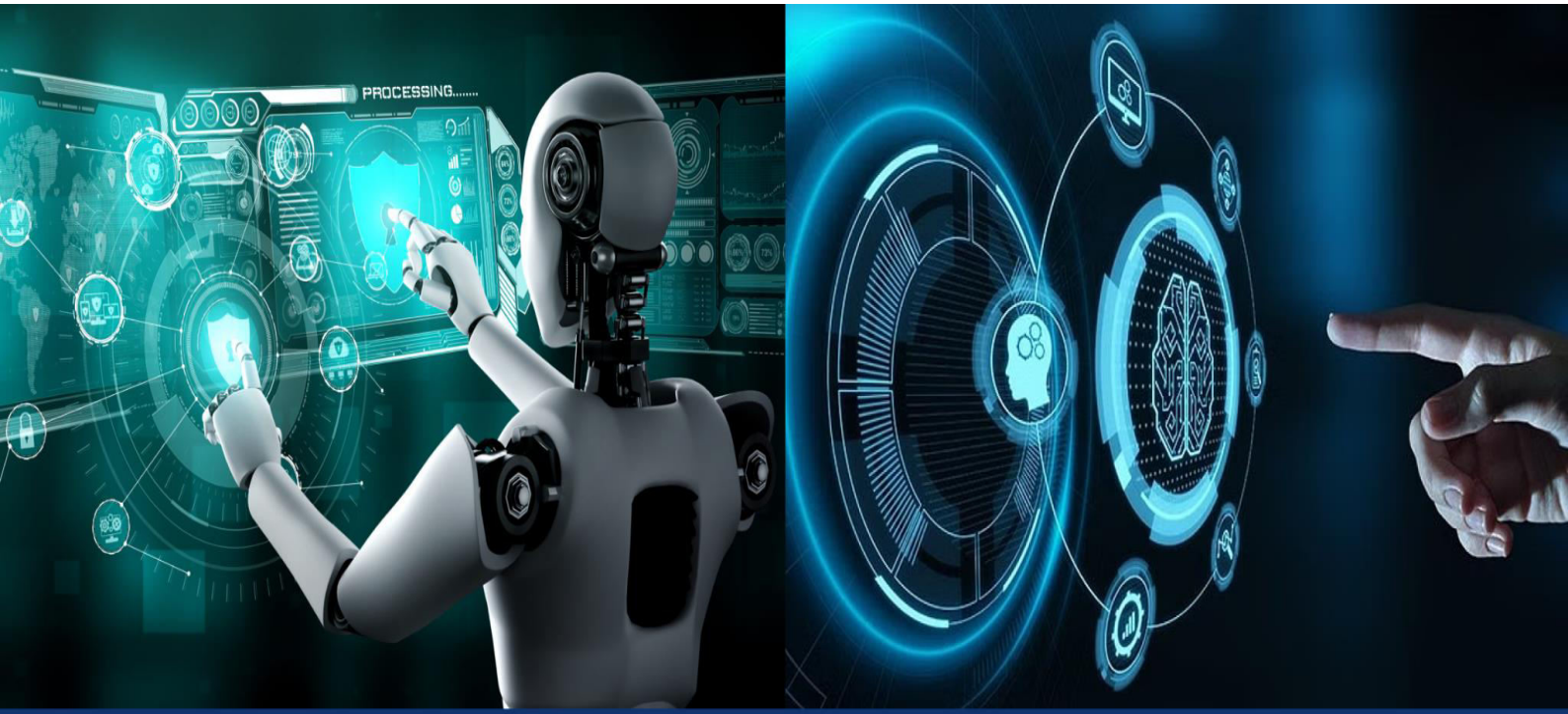
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Hybrid Boosted Forest Approach for Intrusion Detection in High Velocity Networks Using SMOTE and Chi-Square Feature Selection

Mummareddy Bala Sowmya

M. Tech Scholar, Department of CSE, Sir C R Reddy College of Engineering, Eluru, India

ABSTRACT: The rapid expansion of high-velocity networks driven by cloud computing, Internet of Things (IoT) devices, and large-scale enterprise infrastructures has rendered intrusion detection both a critical necessity and a formidable challenge. Conventional Intrusion Detection Systems (IDS) struggle with two persistent limitations: highly imbalanced datasets, in which rare but critical attack classes are underrepresented, and redundant or noisy feature spaces, which increase computational overhead and reduce classification accuracy. This paper proposes a Hybrid Boosted Forest Approach that integrates three complementary techniques. The Synthetic Minority Oversampling Technique (SMOTE) addresses class imbalance by generating synthetic minority-class samples. Chi-Square feature selection eliminates statistically irrelevant attributes, reducing dimensionality and improving efficiency. A boosted forest ensemble combining Random Forest and Gradient Boosting provides robust, scalable classification resilient to overfitting. Experiments conducted on the UNSW-NB15 benchmark dataset validate the proposed approach. The optimised model achieves 98.30% accuracy, precision of 98.30%, recall of 98.30%, F1-score of 98.29%, and an RMSE of 0.1305, outperforming existing methods. These results demonstrate the framework's suitability for real-time deployment in high-velocity network environments.

KEYWORDS: Intrusion Detection System; Random Forest; Gradient Boosting; SMOTE; Chi-Square Feature Selection; UNSW-NB15; Network Security; Ensemble Learning

I. INTRODUCTION

The proliferation of Internet of Things (IoT) devices and the widespread adoption of cloud-centric architectures have exponentially expanded the attack surface of modern networks. As network environments grow in scale and complexity, the inadequacy of traditional Intrusion Detection Systems (IDS) has become increasingly apparent. Signature-based IDS, which match observed traffic patterns against predefined rule sets, are inherently limited in their ability to detect zero-day exploits and novel attack vectors [1].

Machine learning (ML) techniques have emerged as a promising alternative, offering the capacity to learn from historical traffic data and identify intrusions through behavioural anomaly detection. Among ensemble methods, Random Forest (RF) and Gradient Boosting (GB) have demonstrated superior classification performance. However, two persistent challenges constrain the effectiveness of ML-based IDS in practice: class imbalance and high-dimensional feature redundancy.

In publicly available network datasets, benign traffic overwhelmingly dominates over malicious traffic, often by a factor exceeding 100:1. Standard ML models trained on such distributions develop a systematic bias toward the majority class, frequently failing to detect rare but high-impact attack categories such as User-to-Root (U2R) and Remote-to-Local (R2L) intrusions. Furthermore, network traffic datasets typically contain dozens of attributes, many of which are statistically irrelevant to classification and introduce noise that degrades model performance.

This paper proposes a Hybrid Boosted Forest framework that simultaneously addresses these challenges through a structured pre-processing and modelling pipeline. The contributions of this work are as follows:

(i) Application of SMOTE to correct class imbalance and improve minority-class detection sensitivity.
(ii) Deployment of Chi-Square statistical filtering to identify and retain only the most discriminative network attributes.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

(iii) Development of a Boosted Forest ensemble that fuses Random Forest and Gradient Boosting, achieving complementary reduction of variance and bias.
(iv) Hyper parameter optimisation via Grid Search and evaluation on the UNSW-NB15 benchmark dataset.
The remainder of this paper is structured as follows: Section II reviews related work; Section III describes the proposed methodology; Section IV presents the dataset; Section V details pre-processing; Section VI describes feature selection; Section VII reports experimental results; Section VIII concludes the paper.

II. LITERATURE SURVEY

A study on Intrusion Detection Systems (IDS) highlights that IDS are broadly categorized into signature-based and anomaly-based approaches, each with distinct advantages and limitations. Signature-based systems are highly effective in detecting known attack patterns but fail to identify novel or unknown threats. In contrast, anomaly-based systems establish a baseline of normal network behaviour and flag deviations from this baseline; however, they often produce higher false-positive rates, especially in dynamic environments [2]. This distinction underscores the need for more adaptive and intelligent detection mechanisms.

A study on machine learning approaches in IDS demonstrates that the adoption of machine learning techniques has significantly enhanced anomaly-based detection capabilities. Decision Trees provide interpretable classification rules but are prone to overfitting. Support Vector Machines (SVM) are effective in handling high-dimensional feature spaces but suffer from scalability issues when applied to large datasets. Artificial Neural Networks (ANN) are capable of capturing complex non-linear relationships; however, they require large labeled datasets and lack interpretability [3]. Ensemble methods have emerged as a dominant paradigm in IDS. Random Forest reduces variance through bootstrap aggregation by constructing multiple independent decision trees and combining their outputs [4]. Gradient Boosting improves model performance by reducing bias through sequential error correction, where new models are trained on the residuals of previous ones [5]. Hybrid models that combine Random Forest and Gradient Boosting approaches have shown improved performance compared to individual methods.

A study on feature selection and class imbalance in IDS emphasizes their critical role in improving detection efficiency. Feature selection techniques help remove redundant and irrelevant attributes, thereby enhancing model performance. The Chi-Square test is widely used as a computationally efficient filter-based method that measures the statistical dependence between features and class labels [6]. Other methods such as Information Gain and Principal Component Analysis (PCA) are also used; however, the Chi-Square test is preferred for categorical classification problems. Additionally, class imbalance is a significant challenge in intrusion detection datasets, where the majority class dominates the minority attack classes. Without proper handling, classifiers tend to be biased toward the majority class, resulting in poor detection of rare attacks [7]. The Synthetic Minority Oversampling Technique (SMOTE) addresses this issue by generating synthetic samples for the minority class through interpolation, reducing overfitting compared to simple duplication methods [8].

A study on benchmark datasets and research gaps in IDS reveals that many previous works rely on datasets such as KDD Cup 1999 and NSL-KDD. These datasets contain outdated traffic patterns and do not accurately represent modern cyber threats. To overcome these limitations, the UNSW-NB15 dataset was developed using the IXIA Perfect Storm tool at the Australian Centre for Cyber Security, providing a more realistic and diverse representation of modern network traffic [9]. Despite these advancements, several research gaps remain. Existing models often fail to address both class imbalance and feature redundancy simultaneously. Many studies continue to depend on outdated benchmark datasets, and there is limited exploration of hybrid ensemble models combining Random Forest and Gradient Boosting on modern datasets such as UNSW-NB15. Addressing these gaps is essential for developing more robust and effective intrusion detection systems.

III. METHODOLOGY

The proposed methodology is designed as a structured and systematic pipeline for effective intrusion detection in high-velocity network environments using the UNSW-NB15 dataset, which contains modern network traffic with diverse and realistic attack categories. The dataset is selected due to its ability to represent contemporary cyber threats and its suitability for evaluating machine learning-based intrusion detection systems. Initially, data pre-processing is performed to enhance the quality and reliability of the dataset. Missing values present in both numerical and categorical attributes

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

are handled using median imputation for numerical features and mode imputation for categorical features, ensuring robustness against skewed distributions.

Subsequently, outlier detection and removal is carried out using the Interquartile Range (IQR) method to eliminate extreme values that may distort the learning process. To address the significant issue of class imbalance inherent in network intrusion datasets, the Synthetic Minority Oversampling Technique (SMOTE) is applied to generate synthetic samples for minority attack classes, thereby improving the model's ability to detect rare intrusions. Following pre-processing, feature selection is performed using the Chi-Square statistical test to evaluate the dependency between each feature and the target class label. Based on this analysis, the top five most significant features are selected, effectively reducing dimensionality and improving computational efficiency without compromising predictive performance.

The core of the proposed framework is a hybrid ensemble model referred to as the Boosted Forest, which integrates Random Forest and Gradient Boosting algorithms. Random Forest operates using a bagging approach to reduce variance by aggregating predictions from multiple decision trees, while Gradient Boosting enhances performance by sequentially minimizing prediction errors and reducing bias. The combination of these two techniques enables the model to achieve both stability and high accuracy.

To further enhance performance, hyper parameter optimization is conducted using Grid Search in conjunction with five-fold cross-validation, ensuring optimal parameter selection and generalization capability. Finally, the model is evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and Root Mean Square Error (RMSE). The experimental results demonstrate that the proposed approach achieves an accuracy of 98.30 percent, along with strong performance across all evaluation metrics, confirming its effectiveness and reliability for intrusion detection in modern network environments.

The overall working flow of the proposed research methodology is illustrated below. It presents a structured pipeline starting from data acquisition and pre-processing, followed by feature selection and model development. The process further includes hyper parameter optimization and performance evaluation to ensure accuracy and robustness. This flow provides a clear representation of how the proposed system achieves efficient and reliable intrusion detection.

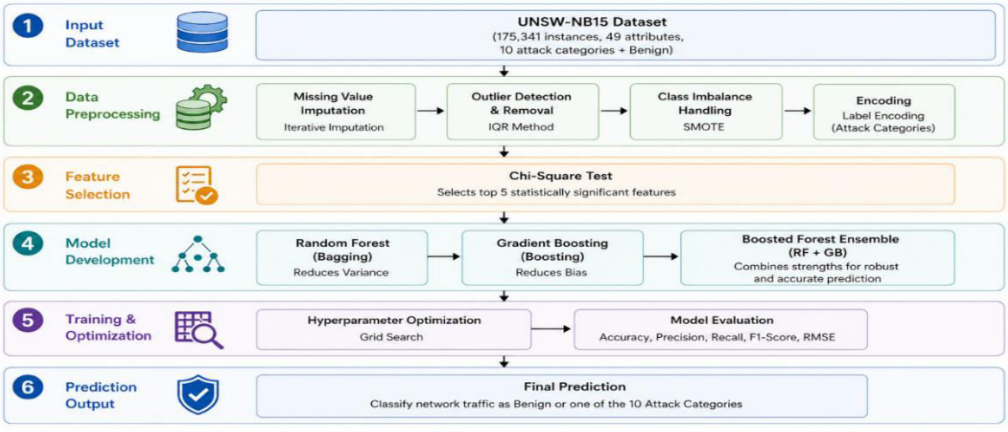


Fig. 1: Workflow

3.1 Dataset Description

The study utilizes the UNSW-NB15 dataset, a widely used benchmark for network intrusion detection generated by the Australian Centre for Cyber Security (ACCS) using the IXIA Perfect Storm tool to simulate both normal user activity and nine categories of cyber-attacks in a controlled environment. The dataset contains 2.54 million records with 49 features, comprising basic connection attributes, content-level features, and traffic- and time-based features. The nine attack categories represented are Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, and Worms. UNSW-NB15 was selected over earlier benchmarks such as KDD99 and NSL-KDD because it incorporates modern low-footprint attack patterns, including Advanced Persistent Threats, which are absent from older datasets. Its

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

realistic class distribution, characterized by severe imbalance and rare attack categories, also makes it an appropriate benchmark for evaluating class-balancing strategies.

3.2 Data Pre-processing

Data pre-processing is a critical step in the proposed framework, aimed at enhancing the quality, consistency, and reliability of the dataset before model training. Network traffic datasets commonly contain missing values, extreme outliers, and severe class imbalances, each of which can negatively impact model performance if not properly addressed. In this study, missing values were identified across 23 of the 49 available features, accounting for approximately 47% of the feature space. Numerical features exhibiting right-skewed distributions, such as destination jitter and source time-to-live, were imputed using the median, owing to its robustness against the extreme values characteristic of high-velocity network traffic. Categorical features, including service type and protocol, were imputed using the mode to preserve the most probable protocol behaviour.

To address the presence of statistical outliers, the Interquartile Range method was applied using the standard 1.5×IQR boundary. The resulting refined dataset represents the statistically consistent core of the original distribution, free from extreme noise that would otherwise distort gradient-based calculations within the model. To ensure that performance on this refined subset remains generalisable, five-fold cross-validation was applied during model evaluation. To address the issue of class imbalance, the Synthetic Minority Over-Sampling Technique is applied following outlier removal to generate synthetic minority-class samples through k-nearest-neighbour interpolation in the feature space. This ensures that the model receives sufficient exposure to rare attack categories, such as Worms and Shellcode, without duplicating existing records, thereby reducing model bias and improving classification fairness across all classes.

3.3 Feature Selection Using Chi-Square Analysis

Following pre-processing, a statistical feature selection procedure is applied to identify the most relevant network attributes for intrusion detection. In many large-scale datasets, the presence of redundant and irrelevant features can negatively affect model performance and increase computational complexity. To address this, the Chi-Square test is employed to measure the statistical dependency between each feature and the target class label. The Chi-Square statistic is formally defined as:

$$\chi^2 = \Sigma \left[(O_i - E_i)^2 / E_i\right]$$

where O_i denotes the observed frequency and E_i denotes the expected frequency under the assumption of independence between the feature and the class. Features yielding a high Chi-Square value demonstrate a strong association with class membership and are retained for model training. Based on this evaluation, five features were selected, namely source time-to-live, flow identifier, source TCP window size, destination load, and destination window size. These attributes capture network-level and transport-level characteristics that serve as reliable discriminators between normal and malicious traffic, thereby reducing dimensionality and enhancing the overall efficiency of the learning process.

TABLE II. Top-5 Features Selected by Chi-Square Analysis

Feature	Chi-Square Score
sttl (Source Time-to-Live)	19,474.82
id (Flow Identifier)	15,452.99
swin (Source TCP Window)	10,618.76
dload (Destination Load)	10,589.03
dwin (Destination Window)	9,833.44

3.4 Boosted Forest Model Design

The proposed framework adopts a hybrid ensemble learning strategy, referred to as the Boosted Forest model, to improve predictive performance while maintaining strong generalization capability. Ensemble learning is widely recognized for its ability to enhance model accuracy by combining multiple learning algorithms, and the proposed design integrates two complementary paradigms, namely Random Forest and Gradient Boosting, as constituent components.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Random Forest serves as the stabiliser layer of the ensemble, employing bootstrap aggregation to construct multiple independent decision trees, each trained on a random subset of data and features. The final prediction is obtained as the majority vote across all trees, a parallel architecture that reduces variance and confers robustness against the noise inherent in network traffic data. Gradient Boosting, on the other hand, functions as the precision layer by constructing trees sequentially, where each successive tree is fitted to the negative gradient of the loss function with respect to the current ensemble prediction. This iterative self-correction mechanism enables the detection of subtle, low-frequency attack patterns that parallel ensemble methods may overlook.

The Boosted Forest integrates these two paradigms through a hybrid fusion mechanism, wherein Random Forest provides the baseline prediction and Gradient Boosting refines it by correcting residual errors. This combination simultaneously mitigates variance through bagging and bias through sequential error correction, yielding superior generalisation relative to either method applied independently. The overall ensemble design reduces the risk of overfitting while maintaining strong predictive performance, making it suitable for real-world network security applications.

3.5 Hyper parameter Optimisation

To further enhance the predictive capability of the proposed model, a systematic hyper parameter optimisation procedure is incorporated into the framework. Grid Search with five-fold cross-validation is applied to exhaustively explore the hyper parameter space, ensuring that the optimal model configuration is identified in a structured and reproducible manner. The parameters subjected to tuning include the number of estimators, maximum tree depth, minimum samples required per split, and the learning rate. The optimal configuration is selected based on the F1-score, ensuring balanced and reliable performance across all classes, including the minority attack categories that are most challenging to classify accurately.

IV. RESULTS AND DISCUSSION

4.1 Performance Metrics

The performance of the proposed boosted forest model is evaluated using standard classification metrics to ensure a comprehensive assessment of its predictive capability. The model is initially assessed without hyper parameter optimization, yielding an accuracy of 97.96 percent, a precision of 0.98, a recall of 0.95, and an f1-score of 0.97. Following grid search optimization, performance improved across all metrics, achieving an accuracy of 98.30 percent, precision of 98.30 percent, recall of 98.30 percent, and an f1-score of 98.29 percent. Notably, the recall for the minority benign class increased from 0.95 to 0.983, and the RMSE decreased from 0.1428 to 0.1305, reflecting tighter probabilistic calibration. These results confirm the robustness and consistency of the proposed framework across all evaluation criteria.

Table 1. Initial and Optimised Model Performance Metrics

Configuration	Accuracy	Precision	Recall	F1-Score
Initial	97.96%	0.98	0.95	0.97
Optimised	98.30%	98.30%	98.30%	98.29%

4.2 Comparative Analysis

The proposed Boosted Forest model is compared against existing intrusion detection methods reported in the literature. The proposed approach achieves the highest accuracy of 98.30 percent and F1-score of 98.29 percent among all compared methods. Bindra and Sood reported 96 percent accuracy using Random Forest with K-fold cross-validation, while Chaganti et al. achieved an equivalent 96 percent through PSO-optimised Deep Neural Networks. More substantial improvements are observed over Sarkar et al., whose ensemble with data augmentation on NSL-KDD yielded 89.32 percent, and Bose et al., whose BiLSTM with attention mechanism achieved 86 percent on the In-SDN dataset. These comparisons collectively demonstrate the superior generalisation capability of the proposed framework across varied experimental conditions.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Reference	Method	Accuracy
Bindra & Sood [12]	Random Forest + K-fold CV	96%
Sarkar et al. [10]	Ensemble + Data Augmentation (NSL-KDD)	89.32%
Chaganti et al. [13]	PSO + Deep Neural Network	96%
Bose et al. [11]	BiLSTM + Attention (In-SDN)	86%
Proposed	**Boosted Forest + SMOTE + Chi-Square + GridSearch**	**98.30%**

4.3 Discussion

The superior performance of the optimised boosted forest model can be attributed to three principal factors. First, smote equalised the class distribution, enabling the gradient boosting component to develop sharp decision boundaries for rare attack categories that were previously ignored by majority-biased models. Second, chi-square feature selection reduced input dimensionality from 49 to 5 features, substantially decreasing the search space for each tree split and accelerating convergence during optimisation. The selected attributes, namely source time-to-live, flow identifier, source tcp window size, destination load, and destination window size, with respective chi-square scores of 19,474.82, 15,452.99, 10,618.76, 10,589.03, and 9,833.44, capture the most discriminative network-level and transport-level characteristics for distinguishing normal from malicious traffic. Third, the architectural fusion of random forest and gradient boosting achieves complementary error reduction, where random forest suppresses variance through democratic aggregation while gradient boosting eliminates residual bias through sequential specialisation. Statistical robustness was confirmed through five-fold cross-validation, demonstrating consistent performance across all data partitions. The precision-recall balance of 0.983 indicates that the model does not sacrifice false-negative rate for false-positive reduction, a critical requirement in security-sensitive deployments where an undetected intrusion carries a significantly higher cost than a false alarm.

V. CONCLUSION

This paper presented a hybrid boosted forest approach for intrusion detection in high-velocity networks, integrating smote-based class balancing, chi-square feature selection, and a random forest–gradient boosting ensemble optimised via grid search. Evaluated on the unsw-nb15 dataset, the model achieved 98.30 percent accuracy, 98.30 percent precision, 98.30 percent recall, 98.29 percent f1-score, and an rmse of 0.1305, surpassing all compared baseline methods. The proposed framework addresses key challenges in network traffic analysis, including class imbalance, high dimensionality, and the detection of rare attack categories, thereby offering a balanced solution that combines accuracy, efficiency, and generalisability essential for modern intrusion detection systems. Future research directions include adversarial robustness testing through adversarial training protocols, federated learning integration for privacy-preserving collaborative intrusion detection across distributed organisations, and extension to streaming data environments for real-time incremental model updates.

REFERENCES

[1] A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour, and H. Janicke (2019) A novel hierarchical Intrusion detection system based on decision tree and rules-based models. In: 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), Santorini Island, Greece, Greece, 29–31 May 2019.

[2] Saber M, Chadli S, Emharraf M, El Farissi I (2015) Modeling and implementation approach to evaluate the intrusion detection system. In: International conference on networked systems, pp 513–517.

[3] Lin W-C, Ke S-W, Tsai C-F (2015) CANN: an intrusion detection system based on combining cluster centers and nearest neighbors. Knowl based Syst 78:1321.

[4] Zhang J, ZulkernineM (2006) A hybrid network intrusion detection technique using random forests. In: First international conference on availability, reliability and security(ARES'06),2006,p8.

[5] DhaliwalSS,NahidA-A,AbbasR(2018).AneffectiveintrusiondetectionsystemusingXGBoost. Information9(7):149

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details