Taylor & Francis
Taylor & Francis Group

Information Security Journal: A Global Perspective

ISSN: 1939-3555 (Print) 1939-3547 (Online) Journal homepage: www.tandfonline.com/journals/uiss20

A hybrid machine learning model with improved feature set for DDoS attack detection under bigdata perspective

Radhika P & S. Kamalakkannan

To cite this article: Radhika P & S. Kamalakkannan (23 Nov 2025): A hybrid machine learning model with improved feature set for DDoS attack detection under bigdata perspective, Information Security Journal: A Global Perspective, DOI: 10.1080/19393555.2025.2589766

To link to this article: https://doi.org/10.1080/19393555.2025.2589766

Published online: 23 Nov 2025.

INFORMATION SECURITY JOURNAL: A GLOBAL PERSPECTIVE
https://doi.org/10.1080/19393555.2025.2589766

A hybrid machine learning model with improved feature set for DDoS attack detection under bigdata perspective

Radhika P and S. Kamalakkannan

Department of Computer Applications, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies, Chennai, India

ABSTRACT
The rapid expansion of internet-connected devices and the surge in digital data generation have significantly increased the risk and complexity of Distributed Denial of Service (DDoS) attacks, posing critical cybersecurity challenges. Traditional detection systems struggle to effectively analyze large-scale, high-dimensional network data while maintaining accuracy and robustness. This research addresses this gap by proposing a novel hybrid machine learning model tailored for DDoS attack detection under the big data paradigm. The primary objective is to enhance detection accuracy, scalability, and robustness against outliers through an improved feature engineering and classification approach. The methodology incorporates a robust normalization process combining Median Absolute Deviation (MAD) and quantile-based Tanh estimation to ensure data consistency and resilience to anomalies. To manage large-scale data efficiently, the system leverages the MapReduce framework for parallel processing, enabling scalable feature extraction that includes improved entropy-based metrics and statistical descriptors. A hybrid classification model is developed by integrating an Improved Support Vector Machine (ISVM) with Neural Networks, utilizing a novel Weighted Exponential Inverse Laplacian kernel to capture complex nonlinear interactions. The proposed ISVM+NN hybrid model achieves the highest detection accuracy of 0.927, significantly outperforming traditional methods such as SVM (0.877), NN (0.858), and others in effectively identifying DDoS attacks.

ARTICLE HISTORY
Received 12 November 2024
Accepted 11 November 2025

KEYWORDS
Big data; DDoS attack detection; machine learning; MapReduce framework; neural network

1. Introduction

The extensive use of internet-connected devices and the rapid increase in digital data have driven substantial progress in worldwide communication and business. Alongside these benefits, the digital revolution has introduced vulnerabilities, prominently DDoS attacks (Afolabi & Aburas, 2022; Ali et al., 2022). These attacks inundate target systems with malicious data, resulting in downtime that leads to financial losses, reputational harm, and operational disruptions for affected organizations. With the evolution of cyber adversaries, the urgency for robust and adaptive detection mechanisms becomes more pronounced. Detecting and mitigating DDoS attacks (Awan et al., 2021; Khempetch & Wuttidittachotti, 2021) is challenging due to the sheer volume and complexity of real-time network data. Traditional methods like rule-based and signature-based detection, effective in certain scenarios, struggle with the dynamic and varied nature of modern DDoS attacks (Elsaeidy et al., 2021) – volumetric attacks overwhelm the network bandwidth, protocol attacks exploit vulnerabilities, and application layer attacks mimic legitimate user behavior, evading conventional detection methods (Yu et al., 2021).

The advent of big data (Azmi et al., 2021; Singhal et al., 2020) technologies has revolutionized cybersecurity analytics, offering unparalleled opportunities to analyze massive datasets in near real-time. In DDoS attack detection, technologies like distributed storage systems and parallel processing frameworks (e.g., MapReduce) play a pivotal role (Gumaste et al., 2020). These tools enable efficient collection, storage, and analysis of large volumes of data, thereby facilitating timely detection and response to evolving threats (Chen et al., 2020; Cheng et al., 2020). By leveraging scalable infrastructure and advanced analytics

CONTACT S. Kamalakkannan kannan.scs@vistas.ac.in Professor, Department of Computer Applications, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies, Chennai, India

techniques, organizations can enhance their ability to proactively detect and mitigate DDoS attacks.

Traditional approaches to DDoS attack detection (Tan et al., 2020) primarily revolve around ML (Seifousadati et al., 2021; Tuan et al., 2020) and DL algorithms. ML algorithms (Shendi et al., 2020), such as decision trees, k-nearest neighbors, are valued for their ability to classify historical data. SVMs, for instance, excel in high-dimensional space but may struggle with large-scale datasets and non-linear relationships (Patil et al., 2020; Sahoo et al., 2020). Existing approaches on DDoS detection have limitations like restricted scalability and worst robustness to noise and outliers, specifically focused on large-scale, heterogeneous network traffic (Adedeji et al., 2023). Many conventional ML models focused more on static and simplistic feature sets which decline its function to represent the modern attack's dynamic behavior. Moreover, traditional normalization approaches like min-max or z-score were subtle to anomalies, which leads to poor performance during pre-processing (Mutholib et al., 2025). Classifiers such as existing SVMs and decision trees frequently face difficulties with random, high-dimensional patterns in big data. These challenges motivate the need for a hybrid model for efficient DDoS attack detection. This research overcomes these drawbacks by proposing a robust normalization approach, scalable feature extraction by MapReduce and a new hybrid classification model (ISVM+NN) by WEIL kernel. This comprehensive system is developed to enhance strength, detection accuracy and applicability to big data environments. This paper introduces a novel approach for DDoS attack detection from a big data perspective, with the major contributions given below:

- Proposing an improved normalization technique that combines Tanh estimation and MAD to improve resilience to outliers by transforming the network traffic information into common plane.
- Proposes an improved entropy-based feature, in which the innovation lies in the computation of weight factors to enhance the discriminative power of entropy features, thereby improving feature relevance and detection accuracy in DDoS attack detection.
- Introduces a Weighted Expo Inverse Laplacian Kernel for SVM, termed here as ISVM, which has an enhanced ability to capture complex and non-linear feature relationships. The ISVM is then combined with Neural Networks (NN) to form a robust hybrid model that significantly improves the accuracy and reliability of DDoS attack detection. This enhancement leads to more accurate identification and classification of diverse DDoS attack patterns.

The remainder of this paper is structured to address the research objectives comprehensively. Section 2 reviews existing literature and methodologies pertinent to DDoS attack detection. Section 3 outlines the methodology, encompassing data acquisition, preprocessing techniques, advanced feature extraction methods, and the hybrid SVM-NN model architecture. Section 4 presents empirical results and discussions, evaluating the model's performance through rigorous experimentation and comparison with benchmark datasets and existing methods. Finally, Section 5 concludes with insights derived from the proposed model.

2. Literature review

Saravanan and Balasubramanian (2024) has introduced UASDAC, a scalable data pipeline designed to detect DDoS traffic in real-time from IoT devices amidst concept drift. The system included an online network stream collector for data aggregation, an analyzer with an unsupervised drift detector, and a repository for subsequent analysis. Utilizing big data technologies, UASDAC adapted to concept drift using an efficient retraining technique. Evaluation with NSL-KDD and IoT23 datasets demonstrated UASDAC's high accuracy in identifying DDoS traffic, achieving 99.7% to 99.9% accuracy.

Alslman et al. (2024) have employed a DDoS attack detection model using Apache Spark with the CIC-DDOS2019 dataset. They employed RF and XGBoost as foundational algorithms, integrating them into a stacked ensemble model to enhance detection accuracy. By leveraging Apache Spark's parallel processing capabilities, the study significantly reduced training time while achieving high accuracy (99.94%). However, this approach necessitated increased RAM usage, prompting recommendations for hardware enhancements to optimize performance. Comparative analyses have shown that the model excels in both accuracy and efficiency for detecting DDoS attacks compared to other methods.

Alhasawi and Alghamdi (2024) has focused on decentralized DDoS attack detection in IoT networks, a Federated Learning-based approach named FL-DAD was introduced. FL-DAD utilized CNN to analyze data locally, ensuring data privacy without requiring central data collection. This decentralized method aimed to address the challenges of traditional centralized detection approaches in diverse IoT environments. Evaluation using the CICIDS2017 dataset demonstrated that FL-DAD surpassed traditional methods, showcasing federated learning's potential to enhance intrusion detection systems in large-scale IoT networks by optimizing security and detection efficiency.

Oyucu et al. (2023) have focused on DDoS detection in SDN-based SCADA systems, and an Ensemble Learning framework was proposed. This framework utilized Decision Tree-based models to accurately differentiate between normal network traffic and DDoS attacks. It aimed to address cybersecurity concerns arising from the integration of SDN technology with traditional SCADA systems handling renewable energy sources. The research involved creating and testing ensemble models using data from a simulated experimental network setup. By optimizing performance over feature selection as well as tuning hyperparameters, the study demonstrated improved accuracy and effectiveness in DDoS attacks detection within SDN-based SCADA structures.

Balasubramaniam et al. (2023) has proposed DDoS attack detection in cloud computing, an optimized approach was developed using the GHLBO algorithm. This algorithm efficiently trained a DSA to detect attacks. The method involved feature fusion using a DMN with an overlap coefficient as well as data augmentation through oversampling. GHLBO was created by integrating gradient descent and HLBO. Evaluation using performance metrics such as TNR (0.909), TPR (0.909) and testing accuracy yielded values (0.917), representative of the effectiveness of the suggested method in DDoS attack detection in cloud environments.

Kumar et al. (2022) has proposed a distributed Intrusion Detection System (IDS) using fog computing to detect DDoS attacks in blockchain-enabled IoT networks. This approach aimed to mitigate challenges such as privacy, safety, and single points of failure inherent in centralized IoT architectures. Integrating blockchain technology addressed these issues while enhancing the reliability of IoT applications. ML methods, specifically RF and optimized gradient tree boosting with XGBoost, underwent training and evaluation on distributed fog nodes utilizing the BoT-IoT dataset. The evaluation output highlighted that XGBoost showed strong performance in detecting binary attacks, whereas RF performed better in scenarios involving multiple types of attacks. Notably, RF demonstrated quicker training as well as testing times on spread fog nodes compared with XGBoost, emphasizing its efficiency for IoT deployments in real-world applications.

Akgun et al. (2022) has focused on cybersecurity, a new intrusion detection model for DDoS attacks based on deep learning was proposed. The model incorporated preprocessing steps and utilized DNN, CNN and LSTM for detection. The evaluation process included testing multiple models on the CIC-DDoS2019 dataset, utilizing methods like feature elimination, random subset selection, and normalization. The CNN-based inception model stood out with exceptional performance, achieving 99.99% accuracy for binary detection and 99.30% for multiclass detection. Additionally, it demonstrated efficient inference times across different test data sizes compared to baseline models, underscoring its effectiveness in real-time applications. Overall, the IDS system coupled with preprocessing methods outperformed existing studies in intrusion detection effectiveness.

Najafimehr et al. (2022) have proposed a hybrid machine learning approach for detecting unprecedented DDoS attacks. This method integrated both supervised and unsupervised algorithms to enhance detection capabilities beyond recognized attack patterns. Initially, a clustering algorithm segregated abnormal traffic from normal data using flow-based characteristics. Subsequently, a classification algorithm utilized statistical measures to label these clusters. The approach was evaluated using the CICIDS2017 dataset for training and testing on the newer CICDDoS2019 dataset. Findings indicated that the Positive Likelihood Ratio (LR+) of the proposed method was about 198% higher than that of traditional machine learning classification algorithms, demonstrating its effectiveness in detecting previously unidentified malicious traffic.

Dasari and Devarakonda (2022) has focused on detecting DDoS attacks using machine learning classification algorithms. It utilized the CIC-DDoS2019 dataset, which encompasses 11 types of DDoS attacks across 87 features each. Algorithms such as Logistic Regression, DT, RF, AdaBoost, GB, KNN, and Naive Bayes were employed to classify these attacks. Assessment metrics were used to evaluate the performance of each classifier. Findings revealed that AdaBoost and Gradient Boost achieved the highest classification accuracy, followed by Logistic Regression, KNN, and Naive Bayes. In contrast, DT and RF exhibited lower classification effectiveness in this context.

Chaudhari et al. (2024) has introduced a DDoS attack detection model utilizing swarm optimization-based feature selection and the Random Forest classifier. By integrating a modified GWO algorithm with SGD for feature selection, the model achieves high accuracy in identifying attacks. Tested on the CICIDS2017 dataset, the approach outperforms existing methods, achieving up to 99.8% accuracy with a reduced feature set. This established the efficiency of swarm optimization techniques in improving the efficiency and accuracy of DDoS detection systems.

Wang et al. (2024) have deployed, a unique network architecture dubbed DDoS-MSCT is proposed, which combines a transformer with a multiscale convolutional neural network. A local feature extraction module (LFEM) and a global feature extraction module (GFEM) make up the DDoS-MSCT block, which is introduced by the DDoS-MSCT architecture. In order to improve the receptive field and simultaneously capture multiscale characteristics, the LFEM uses convolutional kernels of various sizes in conjunction with dilated convolutions. However, in order to address global features, the GFEM is used to capture long-range dependencies.

Ullah et al. (2024) have suggested an intrusion detection system for imbalanced network traffic (IDS-INT) that uses transformer-based transfer learning. Initially, comprehensive details about every kind of attack are obtained via descriptions of network interactions, which comprise host information, reference, attack type, network nodes, etc. Second, using their semantic anchors, the transformer-based transfer learning approach is designed to learn precise feature representation. Third, to detect minority attacks and balance anomalous traffic, the Synthetic Minority Oversampling Technique (SMOTE) is used. Fourth, deep features are extracted from balanced network traffic using the Convolution Neural Network (CNN) model. Lastly, the CNN-Long Short-Term Memory (CNN-LSTM) model hybrid technique is created to identify various attack types from the deep characteristics. A thorough review of current literature on DDoS attack detection has identified various critical challenges and opportunities, as outlined in Table 1.

2.1. Problem statement

Existing methodologies, including conventional techniques such as statistical methods, ML algorithms like RF (Dasari & Devarakonda, 2022; Kumar et al., 2022), and big data frameworks, exhibit strengths in achieving high detection accuracy using advanced ML and big data techniques such as ensemble learning, federated learning (Alhasawi & Alghamdi, 2024), and DNNs (Akgun et al., 2022). Traditional detection systems often struggle to maintain accuracy and scalability when

Table 1. Features and challenges of existing systems.

Author [Citation]	Methodology	Features	Challenges	Dataset used	Performance Metrics
Saravanan and Balasubramanian (2024)	UASDAC	Achieves high accuracy in identifying DDoS attacks under concept drift scenarios using NSL-KDD and IoT23 datasets	Further validation of the effectiveness of UASDAC is necessary across a broader spectrum of diverse and dynamic IoT environments to ensure its robustness.	Benchmark dataset NSL-KDD and the latest IoT dataset IoT23	Accuracy range of 99.7% to 99.9%.
Alslman et al. (2024)	RF, XGBoost	To achieve strong DDoS attack detection performance, the suggested ensemble model combines the RF and XGBoost algorithms, utilizing their complementing advantages.	Requires substantial RAM resources due to Apache Spark's intensive processing demands, potentially limiting scalability on resource-constrained systems.	CIC-DDOS2019 dataset	Accuracy of (99.94%)
Alhasawi and Alghamdi (2024)	CNN	Protecting data privacy is achieved by processing information locally, thereby eliminating the need for centralised data collection.	May face challenges in coordinating and aggregating model updates across distributed IoT devices, potentially impacting detection accuracy and efficiency.	CICIDS2017 dataset	Accuracy rate is 96%
Oyucu et al. (2023)	Decision Tree	Optimises performance through feature selection and hyperparameter tuning	Potential challenges in scaling and adapting the ensemble learning approach to diverse and dynamic SCADA network environments	Primary dataset	Accuracy rate of 95.2%,
Balasubramaniam et al. (2023)	GHLBO	Enhances detection accuracy through DSA training and feature fusion with a DMN.	Potential complexity in implementing and fine-tuning the GHLBO algorithm for optimal performance across diverse cloud computing environments.	NSL-KDD, UNSW-NB15	99.76% accuracy for UNSWNB15, 98.43% accuracy for CICIDS2017
Kumar et al. (2022)	ML, RF and XGBoost	Integrates blockchain technology to enhance security and reliability in IoT applications	May encounter challenges in scaling and managing distributed fog nodes effectively across large-scale IoT networks	BoT- dataset	Accuracy is 96%
Akgun et al. (2022)	LSTM, CNN, DNN	Demonstrates promising inference times across numerous sizes of test data related to baseline models, indicating efficiency in real-time detection scenarios	Potential complexity and computational resources required for training deep learning models like CNN and LSTM on large-scale datasets	CIC-DDoS2019 dataset.	Accuracy is 98%
Najafimehr et al. (2022)	ML	Achieves significantly higher Positive Likelihood Ratio (LR+) compared to traditional ML classification algorithms	May require complex parameter tuning and computational resources due to the hybrid nature of the approach	CICIDS2017 dataset	Accuracy is 96%
Dasari and Devarakonda (2022)	Logistic Regression, DT, RF, AdaBoost, GB, KNN, and Naive Bayes	Provides a thorough evaluation of multiple ML algorithms for DDoS attack detection, offering insights into performance across diverse attack scenarios.	Potential limitations include dataset representativeness and generalizability to real-world DDoS attack scenarios.	CICDDoS2019 dataset	Accuracy is 98%
Chaudhari et al. (2024)	GWO	Integrates SGD into the feature selection process, improving the model's ability to identify critical features for attack detection, thereby improving overall performance.	May require additional computational resources and expertise for implementing and fine-tuning the swarm optimisation algorithms effectively.	IoT DDoS dataset	Accuracy is 94%
Wang et al. (2024)	DDoS-MSCT	Achieving exceptionally high detection accuracy	Increased computational overhead and reduced efficiency in real-time	CIC-DDoS2019 dataset,	Accuracy is 99%
Ullah et al. (2024)	CNN-LSTM	Improves detection accuracy using transformer-based transfer learning and SMOTE-enhanced CNN-LSTM models.	Due to complex attributes and data imbalance issues.	UNSW NB15, CIC-IDS2017, and NSL-KDD.	99.21% Accuracy

processing high-dimensional data and are particularly vulnerable to performance degradation due to noise, outliers, and evolving attack patterns. Moreover, many existing approaches cannot efficiently process and analyze massive volumes of data generated in real-world networks, limiting their effectiveness in big data contexts. Therefore, there is a critical need for a robust, scalable, and intelligent DDoS detection framework that can handle large-scale datasets, effectively model nonlinear relationships, and maintain high detection accuracy while minimizing false alarms.

3. Proposed DDoS attack detection model under big data perspective

Proposing a DDoS attack detection model under big data involves designing a system capable of identifying this malicious attack in large-scale, high-volume data environments. The proposed work is outlined as follows:

- The initial step involves acquiring the necessary data for developing and testing a DDoS attack detection model.
- Subsequently, to ensure the quality and consistency of the acquired data, improved normalization techniques are implemented as a part of the preprocessing phase.
- For big data handling, the MapReduce framework is used, where the feature extraction process is performed. Features extracted include raw attributes, advanced entropy-based measures tailored for big data environments, and statistical descriptors aimed at capturing various aspects of network traffic dynamics.
- Finally, a hybrid ML model is designed for DDoS attack detection, which integrates the strengths of ISVM algorithms and NN. SVMs are chosen for their ability to classify complex data patterns and anomalies effectively, while NNs complement this by capturing nonlinear relationships and dependencies within the extracted features. By combining these approaches, the goal is to enhance the model's detection accuracy, sensitivity, which proves the robustness against evolving DDoS tactics. Figure 1 illustrates the entire process of detecting DDoS attacks.

3.1. Data acquisition

In this paper, data acquisition is the initial. According to this work, benchmark datasets *D* is

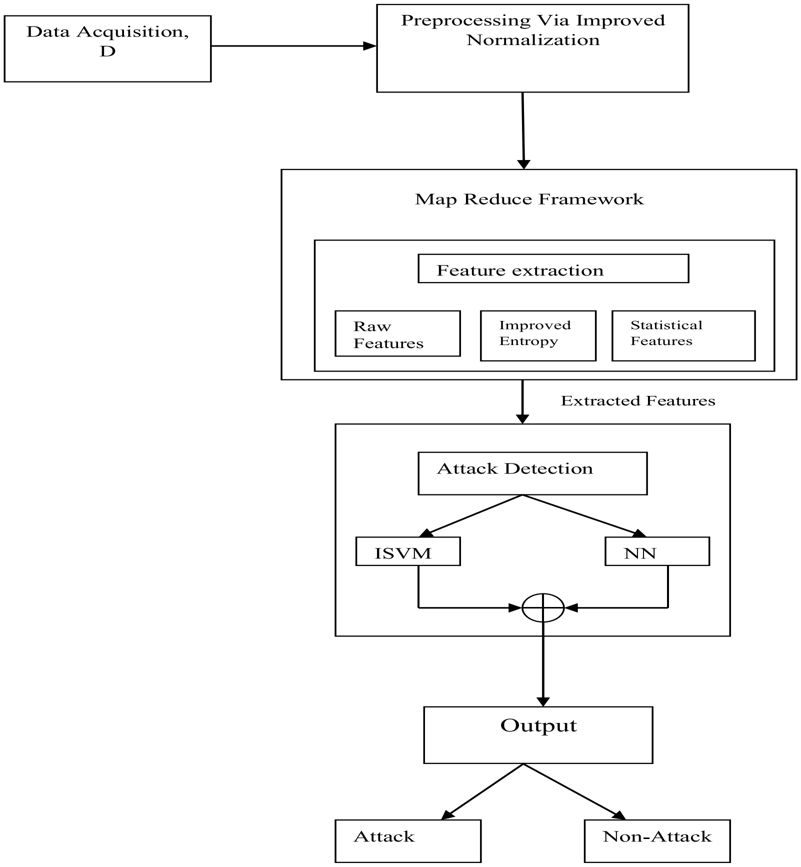


Figure 1. Overall process of this proposed model.

considered. This dataset is pivotal for training and validating the efficacy of the detection model.

3.2. Preprocessing via improved normalization

The preprocessing phase is essential for preparing the data effectively for DDoS attack detection. It includes various steps focused on enhancing the data quality before inputting it into the detection model. This phase ensures that the input data, D is cleaned, normalized, and formatted to facilitate accurate detection of anomalous patterns indicative of DDoS attacks. This work proposes an improved normalization method for preprocessing the input data.

The conventional data normalization process often employs Min-Max normalization (Khalifa et al., 2013), a simple yet widely used method to scale data within a fixed range, typically [0, 1]. Eq. (1) defines the Min-Max normalization formula. where, D_k signifies the raw data, $Max(D)$ and $Min(D)$ denotes the maximum and minimum values of the data, D, respectively.

$$D_k^N = \frac{D_k - Min(D)}{Max(D) - Min(D)} \tag{1}$$

Min-Max normalization is straightforward but sensitive to outliers, which can skew normalized values. It also lacks robustness in handling varied data distributions, potentially reducing its effectiveness in accurately representing complex datasets. Therefore, while easy to implement, its limitations should be considered when choosing normalization methods for data analysis. To overcome the drawbacks of Min-Max normalization, an advanced approach leveraging MAD (Khalifa et al., 2013) and adaptive scaling techniques is proposed.

$$MAD = median(|D_k - median|) \tag{2}$$

The improved normalization (Siddiqi & Pak, 2021) method incorporates MAD and a modified tanh function for enhanced robustness of the model by using Eq. (3). Further normalization to a [0, 1] range is achieved by adjusting $\left(D_k^N\right)'$ based on quartiles as per Eq. (4). where $Q_1\left(D_k^{N'}\right)$ and $Q_3\left(D_k^{N'}\right)$ denotes the 25th and 75th quartiles of $D_k^{N'}$.

$$\left(D_k^N\right)' = \left[\left[0.5\left\{\tanh\left(0.01\left(\frac{D_k - \mu(D_k)}{\sigma(D_k)}\right)\right) + 1\right\}\right] + \left[\frac{D_k - Median}{MAD}\right]\right] \tag{3}$$

$$D_{K_{new}}^N = \frac{\left(D_k^N\right)' - Q_1\left(D_k^{N'}\right)}{Q_3\left(D_k^{N'}\right) - Q_1\left(D_k^{N'}\right)} \tag{4}$$

By effectively scaling and adjusting data, these techniques optimize the detection of anomalies, ensuring that both normal and anomalous patterns are distinguished more accurately. Finally, the preprocessed data is signified as D^P.

3.3. Feature extraction under big data perspective

In this work, Map Reduce framework is used to handle the bigdata, which includes the process of feature extraction from the preprocessed data, D^P under the mapper and reducer phases.

3.3.1. MapReduce framework

The MapReduce paradigm is employed to manage the voluminous preprocessed data effectively. This framework divides the preprocessed dataset into discrete mapper phases. Each mapper phase undertakes the extraction of essential features crucial for DDoS attack detection. These features encompass a spectrum including raw features, improved entropy-based features, and statistical features. The scalable and parallel processing of large-scale network data, making the system suitable for real-time DDoS detection under big data constraints. Figure 2 depicts the MapReduce framework model.

3.3.1.1. Mapper phase. In the initial stage of the MapReduce workflow, the preprocessed data D^P is distributed across multiple mapper tasks. Each mapper is assigned a subset of the data, tasked with extracting specific types of features like raw features, improved entropy-based features and statistical features, which are explained as follows:

3.3.1.2. Raw features F_R. These features capture fundamental metrics derived directly from input data, D

3.3.1.3. Improved entropy-based features F_{IE}. Improved entropy calculations are applied to capture the information content and patterns within

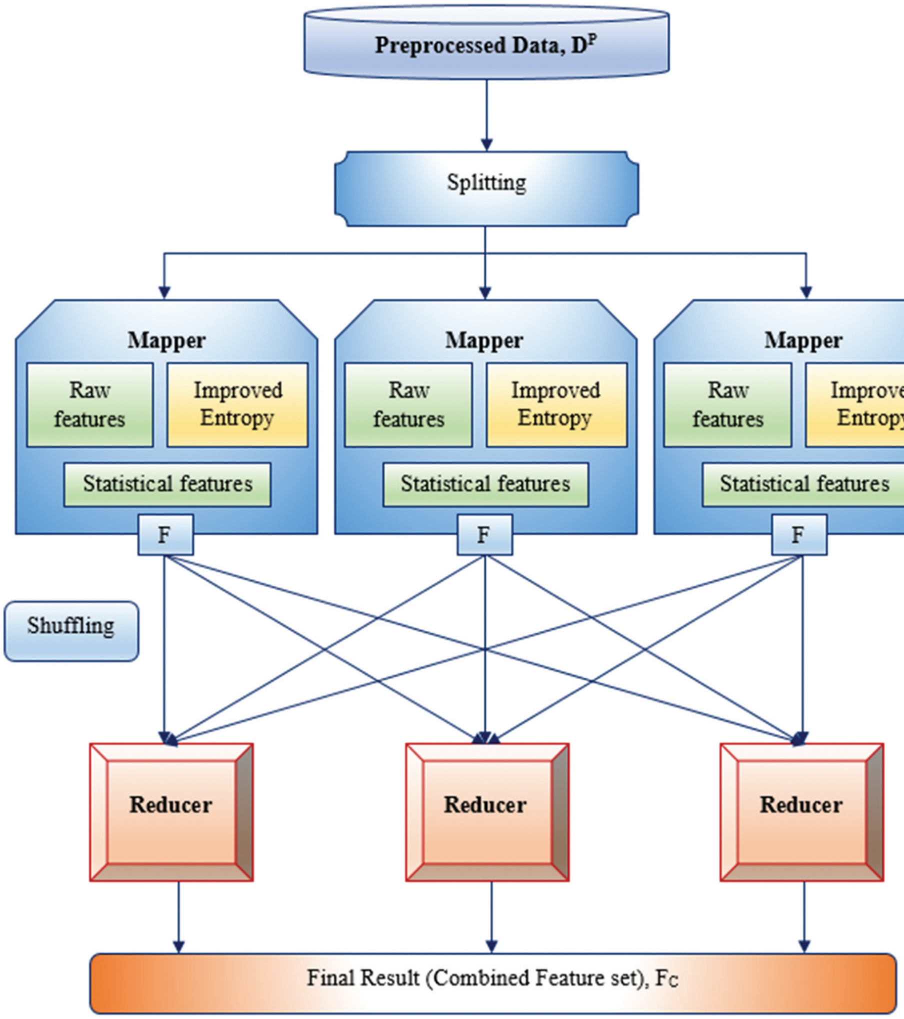


Figure 2. Representation of MapReduce framework.

the data subsets more accurately. This advanced approach mitigates the limitations of conventional entropy methods, ensuring robust feature representation suitable for anomaly detection in complex datasets.

The conventional approach to entropy calculation, represented by Deng entropy (Cui et al., 2019), is given by Eq. (5), where m denotes the mass function defined over the frame of discernment $X$and A represents a focal element of m.

$$E(m) = -\sum_{A \subseteq X} m(A) \log_2 \frac{m(A)}{2^{|A|} - 1} \tag{5}$$

Conventional entropy methods, while useful for quantifying information uncertainty, face significant limitations in anomaly detection. They typically require discretization or binary representation of continuous variables, which can introduce inaccuracies and bias in entropy estimation. The discrete nature of entropy calculations may overlook subtle patterns or fluctuations in data, impacting the detection model's sensitivity to anomalous behaviors. Therefore, while entropy provides insights into data complexity, its application in anomaly detection contexts must consider these constraints to ensure accurate and effective detection of anomalies in diverse datasets.

To overcome the drawbacks of conventional entropy methods, an improved entropy-based feature (Ouyang et al., 2013) is proposed as per Eq. (6). where $\bar{H}$ signifies the entropy value adjusted using weights w_1 and w_2 to amplify the discriminative power of entropy-based features. w_1 and w_2 are the weighting factors computed based on the entropy value. By using Eq. (7) and Eq. (8) are used to enhance the feature relevance and discriminative capability. The enhanced entropy $H_j$is calculated as per Eq. (9). The parameter b and e value is taken as 2 and 10, respectively $p(x_i)$ denote the probability that a state x_i occurs, m and denotes the mass function defined on the frame of discernment.

$$Improved\ Entropy = \begin{cases} (1-\bar{H})w_1 + \bar{H}\cdot w_2, & H_j < 1 \\ 0, & H_j = 1 \end{cases} \tag{6}$$

$$w_1 = \left[\frac{1-H_j}{\sum_{j=1}^{n}\left(1-H_j\right)}\right] \tag{7}$$

$$w_2 = \left[\frac{\left(1+\bar{H}-H_j\right) * \ln(1+e^{H_j-1})}{\left(\frac{2}{\pi}\right)^{0.5}\cdot e^{H_j-1}}\right] \tag{8}$$

$$H_j = \left[-\sum_{A\subseteq\theta} m(A)\log_2\left[\frac{m(A)+p(x_i)\log_b p(x_i)}{2(2^{|A|}-1)}\right]\cdot e^{\frac{|A|-1}{|S|}}\right] \tag{9}$$

The improved entropy method offers several advantages over traditional approaches. By enhancing the discriminative power through a more comprehensive consideration of subset relationships, it effectively filters out the irrelevant fluctuations and disturbances while emphasizing significant patterns in the data. Moreover, it builds upon the foundational benefits of Deng entropy while addressing limitations associated with continuous variables, thereby improving the accuracy and reliability of feature extraction for anomaly detection tasks. This advancement ensures a more robust approach to identifying anomalies in complex datasets.

3.3.1.4. Statistical features, F_S.

Additionally, statistical features (Toptaş & Hanbay, 2021) such as mean, median, and standard deviation are computed within the mapper phase. These metrics provide insights into the distribution and variability of data points, thereby enriching the feature set utilized in DDoS attack detection.

3.3.1.5. Mean.

The mean, or average, is a measure of central tendency that provides insight into the typical value of a dataset. Within each mapper phase, the mean is computed by adding all data points x_i and dividing by the total number n of data points using Eq. (10). This calculation is performed independently by each mapper on its subset of the preprocessed data. The mean helps in understanding the overall trend or baseline behavior of the network traffic characteristics.

$$Mean = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

3.3.1.6. Median.

The median represents the middle value of a dataset when sorted in either ascending or descending order. In the mapper phase, each mapper sorts its portion of data and determines the median based on its subset's size. For an odd number of data points n, the median M is the middle element. For an even number of data points n, the median M indicates the average of the two middle elements. Computing the median in this distributed manner ensures robustness against extreme values and enhances the model's resilience to anomalies typical of DDoS attacks.

3.3.1.7. Standard Deviation.

It quantifies the dispersion of data points from the mean. It quantifies the amount of variation or spread within a dataset. In the mapper phase, each mapper calculates the standard deviation for its subset of data points using Eq. (11).

$$Standard\ Deviation = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - Mean\right)^2} \tag{11}$$

Here x_i are the individual data points, *Mean* is the average is calculated earlier, and n indicates the total number of data points. This metric provides critical insights into the data's volatility and can help detect unusual patterns that might signify DDoS attacks.

Therefore, the total features extracted in the mapper phase are represented as $F = [F_R\ F_{IE}\ F_S]$.

3.3.1.8. Shuffling to reducer phase.

Upon completion of the mapper tasks, intermediate key-value pairs are shuffled and transferred to the reducer phase based on key similarity. This data shuffling process consolidates related information across all mappers, preparing it for further aggregation and analysis in subsequent stages.

3.3.1.9. Reducer phase. In the Reduce phase, the framework aggregates and consolidates the intermediate outputs from all mappers. This phase involves: The Reducer tasks merge and organize the extracted features from different mappers. Features, F across various categories (raw, entropy-based, statistical) are combined to create a unified set of comprehensive features, F_C.

3.3.1.10. Final output. The final output of the Reduce phase is a cohesive collection of features derived from the preprocessed data. These features F_C are now prepared for subsequent stages of analysis and detection, such as classification using the hybrid SVM and NN model.

The application of the MapReduce framework in the feature extraction phase for DDoS attack detection ensures efficient handling and processing of vast amounts of network traffic data. By leveraging distributed computing capabilities and parallel processing, MapReduce enables comprehensive feature extraction necessary for accurate detection of anomalous network behavior indicative of DDoS attacks. This approach not only enhances detection accuracy but also supports the scalability and real-time responsiveness required in modern network security applications.

3.3.2. The procedure of MapReduce framework with an example

Considering the representation of input data as "case_001," which includes feature vectors in a time series sequence, each of which represents the features at a particular time step. Here, the MapReduce framework is demonstrated with the example that follows.

Step 1: Consider the input with three-time steps as case_001: [0.998, 0.997], [0.625, 0.117] [0.370, 0.997].

Step 2: Apply the mapper function, then the obtained output is (case_001, [0.998, 0.997]); (case_001, [0.625, 0.117]); (case_001, [0.370, 0.997]).

Step 3: After the generation of key-value pairs, the shuffle and sort function take place. Then, the obtained output is ["case_001"→ [0.998, 0.997], [0.625, 0.117], [0.370, 0.997].

Step 4: Then the reducer phase receives a list of feature vectors for the key "case_001" and consolidates the data to produce the final result. Here, the reducer phase receives input as [0.998, 0.997], [0.625, 0.117], [0.370, 0.997]. Then, the resultant output of the reducer phase is ("case_001," [0.998, 0.997], [0.625, 0.117], [0.370, 0.997]).

Finally, this kind of feature sequence is given as input to the ISVM and NN classifiers. The output of the model is a multi-class label (0, 1, or 2), where in Dataset 1, labels represent benign, FTP brute-force, and SSH brute-force attacks, respectively, and in Dataset 2, they indicate benign, LOIC-UDP, and HOIC DDoS attacks. This integrated approach ensures accurate, scalable, and robust detection of diverse DDoS threats in real-time network scenarios. Figure 3 illustrates the workflow of the MapReduce framework using mathematical examples.

3.4. Attack detection phase

The proposed hybrid detection model integrates the ISVM with an NN. The model architecture involves each classifier independently processing input data F_C and generating classification scores. These scores are averaged to form the final decision on the presence of a DDoS attack, aiming to mitigate bias and variance and thereby improve the system's robustness.

The hybrid model categorizes instances into three classes based on the specific dataset used for training and validation. For example, in Dataset 1, Class 0 signifies benign (normal) network traffic, Class 1 indicates FTP Brute Force attacks, and Class 2 denotes SSH Brute Force attacks. In Dataset 2, Class 0 represents benign traffic, Class 1 signifies DDOS attack LOIC-UDP, and Class 2 denotes DDOS attack HOIC. These class labels provide critical insights for network administrators and security analysts to understand detected activities and respond effectively to potential threats. This approach enables tailored responses based on the nature and severity of detected network anomalies.

3.4.1. Improved SVM classifier

The input data to the improved SVM is the extracted features F_C. SVMs (Jiang et al., 2007)

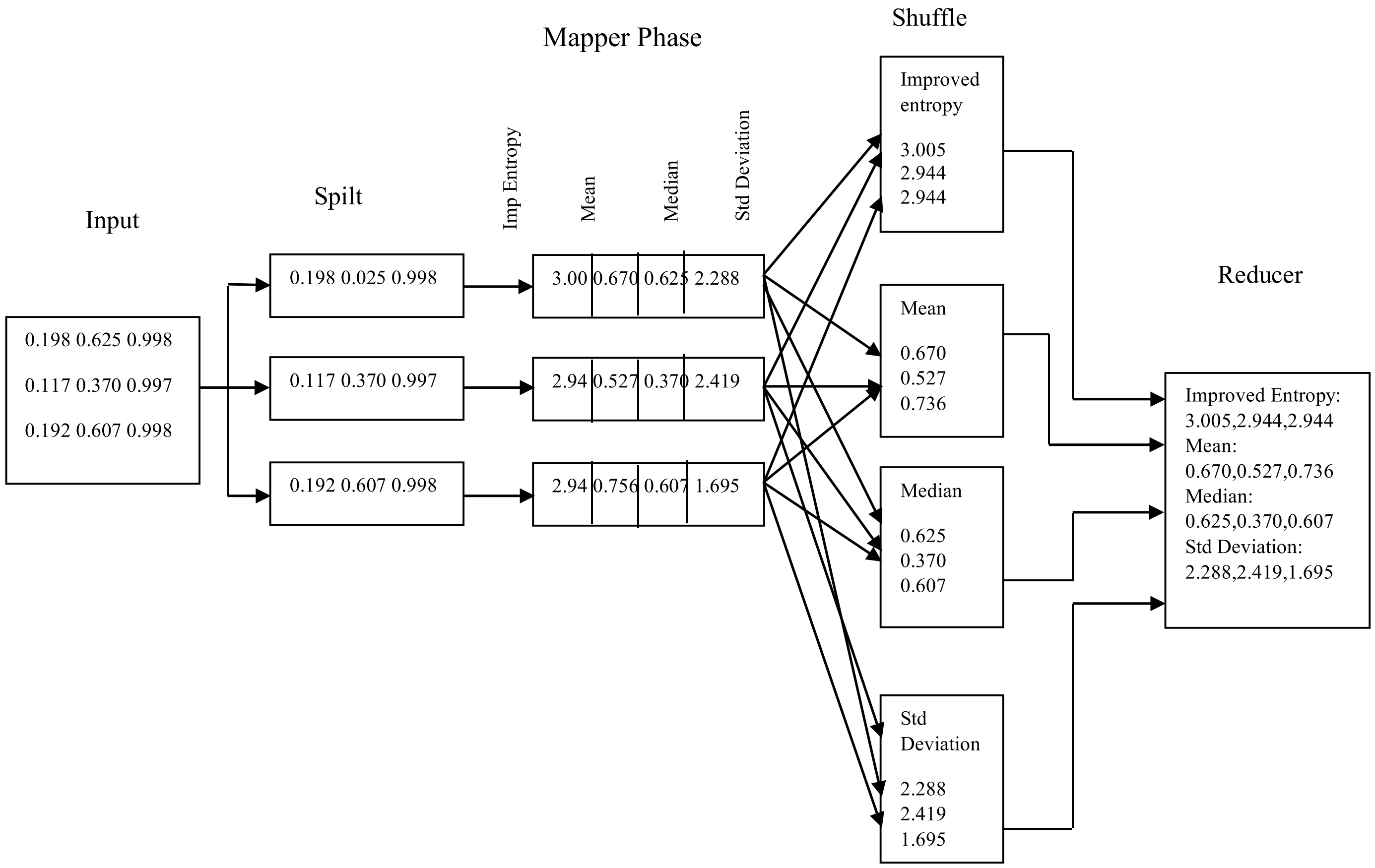


Figure 3. Workflow of MapReduce framework.

are versatile in employing kernels. They operate in a higher-dimensional feature space, aiming to minimize structural risk by identifying a hyperplane with the largest margin between classes. This hyperplane is crucial for effectively separating different classes of data, particularly in scenarios where the feature space is non-linearly separable. The choice of kernel function significantly influences the SVM's classification performance. It enhances classification accuracy and generalization by incorporating a custom kernel (WEIL) that captures complex, non-linear relationships in network traffic data. The decision function for the conventional SVM is expressed in Eq. (12).

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i K(X_i, F_C) + b \tag{12}$$

Where α_i denotes the coefficients determined during the training phase, y_i represents the class labels corresponding to the support vectors X_i, $K(X_i,\ F_C)$ denotes the kernel function that computes the similarity between the support vectors X_i and the input feature F_C, $b$and represents the bias term.

However, despite its strengths, the conventional SVM has several drawbacks. Such as the traditional SVM kernel $K(X_i,\ F_C) = (\Phi(X_i) \cdot \Phi(F_C))$ is symmetric and satisfies Mercer's condition, and treats all input features equally. This symmetric treatment may ignore sequential or contextual information embedded in the data, which is crucial for certain applications. Also, it may struggle to capture complex relationships between variables in high-dimensional spaces. This limitation can affect its ability to generalize well to unseen data and could potentially reduce classification accuracy in intricate data distributions.

To address these limitations, an Improved SVM classifier is proposed in this work. The Improved SVM incorporates a novel kernel function known as the Weighted Expo Inverse Laplacian kernel, designed to enhance the SVM's performance in capturing non-linear relationships between features and target variables. The Weighted Expo Inverse Laplacian kernel better focuses on the most informative portions of the data by taking into consideration the variable significance of each feature, in contrast to typical kernels that treat all features equally. Because it can improve the precision and stability of machine learning models in big data settings, it is suitable for large-scale, real-time cybersecurity applications.

The proposed kernel function (Gaye et al., 2021) $K(X_i, F_C)_{new}$ is defined as per Eq. (13) to Eq. (15), where σ represents a parameter controlling the smoothness of the kernel, β and represents a parameter that adjusts the contribution of each feature, which is expressed in Eq. (16).

$$K(X_i, F_C)_{new} = e^{\left(\frac{-\|X_i - F_C\|}{2\sigma^2}\right)} + \frac{1}{e^{\left(\frac{-\|X_i - F_C\|}{\sigma}\right)}} \tag{13}$$

$$K(X_i, F_C)_{new} = \frac{e^{\left(\frac{-\|X_i - F_C\|}{2\sigma^2}\right)} * e^{\left(\frac{-\|X_i - F_C\|}{\sigma}\right)} + 1}{e^{\left(\frac{-\|X_i - F_C\|}{\sigma}\right)}} \tag{14}$$

$$K(X_i, F_C)_{new} = \left(\frac{e^{\frac{\|X_i - F_C\|(-1-2\sigma)}{2\sigma^2}} + 1}{e^{\left(\frac{-\|X_i - F_C\|}{\sigma}\right)}}\right) * \beta \tag{15}$$

Where, β represents a feature-wise weighting parameter (i.e., each input feature gets its own adaptive weight). Its main role is to modulate the kernel's sensitivity to variations across different features. It allows some dimensions to contribute more to similarity computation based on statistical importance.

$$\beta = \left[\frac{(\exp(X_i))}{\sum_j \exp(X_j)} + \frac{X_i}{|X_i| + 1}\right] \tag{16}$$

The performance the kernel functions such as Laplacian, Exponential, RBF, and the suggested Weighted Exponential Inverse Laplacian kernel functions under varying the input Xin the range $X \rightarrow -10, 10$, is visually represented in Figure 4. By providing smoother, more dependable similarity scores for distant data points and effectively balancing decay, the proposed Weighted Exponential Inverse Laplacian kernel function outperforms other kernel functions in managing outliers and detecting long-range feature dependencies in complex data settings. It also helps to ensure that the detection process stays effective even with the large amount of input data.

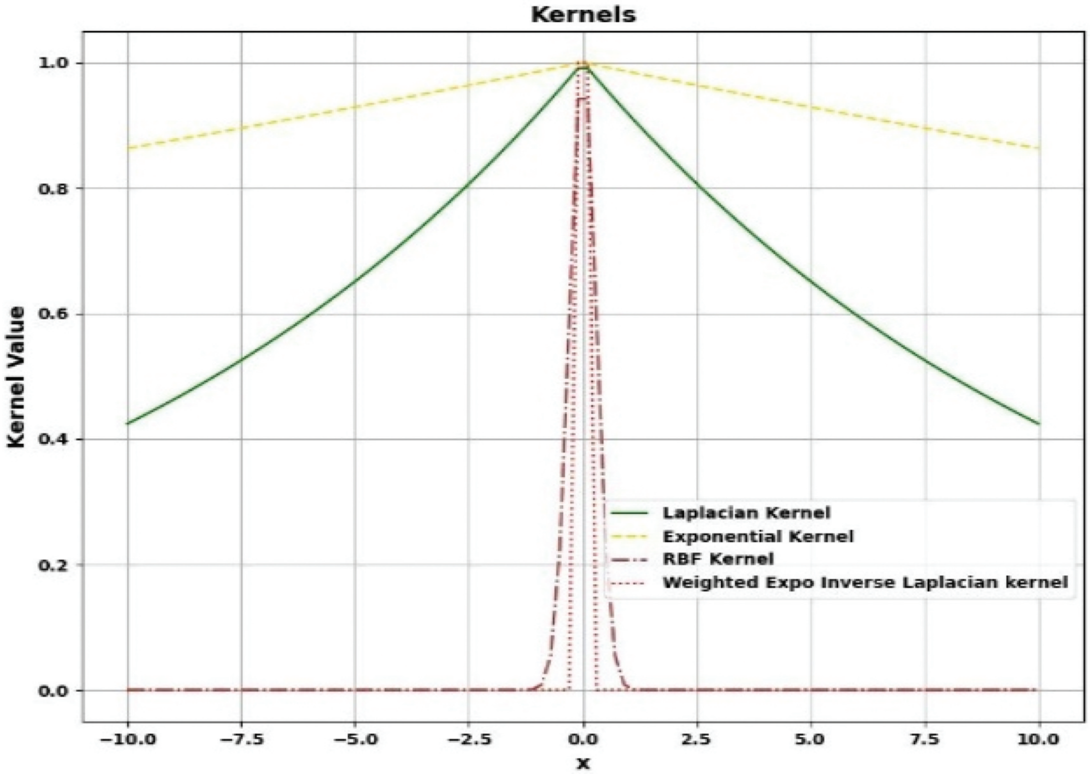


Figure 4. Comparison of kernel functions.

The Improved SVM, leveraging the Weighted Expo Inverse Laplacian kernel, brings several advantages over conventional SVM methods. This kernel enhances the SVM's ability to model intricate non-linear relationships between features and target variables more effectively. This capability is crucial for improving detection accuracy, particularly in scenarios involving complex data patterns. Moreover, the improved SVM kernel contributes to enhanced generalization and accuracy by effectively fitting the training data and mapping data points into a higher-dimensional space where they can be more distinctly separated. This improvement in generalization ensures robust performance when applied to real-world datasets, providing reliable detection and classification of complex patterns with higher accuracy. These advantages make the improved SVM a powerful tool for tackling challenging classification tasks in diverse applications. Here, the output score obtained in this ISVM model is represented as Y_{SVM}. Figure 5 depicts the improved SVM classifier model (Moradibaad & Mashhoud, 2018). The hyperparameters of classifiers are tabulated in Table 2.

3.4.2. NN

In the proposed attack detection system, an NN (Jordanov et al., 2018) classifier plays a pivotal role as part of a hybrid machine learning model alongside an Improved Support Vector Machine (SVM). Unlike traditional SVMs, which excel in finding optimal hyperplanes in high-dimensional feature spaces, NNs are designed to handle complex, non-linear relationships within the data.

A NN classifier is conceptualized as a parallel computing system comprising numerous interconnected simple processors, often referred to as neurons. This architecture allows NNs to process and learn from large datasets, making them suitable for tasks that involve intricate pattern recognition and classification, such as attack detection in cybersecurity. It provides deep learning capabilities that adaptively refine classification boundaries based on feature patterns learned from ISVM outputs.

The multilayered perceptron is a prominent type of NN used in the proposed system. It consists of multiple layers of neurons interconnected through weighted connections, which is shown in Figure 6. Each neuron in one layer is connected to neurons in the subsequent layer, allowing for the propagation of

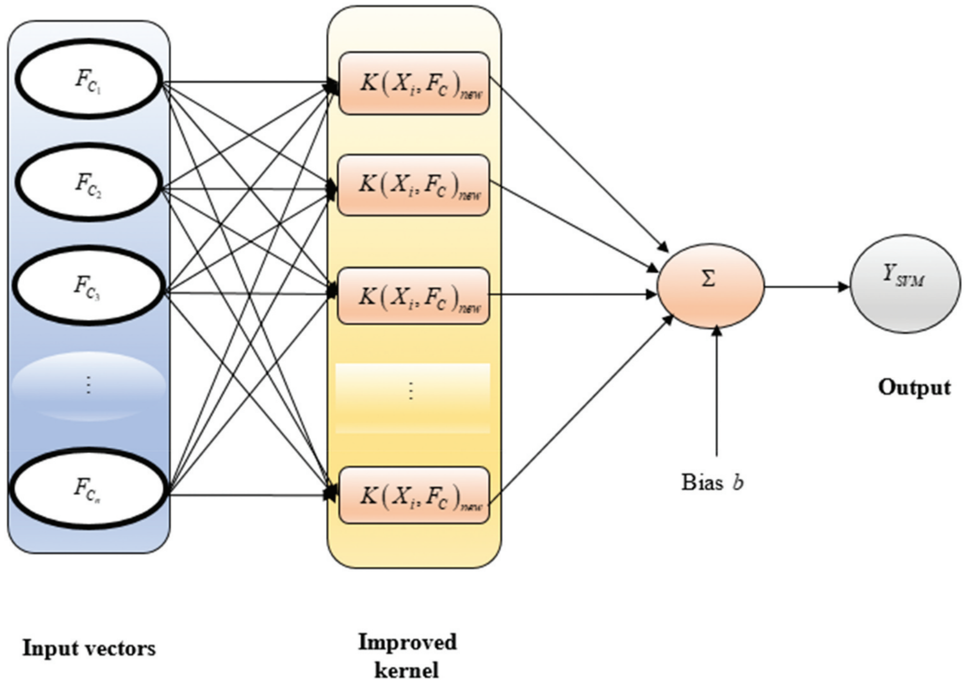


Figure 5. Improved SVM classifier.

information through the network. This architecture is well-suited for handling nonlinear data and typically includes input, hidden, and output layers.

In attack detection, the NN classifier is trained on a dataset that includes features extracted using methods such as Improved SVM and statistical analysis. These features F_C are fed into the NN model, which learns to distinguish among different types of attacks based on patterns and relationships in the data. The output score of the NN classifier is signified as Y_{NN}.

The NN classifier brings significant advantages to attack detection in the proposed system. It excels in capturing intricate nonlinear relationships within data, enabling the detection of subtle attack patterns

Table 2. Hyperparameters of classifiers.

Methods	Hyperparameters

SVM	C=1.0
	kernel=rbf
	degree=3
	gamma='scale'
	coef0=0.0
	shrinking=True
	probability=False
	cache_size=200
	class_weight=None

Improved SVM	C=1.0
	kernel=improved kernel
	degree=3
	gamma='scale'
	coef0=0.0,
	shrinking=True
	probability=False
	cache_size=200
	class_weight=None

NN	Sequential model Dense layer Units-100 Loss – sparse categorical cross entropy Optimizer -rmsprop Metrics – accuracy Batchsize – 1000 Epochs – 50

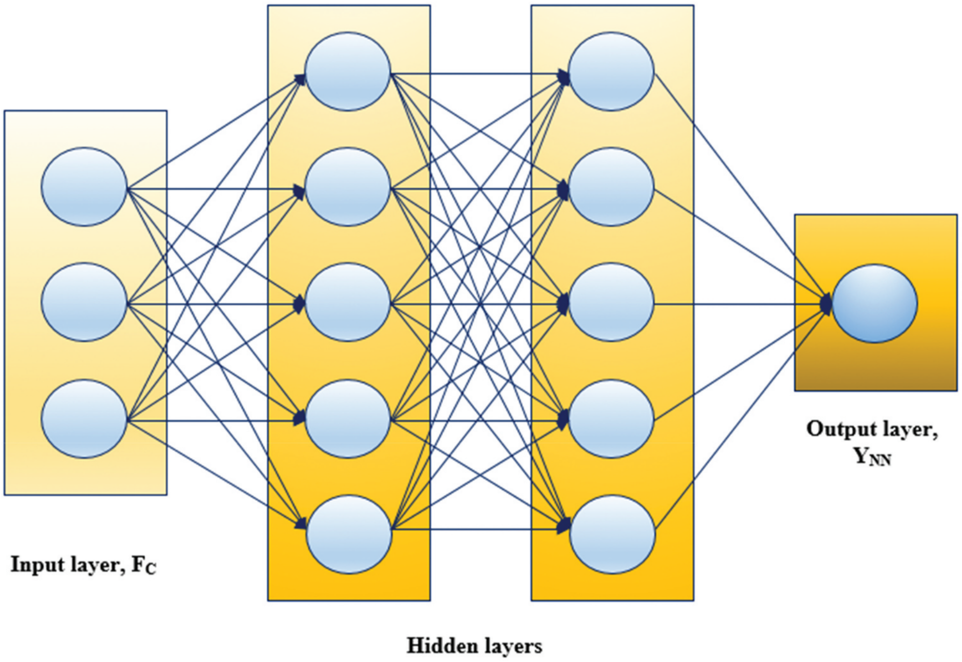


Figure 6. NN classifiers.

that linear models may overlook. Operating in higher-dimensional spaces with multiple layers, NNs effectively handle complex feature extraction and classification during attack detection. Additionally, their capability to learn from large-scale datasets and adapt to new attack patterns enhances the overall robustness and accuracy of the detection system against evolving cyber threats.

Therefore, the integration of an NN classifier alongside the Improved SVM in the proposed attack detection system leverages the strengths of both models. This hybrid approach aims to enhance the overall detection capability, providing a comprehensive solution for identifying and mitigating various types of cyberattacks effectively.

4. Results and discussion

4.1. Simulation procedure

Python 3.7 was utilized to simulate the suggested DDoS attack detection method from a big data perspective. The processes were carried out on a system that had a "Intel® Core™ i5-4210 U CPU @ 1.70 GHz and 8.00 GB of RAM." The Hadoop 3.4.1 MapReduce framework was used to implement the model. The MapReduce paradigm was implemented using the Python multiprocessing package, which enables parallel processing on multicore systems. While this may not utilize distributed computing frameworks such as Hadoop or Apache Spark, it effectively models the MapReduce design pattern on a shared-memory architecture, suitable for medium-to-large datasets within a single machine. Furthermore, the analysis of DDoS attack detection was performed using both the Dataset 1: DDoS evaluation dataset (https://www.unb.ca/cic/datasets/ddos-2019.html) and the Dataset 2: UNSW-NB15 dataset (https://research.unsw.edu.au/projects/unsw-nb15-dataset).

4.2. Dataset1 description

This dataset is a subset of the CICDDoS2019 dataset, focusing on brute-force attack detection scenarios. It comprises a total of 100,000 samples categorized into three labels: Benign (40,000 samples), FTP BruteForce (30,000 samples), and SSH BruteForce (30,000 samples). The dataset features a mix of benign network traffic and the advanced prevalent DDoS attacks, providing a close approximation of real-world PCAP data. It incorporates a range of current reflected DDoS attack techniques, including “PortMap, NetBIOS, LDAP, MSSQL, UDP, UDP-Lag, SYN, NTP, DNS, and SNMP.” In the training phase, 12 different DDoS attack scenarios were implemented, including “NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDP-Lag, WebDDoS, SYN, and TFTP.” The testing phase involved seven attacks, including “PortScan, NetBIOS, LDAP, MSSQL, UDP, UDP-Lag, and SYN.”

4.3. Dataset2 description

This dataset is a filtered subset of the UNSW-NB15 dataset, tailored to assess DDoS attack detection capabilities. It contains 71,730 total samples, distributed across Benign traffic (40,000 samples), DDoS HOIC attack (1,730 samples), and DDoS LOIC-UDP attack (30,000 samples). The raw network packets for this dataset were generated using the “IXIA Perfect Storm tool in the Cyber Range Lab at UNSW Canberra,” producing a combination of genuine, contemporary network activities and synthetic, current attack behaviors. The dataset includes nine diverse attack types: “fuzzers, analysis, backdoors, DoS, exploits, generic, reconnaissance, shell code, and worms.” Twelve algorithms were developed, and both the Argus and Bro-IDS tools were used to generate a total of 49 features along with the class label. The testing and training data of both datasets are presented in Table 3.

4.4. Performance analysis

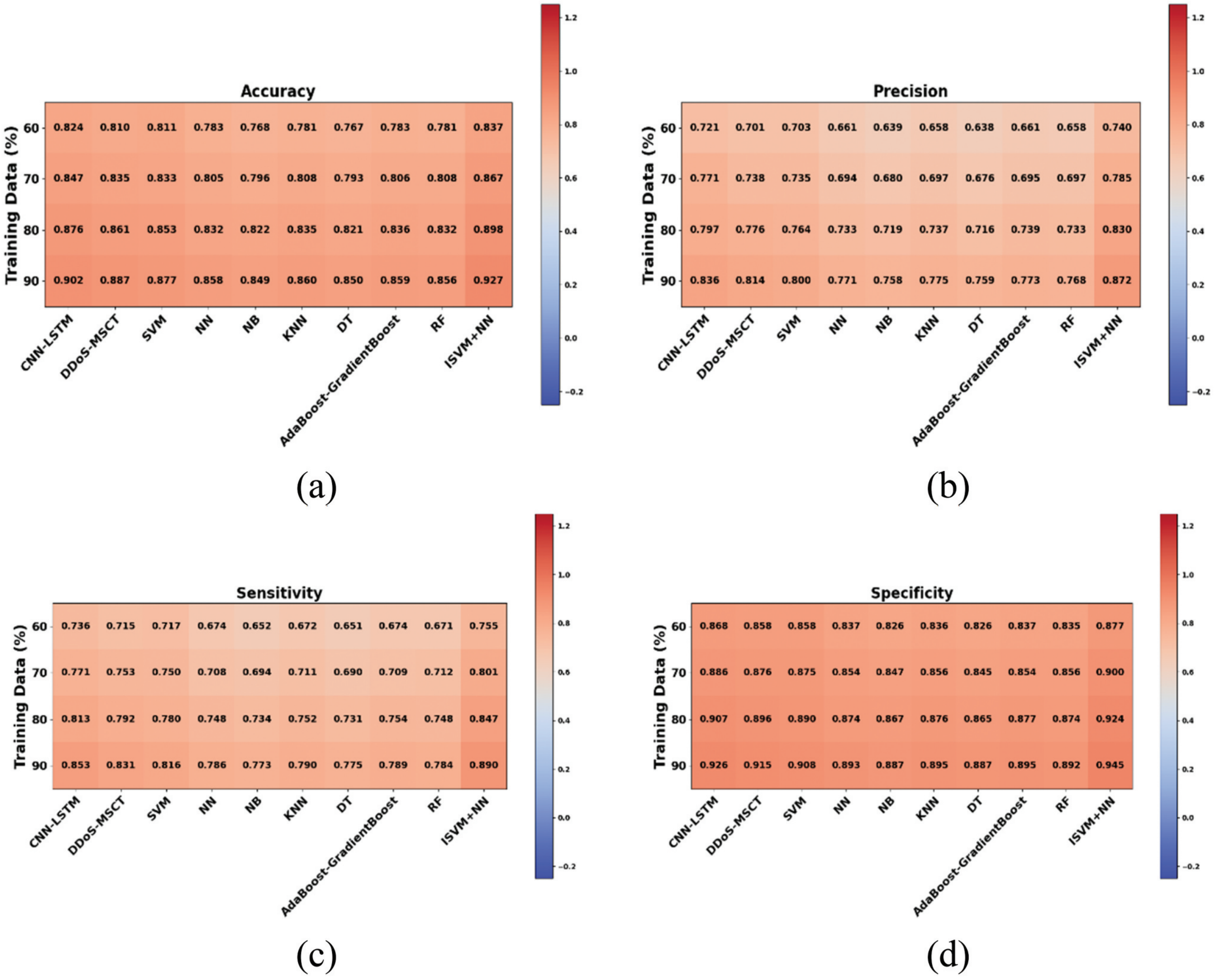
In comparison to traditional methods, a thorough analysis was conducted to evaluate the efficacy of the suggested DDoS attack detection method within a big data framework. The assessment employed a broad range of evaluation measures, including “Sensitivity, False Discovery Rate (FDR), Negative Predictive Value (NPV), Specificity, F-measure, False Negative Rate (FNR), Precision, False Positive Rate (FPR), Matthews Correlation Coefficient (MCC), and Accuracy,” to offer a thorough analysis of the ISVM+NN method’s performance. This evaluation compared the ISVM+NN approach with both state-of-the-art techniques, like AdaBoost-Gradient Boost (Dasari & Devarakonda, 2022), RF (Chaudhari et al., 2024), and DDoS-MSCT, CNN-LSTM, and traditional classifiers such as SVM, NN, NB, KNN, and DT. This comprehensive analysis was carried out using two distinct datasets: the DDoS evaluation and the UNSW-NB15 dataset.

4.5. Evaluation of positive, negative, and other metrics for Dataset 1

To evaluate the effectiveness of ISVM+NN for DDoS attack detection from a big data perspective, a comprehensive examination of positive measures is conducted for dataset 1. Figure 7 illustrates a comparative assessment of the ISVM+NN strategy against several established techniques, including SVM, NN, NB, KNN, DT, AdaBoost-Gradient Boost (Dasari & Devarakonda, 2022), DDoS-MSCT, CNN-LSTM and RF (Chaudhari et al., 2024). This comparative analysis focuses on the ability of these models to accomplish high positive metric values, which are crucial for the effective detection of DDoS attacks. A model’s performance in DDoS attack detection is significantly reflected through these positive metrics. Therefore, the analysis aims to demonstrate how well the ISVM+NN performs in comparison to these traditional methods, thereby highlighting its effectiveness in managing and mitigating DDoS threats in a big data environment. As shown in Figure 7(a), the ISVM+NN strategy consistently achieves higher accuracy compared to the other methods. At a 60% training data, the ISVM+NN strategy’s accuracy of 0.837

Table 3. Samples on each class label in training and testing.

	Dataset 1		Dataset 2	
Training Percentage	Training Data	Testing Data	Training Data	Testing Data
60	60000	40000	43038	28692
70	70000	30000	50211	21519
80	80000	20000	57384	14346
90	90000	10000	64557	7173



(a)

(b)

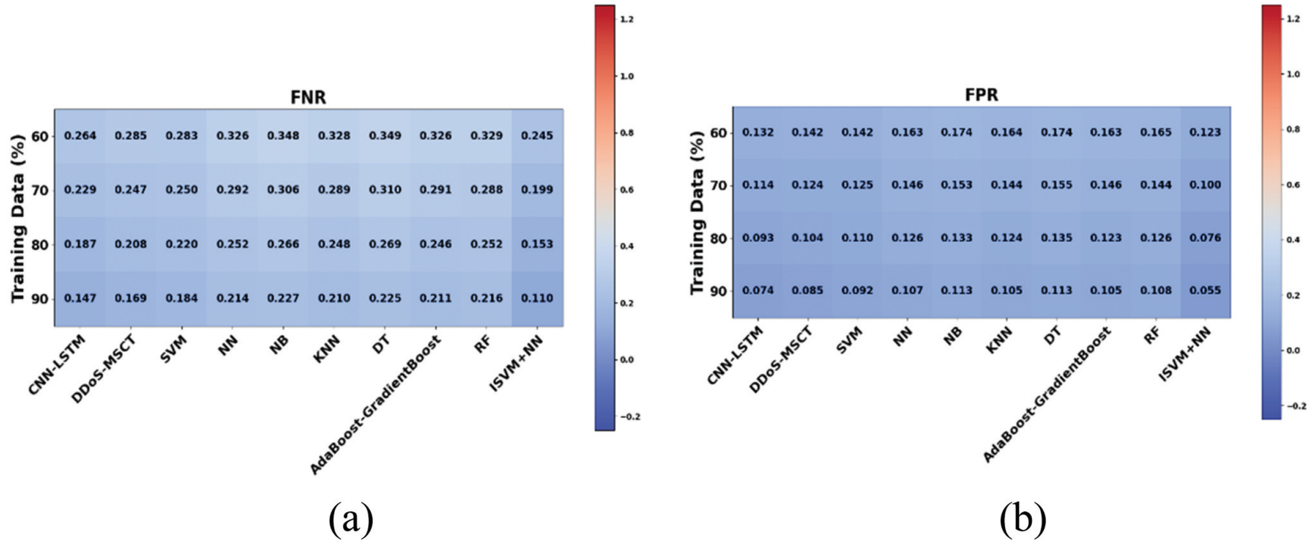
(c)

(d)

Figure 7. Positive performance metrics comparison: ISVM+NN vs. Conventional methods on Dataset1.

surpasses that of SVM (0.811), NN (0.783), NB (0.768), KNN (0.781), DT (0.767), AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) (0.783), and RF (Chaudhari et al., 2024) (0.781). As the percentage of training data rises, this pattern persists. With an accuracy of 0.867 at 70%, the ISVM+NN approach once again outperforms all other methods, which vary from 0.805 for NN to 0.833 for SVM. By the time 80% of the data is used for training, the ISVM+NN strategy's accuracy reaches 0.898, a significant improvement over SVM, NN, and other methods. Most notably, with 90% of training data, the ISVM+NN strategy achieves the highest accuracy of 0.927, surpassing SVM (0.877), NN (0.858), and all other methods, including AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) (0.859) and RF (Chaudhari et al., 2024) (0.856). At 90% of training data, the ISVM+NN strategy accomplishes a specificity of 0.877, which is the highest among all the models compared. In contrast, the specificity of other models such as SVM, NN, and KNN is lower, at 0.908, 0.893 and 0.895, respectively, indicating that the ISVM+NN strategy is more effective in detecting DDoS attacks.

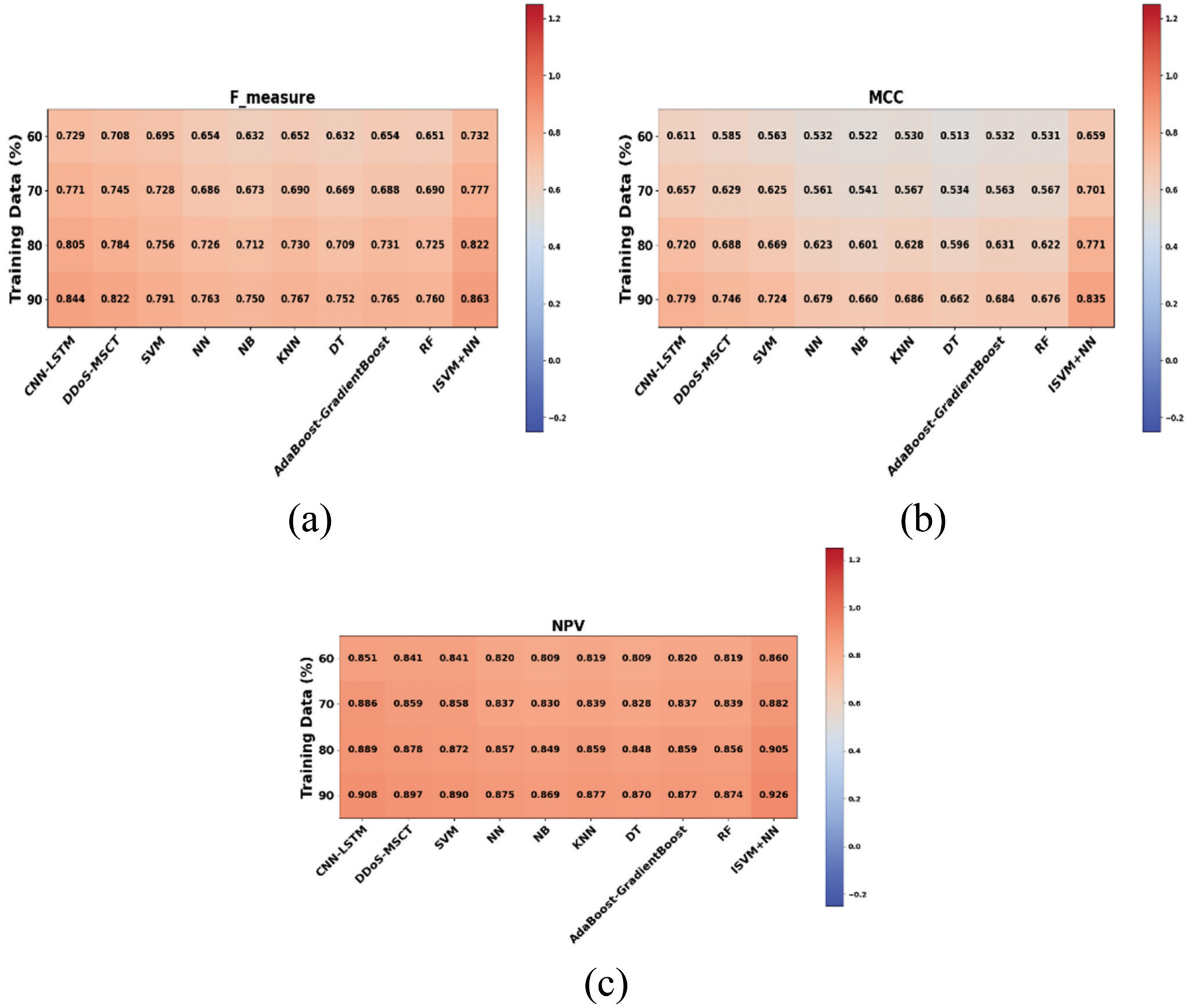
A comprehensive assessment of negative and other measures is shown in Figures 8 and 9 to further confirm the efficacy of the ISVM+NN technique for DDoS attack detection. This analysis compares the performance of the ISVM+NN strategy against several established models, including SVM, NN, NB, KNN, DT, AdaBoost-GradientBoost (Dasari & Devarakonda, 2022), DDoS-MSCT, and RF (Chaudhari et al., 2024).



(a)

(b)

Figure 8. Negative performance metrics comparison: ISVM+NN vs. Conventional methods on Dataset1.



(a)

(b)

(c)

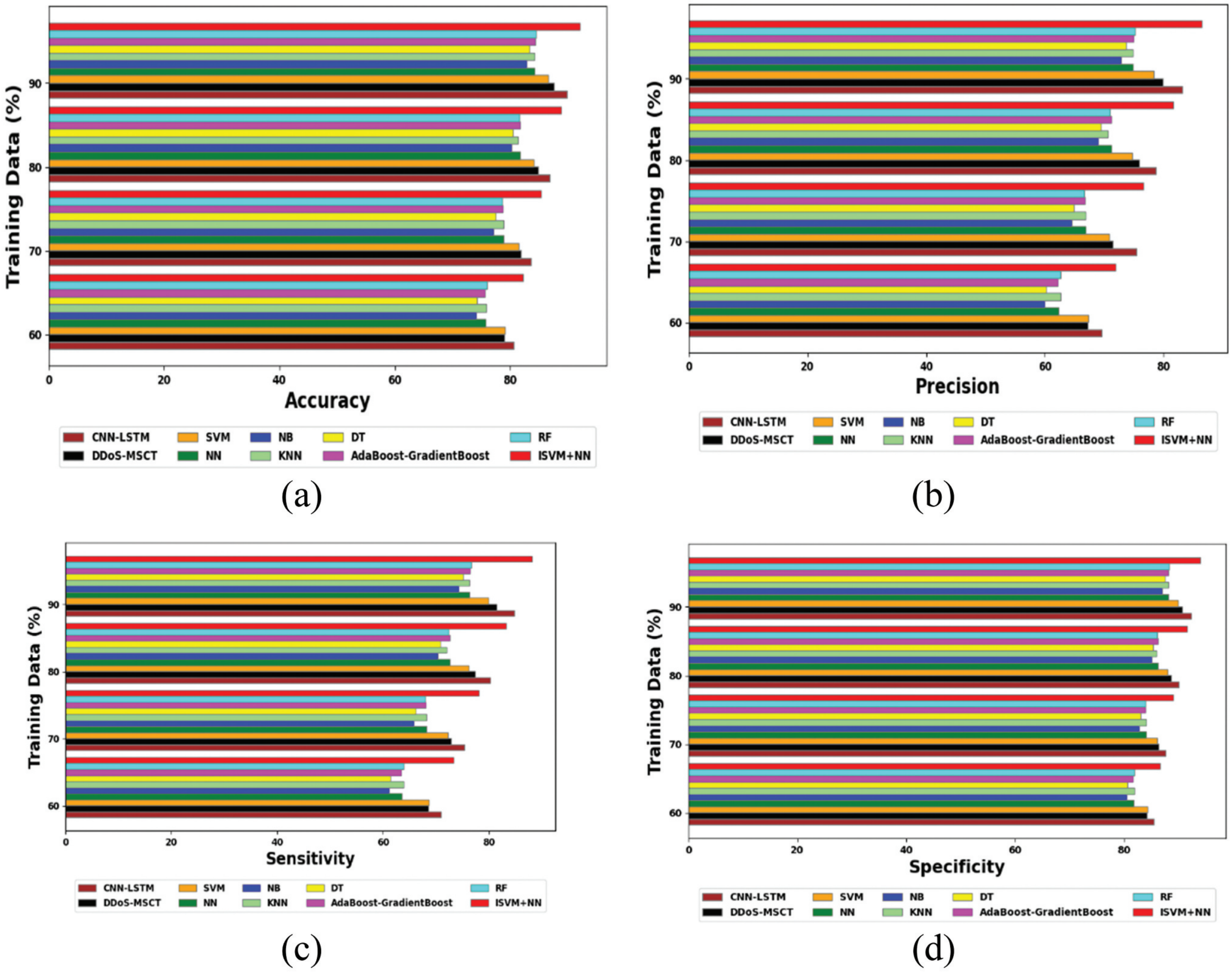
Figure 9. Other performance metrics comparison: ISVM+NN vs. Conventional methods on Dataset1.

For effective DDoS attack detection, the ISVM+NN model should not only achieve higher values for these positive metrics but also demonstrate reduced negative metric ratings. A successful DDoS detection model needs to balance the achievement of high other metric values with minimizing negative outcomes. As shown in Figure 7, the ISVM+NN model consistently achieves the lowest FNR compared to conventional strategies. For example, with 60% training data, the FNR for the ISVM+NN approach is 0.245, which is lower than SVM's 0.283 and other models' FNR ranging from 0.326 to 0.349. As the training data increases to 70%, 80%, and 90%, the ISVM+NN model's FNR decreases to 0.199, 0.153, and 0.110, respectively, further outperforming the other models, which have FNR values between 0.184 and 0.306. This consistent reduction in FNR demonstrates that the ISVM+NN approach is more effective in detecting DDoS attacks and minimizing errors across different training data scenarios, proving its robustness in a big data context for DDoS attack detection.

At 90% of the training data, the MCC values for various DDoS attack detection models reveal distinct performance differences. The ISVM+NN approach achieves the highest MCC of 0.835, demonstrating the most effective overall classification performance. In comparison, SVM attains an MCC of 0.724, reflecting a strong but slightly less balanced classification compared to the ISVM+NN model. NN exhibit an MCC of 0.679, which is lower than that of SVM, indicating that NN has less effective detection of DDoS attacks. NB shows an MCC of 0.660, while KNN achieves an MCC of 0.686, both of which are lower than the ISVM+NN model and demonstrate less effectiveness in classifying DDoS attacks. DT have an MCC of 0.662, slightly above NB but still below the ISVM+NN model's performance. AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) scores an MCC of 0.684, and RF (Chaudhari et al., 2024) has an MCC of 0.676, both of which fall short compared to the ISVM+NN approach. Therefore, the ISVM+NN approach substantially surpasses traditional methods in detecting DDoS attacks, showing superior performance across positive, negative, and other critical metrics. The ISVM+NN strategy outperforms traditional methods due to three key advancements: Improved Normalization techniques, Improved Entropy-Based Features, and the integration of a Hybrid Model. Together, these innovations contribute to superior detection capabilities and overall performance.

4.6. Evaluation of positive, negative, and other metrics for dataset 2

The comparative evaluation of the ISVM+NN strategy against SVM, NN, NB, KNN, DT, AdaBoost-GradientBoost (Dasari & Devarakonda, 2022), DDoS-MSCT, CNN-LSTM and RF (Chaudhari et al., 2024) for DDoS attack detection under a big data framework using Dataset2 is presented. This evaluation comprehensively analyzes the performance of these methods for positive, negative, and other metrics, as illustrated in Figures 10–12. The figures provide a detailed visualization of the ISVM+NN strategy's effectiveness compared to traditional methodologies, highlighting its superior performance across various critical metrics essential for robust DDoS attack detection. In terms of precision, when the training data is set at 60%, the ISVM+NN strategy achieves a precision of 0.7192, which is higher than SVM (0.672), NN (0.622), NB (0.599), KNN (0.626), DT (0.601), AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) (0.621), DDoS-MSCT (0.529) and RF (Chaudhari et al., 2024) (0.626). With 70% training data, the ISVM+NN strategy's precision improves to 0.765, and with 80% training data, it achieves an even higher precision of 0.816, consistently outperforming SVM (0.746), NN (0.711), NB (0.689), KNN (0.705), DT (0.694), AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) (0.711), and RF (Chaudhari et al., 2024) (0.709). Finally, with 90% training data, the ISVM+NN strategy reaches the highest precision of 0.864, significantly higher than SVM (0.783), NN (0.748), NB (0.727), DT (0.736), AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) (0.749), and RF (Chaudhari et al., 2024) (0.752). As illustrated in Figure 10(b), the ISVM+NN strategy consistently outperforms the conventional methodologies across all training data proportions. Specifically, the FPR for the ISVM+NN approach decreases from 0.133 at 60% training data to 0.059 at 90%, demonstrating a significant reduction in error values as the amount of training data increases. In contrast, the FPR for traditional methods such as SVM, NN, NB,



(a)

(b)

(c)

(d)

Figure 10. Positive performance metrics comparison: ISVM+NN vs. traditional strategies on Dataset2.

KNN, DT, AdaBoost-GradientBoost (Dasari & Devarakonda, 2022), DDoS-MSCT and RF (Chaudhari et al., 2024) either remains relatively constant or shows less pronounced improvements with increased training data.

When utilizing 90% of training data for DDoS attack detection, the ISVM+NN accomplishes the maximum F-measure of 0.855, significantly surpassing all other techniques. Traditional methods such as SVM, NN, KNN, AdaBoost-GradientBoost (Dasari & Devarakonda, 2022), and RF (Chaudhari et al., 2024) algorithms show lower F-measure values, with SVM achieving the highest among them at 0.775, followed by NN and KNN at 0.740, and AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) and RF (Chaudhari et al., 2024) at 0.742 and 0.744, respectively. NB and DT perform slightly worse, with F-measures of 0.720 and 0.729, respectively. This substantial difference demonstrates that the ISVM+NN method not only detects DDoS attacks more effectively but also achieves a better balance between precision and recall compared to traditional approaches. Therefore, the ISVM+NN approach outperforms traditional approaches across various metrics, achieving superior values. These enhancements are largely due to the Improved Normalization techniques, Entropy-Based Features, and the Hybrid Model utilized in the approach. By integrating these advanced methods, the ISVM+NN strat-

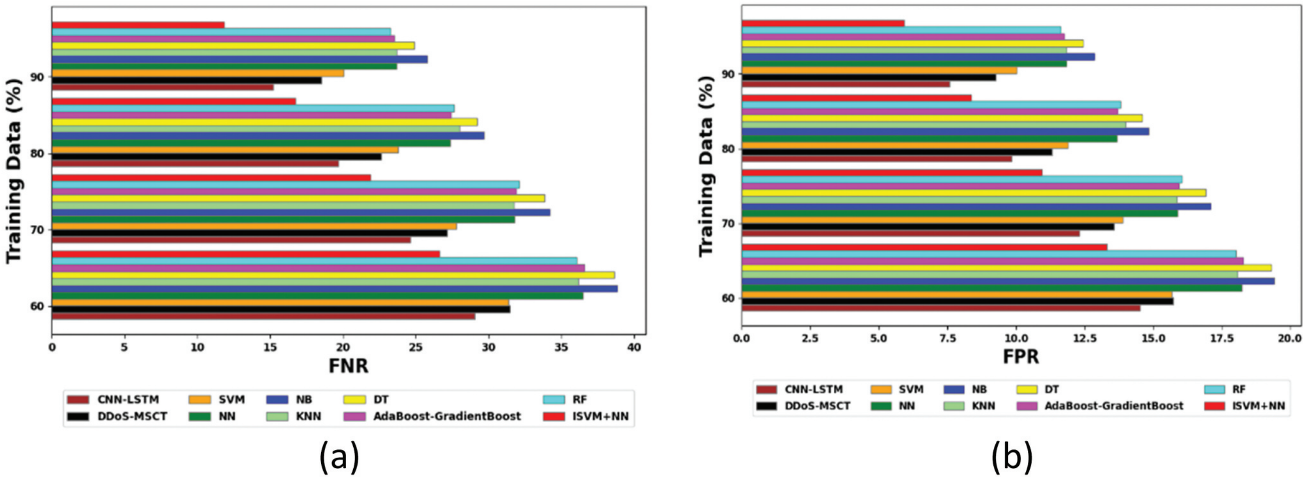


Figure 11. Negative performance measures comparison: ISVM+NN vs. conventional strategies on Dataset2.

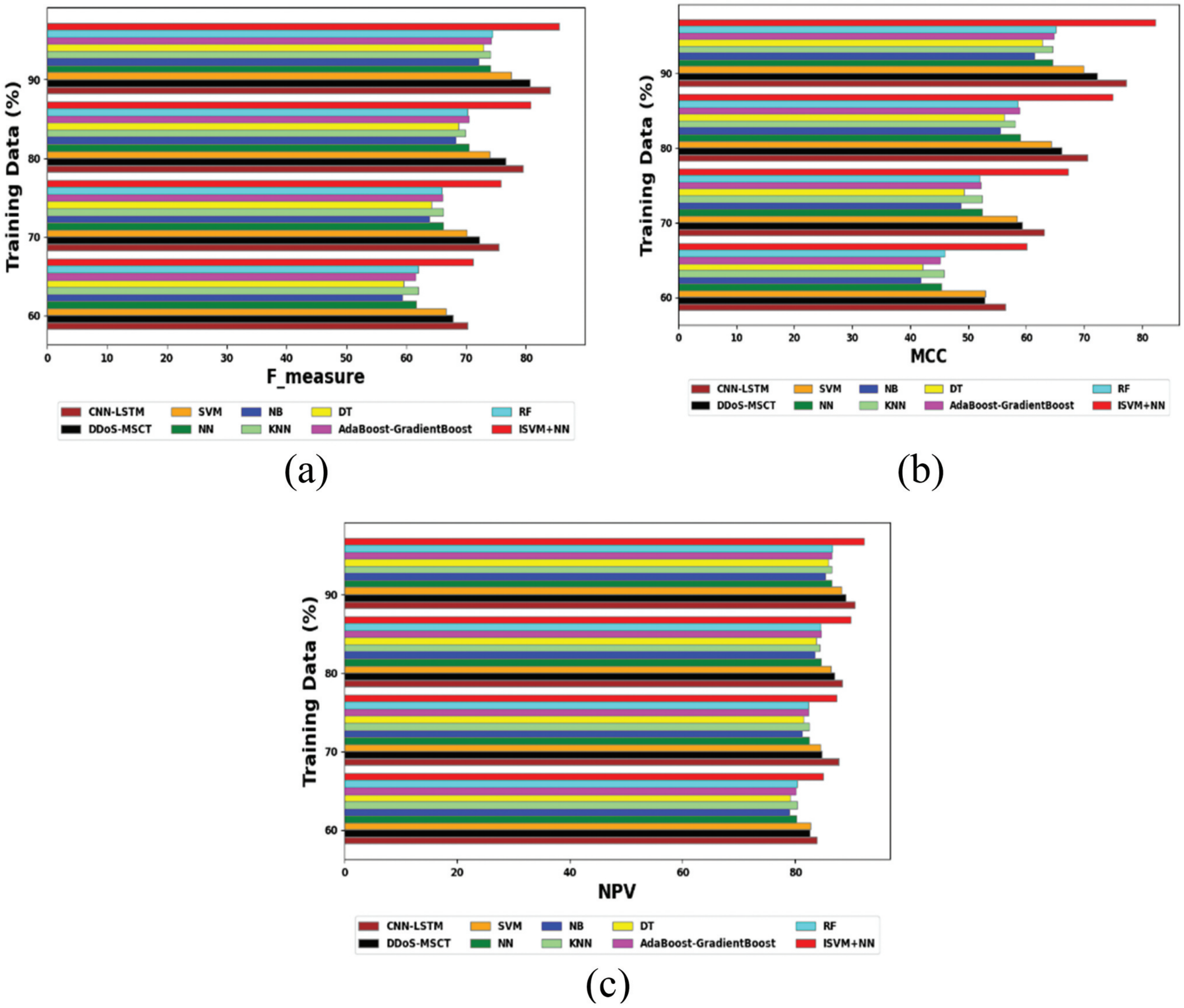
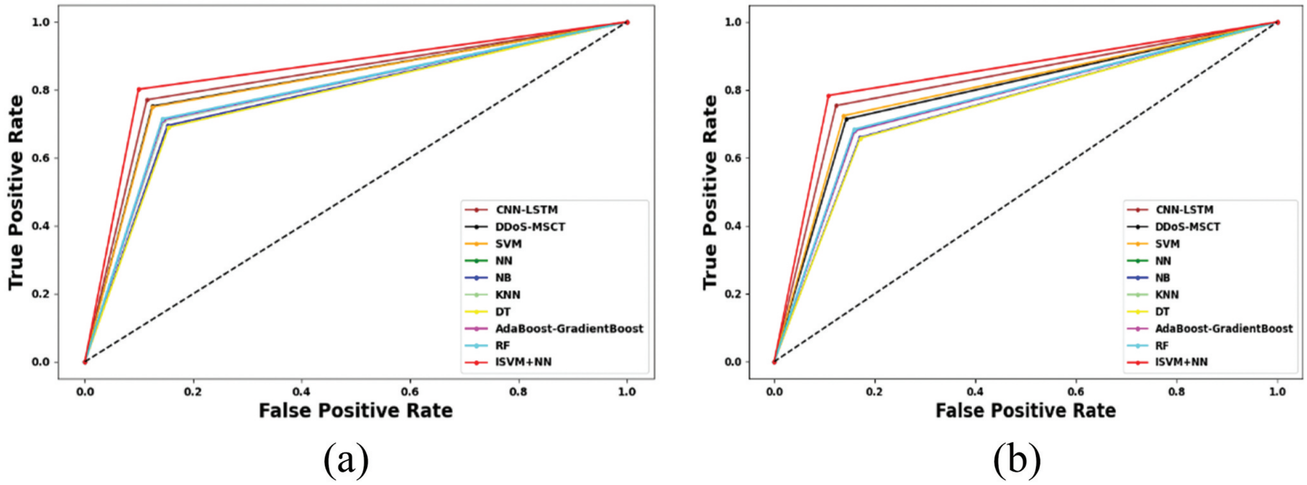


Figure 12. Other performance measures comparison: ISVM+NN vs. traditional techniques on Dataset2.



(a)

(b)

Figure 13. Roc curve analysis on ISVM+NN and traditional methods a) Dataset1 and b) Dataset2.

egy provides a more effective and reliable solution for the detection of DDoS attack.

4.7. ROC analysis using Dataset1 and Dataset2

A ROC curve is a visual representation of a binary classification model's performance that plots the TPR versus the FPR at different thresholds. The AUC measures how well the model distinguishes between positive and negative classes, with higher values signifying better performance. To comprehensively assess the effectiveness of the ISVM+NN strategy for DDoS attack detection within a big data framework, an ROC curve analysis was conducted to compare its performance against several conventional methodologies. This analysis is illustrated in Figures 13(a) and 13(b), which depict the ROC curves for the ISVM +NN model alongside SVM, NN, NB, KNN, DT, DDoS-MSCT, CNN-LSTM, AdaBoost-GradientBoost (Dasari & Devarakonda, 2022), and RF (Chaudhari et al., 2024) using Dataset1 and Dataset2, respectively. While examining Dataset1, the ISVM +NN strategy achieved a remarkable TPR of 0.962 when the FPR was set at 1.0. In contrast, the conventional methods showed lower TPR values, with SVM achieving 0.951, NN at 0.948, NB at 0.937, KNN at 0.924, DT at 0.917, AdaBoost-GradientBoost (Dasari & Devarakonda, 2022) at 0.908, and RF (Chaudhari et al., 2024) at 0.910. This demonstrates that the ISVM +NN model not only outperforms these conventional methods but also maintains the highest TPR under the same FPR condition, highlighting its superior capability in correctly identifying attack instances. For Dataset2, the ISVM+NN strategy consistently achieved the maximum TPR values, illustrating its robust performance across different datasets.

4.8. Ablation analysis of ISVM+NN based attack detection for Dataset1 and Dataset2

To comprehensively evaluate the effectiveness of the ISVM+NN model for DDoS attack detection, an ablation study was performed to analyze the contributions of the various components and techniques employed in the model. This study compares the performance of full ISVM+NN model, which includes advanced feature extraction, improved normalization methods, and enhanced entropy-based features, against alternative configurations to isolate and understand the impact of each component. This evaluation involves comparing the ISVM+NN model with a version that lacks feature extraction, a version that uses conventional normalization techniques, and a version that relies on conventional entropy-based features. This comprehensive comparison is essential for understanding the impact of each component of the ISVM +NN approach on DDoS attack detection performance. The results of this ablation evaluation utilizing Dataset1 and Dataset2 are summarized in Tables 4 and 5, providing a detailed comparison of how each model variant performs under a big data perspective. The ablation study's sensitivity analysis, using 90% of the training data, reveals that the ISVM+NN model achieves a high sensitivity of 0.890, successfully

Table 4. Ablation study of ISVM+NN, model without extraction of feature, model with traditional normalization and model with traditional entropy-based features for Dataset1.

Metrics	Proposed without preprocessing	Proposed without extraction of feature	Proposed with tarditional normalization	Proposed with traditional entropy-based features	Proposed
Accuracy	83.08%	82.49%	83.16%	83.61%	92.66%
Sensitivity	74.63%	73.73%	74.73%	75.41%	88.98%
Specificity	87.31%	86.87%	87.37%	87.71%	94.49%
Precision	73.87%	72.26%	73.24%	73.90%	87.21%
F_measure	72.39%	71.52%	72.49%	73.15%	86.32%
MCC	62.28%	61.63%	62.10%	63.12%	83.48%
NPV	85.57%	85.13%	85.62%	85.95%	92.60%
FPR	12.69%	13.13%	12.63%	12.29%	5.51%
FNR	25.37%	26.27%	25.27%	24.59%	11.02%

detecting DDoS attacks. In comparison, models lacking feature extraction (0.184), employing conventional normalization (0.214), or using standard entropy-based features (0.227) demonstrate significantly lower sensitivities. This underscores the effectiveness of the advanced techniques incorporated in the ISVM+NN model for DDoS attack detection. In the ablation study using Dataset2, the FPR metric indicates that the ISVM+NN model outperforms other methods with a low FPR of 0.059. In comparison, the model without feature extraction has an FPR of 0.134, the model with conventional normalization has an FPR of 0.133, and the model with traditional entropy-based features has an FPR of 0.124. This specifies that the ISVM+NN is the most effective at detecting DDoS attacks, resulting in fewer errors and a more reliable attack detection system.

4.9. Statistical evaluation of proposed and conventional models in terms of accuracy for Dataset1 and Dataset2

To provide a comprehensive evaluation of the ISVM +NN strategy for DDoS attack detection, a detailed statistical assessment was conducted. This assessment involved comparing the ISVM+NN strategy with several well-established models, including SVM, NN, NB, KNN, DT, AdaBoost-GradientBoost (Dasari & Devarakonda, 2022), DDoS-MSCT, CNN-LSTM and RF (Chaudhari et al., 2024). The comparison utilized two Datasets, Dataset1 and Dataset2, and the outcomes are offered in Tables 6 and 7. In the statistical evaluation of accuracy for DDoS attack detection on Dataset1, the ISVM+NN model outperforms all other methods. Specifically, the ISVM+NN model achieves an accuracy of 0.882 on the mean statistical metric, which outperforms all other evaluated models. This indicates that, on average, the ISVM+NN model successfully detects the DDoS attack. In comparison, the SVM has an accuracy of 0.844, reflecting its effectiveness but showing that it falls short of the ISVM+NN model's performance. The NN follows with an accuracy of 0.819, which is slightly lower than SVM's performance. The NB model shows an accuracy of 0.809, indicating that it is less effective in detecting DDoS attacks compared to SVM and the ISVM+NN model. The KNN model and DT model both have mean accuracies of 0.821 and 0.808, respectively, representing performance levels similar to NN and slightly below SVM. Additionally, the AdaBoost-GradientBoost (Dasari & Devarakonda, 2022), DDoS-MSCT, CNN-LSTM and RF (Chaudhari et al., 2024)

Table 5. Ablation study on ISVM+NN, model without feature extraction, model with traditional normalization and model with traditional entropy-based features for Dataset2.

Metrics	Proposed without preprocessing	Proposed without extraction of feature	Proposed with traditional normalization	Proposed with traditional entropy-based features	Proposed
Accuracy	81.72%	82.12%	82.26%	83.46%	92.11%
Sensitivity	72.58%	73.19%	73.39%	75.19%	88.17%
Specificity	86.29%	86.59%	86.70%	87.60%	94.09%
Precision	71.13%	71.73%	71.93%	73.69%	86.41%
F_measure	71.84%	70.99%	71.19%	72.94%	85.53%
MCC	58.86%	59.78%	60.09%	62.79%	82.26%
NPV	84.56%	84.86%	84.96%	85.84%	92.20%
FPR	13.71%	13.41%	13.30%	12.40%	5.91%
FNR	27.42%	26.81%	26.61%	24.81%	11.83%

Table 6. Statistical analysis of proposed and traditional approaches based on accuracy using Dataset 1.

Methods	MIN	MAX	MEAN	MEDIAN	STANDARD DEVIATION
DDoS-MSCT	81.02%	88.71%	84.84%	84.82%	2.87%
CNN-LSTM	82.40%	90.19%	86.22%	86.15%	2.93%
SVM	81.13%	87.72%	84.38%	84.33%	2.43%
NN	78.28%	85.75%	81.94%	81.87%	2.81%
NB	76.80%	84.88%	80.89%	80.93%	3.01%
KNN	78.12%	86.03%	82.10%	82.12%	2.96%
DT	76.74%	84.98%	80.77%	80.68%	3.08%
AdaBoost-Gradient Boost	78.28%	85.94%	82.10%	82.10%	2.91%
RF	78.06%	85.59%	81.90%	81.98%	2.80%
ISVM+NN	83.66%	92.66%	88.21%	88.26%	3.36%

models have accuracies of 0.821 and 0.819, respectively, which are comparable to the performance of KNN and NN. Overall, the ISVM+NN model's accuracy of 0.882 not only surpasses the accuracies of traditional methods but also highlights its superior effectiveness in detecting DDoS attacks. The lower accuracies of the other models, which range from 0.808 to 0.844, underscore the ISVM+NN model's robustness and its ability to consistently perform better in identifying DDoS attack instances. Table 6 shows that the ISVM+NN model realizes the highest accuracy of 0.921 (maximum statistical metric) for Dataset2 in DDoS attack detection, surpassing all evaluated methods. In comparison, the SVM model reaches an accuracy of 0.866, while NN and KNN both achieve 0.842. The NB model has an accuracy of 0.828, and the DT model scores 0.834. AdaBoost-Gradient Boost (Dasari & Devarakonda, 2022) and RF (Chaudhari et al., 2024) models achieve accuracies of 0.843 and 0.845, respectively, which are lower than those of the ISVM+NN model and SVM.

4.10. Analysis on T-test for Dataset1 and Dataset2

A T-test is a statistical test employed to assess whether there are significant differences between the means of two groups or between a sample mean and a known value. It evaluates if observed differences are likely due to chance or reflect actual differences in the population. The T-test analysis for the ISVM+NN model, in comparison to SVM, NN, NB, KNN, DT, DDoS-MSCT, CNN-LSTM, AdaBoost-Gradient Boost (Dasari & Devarakonda, 2022), and RF (Chaudhari et al., 2024), for DDoS attack detection from a big data perspective is detailed in Table 8 for Dataset1 and Dataset2, respectively. For both Datasets, deep learning models like CNN-LSTM and DDoS-MSCT exhibit high p-values (CNN-LSTM: 0.4701 and 0.5376; DDoS-MSCT: 0.2354 and 0.2297), indicating no statistically significant variation in their results and suggesting consistent performance. In contrast, traditional machine learning models such as Decision Tree (DT) and Naive Bayes (NB) show very low p-values (DT: 0.0301 and 0.0297; NB: 0.0307 and 0.0236), implying statistically significant differences and potentially higher sensitivity to data variations. With p-values typically below 0.06, models like as SVM, NN, KNN, AdaBoost-Gradient Boost, and RF fall into a moderate range and show substantial significance in performance differences. The p-values for Random Forest (RF) and Neural Networks (NN) are valuable because they are near the 0.05 cutoff, indicating slightly significant differences. Overall, the T-test results confirm that the ISVM+NN model achieves a statistically significant improvement over the traditional approaches tested for DDoS attack detection in both Datasets.

Table 7. Statistical analysis of suggested and conventional approaches based on accuracy for Dataset 2.

Methods	MIN	MAX	MEAN	MEDIAN	STANDARD DEVIATION
DDoS-MSCT	79.02%	87.65%	83.38%	83.42%	3.23%
CNN-LSTM	80.64%	89.88%	85.25%	85.24%	3.47%
SVM	79.10%	86.63%	82.84%	82.82%	2.82%
NN	75.71%	84.21%	80.12%	80.29%	3.18%
NB	74.11%	82.83%	78.59%	78.70%	3.26%
KNN	75.92%	84.21%	80.08%	80.09%	3.06%
DT	74.25%	83.42%	78.92%	79.00%	3.42%
AdaBoost-Gradient Boost	75.62%	84.33%	80.11%	80.24%	3.26%
RF	75.97%	84.49%	80.16%	80.10%	3.19%
ISVM+NN	82.25%	92.11%	87.16%	87.14%	3.69%

Table 8. Comparison of ISVM+NN and Conventional methods using both Datasets 1&2: T-Test results.

Methods	T test *p* value For Dataset 1	T test *p* value for Dataset 2
DDoS-MSCT	.235414	.229709
CNN-LSTM	.47014	.53764
SVM	.161112	.158278
NN	.048121	.046499
NB	.0307	.023558
KNN	.056084	.043043
DT	.030128	.029675
AdaBoost-Gradient Boost	.054943	.047683
RF	.046781	.04749

4.11. Friedman test evaluation for Dataset1 and Dataset2

Using rank-based data, the Friedman Test is a non-parametric statistical technique for determining differences between several related groups. It is commonly employed when comparing treatments or conditions in repeated measures or matched groups without assuming normal distribution. Table 9 shows the results of the Friedman Test, which was used to compare the ISVM+NN method against traditional DDoS attack detection techniques from a big data perspective using Datasets 1 and 2, respectively.

The Friedman chi-square p-values presented for DDoS attack detection across two Datasets evaluate the statistical significance of performance differences between the proposed ISVM+NN hybrid model and existing classifiers. When compared to the suggested model, the high p-values for CNN-LSTM (0.4418 and 0.5215) and DDoS-MSCT (0.2114 and 0.2170) indicate that there are no appreciable performance changes, suggesting comparatively steady and consistent behavior. The suggested approach is more effective than conventional classifiers, as evidenced by lower p-values for Decision Tree (DT) (0.0452 and 0.0554) and Naive Bayes (NB) (0.0629 and 0.0276), which indicate statistically significant performance differences. In contrast, classifiers such as DDoS-MSCT, SVM, KNN, and AdaBoost-Gradient Boost yield higher p-values (generally >0.15), suggesting their performance differences relative to the proposed model are not statistically significant. Overall, the results imply that while some traditional classifiers show noticeable performance gaps, the proposed ISVM+NN model demonstrates statistically robust superiority in performance across diverse classifiers, especially where p-values fall below the common threshold of 0.05. The extremely low p-values across all methods confirm that the ISVM+NN model's performance is statistically superior to the conventional techniques in the context of DDoS attack detection for Dataset2.

4.12. Analysis on Wilcoxon test for Dataset1 and Dataset2

As a non-parametric statistical test, the Wilcoxon Test compares two related samples to assess whether their population mean ranks are significantly different. It is often applied when data do not meet the assumptions of parametric tests like the paired t-test, especially when the data are not normally distributed. The Wilcoxon Test analysis presented in Table 10 provides a detailed

Table 9. Comparison of ISVM+NN and Conventional methods using Dataset1: Friedman test results.

Proposed Vs	Friedman chisquare *p* value for Dataset 1	Friedman chi square *p* value for Dataset 2
DDoS-MSCT	.211443	.216979
CNN-LSTM	.441761	.521473
SVM	.184626	.183617
NN	.101767	.148762
NB	.0629	.027619
KNN	.164205	.104209
DT	.045157	.055406
AdaBoost-GradientBoost	.162872	.164007
RF	.071558	.151079

Table 10. Performance comparison of methods using the Wilcoxon test on Datasets 1 and 2.

Methods	Wilcoxon *p* value for Dataset 1	Wilcoxon *p* value for Dataset 2
DDoS-MSCT	.192123	.20516
CNN-LSTM	.452617	.501927
SVM	.134144	.19564
NN	.090633	.119238
NB	.063969	.026992
KNN	.121513	.067457
DT	.035062	.045955
AdaBoost-GradientBoost	.093392	.189692
RF	.071971	.138748

comparison of the ISVM+NN method against traditional DDoS attack detection techniques, including SVM, NN, NB, KNN, DT, DDoS-MSCT, CNN-LSTM, AdaBoost-Gradient Boost (Dasari & Devarakonda, 2022), and RF (Chaudhari et al., 2024), using Dataset1 and Dataset2. DDoS-MSCT exhibits comparatively strong p-values (0.1921 for Dataset 1 and 0.2052 for Dataset 2), suggesting that there is no discernible performance variance among the approaches. Similarly, CNN-LSTM shows the least statistically significant differences and potentially stable performance across Datasets, as evidenced by its highest p-values (0.4526 and 0.5019). The lowest p-values, on the other hand, are displayed by conventional classifiers like Decision Tree (DT) and Naive Bayes (NB) (DT: 0.0351 and 0.0460; NB: 0.0640 and 0.0270), suggesting statistically significant differences and possible sensitivity to Dataset modifications. The p-values of other models, such as SVM, NN, KNN, AdaBoost-GradientBoost, and Random Forest (RF), are typically above 0.06 but below 0.20, indicating significant levels of variation. The consistently lower p-values for the ISVM+NN model across both Datasets underscore its effectiveness in enhancing DDoS attack detection performance, demonstrating a significant advancement over traditional techniques in the context of big data environments.

4.13. Comparison of DDoS detection classifiers across two Datasets using confidence intervals (CI) based on accuracy

The performance evaluation of various classifiers across two Datasets demonstrates the effectiveness of the proposed ISVM+NN hybrid model, particularly in terms of detection accuracy is tabulated in Table 11. For Dataset 1, the 95% confidence interval (CI) for ISVM+NN ranges from 0.844 to 0.912, notably outperforming traditional classifiers such as SVM (CI: 0.817–0.865), NN (0.788–0.845), CNN-LSTM also performs well with a CI of 0.8357–0.8887 and Random Forest (0.793–0.844). Similarly, in Dataset 2, ISVM+NN maintains superior performance with a 95% CI of 0.830 to 0.905, compared to SVM (0.797-–0.853), NN (0.772–0.830), CNN-LSTM (0.8138–-0.8839) and AdaBoost-Gradient Boost (0.764–0.830). Traditional classifiers such as SVM, NN, NB, and ensemble methods like AdaBoost-GradientBoost and Random Forest (RF) show moderate performance, with narrower intervals, reflecting comparatively lower confidence in extreme cases. Overall, the ISVM+NN model consistently outperforms others, followed by DDoS-MSCT, emphasizing the effectiveness of hybrid and deep learning-based approaches in handling complex DDoS traffic patterns.

4.14. Proposed ISVM+NN model using different kernel functions across two Datasets

Table 12 evaluates the effectiveness of the proposed Weighted Expo Inverse Laplacian (WEIL) kernel against traditional non-linear kernels RBF, Exponential, and Laplacian within the ISVM+NN hybrid model, across two benchmark Datasets. In every performance metric, the suggested WEIL kernel continuously beats the other kernels. It has the lowest false positive rate (0.0551) and false negative rate (0.1102) in Dataset 1, as well as the greatest precision (0.9266), sensitivity (0.8898), specificity (0.9449), F1-score (0.8632), and MCC (0.8348). The WEIL kernel also performs best in Dataset 2, identifying true positives and true negatives with an accuracy of 0.9211, sensitivity of 0.8817, and specificity of 0.9409. These results empirically demonstrate that the WEIL kernel enhances the model’s ability to capture complex nonlinear relationships more effectively than standard

Table 11. Confidence intervals using two Datasets based on accuracy.

Classifiers	Lower bound	Higher Bound
	Dataset 1	
DDoS-MSCT	0.816472	0.874125
CNN-LSTM	0.835707	0.888742
SVM	0.816833	0.865158
NN	0.788356	0.844892
NB	0.781631	0.837057
KNN	0.794351	0.847558
DT	0.773772	0.835628
AdaBoost-GradientBoost	0.788544	0.846
RF	0.793452	0.843892
ISVM+NN	0.844227	0.912325
	Dataset 2	
DDoS-MSCT	0.797447	0.862184
CNN-LSTM	0.813776	0.883855
SVM	0.797	0.853471
NN	0.772198	0.829836
NB	0.748812	0.814187
KNN	0.766463	0.827792
DT	0.750541	0.819315
AdaBoost-GradientBoost	0.764051	0.830301
RF	0.766322	0.830324
ISVM+NN	0.830469	0.904863

kernels, providing a stronger theoretical justification for its use in DDoS detection tasks.

4.15. Computational time analysis

According to Table 13, the suggested ISVM+NN hybrid model outperforms other conventional and ensemble-based techniques in terms of computational time evaluation across two Datasets. The lightweight processing capabilities of ISVM+NN is demonstrated by its lowest computation times, which are 83.02 seconds on Dataset 1 and 82.72 seconds on Dataset 2. On the other hand, models like Random Forest and AdaBoost-Gradient Boost show far longer computation times more than 90 seconds on both Datasets, which suggests higher processing overhead. Traditional classifiers such as SVM, KNN, and DT show moderate performance with times ranging from 85 to 88 seconds, while the DDoS-MSCT model also exhibits relatively higher times (~90 seconds). Overall, the ISVM+NN approach is very appropriate for real-time identification of DDoS in big data situations since it guarantees computational economy in addition to high detection accuracy.

Table 12. Different kernel functions across two Datasets.

Metrics	Model with RBF kernel	Model with exponential kernel	Model with Laplacian kernel	Proposed
		Dataset 1		
Accuracy	89.77%	87.86%	88.80%	92.66%
Sensitivity	84.65%	81.78%	83.20%	88.98%
Specificity	92.33%	90.89%	91.60%	94.49%
Precision	82.97%	80.15%	81.55%	87.21%
F_measure	83.80%	80.96%	82.37%	86.32%
MCC	76.99%	72.68%	74.81%	83.48%
NPV	90.48%	89.07%	89.77%	92.60%
FPR	7.67%	9.11%	8.40%	5.51%
FNR	15.35%	18.22%	16.80%	11.02%
		Dataset 2		
Accuracy	89.02%	87.75%	87.19%	92.11%
Sensitivity	83.52%	81.61%	80.78%	88.17%
Specificity	91.77%	90.81%	90.40%	94.09%
Precision	81.86%	79.99%	79.18%	86.41%
F_measure	82.68%	80.80%	79.97%	85.53%
MCC	75.29%	72.43%	71.18%	82.26%
NPV	89.92%	88.99%	88.58%	92.20%
FPR	8.23%	9.19%	9.60%	5.91%
FNR	16.48%	18.39%	19.22%	11.83%

Table 13. Computational time analysis on Dataset 1 and Dataset 2.

Methods	Computation Time(s) (Dataset 1)	Computation Time (s) (Dataset 2)
DDoS-MSCT	90.8274	89.81647
CNN-LSTM 88.12849	88.12849	86.19174
SVM	86.0024	85.8748
NN	85.4422	84.98201
NB	89.6640	88.80174
KNN	87.9017	86.10127
DT	88.5152	87.2978
AdaBoost-Gradient Boost	93.3274	91.1849
RF	92.91274	90.8217
ISVM+NN	83.01753	82.7216

4.16. Discussion

According to the experimental results, the suggested Hybrid Machine Learning Model for DDoS attack detection under a big data paradigm performs noticeably better than both conventional and some recent classification techniques. This model combines an Improved Support Vector Machine (ISVM) with Neural Networks (NN) and uses an enhanced feature set. The model effectively detects malicious traffic while reducing false alarms, as evidenced by its greatest accuracy of 92.65% and strong performance across other crucial parameters like sensitivity (88.98%) and specificity (94.49%). The integration of the Weighted Exponential Inverse Laplacian (WEIL) kernel in ISVM enhances its ability to model complex, nonlinear relationships, while the hybrid design with NN improves generalization and decision boundaries. The advanced normalization technique combining MAD and Tanh estimation contributes to robustness against outliers, and the use of the MapReduce framework ensures scalability for large-scale network traffic analysis. Compared to baseline models and conventional preprocessing methods, the proposed model shows consistent improvements, reflecting its suitability for real-time, high-volume cybersecurity applications. However, further comparative studies with cutting-edge deep learning methods are necessary to fully establish its position in the current research landscape.

5. Conclusion

In conclusion, this study developed and demonstrated a comprehensive framework for enhancing DDoS attack detection systems by leveraging advanced machine learning models and refined data preprocessing techniques. Beginning with the selection and utilization of a benchmark Dataset that accurately represented various DDoS attack patterns, the research validated the efficacy of the detection model in distinguishing normal network behaviors from anomalous activities associated with DDoS attacks. The preprocessing phase played a crucial role in optimizing data quality, where an Improved Normalization approach effectively scaled and adjusted input data, enhancing the model's sensitivity to subtle deviations indicative of potential attacks. Using the MapReduce framework for feature extraction, the model extracted diverse features and integrated advanced statistical measures, enabling nuanced pattern recognition and accurate classification of network activities. The hybrid machine learning model, combining ISVM and NN classifiers, proved effective in aggregating outputs and robustly identifying specific attack types with higher accuracy and improved generalization across Datasets. This study contributes to advancing cybersecurity by presenting a robust framework that addresses critical challenges in data analysis and attack detection, paving the way for enhanced defenses against sophisticated DDoS threats in network environments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Abbreviations

CNN	Convolutional Neural Network
DDoS	Distributed Denial of Service
DL	Deep Learning
DMN	Deep Maxout Network
DNN	Deep Neural Network
DSA	Deep Stacked Autoencoder
DT	Decision Tree
GB	Gradient Boosting

GHLBO	Gradient Hybrid Leader Optimization
GWO	Grey Wolf Optimization
HLBO	Hybrid Leader-Based Optimization
ISVM	Improved Support Vector Machine
KNN	K-Nearest Neighbour
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
ML	Machine Learning
NN	Neural Network
NB	Naïve Bayes
RF	Random Forest
RNN	Recurrent Neural Network
SCADA	Supervisory Control and Data Acquisition
SDN	Software Defined Networking
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
UASDAC	Unsupervised Adaptive Scalable DDoS Attack Classification
XGBoost	Xtreme Gradient Boosting

References

Adedeji, K. B., Abu-Mahfouz, A. M., & Kurien, A. M. (2023). DDoS attack and detection methods in internet-enabled networks: Concept, research perspectives, and challenges. *Journal of Sensor and Actuator Networks*, *12*(4), 51. https://doi.org/10.3390/jsan12040051

Afolabi, H. A., & Aburas, A. A. (2022). RTL-DL: A hybrid deep learning framework for DDoS attack detection in a big data environment. *International Journal of Computer Networks & Communications*, *14*(6), 51–66. https://doi.org/10.5121/ijcnc.2022.14604

Akgun, D., Hizal, S., & Cavusoglu, U. (2022). A new DDoS attacks intrusion detection model based on deep learning for cybersecurity. *Computers & Security*, *118*, 102748. https://doi.org/10.1016/j.cose.2022.102748

Alhasawi, Y., & Alghamdi, S. (2024). Federated learning for decentralized DDoS attack detection in IoT networks. *IEEE Access*, *12*, 42357–42368. https://doi.org/10.1109/ACCESS.2024.3378727

Ali, M. H., Jaber, M. M., Abd, S. K., Rehman, A., Awan, M. J., Damaševičius, R., & Bahaj, S. A. (2022). Threat analysis and distributed denial of service (DDoS) attack recognition in the internet of things (IoT). *Electronics*, *11*(3), 494. https://doi.org/10.3390/electronics11030494

Alslman, Y., Khalil, A., Younisse, R., Alnagi, E., Al-Saraireh, J., & Ghnemat, R. (2024). DDoS attack-detection approach based on ensemble models using Spark. *Jordanian Journal of Computers and Information Technology*, *10*(2), 1. https://doi.org/10.5455/jjcit.71-1694806966

Awan, M. J., Farooq, U., Babar, H. M. A., Yasin, A., Nobanee, H., Hussain, M., Zain, A. M., & Zain, A. M. (2021). Real-time DDoS attack detection system using big data approach. *Sustainability*, *13*(19), 10743. https://doi.org/10.3390/su131910743

Azmi, M. A. H., Foozy, C. F. M., Sukri, K. A. M., Abdullah, N. A., Hamid, I. R. A., & Amnur, H. (2021). Feature selection approach to detect DDoS attack using machine learning algorithms. *JOIV: International Journal on Informatics Visualization*, *5*(4), 395–401. https://doi.org/10.30630/joiv.5.4.734

Balasubramaniam, S., Vijesh Joe, C., Sivakumar, T. A., Prasanth, A., Satheesh Kumar, K., Kavitha, V., & Dhanaraj, R. K. (2023). Optimization enabled deep learning-based DDoS attack detection in cloud computing. *International Journal of Intelligent Systems*, *2023*(1), 2039217. https://doi.org/10.1155/2023/2039217

Chaudhari, S. S., Sparshika, T. N., Preethi, D., Muttur, C. S., & Shadaksharaiah, A. M. (2024). Sand cat swarm intelligent based random forest approach for DDoS attack detection in IoT network scenario using NS3. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1–6). IEEE.

Chen, J., Tang, X., Cheng, J., Wang, F., & Xu, R. (2020). DDoS attack detection method based on network abnormal behaviour in big data environment. *International Journal of Computational Science & Engineering*, *23*(1), 22–30. https://doi.org/10.1504/IJCSE.2020.110182

Cheng, J., Liu, Y., Tang, X., Sheng, V. S., Li, M., & Li, J. (2020). DDoS attack detection via multi-scale convolutional neural network. *Computers, Materials and Continua*, *62*(3), 1317–1333. https://doi.org/10.32604/cmc.2020.06177

Cui, H., Liu, Q., Zhang, J., & Kang, B. (2019). An improved Deng entropy and its application in pattern recognition. *IEEE Access*, *7*, 18284–18292. https://doi.org/10.1109/ACCESS.2019.2896286

Dasari, K. B., & Devarakonda, N. (2022). Detection of DDoS attacks using machine learning classification algorithms. *International Journal of Computer Network and Information Security*, *14*(6), 89. https://doi.org/10.5815/ijcnis.2022.06.07

Elsaeidy, A. A., Jamalipour, A., & Munasinghe, K. S. (2021). A hybrid deep learning approach for replay and DDoS attack detection in a smart city. *IEEE Access*, *9*, 154864–-154875. https://doi.org/10.1109/ACCESS.2021.3128701

Gaye, B., Zhang, D., & Wulamu, A. (2021). Improvement of support vector machine algorithm in big data background. *Mathematical Problems in Engineering*, *2021*(1), 1–9. https://doi.org/10.1155/2021/5594899

Gumaste, S., Narayan, D. G., Shinde, S., & Amit, K. (2020). Detection of DDoS attacks in OpenStack-based private cloud using Apache Spark. *Journal of Telecommunications and Information Technology*, *4*(2020), 62–71. https://research.unsw.edu.au/projects/unsw-nb15-Dataset

Jiang, Y., Li, Z., Zhang, L., & Sun, P. (2007, June). An improved SVM classifier for medical image classification.

In *International conference on rough sets and intelligent systems paradigms* (pp. 764–773). Springer.

Jordanov, I., Petrov, N., & Petrozziello, A. (2018). Classifiers accuracy improvement based on missing data imputation. *Journal of Artificial Intelligence and Soft Computing Research*, *8*(1), 31–48. https://doi.org/10.1515/jaiscr-2018-0002

Khalifa, A. B., Gazzah, S., & BenAmara, N. E. (2013). Adaptive score normalization: A novel approach for multimodal biometric systems. *International Journal of Computer and Information Engineering*, *7*(3), 376–384.

Khempetch, T., & Wuttidittachotti, P. (2021). DDoS attack detection using deep learning. *IAES International Journal of Artificial Intelligence*, *10*(2), 382. https://doi.org/10.11591/ijai.v10.i2.pp382-388

Kumar, R., Kumar, P., Tripathi, R., Gupta, G. P., Garg, S., & Hassan, M. M. (2022). A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network. *Journal of Parallel and Distributed Computing*, *164*, 55–68. https://doi.org/10.1016/j.jpdc.2022.01.030

Moradibaad, A., & Mashhoud, R. J. (2018). Use dimensionality reduction and SVM methods to increase the penetration rate of computer networks. *arXiv preprint arXiv: 1812.03173*.

Mutholib, A., Rahim, N. A., Gunawan, T. S., & Kartiwi, M. (2025). Trade-space exploration with data preprocessing and machine learning for satellite anomalies reliability classification. *IEEE Access*, *13*, 35903–35921. https://doi.org/10.1109/ACCESS.2025.3543813

Najafimehr, M., Zarifzadeh, S., & Mostafavi, S. (2022). A hybrid machine learning approach for detecting unprecedented DDoS attacks. *Journal of Supercomputing*, *78*(6), 8106–8136. https://doi.org/10.1007/s11227-021-04253-x

Ouyang, S., Liu, Z. W., Li, Q., & Shi, Y. L. (2013). A new improved entropy method and its application in power quality evaluation. *Advanced Materials Research*, *706*, 1726–1733. https://doi.org/10.4028/www.scientific.net/AMR.706-708.1726

Oyucu, S., Polat, O., Türkoğlu, M., Polat, H., Aksöz, A., & Ağdaş, M. T. (2023). Ensemble learning framework for DDoS detection in SDN-based SCADA systems. *Sensors*, *24*(1), 155. https://doi.org/10.3390/s24010155

Patil, N. V., Rama Krishna, C., & Kumar, K. (2020). S-DDoS: Apache Spark based real-time DDoS detection system. *Journal of Intelligent & Fuzzy Systems*, *38*(5), 6527–6535. https://doi.org/10.3233/JIFS-179733

Sahoo, K. S., Tripathy, B. K., Naik, K., Ramasubbareddy, S., Balusamy, B., Khari, M., & Burgos, D. (2020). An evolutionary SVM model for DDoS attack detection in software defined networks. *IEEE Access*, *8*, 132502–132513. https://doi.org/10.1109/ACCESS.2020.3009733

Saravanan, S., & Balasubramanian, U. M. (2024). UASDAC: An unsupervised adaptive scalable DDoS attack classification in large-scale IoT network under concept drift. *IEEE Access*, *12*, https://doi.org/10.1109/ACCESS.2024.3397512.

Seifousadati, A., Ghasemshirazi, S., & Fathian, M. (2021). A machine learning approach for DDoS detection on IoT devices. *arXiv preprint arXiv: 2110.14911*.

Shendi, M. M., Elkadi, H. M., & Khafagy, M. H. (2020). Real-time attacks detection model and platform using big data and machine learning. *International Journal of Scientific and Technology Research*, *9*(9), 108–116.

Siddiqi, M. A., & Pak, W. (2021). An agile approach to identify single and hybrid normalization for enhancing machine learning-based network intrusion detection. *IEEE Access*, *9*, 137494–137513. https://doi.org/10.1109/ACCESS.2021.3118361

Singhal, S., Medeira, P. A., Singhal, P., & Khorajiya, M. (2020). Detection of application layer DDoS attacks using big data technologies. *Journal of Discrete Mathematical Sciences and Cryptography*, *23*(2), 563–571. https://doi.org/10.1080/09720529.2020.1729505

Tan, L., Pan, Y., Wu, J., Zhou, J., Jiang, H., & Deng, Y. (2020). A new framework for DDoS attack detection and defense in SDN environment. *IEEE Access*, *8*, 161908–161919. https://doi.org/10.1109/ACCESS.2020.3021435

Toptaş, B., & Hanbay, D. (2021). Retinal blood vessel segmentation using pixel-based feature vector. *Biomedical Signal Processing and Control*, *70*, 103053. https://doi.org/10.1016/j.bspc.2021.103053

Tuan, T. A., Long, H. V., Son, L. H., Kumar, R., Priyadarshini, I., & Son, N. T. K. (2020). Performance evaluation of botnet DDoS attack detection using machine learning. *Evolutionary Intelligence*, *13*(2), 283–294. https://doi.org/10.1007/s12065-019-00310-w

Ullah, F., Ullah, S., Srivastava, G., & Lin, J. C. W. (2024). IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic. *Digital Communications and Networks*, *10*(1), 190–204. https://doi.org/10.1016/j.dcan.2023.03.008

Wang, B., Jiang, Y., Liao, Y., & Li, Z. (2024). DDoS-MSCT: A DDoS attack detection method based on multiscale convolution and transformer. *IET Information Security*, *2024*(1), 1056705. https://doi.org/10.1049/2024/1056705

Yu, X., Yu, W., Li, S., Yang, X., Chen, Y., & Lu, H. (2021). Web DDoS attack detection method based on semisupervised learning. *Security and Communication Networks*, *2021*(1), 1–10. https://doi.org/10.1155/2021/9534016