VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2025, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 13, Number 2, April-June 2025 pp: 01-12

Supervised Learning Approach for Intrusion Detection in Unbalanced Network Traffic

Zeeshan Ali [1*], **Adnan Akram** [2], **Naeem Aslam** [3], **Muhammad Saeed Khurram** [3]

[1]Department of Information Engineering and Computer Science and Mathematics, University of L'Aquila, Italy; [2]Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan; [3]Department of Computer Science, National College of Business Administration and Economics Lahore, Pakistan

Keywords: *Intrusion Detection System, Machine Learning, Cyber Security, Artificial Intelligence, Support Vector Machine, Unbalanced Dataset, Synthetic Minority Oversampling Technique.*

Journal Info:
Submitted:
April 6, 2025
Accepted:
April 28, 2025
Published:
May 3, 2025

Abstract Intrusion detection systems (IDS) serve as critical sentinels in network security, assuming a paramount role in identifying and mitigating potential threats. With the evolution of our digital landscape, robust and productive intrusion detection mechanisms have become increasingly imperative. The significance of IDS lies in their ability to safeguard network resources' integrity, confidentiality, and availability. In an era where cyber threats constantly evolve in complexity and scale, IDS serves as the front line of defence, tirelessly monitoring network traffic to pinpoint suspicious activities and mitigate potential security breaches. To address the class imbalance problem, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to pre-process the CIC-IDS 2017 and NSL-KDD 2009 datasets. Advanced machine learning technique is harnessed to enhance IDS capabilities, specifically through utilising Support Vector Machines (SVM) for subsequent classification tasks. The experimental outcomes on both datasets unveil exceptional accuracy of 99% and performance across multiple intrusion types, underscoring the effectiveness of our SVM-based approach in strengthening IDS.

***Corresponding author email address:** zeeshan.ali@graduate.univaq.it
DOI: 10.21015/vtse.v13i2.2116

1 Introduction

The proliferation of computer networks and the increasing dependence on information technology have amplified the need for robust security measures. IDS have emerged as critical components in ensuring the integrity and availability of computer systems [1] by detecting and responding to unauthorized activities. IDS can be categorized into two primary types: Host-based IDS (HIDS) and Network-based IDS (NIDS). HIDS monitors activities within individual hosts, examining system logs, file integrity, and other host-specific information to identify potential security breaches [2]. Conversely, NIDS monitors network traffic, analyzes packets, and identifies suspicious patterns or anomalies that may indicate unauthorized access attempts or malicious activities [3]. The Figure 1 depicts the general NIDS. Furthermore, within IDS techniques, there are two main categories: signature-based and

anomaly-based approaches. Signature-based IDS relies on predefined patterns or signatures representing known attacks, typically obtained through extensive research or by capturing historical attack patterns. However, signature-based IDS may need help to detect novel or unknown attacks that deviate from the established signatures [4]. Anomaly-based IDS, on the other hand, aims to identify deviations from normal behaviour and raise alerts when unusual or suspicious activities occur. This makes it particularly effective in detecting novel or unknown attacks that deviate from the established signatures [5]. In cybersecurity, various attacks severely threaten computer systems' integrity and availability. Distributed Denial of Service (DDoS) attacks, for instance, flood target systems with overwhelming traffic, rendering them inaccessible to legitimate users. Advanced Persistent Threats (APTs) represent sophisticated and persistent attacks, often orchestrated by well-funded adversaries, to breach network defences and gain unauthorized access to sensitive information. Traditional rule-based IDS relies on predetermined patterns and conditions to identify potential security threats in network traffic or system logs, and they have long been used to identify and mitigate security threats. Still, they often struggle to keep pace with the evolving sophistication of cyber-attacks. This is where Machine Learning (ML) and Artificial Intelligence (AI) come into play. ML-based IDS leverages the power of advanced algorithms and computational intelligence to enhance the security posture. Unlike rule-based systems, ML-based IDS can learn and adapt from large datasets, making them well-suited to identify novel and previously unseen threats.

These systems employ various ML techniques, such as supervised, unsupervised, and deep learning (DL), to analyze network traffic, system logs, and other relevant data sources. By doing so, they can autonomously detect unusual patterns, anomalies, or suspicious behaviours that might indicate a cyber intrusion. ML-IDS reduces false positives, enhances threat detection accuracy, and offers real-time responses to security incidents. Furthermore, they allow security teams to stay one step ahead of attackers by continuously learning and adapting to new attack vectors.

Class imbalance is another prevalent challenge encountered in IDS, where the number of instances representing normal behaviour significantly outweighs those depicting attacks. This class imbalance can hinder the detection performance of ML algorithms, as they tend to favour the majority class. To address this issue, the SMOTE has been widely employed to generate synthetic samples, thus balancing the class distribution and improving the effectiveness of classification models [6].

Support Vector Machines (SVM) is a powerful machine learning algorithm known for its effectiveness in classification tasks. SVM works by finding the optimal hyperplane that best separates data points belonging to different classes in a high-dimensional space [7]. It can handle both linear and non-linear classification tasks, making it a versatile choice for intrusion detection. SVM-based IDS models can effectively distinguish between normal network behaviour and anomalous activities, making them a valuable tool in network security.

This paper presents our contributions to addressing the aforementioned challenges as follows:

1. We employed the SMOTE technique to handle class imbalance in two widely used datasets: NSL-KDD 2009 and CIC-IDS 2017. By balancing the class distribution, we aim to improve the performance of the IDS model on these datasets.
2. We introduce SVM as a robust approach for classification in IDS. SVM's ability to find optimal decision boundaries makes it a valuable tool for identifying intrusions in network traffic data.

The subsequent sections of the document are structured in the following manner: Section 2 presents a comprehensive discussion of the relevant and recently published literature. Section 3 explains the proposed model. The results are presented in Section 4. Section 5 serves as the concluding part of the paper.

2 Related Work

NIDS can be viewed as a supervised anomaly detection task and phrased as a classification challenge.

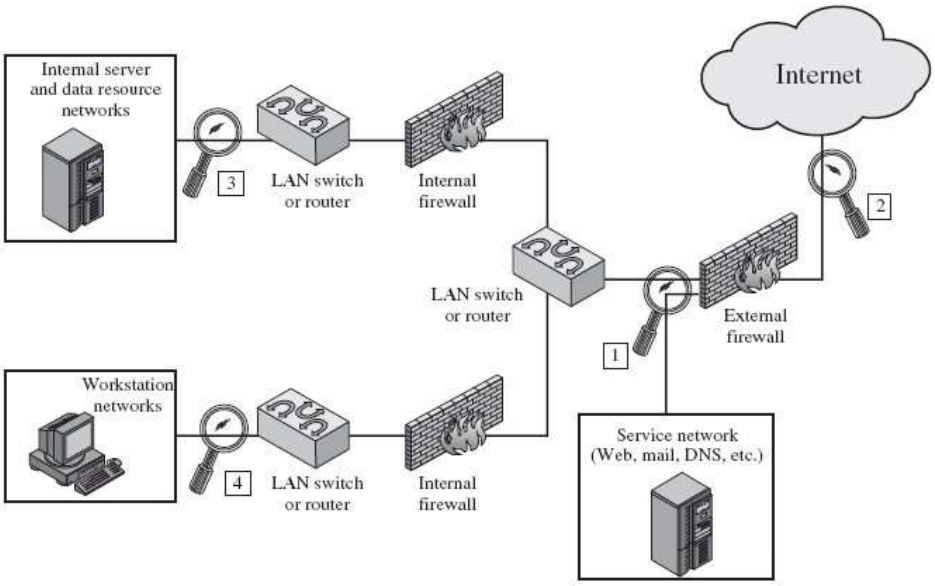

Figure 1. Network-based Intrusion Detection System(NIDS).

Several research papers have been proposed recently to extract characteristic patterns from attack intrusions and effectively distinguish them from typical network traffic. These approaches span a spectrum from conventional statistical techniques to state-of-the-art DL-based methods. Using feature selection is a pre-processing technique in machine learning aimed at reducing computational costs [8]. Its primary objective is to eliminate irrelevant features while simultaneously preserving or improving the performance of IDS. Aslahi-Shahri et al. [9] employed a novel hybrid methodology that integrated a support vector machine with a genetic algorithm. Their primary objective was to enhance the efficiency of feature selection for the KDD CUP 99 [10] dataset. Specifically, this approach aimed to reduce the dimensionality of the dataset, which originally encompassed 41 features, ultimately distilling it into a more manageable set of 10 essential features. Their experiment's results demonstrated their innovative technique's effectiveness by increasing true positive rates while minimizing false positives. In another study by Alazzam et al. [11], they endeavored to refine the feature selection methodology for NIDS by harnessing the power of a Pigeon Inspired Optimizer (PIO). Their technique efficiently reduced feature sets in KDD CUP 99 and NSL-KDD datasets, condensing them from 41 to 7 and 5 features, respectively. Furthermore, this innovative approach showcased notable achievements, including elevated true positive rates and accuracy, while reducing the time required to construct decision trees. Khammassi et al. [12] implemented a methodological framework in their research, where logistic regression was the feature selection technique, supplemented by a trained model to ascertain the most suitable subset for NIDS. This approach yielded commendable results, manifesting as a heightened detection accuracy while utilizing a reduced set of 18 features from the KDD CUP 99 dataset. Nevertheless, it is imperative to acknowledge that the efficacy of feature selection-based methodologies is intricately linked to the selected features' intrinsic quality and semantic relevance. Consequently, these techniques may only ensure optimal outcomes in specific data scenarios, particularly in unbalanced distributions. Moreover, it is worth noting that prevalent feature selection strategies often rely on heuristic principles and evaluation metrics, potentially constraining their ability to effectively learn and capture intricate interrelationships among features in the data. Zhang et al. [13] introduce a novel model

designed to address the challenge of class imbalance in supervised learning. Their approach combines the SMOTE with under-sampling for clustering using a Gaussian Mixture Model (GMM) to tackle the class imbalance problem through resampling effectively. By employing this resampling technique, the model aims to rebalance the dataset and mitigate the impact of class imbalance. Mimura et al. [14] introduced a network interface packet analysis approach that primarily relied on source and destination IP information within Natural Language Processing (NLP). Their investigation revealed that this method could have substantially improved anomaly detection. The authors reported that the best-performing metric for their approach was an F1-score of 98%. In another study, Zhang et al. [15] proposed a Parallel Cross Convolutional Neural Network (PCCNN) to extract feature representations while preventing the neglect of few-sample categories. The PCCNN architecture is intended to overcome the limitations of conventional methods, which often need help handling datasets containing classes with very few samples. By leveraging the PCCNN, the researchers seek to improve the ability to learn meaningful representations from limited data. It is worth noting that these commonly used methods typically rely on labeled data for rebalancing the dataset or learning features. However, in situations where labels are insufficient or scarce, there is a risk of overfitting due to the limited diversity in the training data. J.H Joloudari [16] proposed a model based on Convolutional Neural Network (CNN) and SMOTE to handle imbalanced binary datasets more effectively. Their findings show that the combined CNN-SMOTE achieved 99.08% accuracy on 24 imbalanced datasets. [17] presents a novel IDS framework that combines BERT for contextual feature extraction, SMOTE for handling class imbalance, and an MLP classifier to achieve high accuracy in detecting diverse cyber threats across imbalanced datasets.

While the related work has highlighted various strategies for feature selection, class imbalance mitigation, and using conventional and DL techniques, our paper introduces a novel paradigm for IDS. We harnessed the power of SMOTE to handle the class unbalance and SVM model for classification tasks.

3 Proposed Methodology

The proposed research framework, as depicted in Figure 2, consists of several vital steps to address the classification task. The stages in our proposed model are illustrated in the algorithm 1.

3.1 Dataset Sources

In the pursuit of developing a robust and effective IDS, our research methodology incorporates two prominent publicly available datasets: the CIC-IDS 2017 [18] and the NSL-KDD 2009 [10]. The selection of these datasets is motivated by their widespread recognition within network security and intrusion detection, as well as their diverse characteristics that collectively enable comprehensive model evaluation and validation.

3.1.1 CIC-IDS 2017 Dataset

The dataset is an openly accessible repository that comprises raw pcap network traces collected under controlled conditions over five days. These traces encompass various types of attacks co-occurring with benign network traffic. The dataset incorporates a wide range of cyberattacks, including Denial of Service (DoS), Distributed Denial of Service (DDoS), Infiltration, Brute Force, Port Scan, Web attack, and Botnet activities as illustrated in Table 1. These attacks are examined across a feature set consisting of 79 distinct features.

3.1.2 NSL-KDD 2009 Dataset

Complementing our research arsenal, the dataset is another essential component of our data-driven methodology. This dataset is derived from the original KDD Cup 1999 dataset, which has been refined to address certain limitations and challenges. The NSL-KDD dataset provides a well-balanced representation of network traffic data, classifying it into four primary categories: Normal, DoS (Denial of Service), Probe, U2R (User to Root) and R2L (Root to Local) types. This dataset encompasses a substantial training set with 125,973 records and an additional testing set containing 22,544 records as illustrated in Table 2. Notably, it

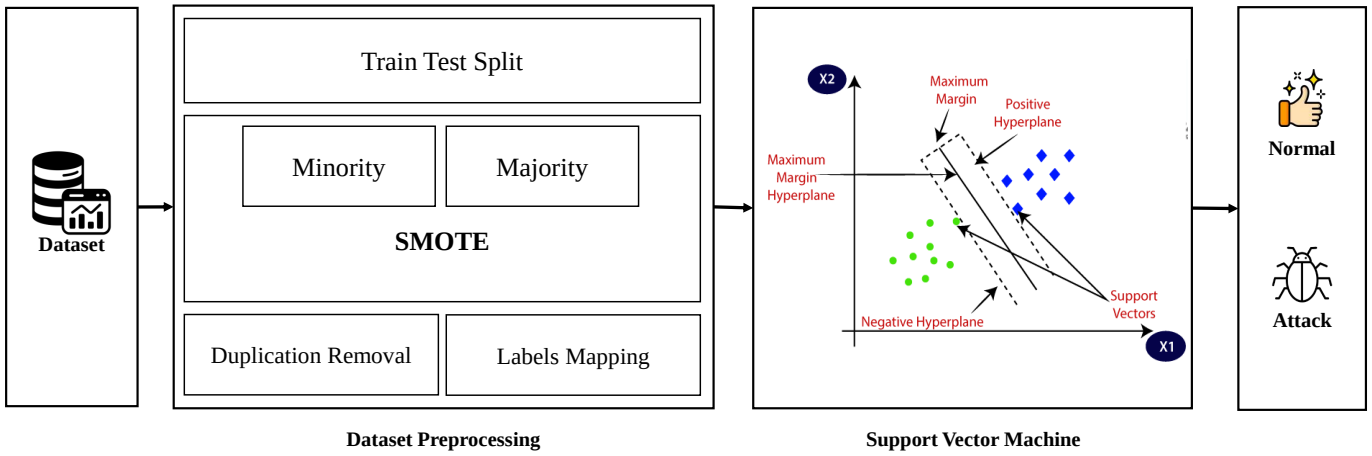

Figure 2. Proposed SVM Model for IDS.

Table 1. CIC-IDS 2017 Dataset

Attacks	Attacks Found	Records
Benign	Benign	2,096,484
Brute Force	FTP-Patator SSH-Patator	9,152
DoS	Golden Eye Hulk Slowhttptest Slowloris Heartbleed Bot	195,712
Web Attacks	Brute Force, SQL Injection, XSS	2,143
Infiltration	Infiltration	36
Port Scan	Port Scan	128,016
DDoS	DDoS	90,819

Table 2. NSL-KDD 2009 Dataset

No	Labels	Training Set	Testing Set
1	Normal	67,343	9,711
2	DoS	45,927	7,460
3	Probe	11,656	2,421
4	R2L	995	2,885
5	U2R	52	67
	Total	125,973	22,544

encompasses 42 distinct features facilitating rigorous evaluation and benchmarking of intrusion detection models.

3.2 Dataset Preprocessing

Duplicate samples are instances in the dataset that are identical in all attributes. Removing duplicates is crucial to avoid biased machine learning models and to ensure that each instance contributes unique information to the analysis. In intrusion detection datasets, labels often represent different types of network traffic or activities, such as normal traffic or various intrusion categories. Label mapping involves transforming these categorical labels into numerical values. Given the potential disparity between the normal and abnormal network data, the SMOTE is employed to rebalance the features, thereby mitigating the impact of class imbalance. This step ensures a balanced representation of different classes, thereby improving the overall effectiveness of the intrusion detection system. It creates synthetic instances in the minority class to balance class distribution, preventing the model from being biased towards the majority class. Additionally, undersampling is applied to reduce the size of the majority class. To evaluate the performance of machine learning models, the preprocessed datasets were divided into training and testing subsets using an 80-20 split. Specifically, 80% of the data was allocated to the training set, while the

remaining 20% was reserved for testing.

3.3 Proposed Model

We leverage the effectiveness of SVM as the core classification approach. SVM, known for its robust classification capabilities, distinguishes between normal network behaviour and anomalous activities, enhancing the system's ability to detect and respond to potential security breaches. By harnessing the power of SVM, we aim to improve the accuracy and reliability of our IDS model in identifying and mitigating network intrusions. This choice aligns to enhance the security posture of computer systems in the face of evolving cyber threats.

Our proposed model represents a holistic approach to data analysis and classification, combining data preprocessing and a powerful classification algorithm to yield accurate and reliable results. By following this comprehensive research framework, our study aims to leverage the strengths of SMOTE for class balance SVM for classification, thereby contributing to improved performance and robustness in network intrusion detection. This underscores our commitment to addressing the research problem effectively and advancing the state-of-the-art in this domain.

3.4 System Configuration

In this study, we conducted our ML experiments on a high-performance computing platform with a carefully configured system as show in Table 3.

Table 3. System Configuration

Component	Specification
GPU	NVIDIA A40
Environment	Jupyter Notebook
Programming Language	Python v3.10.9
RAM	206.09 GB
GPU Memory	45.00 GB
Disk	115.00 GB

4 Results

We formally describe the results obtained from our experimental analysis on two distinct datasets: CIC-IDS 2017 and NSL-KDD 2009. These datasets were subjected to rigorous evaluation and analysis to assess the performance of various machine-learning models and intrusion detection techniques. The outcomes of these evaluations will be elucidated in the subsequent sections.

4.1 Evaluation Metrics

The proposed approach was subjected to a comprehensive evaluation using five distinct metrics:

4.1.1 Accuracy

Accuracy is a widely used metric for classification tasks and represents the overall effectiveness of the classification process. It is the ratio of correctly classified instances (True Positives and True Negatives) to the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

4.1.2 Precision

Precision is a metric that measures the accuracy of positive predictions made by the classification model. It is the True Positives (TPs) ratio to the sum of True Positives and False Positives (FPs). In other words, precision tells us the percentage of instances predicted as positive (abnormal) that were correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

4.1.3 Recall

Recall, also known as sensitivity or true positive rate, measures the ability of the model to identify all positive instances correctly. It is the True Positives (TPs) ratio to the sum of True Positives and False Negatives (FNs). Recall tells us the percentage of actual positive instances (abnormal) correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

4.1.4 F1-Score

The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance between precision and recall.

Algorithm 1. Algorithm of Proposed Model

procedure Main(Dataset)
    Step 1: Acquire the dataset.
    Step 2: Preprocess the dataset:
        a. Remove duplicate samples.
        b. Map labels to numerical values.
        c. Apply SMOTE for minority oversampling and majority undersampling to balance the dataset.
        e. Split the dataset into training and testing sets.
    Step 3: Build and Train SVM Model:
        b. Define the architecture of the SVM.
        c. Training the SVM model on the training set.
    Step 5: Evaluate the Model:
        a. Predict labels on the test data using the trained classifier.
        b. Calculate evaluation metrics (e.g., accuracy, precision, recall, F1-score).
    Step 6: Analyze and Interpret Results:
        a. Examine the model's performance and interpret the findings.
    Step 7: Conclusion and Discussion:
        a. Summarize the results and implications of the model.
end procedure

The F1-Score is especially useful when dealing with imbalanced datasets, where one class is more dominant than the other.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4.1.5 Confusion Matrix

The confusion matrix in Table 4 provides a comprehensive view of the model's performance by showing the counts of True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs). Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class.

Table 4. Confusion Matrix

Actual	Predicted	
	Positive	**Negative**
Positive	True Positives (TP)	False Negatives (FN)
Negative	False Positives (FP)	True Negatives (TN)

4.2 Experimental Analysis

In this section, we provide an in-depth performance evaluation of our model, leveraging the power of SVM for multiclass and binary classification tasks. We conducted these evaluations using two benchmark datasets: CIC-IDS 2017 and NSL-KDD 2009.

4.2.1 Performance evaluation of CIC-IDS 2017 Dataset

The Table 5 and Table 6 represents the results of the binary and multi-classification model applied to a dataset with seven different network traffic classes. In addition, Figure 3 and 4 present the multiclass and binary confusion matrices, respectively. The model performed exceptionally well, achieving high accuracy, with all classes having accuracy rates between 99% and 100%. Precision was consistently above 97%, indicating low false positives, while recall was mainly above 98%, indicating the model's ability to capture actual positive instances. The F1-Scores, which balance precision and recall, were also notably high, ranging from 98.51% to a perfect 100%. This excellent performance suggests that the model accurately classifies

network traffic into the specified categories. Additionally, the multiclass Receiver Operating Characteristic (ROC) with Area Under the Curve (AUC) is presented in Figure 5. It is a metric that quantifies the overall discriminative power of a model in distinguishing between the classes across different thresholds. In the ROC-AUC curve, the y-axis represents the True Positive Rate (TPR) ranging from 0.995 to 1.000, while the x-axis represents the False Positive Rate (FPR) ranging from 0.0 to 1.0. Also, the multiclass precision-recall is presented in Figure 6. In the precision-recall curve, the y-axis represents precision, which is the proportion of true positive samples out of all predicted positive instances in each class, and the x-axis represents recall, which is the proportion of true positive samples (correctly classified instances) out of all actual positive instances, ranging from 0.97 to 1.00. The curve's trajectory shows how precision and recall change as the decision threshold varies.

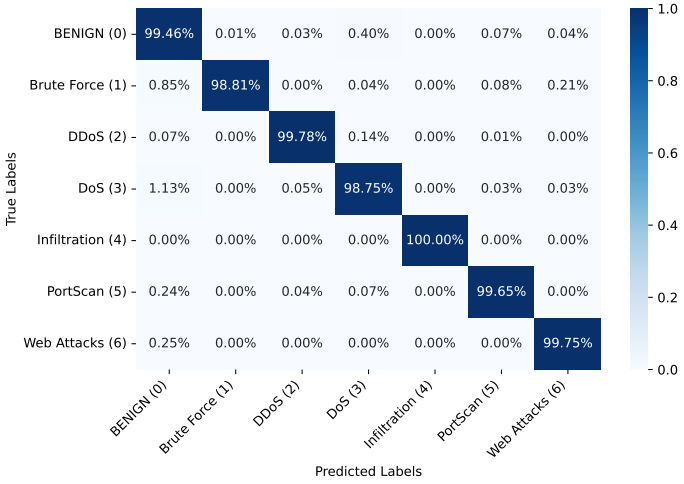

Figure 3. Confusion Matrix of Multiclass Classification on CIC-IDS 2017 Dataset.

4.2.2 Performance evaluation of NSL-KDD 2009 Dataset

Table 7 and Table 8 provides a comprehensive overview of the model's performance in both binary and multiclass classifications using the NSL-KDD 2009 dataset, which encompasses five distinct network traffic classes. Figure 5 and Figure 6 present the

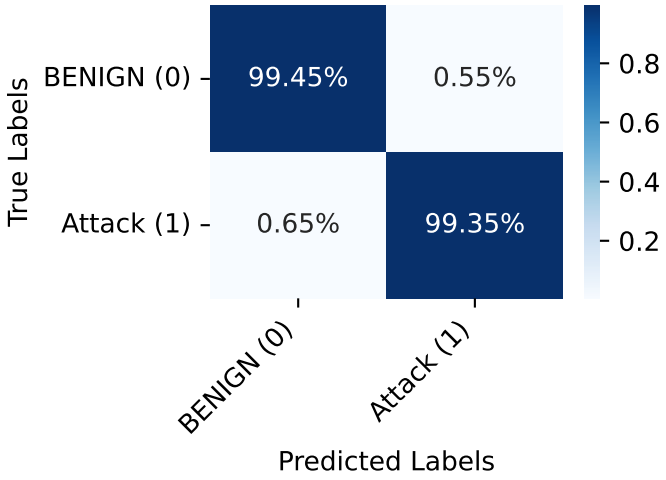

Figure 4. Confusion Matrix of Binary Classification on CIC-IDS 2017 Dataset.

corresponding multiclass and binary confusion matrices, enriching our understanding of the model's performance. Notably, our model demonstrates outstanding accuracy across various attack classes. For DoS attacks, it achieves an impressive accuracy of 99.83%, showcasing its proficiency in correctly identifying such attacks. Furthermore, the high precision of 99% in this category indicates that it is overwhelmingly accurate when our model predicts DoS attacks. Moving on to the Probe class, our model excels with a 98.47% accuracy, and its precision and recall both stand at 99%. This suggests that it consistently identifies Probe attacks with high confidence. In the case of R2L attacks, the model performs admirably, achieving an accuracy of 97.06% and maintaining high precision, recall, and F1 score, all at 97%. These results highlight the model's effectiveness in identifying instances of unauthorized access and classifying them correctly. For the U2R class, although the precision is relatively lower at 89%, indicating some false positives, the recall stands at a respectable 96%, suggesting that it can detect most U2R attacks. The F1-score of 92% in this category shows a reasonable balance between precision and recall, indicating good overall performance. It is worth noting that the high precision and recall values for the other classes demonstrate the model's suitability for detecting and classifying various types

Table 5. Performance Evaluation (Multiclass Classification) on CIC-IDS 2017 Dataset

Labels	Accuracy	Precision	Recall	F1-Score
BENIGN	99.00	99.44	99.46	99.45
Brute Force	99.00	99.44	98.81	99.13
DoS	99.00	98.89	98.75	98.82
Web Attacks	100.00	97.30	99.75	98.51
Infiltration	100.00	100.00	100.00	100.00
Port Scan	100.00	99.58	99.65	99.62
DDoS	100.00	99.81	99.78	99.80

Table 6. Performance Evaluation (Binary Classification) on CIC-IDS 2017 Dataset

Labels	Precision	Recall	F1-Score	Accuracy
BENIGN	99.24	99.45	99.35	99.45
Attack	99.53	99.35	99.44	99.35
Overall Accuracy				99.40
Macro Average	99.39	99.40	99.40	
Weighted Average	99.40	99.40	99.40	

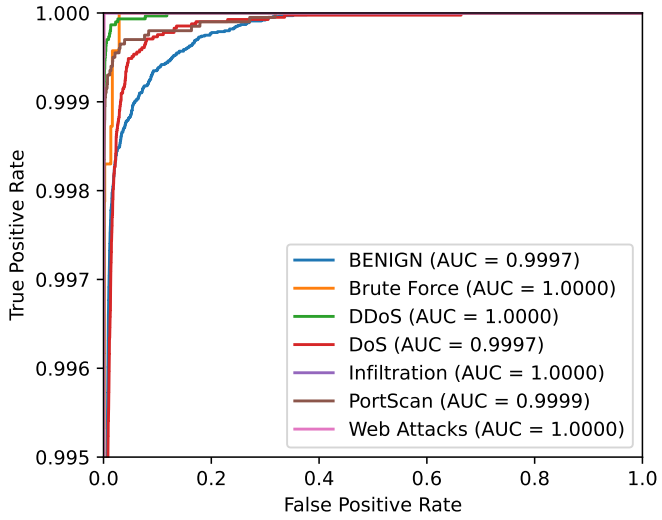

Figure 5. ROC-AUC Curve of CIC-IDS 2017 Dataset.

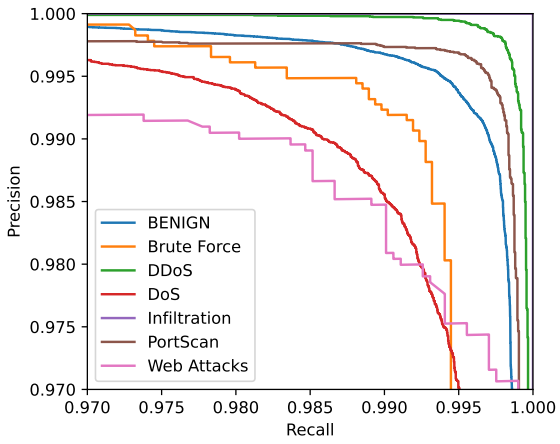

Figure 6. Class-wise Precision-Recall Curves of CIC-IDS 2017 Dataset.

of network attacks with a high degree of accuracy. Additionally, the multiclass ROC-AUC is presented in Figure 9. The y-axis represents the TPR ranging from 0.9970 to 1.00, while the x-axis FPR ranges from 0.0 to 1.0. Also, the multiclass precision-recall is presented in Figure 10. The y-axis represents precision ranges from 0.90 to 1.0, and the x-axis represents recall ranges from 0.50 to 1.00.

5 Conclusion

This paper has introduced a novel approach to enhancing the capabilities of IDS by incorporating SVM in conjunction with meticulous preprocessing for classification tasks. The ever-evolving digital landscape necessitates robust and effective IDS mechanisms to

Table 7. Performance Evaluation (Multiclass Classification) on NSL-KDD 2009 Dataset

Labels	Accuracy	Precision	Recall	F1-Score
BENIGN	99.40	99.00	99.00	99.00
DoS	99.83	99.00	98.00	99.00
Probe	98.47	100.00	99.00	99.00
R2L	97.06	97.00	97.00	97.00
U2R	96.00	89.00	96.00	92.00

Table 8. Performance Evaluation (Binary Classification) on NSL-KDD 2009 Dataset

Labels	Precision	Recall	F1	Accuracy
BENIGN	99.40	99.41	99.41	99.41
Attack	99.36	99.36	99.36	99.36
Accuracy				99.38
Mac. Average	99.38	99.38	99.38	
Weight. Average	99.38	99.38	99.38	

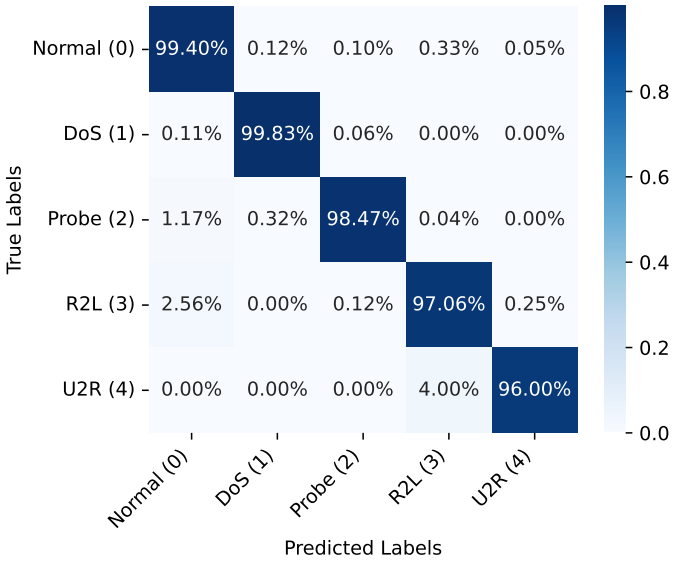

Figure 7. Confusion Matrix of Multiclass Classification on NSL-KDD 2009 Dataset

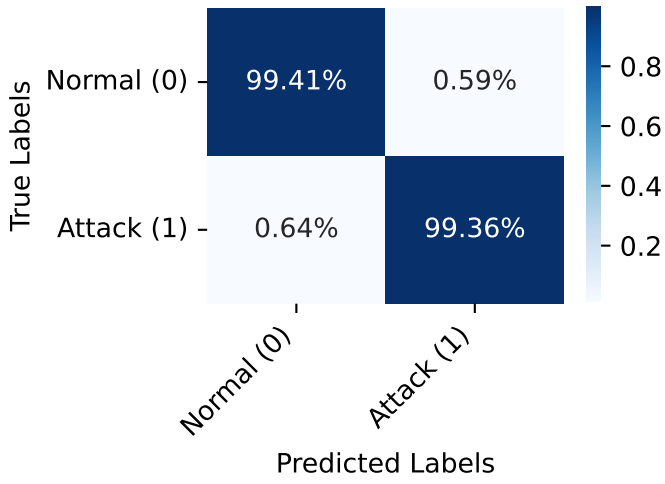

Figure 8. Confusion Matrix of Binary Classification on NSL-KDD 2009 Dataset

safeguard network resources' integrity, confidentiality, and availability. As cyber threats evolve in complexity and scale, IDS serves as the front line of defence, tirelessly monitoring network traffic to pinpoint suspicious activities and mitigate potential security breaches. The methodology presented in this paper has been rigorously evaluated using two prominent and diverse datasets, CIC-IDS 2017 and NSL-KDD 2009, each presenting unique challenges. To address the challenge of class imbalance in these datasets, meticulous preprocessing and the application of the Synthetic Minority Over-sampling Technique (SMOTE) were employed. The experimental results on the CIC-IDS 2017 dataset have demonstrated exceptional performance enhancements, with an impressive accuracy of 99%, delivering flawless precision, recall, and

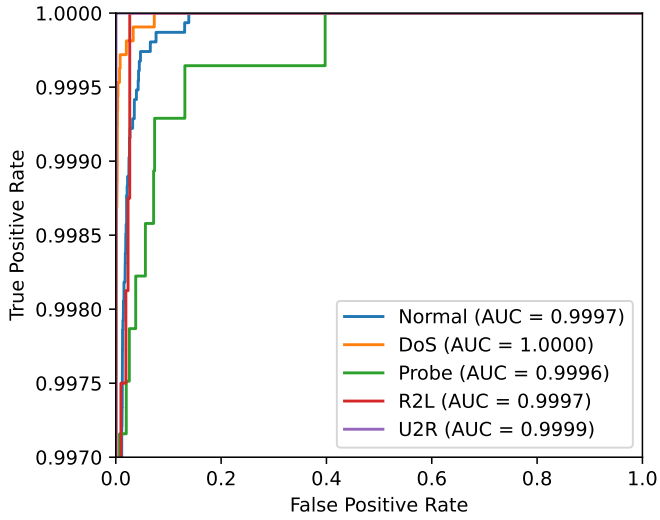

Figure 9. ROC-AUC Curve of NSL-KDD 2009 Dataset.

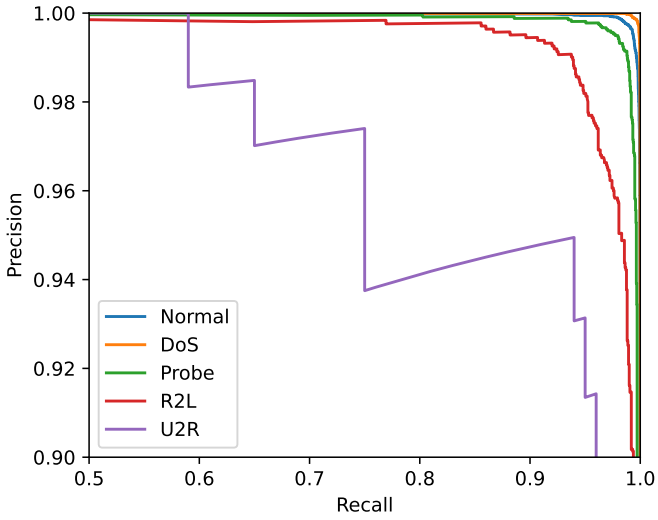

Figure 10. Class-wise Precision-Recall Curves of NSL-KDD 2009 Dataset.

F1-scores for various intrusion types. Similarly, on the NSL-KDD 2009 dataset, our approach has exhibited outstanding results, achieving an accuracy of 99.83%. These findings underscore the efficacy of the proposed SVM-based approach in enhancing intrusion detection systems across diverse datasets. In an era where the imperative for resilient intrusion detection systems remains undeniably clear, the approach presented in this paper represents a significant stride towards fortifying our digital defences. By harnessing state-of-the-art machine learning techniques, we can better adapt to the ever-changing landscape of cyber threats, providing organizations and individuals with a more robust and effective means of safeguarding their network assets. As the digital realm continues to evolve, research and innovation in intrusion detection systems will remain crucial in ensuring the security and stability of our interconnected world.

Author Contributions

Zeeshan Ali: Supervision,Idea, Methodology.**Adnan Akram:** Writing- Original draft preparation.**Naeem Aslam:** Analysis. **Muhammad Saeed Khurram:** Simulation work, Software, Validation.

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] H. Nandanwar and R. Katarya, "Deep learning enabled intrusion detection system for industrial iot environment," *Expert Systems with Applications*, vol. 249, p. 123808, 2024.

[2] Z. Wang and Y. Zhu, "A centralized hids framework for private cloud," in *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 115–120, 2017.

[3] T. Zhang and S. Bao, "A novel deep neural network model for computer network intrusion detection considering connection efficiency of network systems," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 962–965, 2022.

[4] S. Jin, J.-G. Chung, and Y. Xu, "Signature-based intrusion detection system (ids) for in-vehicle can bus network,"

in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2021.

[5] S. Bhadauria and T. Mohanty, "Hybrid intrusion detection system using an unsupervised method for anomaly-based detection," in *2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 1–6, 2021.

[6] V. Surya and M. M. Selvam, "An effective machine learning approach for lot intrusion detection system based on smote," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pp. 905–911, 2022.

[7] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[8] S. Hafeez and N. Kathirisetty, "Effects and Comparison of different Data pre-processing techniques and ML and deep learning models for sentiment analysis: SVM, KNN, PCA with SVM and CNN," in *2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*, pp. 1–6, 2022.

[9] B. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M. J. Golkar, and A. Ebrahimi, "A hybrid method consisting of ga and svm for intrusion detection system," *Neural computing and applications*, vol. 27, pp. 1669–1676, 2016.

[10] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, pp. 1–6, Ieee, 2009.

[11] H. Alazzam, A. Sharieh, and K. E. Sabri, "A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer," *Expert systems with applications*, vol. 148, p. 113249, 2020.

[12] C. Khammassi and S. Krichen, "A ga-lr wrapper approach for feature selection in network intrusion detection," *computers & security*, vol. 70, pp. 255–277, 2017.

[13] H. Zhang, L. Huang, C. Q. Wu, and Z. Li, "An effective convolutional neural network based on smote and gaussian mixture model for intrusion detection in imbalanced dataset," *Computer Networks*, vol. 177, p. 107315, 2020.

[14] M. Mimura and H. Tanaka, "Reading network packets as a natural language for intrusion detection," in *Information Security and Cryptology–ICISC 2017: 20th International Conference, Seoul, South Korea, November 29-December 1, 2017, Revised Selected Papers 20*, pp. 339–350, Springer, 2018.

[15] Y. Zhang, X. Chen, D. Guo, M. Song, Y. Teng, and X. Wang, "Pccn: parallel cross convolutional neural network for abnormal network traffic flows detection in multi-class imbalanced network traffic flows," *IEEE Access*, vol. 7, pp. 119904–119916, 2019.

[16] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective class-imbalance learning based on smote and convolutional neural networks," *Applied Sciences*, vol. 13, no. 6, p. 4006, 2023.

[17] Z. Ali, W. Tiberti, A. Marotta, and D. Cassioli, "Empowering network security: Bert transformer learning approach and mlp for intrusion detection in imbalanced network traffic," *IEEE Access*, vol. 12, pp. 137618–137633, 2024.

[18] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization.," *ICISSp*, vol. 1, pp. 108–116, 2018.