International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol.2, No.2, June 2023, pp. 11~20
DOI: 10.59461/ijdiic.v2i2.58

A Big Data Analytical Framework for Intrusion Detection Based On Novel Elephant Herding Optimized Finite Dirichlet Mixture Models

V. Suresh Kumar [1]
[1]Department of Information Technology, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College,
Avadi, Chennai, India

Article Info

Article history:

Received May 19, 2023
Revised June 14, 2023
Accepted June 16, 2023

Keywords:

Big Data
Cybersecurity
Intrusion Detection System (IDS)
Elephant Herding Optimized Finite Dirichlet Mixture Models (EHO-FDMM)

ABSTRACT

For the purpose of identifying a wide variety of hostile activity in cyberspace, an Intrusion Detection System (IDS) is a crucial instrument. However, traditional IDSs have limitations in detecting zero-day attacks, which can lead to high false alarm rates. To address this issue, it is crucial to integrate the monitoring and analysis of network data with decision-making methods that can identify anomalous events accurately. By combining these approaches, organizations can develop more effective cybersecurity measures and better protect their networks from cyber threats. In this study, we proposed a novel called the Elephant Herding Optimized Finite Dirichlet Mixture Model (EHO-FDMM). This framework consists of three modules: capture and logging, pre-processing, and an innovative IDS method based on the EHO-FDMM. The NSL-KDD and UNSW-NB15 datasets are used to assess this framework's performance. The empirical findings show that selecting the optimum model that accurately fits the network data is aided by statistical analysis of the data. The EHO-FDMM-based intrusion detection method also offers a lower False Alarm Rate (FPR) and greater Detection Rate (DR) than the other three strong methods. The EHO-FDMM and exact interval of confidence bounds were used to create the suggested method's ability to detect even minute variations between legal and attack routes. These methods are based on correlations and proximity measurements, which are ineffective against contemporary assaults that imitate everyday actions.

Corresponding Author:

V. Suresh Kumar
Department of Information Technology
Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College
Avadi, Chennai
India
Email: sureshkumar@veltechmultitech.org

1. INTRODUCTION

The field of Data Science (DS) often uses advanced analytical methods and scientific concepts to draw useful commercial insights from data. By analyzing data to identify patterns and make predictions about what is expected to happen, advanced analytics puts a greater emphasis on forecasting future occurrences. Advanced analytics go beyond basic analytics by offering a deeper understanding of data and helping with the examination of detailed data, whereas basic analytics just give a general description of data, which is what we are interested in [1]. Modern life is increasingly influenced by networks, making cybersecurity a crucial area of study. Anti-virus software firewalls, and Intrusion Detection Systems (IDSs) are the key cyber security tools. These methods defend networks against both internal and external intrusions. An IDS is one of these detection systems that are crucial in ensuring cyber security by keeping track of the hardware and software configurations inside a network [2]. In 1980, the first IDS was suggested. Many mature IDS products have now emerged. However, a lot of IDS continue to emit alerts for low-threat situations often because of their high false alarm rate. This makes security analysts' workloads heavier and raises the possibility that very damaging assaults would go unnoticed. IDSs with greater rates of detection and fewer false alarms have been developed as a consequence of the extensive study. Existing IDSs also have the drawback of being unable to recognize unidentified assaults. Fast-changing network settings result in a steady emergence of new attack types. Therefore, it is essential to create IDSs that can recognize unidentified assaults [3-4].

As a result, cybersecurity is quickly overtaking other pressing problems in contemporary society. Monitoring and analysis of network traffic data are essential for detecting likely attack trends. Worldwide businesses and IT companies have been putting money into data science to create more sophisticated IDSs that can prevent damaging assaults and provide higher cybersecurity [5]. To analyze, display, and derive insights that might help forecast and halt cyber attacks, big data analytics in security requires the capacity to collect enormous volumes of digital data. It improves our cyber defense posture together with security technology. This idea encompasses a variety of techniques from the domains of computers, statistics, and data technological equipment, including the well-known Machine Learning (ML) technique [6]. However, because of the enormous amount of heterogeneous big data produced by several sources, standard data analytics and shallow ML approaches are worthless and ineffective in dealing with such security risks directly. Notably, classical ML approaches struggle with processing complexity and latency and may be unable to comprehend the complex and time-varying non-linear relationships seen in huge datasets [7]. The main contributions of the paper are,

- To evaluate this technique's dependability for detecting intrusion, we compare it with three other ways and utilize two benchmark datasets: UNSW-NB15 and NSL-KDD. Then, we use z-score normalization for pre-processing the data.
- We also develop a novel EHO-FDMM based on intrusion detection to efficiently detect harmful events in this framework.

The following are the other portions of the study: Pertinent Studies are provided in part 2, the technique is introduced in part 3, the results and discussion are presented in part 4, and the article is concluded in the last part.

2. RELATED WORKS

The question of whether the CRoss-Industry Standard Process for Data Mining (CRISP-DM) is still appropriate for use in data science projects was examined in the publication [8]. They contend that the process model approach still mainly holds if the project is goal-directed and process-driven. However, as DS initiatives get more experimental, the potential directions they might go down become more diversified, necessitating a more adaptable paradigm. The field of supply chain management (SCM) is paying an increasing amount of attention to big data analytics (BDA). The purpose of the investigation [9] was to suggest a categorization of these predicted BDA implications for supply chain demand projections, identify the holes, and give recommendations for future research. Because of various constraints, typical IDS techniques need to be updated and enhanced before they can be used in the Internet of Things (IoT). These constraints include resource-constrained devices, the restricted memory and battery capacity of nodes, and a specialized protocol stack. A lightweight attack detection technique that uses a supervised ML-based Support Vector Machine (SVM) was created in the research [10] to identify an opponent who was trying to inject extraneous data into the IoT network. The method of discovering hostile activity in a network by examining the behavior of network traffic was created in the research [11], referred to as the approach known as network intrusion detection. To identify abnormalities, IDS often makes use of data mining methods. Because spotting abnormalities in high-dimensional network traffic features is a laborious operation, IDS relies heavily on dimensionality reduction as a key component. The study [12], proposed Passban, an intelligent intrusion detection system that can safeguard the Internet of Things devices that are directly linked to it. The suggested system is unique in that it can be installed directly on extremely inexpensive IoT gateways. As a result, it takes full use of the edge computing paradigm to identify cyber risks. IDS are among the most reliable options, particularly those that were developed with the assistance Of Artificial Intelligence (AI). An artificially fully automated IDS for fog security against cyberattacks was proposed in the study [13]. Recurrent neural networks (RNNs) with many layers are used in the proposed model to provide security for fog computing that is located very close to end users and IoT devices. Using collaborative learning and feature selection, the innovative IDS architecture was developed in the research [14]. The CFS-BA heuristic method, which chooses the best subset based on the relationship between features, is presented as the initial stage for dimensionality reduction. In the research [15], the Variational Long Short-Term Memory (VLSTM) approach to learning for skilled discrepancy detection using a reconstituted depiction of features was introduced to address the discrepancy between dimensionality reduction and feature retention in unbalanced Industrial Big Data (IBD).

3. METHOD

This section explains the recommended technique for utilizing the EHO-FDMM to develop efficient IDS, as well as the mathematical aspects of data modeling and estimation utilizing the DMM. Figure 1 depicts the structure of EHO-FDMM.

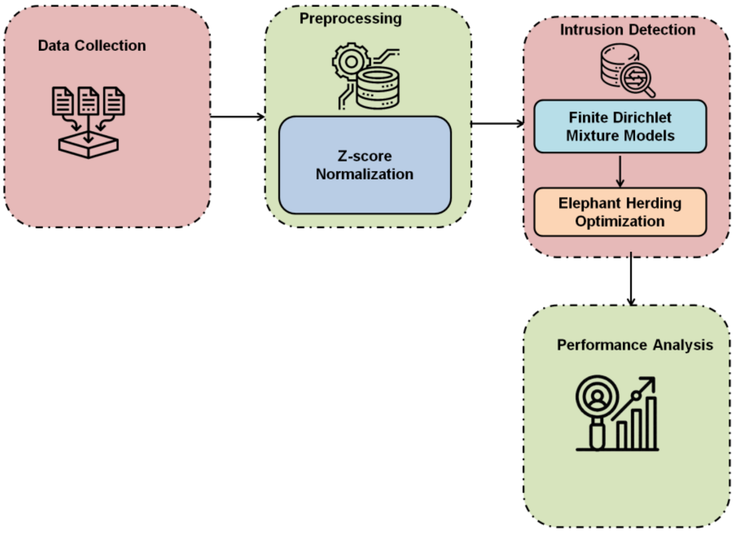

Figure 1. Structure of proposed method

3.1. Dataset

For analyzing the effectiveness of the suggested methodologies, several standalone databases have been obtained using a variety of normal and malicious records, including the KDD CUP 99, NSL-KDD, and UNSWNB15. The KDD CUP 99 dataset has been enhanced using the NSL-KDD dataset. To prevent each classifier from favoring the records with the highest frequency, redundant records were removed from the training and testing sets in the KDD CUP 99 dataset. The NSL-KDD dataset comprises 41 characteristics and a class label for each record, much like this dataset.

The UNSW-NB15 dataset combines real, recent recordings of attacks and normal behavior. Its network packets have a size of 1,450,133 records and have been stored in four CSV files totaling around 100 Gigabytes. Each investigation has 58 characteristics, along with a class name that highlights its high dimensionality diversity. With an average velocity of 5 to 10 MB/s between sources and destinations, it allows for larger data rate transfers via Ethernets, perfectly simulating actual network situations [16].

3.2. Pre-processing using z-score normalization

Z-score normalization is a common pre-processing technique used in intrusion detection systems to standardize the scale of input data. It involves subtracting the mean value of the data and dividing it by the standard deviation. The resulting data has a mean of zero and a standard deviation of one, which makes it easier to compare and analyze.

The equation for z-score normalization is as follows:

$$z = (x - \mu)/\sigma \tag{1}$$

Where z is the standardized value, x is the original value, μ is the mean of the data, and σ is the standard deviation of the data.

To use z-score normalization in an intrusion detection system, the first step is to calculate the mean and standard deviation of the training data for each feature. Then, for each new data point, the z-score is calculated using the above equation. If the z-score falls outside a certain threshold, the system raises an alarm indicating a potential intrusion.

3.3. Intrusion detection using Finite Dirichlet Mixture Model (FDMM)

FDMM is a probabilistic model that assumes that the data comes from a finite number of unknown Gaussian distributions. The number of distributions is not known a priori, but it can be estimated from the data. The model assigns each data point to one of the Gaussian distributions with a probability proportional to its likelihood under that distribution. In intrusion detection, FDMM can be used to detect anomalies in network traffic. To teach the model the features, it may be trained using an inventory of typical network activity of legitimate traffic. Once the model is trained, it can be used to classify new network traffic as either normal or anomalous.

The finite mixture model is an effective and adaptable probabilistic modeling network data that may be viewed as a convex amalgamation of two or more Probability Density Functions (PDFs), whose combined features can roughly mimic any random distributions. Figure 2 depicts a finite mixture of K-component Dirichlet distributions, and it is denoted by

$$(W|\pi,\alpha) = \sum_{j=1}^{L} \pi_j Cjq\ (W|\alpha_j) \tag{2}$$

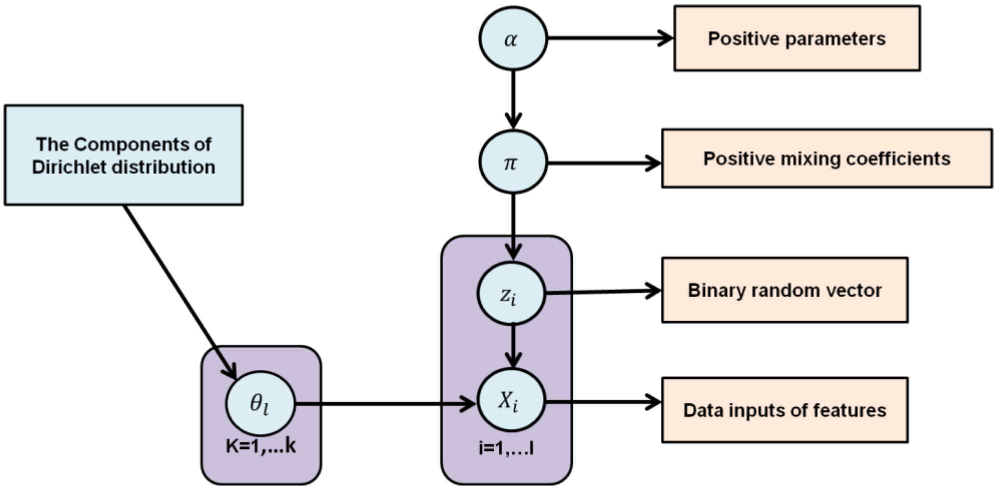

Figure 2. Finite mixture model

Where $\pi = (\pi_1, \dots, \pi_L)$denotes the good summing of the favorable mixing factors 1, $\sum_{j=1}^{L} \pi_{j,}\alpha = (\alpha_1, \dots . \alpha_L), and\ Cjq\ (W|\alpha_j)$ denotes element j's own optimistic values for the Dirichlet dispersion $(\alpha = (\alpha_{j1,\dots\dots}\alpha_{jT}))$as

$$Cjr(W|\alpha_j) = \frac{\Gamma(\sum_{t=1}^{T} \alpha_{jt})}{\prod_{T=1}^{T} \Gamma(\alpha_{jt})} \prod_{t=1}^{T} W_T^{\alpha_{jt-1}} \tag{3}$$

Where $W = (W_1, \dots W_T)$, $\sum_{T=1}^{T} w_T = 1{,}0 \leq W_T \geq 1$ for $W = 1, \dots \dots W$ and, T for the dimension of W. It is important to note the use of a Dirichlet probability as a parent distributed rather than as a before the roles that require to directly represent the data.

If we suppose that the mixed distribution in Equation (2) created a collection of N distinct, equally dispersed (j, j, c) vectors $(W = \{W_{1,\dots\dots}W_M)$, the possible function of the FDMM is

$$O(W|\pi,\alpha) = \prod_{k=1}^{M} \{ \prod_{j=1}^{L} \Pi_j\ Cjq\ (W_k|\alpha_j)\} \tag{4}$$

The latent variable model in Equation (2) is the finite mixture model. As a result, we create a K-dimensional binary random vector $(Yj = \{Y, \dots \dots Y_{jL}\})$ for each vector (W_j), with $Y_{jt} \in \{0,1\}, \sum_{j=1}^{L} and\ Y_{jt} = 1\ if\ \ W_j$ belongs to element j, otherwise 0. The distribution under the condition of Y given combination coefficients $(\pi)$is defined as follows for the latent variables $(Y = \{Y_{1,\dots\dots}Y_M\})$, which are essentially concealed variables that are not mentioned directly in the model.

$$O(Y|\pi) = \prod_{k=1}^{M} \prod_{j=1}^{L} \pi_j^{Y_{kj}} \tag{5}$$

The consequent allocation of a dataset K given the class labels, or the probability functional with latent factors, which may therefore be represented as

$$O(W|\pi,\alpha) = \prod_{k=1}^{M} \prod_{j=1}^{L} Cjq(W_k|\alpha_j) \tag{6}$$

The process of acquiring knowledge of the combination variables, which involves both predicting the settings and choosing the number of elements (L), is a significant issue given some data K and a collection of characteristics C.

The FDMM approach has several advantages over other intrusion detection methods. First, it can detect both known and unknown types of attacks. Second, it can adapt to changing network traffic patterns over time. Third, it has a low false positive rate, which means that it is less likely to classify normal traffic as anomalous. As result, FDMM is a powerful and flexible model that can be used for intrusion detection. Its ability to detect both known and unknown types of attacks and its low false positive rate make it a promising approach for securing computer networks.

3.4. Elephant Herding Optimization (EHO)

EHO is a recent optimization algorithm that mimics the behavior of elephants in a herd. In a herd, elephants cooperate and communicate with each other to achieve a common goal, such as finding food or water. EHO algorithm is based on this concept of cooperation and communication among individuals to solve optimization problems.

In the fundamental EHO algorithm, the separation operation is executed after the update functioning, which establishes the search orientation and local search detail level of the method. The group updating operations and the division operation are the two stages of this procedure.
Establish the elephant community at random, and then split it into n clans, with j elephants living in each group. The location of each elephant in each iteration is specified by Equation (7).

$$w_{new}, dj, i = w_{dj,i} + \alpha . \left(w_{best,dj-} w_{dj,i}\right) . q \tag{7}$$

Equation (8) is used to determine the role of the female matriarch $(, w_{best,dj})$. Equation (9) is used to define the elephant group's nucleus.

$$w_{new}, dj, i = \beta \times w_{center,dj} \tag{8}$$

$$w_{center,dj,c=\frac{1}{m_{dj}} . \sum_{i=1}^{m_{dj}} w_{dj,i,c}} \tag{9}$$

The definition of Equation (10) is the elephant position with the poorest fitness value.

$$w_{worst,dj} = w_{min} + (w_{max} - w_{min} + 1) \times rand \tag{10}$$

EHO is an optimization method that uses the behavior of elephants in a herd to guide the search for the optimal solution. The algorithm is based on the concept of cooperation and communication among individuals, and it is effective in solving a wide range of optimization problems.

3.5. Elephant Herding Optimized-Finite Dirichlet Mixture Model (EHO-FDMM)

The concept of using an Elephant Herding Optimized Finite Dirichlet Mixture Model (EHO-FDMM) in an intrusion detection system involves using an ML algorithm to identify and classify patterns of behavior in network traffic that may indicate an attempted intrusion or attack. The EHOFDMM is a variant of the Dirichlet Mixture Model (DMM), which is a probabilistic model used in machine learning for clustering and classification tasks. The EHO-FDMM uses a swarm intelligence algorithm inspired by the behavior of elephant herds in nature to optimize the DMM's performance.

The EHO-FDMM in intrusion detection can be formulated mathematically using the following equation:

$$q(w|\theta) = \sum l = 1 \wedge L x_l \; q(w|\theta_l) \tag{11}$$

Where $q(w|\theta)$ represents the probability of observing network traffic data w given the model parameters θ, L is the number of mixture components, x_l represents the weight of the l^{th}mixture component, and $q(w|\theta_l)$ represents the probability of observing data w given the parameters of the l^{th} mixture component.

The EHOFDMM optimizes the values of the parameters θ and the weights x_l using a swarm intelligence algorithm inspired by the behavior of elephant herds in nature. This algorithm iteratively adjusts the values of θ and x_l to maximize the likelihood of observing the network traffic data.

The EHO-FDMM can be used in conjunction with other intrusion detection techniques, such as signature-based detection and anomaly-based detection, to provide a more comprehensive and accurate approach to intrusion detection. By leveraging the power of ML and swarm intelligence, the EHO-FDMM has the potential to improve the effectiveness of network security measures and detect new and emerging threats in real-time

4. RESULTS AND DISCUSSION

The datasets utilized to evaluate the proposed approach are covered in this part, followed by the evaluation measures that were used to compare the effectiveness of the suggested strategy in comparison to other methods. The analysis of the variance of the traits selected from the NSL-KDD and UNSW-NB15 datasets is given in this section.

4.1. Analysis of Performance

The proposed EHO-FDMM-based IDS technique was evaluated in several experiments on the two datasets using external evaluation metrics, such as accuracy, Detection Rate (DR), and False Positive Rate (FPR), which depend on the four terms true positive (TP), true negative (TN), false negative (FN), and false positive (FP). The number of real records classed as assaults are denoted by the letters TP, normal records are denoted by the letters TN, attack records are denoted by the letters FN, and normal recordings are denoted by the letters FP. Following is a definition of these measures.

Accuracy is measured as the proportion of all normal and attack records that are properly categorized, or more specifically in Equation (12). Figure 3 and Table 1 compare the EHO-FDMM accuracy with other current techniques.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{12}$$

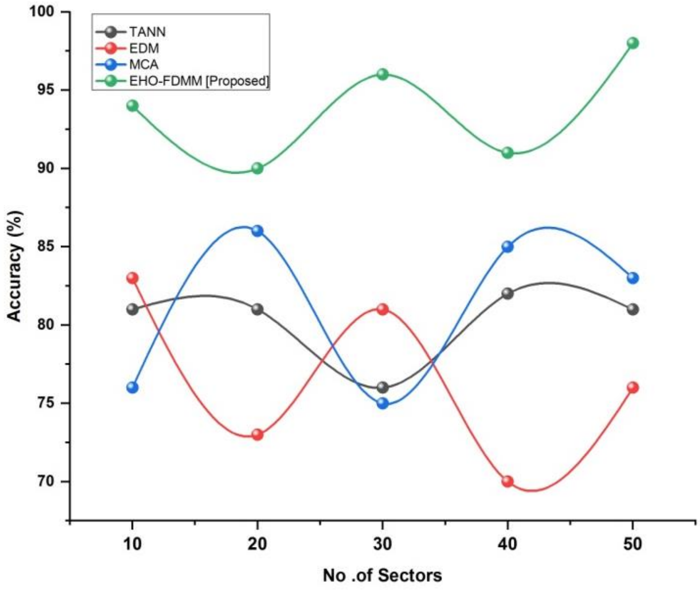

Figure 3. Comparison of accuracy

Table 2. Evaluation of accuracy

No .of Sectors	Accuracy (%)			
	TANN	EDM	MCA	EHO-FDMM [Proposed]
10	81	83	76	94
20	81	73	86	90
30	76	81	75	96
40	82	70	85	91
50	81	76	83	98

The proportion of successfully identified attack recordings is referred to as the Detection Rate (DR), which may be found in Equation (13). The DR of the suggested approach is contrasted with that of other existing methods in Figure 4 and Table 2.

$$DR = \frac{TP}{(TP+FN)} \tag{13}$$

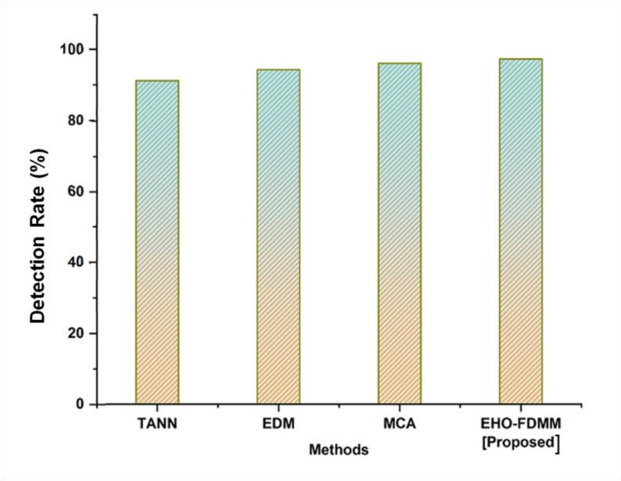

Figure 4. Comparison of DR

Table 2. Evaluation of DR

Methods	Detection rate (%)
TANN	91.2
EDM	94.3
MCA	96.1
EHO-FDMM [Proposed]	97.3

The proportion of records that were mistakenly identified as an attack is denoted by the False Positive Rate (FPR), which can be found in Equation 14. In Figure 5 and Table 3, a comparison is made between the suggested method's FPR and other methods.

$$FPR = \frac{FP}{(FP+TN)} \tag{14}$$

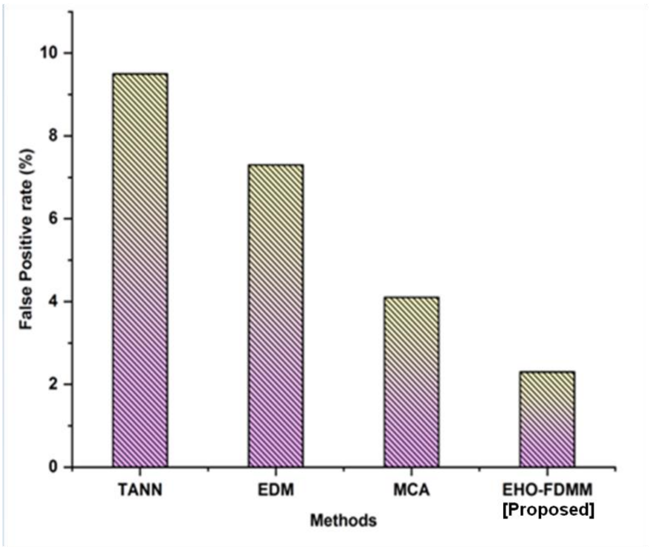

Figure 5. Comparison of FPR

Table 3. Evaluation of FPR

Methods	False Positive rate (%)
TANN	9.5
EDM	7.3
MCA	4.1
EHO-FDMM [Proposed]	2.3

The findings of the efficiency assessment for the EHO-FDMM IDS approach depending on the NSL-KDD dataset, the outcomes of three alternative methods were compared, including the Triangle Area Nearest Neighbours (TANN) [17], Euclidean Distance Map (EDM) [18], and Multivariate Correlation Analysis (MCA) [19], total DRs and FPRs are shown in Table 4.

Table 4. Performance evaluation of four approaches

Methods	Accuracy (%)	Detection rate (%)	False Positive rate (%)
TANN	81	91.2	9.5
EDM	76	94.3	7.3
MCA	83	96.1	4.1
EHO-FDMM [Proposed]	98	97.3	2.3

Since they are more current and offer comparable statistical measures to our EHO-FDMM, these approaches are utilized for comparison with ours. The TANN, EDM, and MCA had respective accuracy of 81%, 76%, and 83%, and DRs of 91.2%, 94.3%, and 96.1%, with FARs of 9.5%, 7.3%, and 4.1%. The EHO-FDMM, in comparison, had superior outcomes with 98% accuracy, 97.3% DR, and 2.3% FPR.

4.2. Datasets Applied to Analysis

The characteristics from the two datasets that were chosen for the performance assessment of the DMM-based IDS approach are provided in Table 5 along with the total DR, accuracy, and FPR scores.

Table 5. Performance assessment of the characteristics chosen from the two datasets

w value	Datasets					
	UNSW-NB15			NSL-KDD		
	FPR	DR	Accuracy	FPR	Accuracy	DR
1.5	9.3	84.2	89.1	3.1	93.1	93.2
2	6.7	87.2	88.2	4.2	93.3	93.3
2.5	7.1	89.3	90.8	2.9	97.2	97.4
3	5.9	93.8	94.4	2.5	97.9	97.9

In the NSL-KDD dataset, as a whole DR and accuracy climbed from 82.2% to 86.8% and 92.1% to 96.7%, accordingly, as the v value continuously grew from 1.5 to 3, whereas the FPR globally decreased from 2.2% to 1.4%.

Similarly, in the UNSW-NB15 dataset, when the v value climbed from 1.5 to 3, the total accuracy and DR climbed from 84.1% to 93.9% and 89.1% to 94.3%, accordingly, while the overall FPR decreased from 9.2% to 5.8%.

The FDMM precisely matches the bounds as it provides a list of chances used to calculate every incident's PDF of every characteristic, which is the main factor that made the EHO-FDMM-based IDS approach perform better than the other techniques. Nevertheless, even though the EHO-FDMM-based IDS approach had the lowest FPR and highest DR on the NSL-KDD dataset, but performed much worse on the UNSW-NB15 due to subtle differences between regular and atypical instances. This demonstrated the intricate, quite normal-looking assault patterns of the present.

5. CONCLUSION

This research covered a proposed scalable framework with three primary modules: data source, pre-processing, and a suggested procedure. To easily manage large-scale settings, the goal of the first component was to detect and gather network information from a database that is distributed while the second module's objective was to handle smaller-scale settings to analyze and filter network data to increase the performance of the suggested technique. The third approach, the EHO-FDMM-based intrusion detection, was developed based on an intrusion detection approach that uses a lower-upper interval of confidence as an indicator to identify abnormal data. The performance assessment of the EHO-FDMM-based intrusion detection system showed that it had been more precise than many other important techniques. In the future, we'll investigate more statistical methods to use them in conjunction to provide a visual tool for analysis, monitoring, and making choices on individual intrusions. We will further expand on this research to integrate the proposed framework's architecture with SCADA and cloud computing platforms.

REFERENCES

[1] P. Tyagi and S. K. Manju Bargavi, “Using Federated Artificial Intelligence System of Intrusion Detection for IoT Healthcare System Based on Blockchain,” *Int. J. Data Informatics Intell. Comput.*, vol. 2, no. 1, pp. 1–10, Mar. 2023, doi: 10.59461/ijdiic.v2i1.42.

[2] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, “Network intrusion detection system: A systematic study of machine learning and deep learning approaches,” *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, Jan. 2021, doi: 10.1002/ett.4150.

[3] H. Liu and B. Lang, “Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey,” *Appl. Sci.*, vol. 9, no. 20, p. 4396, Oct. 2019, doi: 10.3390/app9204396.

[4] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, “Survey of intrusion detection systems: techniques, datasets and challenges,” *Cybersecurity*, vol. 2, no. 1, p. 20, Dec. 2019, doi: 10.1186/s42400-019-0038-7.

[5] F. H. Almasoudy, W. L. Al-Yaseen, and A. K. Idrees, “Differential Evolution Wrapper Feature Selection for Intrusion Detection System,” *Procedia Comput. Sci.*, vol. 167, pp. 1230–1239, 2020, doi: 10.1016/j.procs.2020.03.438.

[6] C. Ieracitano, A. Adeel, F. C. Morabito, and A. Hussain, “A novel statistical analysis and autoencoder driven intelligent intrusion detection approach,” *Neurocomputing*, vol. 387, pp. 51–62, Apr. 2020, doi: 10.1016/j.neucom.2019.11.016.

[7] G. Marín, P. Caasas, and G. Capdehourat, “DeepMAL - Deep Learning Models for Malware Traffic Detection and Classification,” in *Data Science – Analytics and Applications*, Wiesbaden: Springer Fachmedien Wiesbaden, 2021, pp. 105–112. doi: 10.1007/978-3-658-32182-6_16.

[8] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, Aug. 2021, doi: 10.1109/TKDE.2019.2962680.

[9] M. Seyedan and F. Mafakheri, “Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities,” *J. Big Data*, vol. 7, no. 1, p. 53, Dec. 2020, doi: 10.1186/s40537-020-00329-2.

[10] S. U. Jan, S. Ahmed, V. Shakhov, and I. Koo, “Toward a Lightweight Intrusion Detection System for the Internet of Things,” *IEEE Access*, vol. 7, pp. 42450–42471, 2019, doi: 10.1109/ACCESS.2019.2907965.

[11] S. B and M. K, “Firefly algorithm based feature selection for network intrusion detection,” *Comput. Secur.*, vol. 81, pp. 148–155, Mar. 2019, doi: 10.1016/j.cose.2018.11.005.

[12] M. Eskandari, Z. H. Janjua, M. Vecchio, and F. Antonelli, “Passban IDS: An Intelligent Anomaly-Based Intrusion Detection System for IoT Edge Devices,” *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6882–6897, Aug. 2020, doi: 10.1109/JIOT.2020.2970501.

[13] M. Almiani, A. AbuGhazleh, A. Al-Rahayfeh, S. Atiewi, and A. Razaque, “Deep recurrent neural network for IoT intrusion detection system,” *Simul. Model. Pract. Theory*, vol. 101, p. 102031, May 2020, doi: 10.1016/j.simpat.2019.102031.

[14] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, “Building an efficient intrusion detection system based on feature selection and ensemble classifier,” *Comput. Networks*, vol. 174, p. 107247, Jun. 2020, doi: 10.1016/j.comnet.2020.107247.

[15] X. Zhou, Y. Hu, W. Liang, J. Ma, and Q. Jin, “Variational LSTM Enhanced Anomaly Detection for Industrial Big Data,” *IEEE Trans. Ind. Informatics*, vol. 17, no. 5, pp. 3469–3477, May 2021, doi: 10.1109/TII.2020.3022432.

[16] N. Moustafa and J. Slay, “The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set,” *Inf. Secur. J. A Glob. Perspect.*, vol. 25, no. 1–3, pp. 18–31, Apr. 2016, doi: 10.1080/19393555.2015.1125974.

[17] M. Yan, Y. Chen, X. Hu, D. Cheng, Y. Chen, and J. Du, “Intrusion detection based on improved density peak clustering for imbalanced data on sensor-cloud systems,” *J. Syst. Archit.*, vol. 118, p. 102212, Sep. 2021, doi: 10.1016/j.sysarc.2021.102212.

[18] D. Crow, S. Graham, B. Borghetti, and P. Sweeney, “Engaging Empirical Dynamic Modeling to Detect Intrusions in Cyber-Physical Systems,” 2020, pp. 111–133. doi: 10.1007/978-3-030-62840-6_6.

[19] Zhiyuan Tan, A. Jamdagni, Xiangjian He, P. Nanda, and Ren Ping Liu, “A System for Denial-of-Service Attack Detection Based on Multivariate Correlation Analysis,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 447–456, Feb. 2014, doi: 10.1109/TPDS.2013.146.

BIOGRAPHIES OF AUTHORS

V. Suresh Kumar has spent 26 years with Academics as well as 10 years with research. Currently, associated with Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College (Autonomous Institution affiliated to Anna University, Chennai) as Professor in the Department of Information Technology. Earlier he was the Principal, Kottayam Institute of Technology & Science, Kottayam, Kerala. Suresh Kumar has degrees in Electronics and Communication Engineering, Computer science and Engineering. He has his Doctoral Degree from Madurai Kamaraj University - a public state university, located in southern Tamil Nadu, India. MKU is one of the 15 universities in India with the “University with Potential for Excellence” status which was awarded by the University Grants Commission in India. He is a Life Member of ISTE. He can be contacted at email: sureshkumar@veltechmultitech.org