Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 15, Issue. 1, January 2026, pg.66 – 72

Ensemble Learning Based Detection of Mirai Botnet Attacks using Simulated, Generative AI and Hybrid Network Traffic

P.T.Goyal; K.G. Kharade; Kavita S. Oza

[1,2]Assistant Professor, Department of Computer Science, Shivaji University, Kolhapur, India
[3]Associate Professor, Department of Computer Science, Shivaji University, Kolhapur, India
prasad.goyal@gmail.com; kgk_csd@unishivaji.ac.in; kso_csd@unishivaji.ac.in

DOI: https://doi.org/10.47760/ijcsmc.2026.v15i01.006

Abstract: The Mirai botnet is one of the most damaging malware families targeting Internet of Things (IoT) devices. By manipulating weak login ID's and unpatched vulnerabilities, Mirai converts compromised devices into large scale botnets accomplished to launch distributed denial-of-service attacks. This paper presents a complete study & evaluation of multiple machine learning classifiers such as Random Forest, K-Nearest Neighbors (KNN), Naïve Bayes and XGBoost for detecting Mirai botnet traffic. The models are assessed on three datasets namely real Mirai traffic, Generative AI based synthetic traffic, and a hybrid dataset combining both. Performance is measured using accuracy, precision, recall, and F1-score. The results determine that ensemble models such as Random Forest and XGBoost reliably achieve superior performance, mainly when trained on the hybrid dataset, making them extremely appropriate for real world IoT botnet detection systems.

Keywords: AdaBoost, Ensemble Learning, Generative AI, Hybrid Dataset, Mirai Botnet, Random Forest, XGBoost

1. INTRODUCTION

The rapid growth of IoT has transformed industries and daily life; though, it has also introduced significant security weaknesses. IoT devices do not have strong authentication and update mechanisms, making them perfect targets for malware attacks. Among all IoT based malware types, Mirai stands out as one of the most unsafe due to its capacity to form huge botnets that launch devastating DDoS attacks [1].

Network attacks like denial of service and spoofing may also have an effect on devices connected to the internet. Web apps and other software used by IoT devices might be oppressed due to security loopholes. Internet of

© 2026, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

Things device deployment in important settings raises the stakes for cybercrime. If the IoT devices are not sufficiently protected, a cyberattack on these very important facilities might have disastrous consequences. The machines in a botnet are all infected with malware, and they're all linked together to be managed by one or more command and control servers [2].

The constantly changing threat landscape is too much for traditional systems that rely on rule-based methods or basic machine learning (ML) algorithms to handle. Because of this, these systems have a high rate of miss detections and false positives, which can be significant. To improve the precision, effectiveness, and scalability of the battle against the Mirai botnet's activity, novel detection approaches utilizing cutting-edge machine learning techniques are crucial [3].

The goal of the research is to solve the difficulties in identifying activity on the Mirai botnet by offering a unique detection approach based on XGBoost, a sophisticated ensemble learning algorithm [4].

The main goal is to increase the detection system's accuracy, precision, recall, and overall performance metrics while lowering the amount of false positives and false negatives [5].

2. Literature Review

The development of new technologies that improve the efficiency and accessibility of everyday tasks is made possible by the advancement of computer resources. IoT network implementation was made easier by these new technologies. Nevertheless, practically all IoT devices are still vulnerable to hackers due to a lack of security measures. The Mirai Botnet is one of the most dangerous attacks on IoT systems; it is caused by malware that turns connected consumer devices into a botnet for DDoS attacks [6].

Finding IoT bots infected with the Mirai virus is the study's main goal. Monitoring the packet traffic produced by IoT devices over a predetermined amount of time is the basis for mirai detection. Additionally, the suggested approach helps identify the precise trait of Mirai in addition to other viruses of the same type. The main focus is on having an encryption solution to prevent potential attacks on IoT devices [7].

Another main challenge in IoT security research is the incomplete availability of large, labeled, and diverse datasets. While real world datasets capture real attack behavior, they are expensive to collect and are often short of variety. Synthetic datasets generated using Generative AI can extent easily but may not fully stand for real attack characteristics. Hybrid datasets combines real and synthetic traffic shows potential solution by getting better data variety and generalization.

To address the problems of Mirai botnet a variety of supervised and unsupervised machine learning techniques are employed. A hybrid method for detecting botnet attacks is suggested. It helps identify new inbound threats in the traffic by utilizing both controlled and uncontrolled machine learning approaches. It first constructs the technique to achieve that. The results of that strategy should then be entered into a system that separates the data into attack and benign categories [8].

One of the several attacks that target IoT devices is the Mirai botnet, which targets routers and switches, surveillance, web recording devices, and other IoT devices. People have even experienced internet disruptions as a result. Because of this, it is believed to be essential to have a system that uses the SVM model to identify Mirai botnet attacks. When the training dataset's parameter is adjusted correctly, the system can identify Mirai botnet attacks even in the absence of a large amount of training data and produce precise classification results [9].

The study examines recent advancements in machine learning based botnet identification, which can recognize patterns and automatically discern between harmful and benign activity. The findings demonstrate that machine learning can use network data byproduct to identify a variety of botnets [10].

Despite these efforts, limited work has systematically evaluated multiple classifiers across real, generative, and hybrid datasets for Mirai detection. This study addresses this gap.

3. Methodology

This section displays on the whole experimental framework implemented for detecting Mirai botnet attacks using machine learning models on real, generative and hybrid IoT network traffic. The methodology includes dataset preparation, preprocessing, feature engineering, model training, validation and performance evaluation.

Overall Framework

The proposed methodology follows a methodical pipeline consisting of the following main stages:

© 2026, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

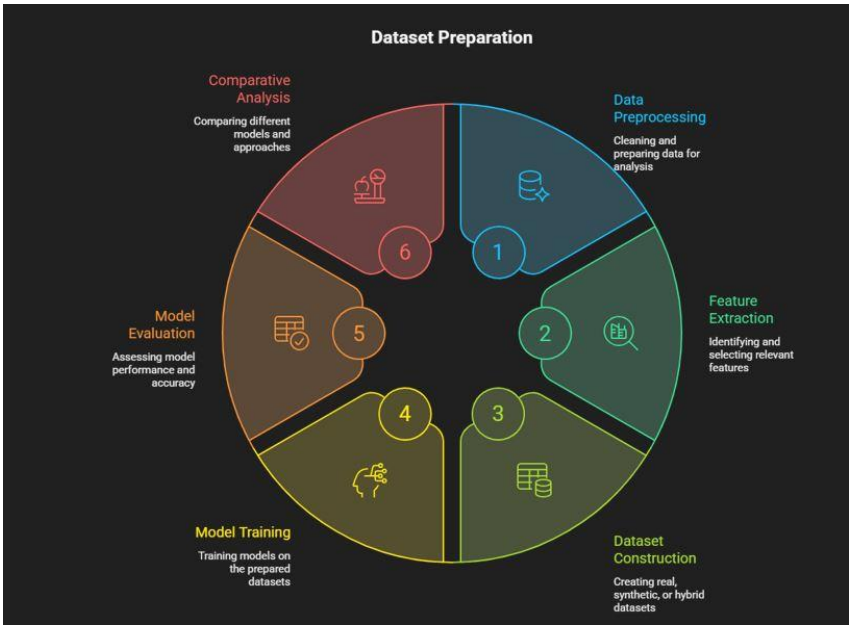


Fig 1: Dataset Preparation

This planned approach ensures reasonable evaluation of all classifiers and dependable comparison of results.

3.1 Dataset Preparation for Simulated Mirai Dataset

During the Mirai dataset generation phase, network traffic exchanged between IoT devices, and the surrounding network environment was captured over a defined time. This traffic included both legitimate communication from normal IoT devices and malicious communication originating from Mirai infected devices. The captured traffic covered commonly used network protocols such as TCP and UDP, which are frequently exploited by botnets for command and control and attack activities. Wireshark records the captured traffic in PCAP (Packet Capture) format. These PCAP files store detailed packet-level information such as source and destination IP addresses, port numbers, packet sizes, timestamps, and protocol types. Such detailed data provides a realistic view of how network traffic behaves under normal conditions as well as during botnet attacks.

3.2 Generative AI Dataset

The dataset contains synthetically generated IoT traffic that imitate real world communication & also includes noise, randomness along with evolving patterns. It simulates realistic but volatile network behavior which makes detection more challenging.

3.3 Hybrid Dataset

The hybrid dataset combines both real Mirai traffic with synthetic traffic, creating a more diverse and realistic dataset. This setup improves model generalization and avoids overfitting to only real or only synthetic patterns.

3.4 Data Preprocessing

Before model training, the datasets were preprocessed to improve data quality and learning effectiveness. Data preprocessing is applied for cleaning and making the data suitable for the training of the ML classifier. Frequently, data contain some missing, invalid and infinite variables. These values should be removed or treated to make data available for an algorithm that can be trained on the data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre processing task such as Importing libraries, datasets, finding missing data, encoding categorical data, splitting dataset into training and test set.

© 2026, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

3.4.1 Data Cleaning

It involves elimination of duplicate records, Handling missing values using suitable imputation, also elimination of inconsistent and corrupted samples.

3.5 Feature Normalization

To stay away from control of high magnitude features, all numeric features were normalized using Min Max Scaling to ensure that each feature contributes evenly to model learning.

3.5.1 Label Encoding

Categorical attributes such as protocol types were converted into numerical values using Label encoding. This enables machine learning algorithms to process categorical data effectively.

3.6. Feature Extraction and Selection

Suitable flow based features were extracted from raw network traffic as shown in below figure.

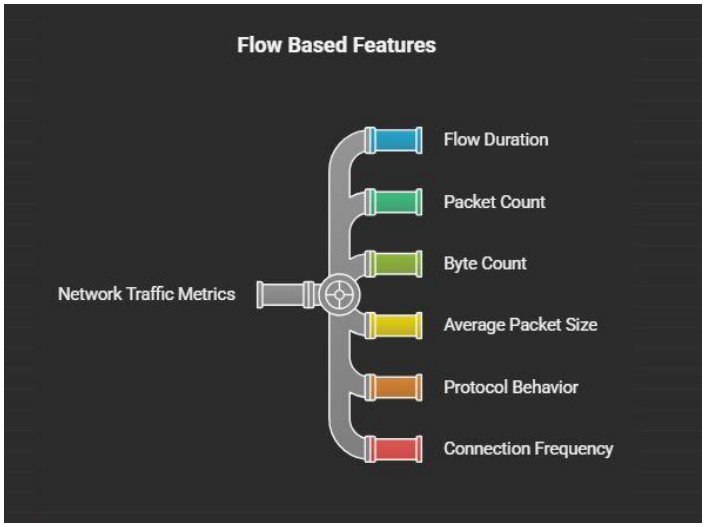


Fig 2: Flow based features

Unnecessary and inappropriate features were removed using Correlation analysis.

This step reduced dimensionality and enhanced computational efficiency without reducing accuracy.

3.7. Construction of Training and Testing Sets

Dataset of Real, Synthetic, Hybrid was divided into:

- Training set:70%
- Testing set: 30%

The split ensures sufficient training data, unbiased performance evaluation. Stratified sampling was applied to maintain class balance across training and testing sets.

3.8. Machine Learning Model Training

Six machine learning classifiers were implemented and evaluated:

Random Forest (RF)

It is an collection of decision trees trained with bootstrap sampling and random feature selection. Hyperparameters such as n_estimators=100 was used. Random Forest (96.85%) attained the highest performance due to its ensemble nature which captures complex feature connections and reduces overfitting.

K-Nearest Neighbors (KNN)

A distance based classifier that allocates class labels based on majority voting among nearest neighbors. Hyperparameters such as n_neighbors=5 was used. KNN (96.7%) performed well because Mirai attacks form

© 2026, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

thick clusters in feature space.

Naïve Bayes (NB)

A probabilistic classifier based on Bayes' theorem assuming feature independence.

It is computationally capable and performs fine on noisy data. Naïve Bayes (96.5%) showed reasonable results since Mirai traffic display strong statistical regularities.

Support Vector Machine (SVM)

A margin based classifier that constructs optimal separating hyperplanes between attack and normal traffic. SVM (96.68%) showed even performance by maximizing class separation margins.

XGBoost

A gradient boosting algorithm that builds trees sequentially to minimize classification errors. Hyperparameters such as eval_metric='logloss' was used to achieve 96.82% accuracy.

AdaBoost

An adaptive boosting technique that emphasizes misclassified samples in consecutive iterations.All models were trained using the same preprocessing pipeline to make sure fairness. AdaBoost (96.68%) performed well but was somewhat affected by overlapping feature distributions.

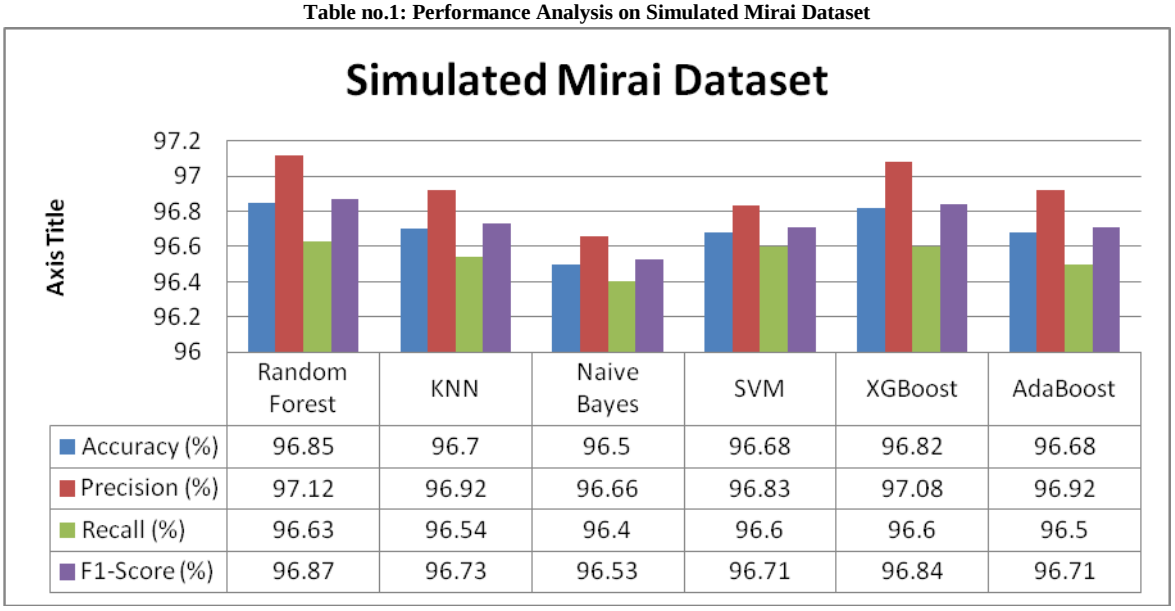
Model Evaluation Metrics

To evaluate performance, the following metrics were used namely: Accuracy: General correctness, Precision – Attack detection performance, Recall – Ability to capture all attacks, F1-Score: Balancing between precision and recall. These are very important in security systems, where failing to detect an attack (false negative) can be terrible.

All models were trained and tested using identical datasets and splits, Hyperparameters were tuned using cross validation, Each experiment was repeated multiple times and averaged, The same evaluation metrics were applied across all datasets. This ensures reproducibility and fair comparison.

4. Results on Mirai Dataset

Table no.1: Performance Analysis on Simulated Mirai Dataset



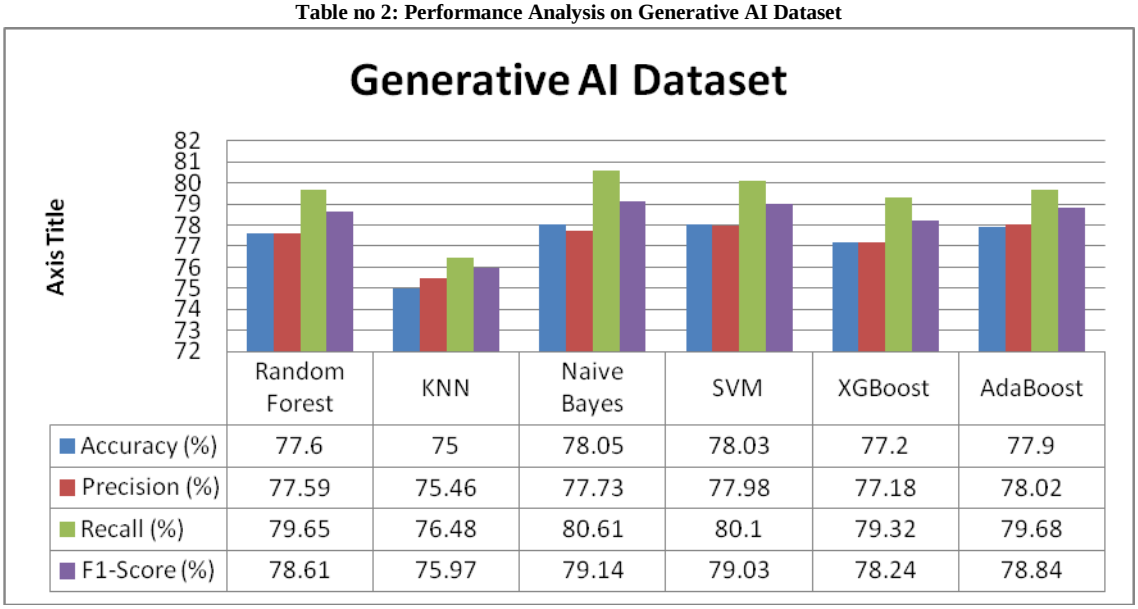
	Random Forest	KNN	Naive Bayes	SVM	XGBoost	AdaBoost
Accuracy (%)	96.85	96.7	96.5	96.68	96.82	96.68
Precision (%)	97.12	96.92	96.66	96.83	97.08	96.92
Recall (%)	96.63	96.54	96.4	96.6	96.6	96.5
F1-Score (%)	96.87	96.73	96.53	96.71	96.84	96.71

© 2026, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

4.1 Results on Generative AI Dataset

In Generative AI Dataset Performance dipped across all models due to noise and randomness. Naïve Bayes and SVM handled uncertainty better because probabilistic reasoning and margin maximization are robust to noise. Random Forest and XGBoost struggled due to inconsistent attack signatures. KNN suffered the most because distance measures become undependable in noisy feature space. AdaBoost was affected because boosting expands noisy misclassified samples.

Table no 2: Performance Analysis on Generative AI Dataset

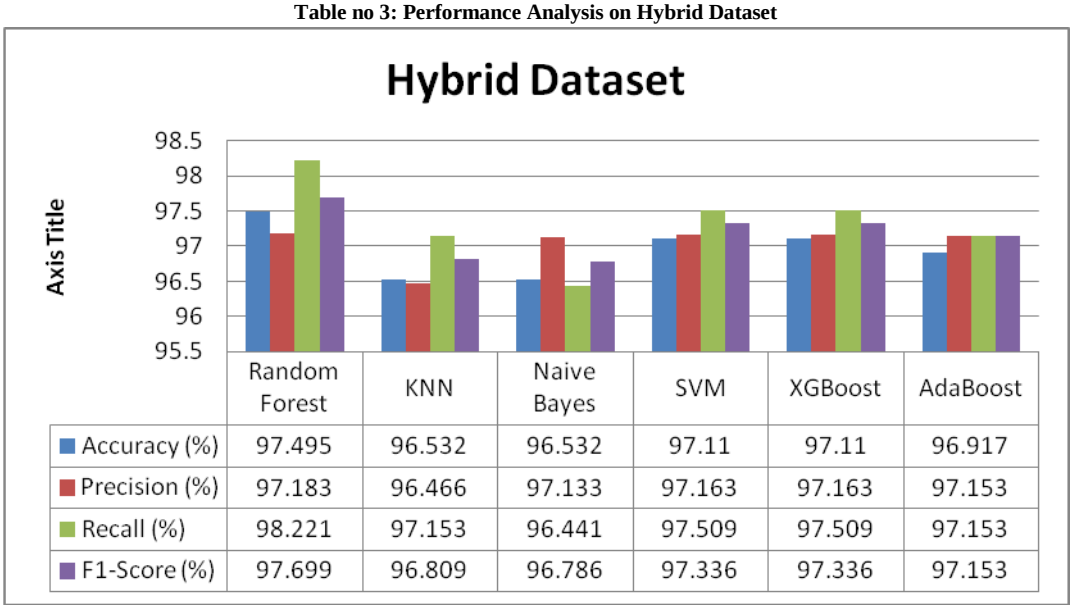


	Random Forest	KNN	Naive Bayes	SVM	XGBoost	AdaBoost
Accuracy (%)	77.6	75	78.05	78.03	77.2	77.9
Precision (%)	77.59	75.46	77.73	77.98	77.18	78.02
Recall (%)	79.65	76.48	80.61	80.1	79.32	79.68
F1-Score (%)	78.61	75.97	79.14	79.03	78.24	78.84

4.2 Results on Hybrid Dataset

All models showed major improvement. Random Forest attained the highest accuracy due to introduction to both real and synthetic patterns. XGBoost and SVM also performed robustly along with improved generalization. KNN fared better as hybrid data created clearer neighborhoods. Naïve Bayes benefited from better class probability estimation. AdaBoost improved because hybrid data reduced extreme noise.

Table no 3: Performance Analysis on Hybrid Dataset



	Random Forest	KNN	Naive Bayes	SVM	XGBoost	AdaBoost
Accuracy (%)	97.495	96.532	96.532	97.11	97.11	96.917
Precision (%)	97.183	96.466	97.133	97.163	97.163	97.153
Recall (%)	98.221	97.153	96.441	97.509	97.509	97.153
F1-Score (%)	97.699	96.809	96.786	97.336	97.336	97.153

Mirai dataset gave the highest performance due to structured attacks. Generative dataset was toughest due to noise.Hybrid dataset achieved the best balance between realism and diversity. Random Forest and XGBoost were the most steady across all datasets.

© 2026, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

5. Conclusion

This research demonstrates that ensemble learning techniques, particularly Random Forest (96.85%) and XGBoost (96.82%) are very efficient for Mirai botnet detection as compared to other algorithms. The use of hybrid datasets considerably improves detection accuracy and generalization. Generative AI dataset alone is not enough for training reliable intrusion detection systems, however when combined with real data it enhances robustness and adaptability. These results strongly support the implementation of ensemble learning with hybrid datasets for scalable IoT security solutions. Future work may determine deep learning architectures, online learning & real time deployment.

References

[1]. Hussain, F., Abbas, S. G., Fayyaz, U. U., Shah, G. A., Toqeer, A., & Ali, A. (2020). Towards a universalfeatures set for IoT botnet attacks detection. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-6). IEEE.

[2]. Injadat, M., Moubayed, A., &Shami, A. (2020). Detecting botnet attacks in IoT environments: An optimized machine learning approach. In 2020 32nd International Conference on Microelectronics (ICM) (pp. 1-4). IEEE

[3]. A. R. S. A. Cruz, R. L. Gomes, and M. P. Fernández, "An Intelligent Mechanism to Detect Cyberattacks of Mirai Botnet in IoT Networks," 2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS), Jul. 2021, doi: 10.1109/dcoss52077.2021.00047.

[4]. A. Woodiss-Field, M. N. Johnstone, and P. Haskell- Dowland,"Towards evaluating the effectiveness of botnet detection techniques," in Communications in computer and information science, 2022, pp. 292–308. doi: 10.1007/978-981-19-0468-4_22.

[5]. M. a. A. Da Cruz, L. R. Abbade, P. Lorenz, S. Mafra, and J. J.P. C. Rodrigues, "Detecting compromised IoT devices through XGBoost," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 15392–15399,Dec. 2023, doi: 10.1109/tits.2022.3187252.

[6]. M. Nakıp and E. Gelenbe, "MIRAI Botnet Attack Detection with Auto-Associative Dense Random Neural Network," 2021 IEEE Global Communications Conference (GLOBECOM),Dec. 2021, doi: 10.1109/globecom46510.2021.9685306.

[7]. S. Das, P. P. Amritha, and K. Praveen, "Detection and Prevention of Mirai Attack," Soft Computing and Signal Processing, pp. 79–88, Jan. 2021, doi: 10.1007/978-981-33-6912-2_8.

[8]. M. G. Desai, Y. Shi, and K. Suo, "A hybrid approach for IoT botnet attack detection," 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Oct. 2021, doi:10.1109/iemcon53756.2021.9623102.

[9]. R. G. Azhari, V. Suryani, R. R. Pahlevi, and A. A. Wardana,"The Detection of Mirai Botnet Attack on the Internet of Things (IoT) Device Using Support Vector Machine (SVM) Model," 2022 10th International Conference on Information and Communication Technology (ICoICT), Aug. 2022, doi: 10.1109/icoict55009.2022.9914830.

[10]. M. Alauthman, "Botnet Detection Techniques: A Comparative Study," Version 1: Received: 5 November 2023/ Approved: 6 November 2023 / Online: 6 November 2023(08:17:49 CET), Nov. 2023, doi: 10.20944/preprints202311.0311.v1.

© 2026, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden