Harnessing DBSCAN and auto-encoder for hyper intrusion detection in cloud computing

Prabu Kaliyaperumal[1], Sudhakar Periyasamy[1], Muthusamy Periyasamy[2], Abinaya Alagarsamy[3]
[1]School of Computer Science and Engineering, Galgotias University, Uttar Pradesh, India
[2]Department of Cyber Security, Paavai Engineering College, Tamil Nadu, India
[3]Department of Information Technology, Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Tamil Nadu, India

Article Info

Article history:

Received Jan 2, 2024
Revised Feb 19, 2024
Accepted Feb 28, 2024

Keywords:

Cloud security
Deep learning
Network security
Unsupervised learning
Zero day attack

ABSTRACT

The widespread availability of internet services has led to a surge in network attacks, raising serious concerns about cybersecurity. Intrusion detection systems (IDS) are pivotal in safeguarding networks by identifying malicious activities, including denial of service (DoS), distributed denial of service (DDoS), botnet, brute force, probe, remote-to-local, and user-to-root attacks. To counter these threats effectively, this research focuses on utilizing unsupervised learning to train detection models. The proposed method involves employing auto-encoders (AE) for attack detection and density-based spatial clustering of applications with noise (DBSCAN) for attack clustering. By using preprocessed and unlabeled normal network traffic data, the approach enables the identification of unknown attacks while minimizing the impact of imbalanced training data on model performance. The auto-encoder method utilizes the reconstruction error as an anomaly detection metric, while DBSCAN employs a density-based approach to identify clusters, manage noise, accommodate diverse shapes, automatically determine cluster count, ensure scalability, and minimize false positives. Tested on standard datasets such as KDDCup99, UNSW-NB15, CICIDS2017, and CSE-CIC-IDS2018, this proposed model achieves accuracies exceeding 98.36%, 98.22%, 98.45%, and 98.51%, respectively. These results demonstrate the effectiveness of unsupervised learning in detecting and clustering novel intrusions while managing imbalanced data.

Corresponding Author:

Prabu Kaliyaperumal
School of Computer Science and Engineering Galgotias University
Greater Noida, Uttar Pradesh-203201, India
Email: k.prabu@galgotiasuniversity.edu.in

1. INTRODUCTION

The internet's extensive growth has brought forth cybersecurity challenges. Diverse data sources, including network packets, logs, and traffic statistics, contribute to intrusion detection [1]. This defensive mechanism analyzes web traffic for threats, comparing patterns to known signatures and unknown behavior. Upon detection, it alerts administrators for preventive action, addressing the evolving threat landscape. It is regarded as the second security gateway [2], following the firewall, it plays a crucial role in network security. An intrusion detection systems (IDS) is a comprehensive security tool safeguarding network. It monitors traffic, analyzes data, detects threats, issues alert, offers administrative controls and provides security reports to enhance cybersecurity. Machine learning and deep learning methods excel in anomaly detection [3], with datasets categorized into supervised and unsupervised learning. Supervised methods achieve accuracy but demand extensive labeled data, limiting accessibility and novelty detection. Unsupervised methods, like the proposed approach, excel in identifying abnormal traffic without labeled data, allowing detection of unknown attacks. The impact of training data quality is crucial, emphasizing the need for diverse and representative datasets in unsupervised learning-based anomaly detection to ensure effective model performance amid inherent traffic feature differences. Current methods [4] often neglect the influence of complex data patterns and diverse feature distributions in training data on unsupervised learning-based detection performance.

Supervised learning in abnormal traffic detection deploys machine learning algorithms to classify data, identifying and flagging unusual patterns in the dataset [5], [6]. Fan *et al.* [7] introduced the random forest-support vector machine (RF-SVM-IL) model for distributed denial of service (DdoS) attack detection, employing RF and SVM for dual traffic data classification. They integrated an incremental learning algorithm to filter additional input samples, reducing the computational load and enhancing the model's ability to accurately classify traffic, particularly when confronted with substantial attack volumes. Saeed and Jameel [8] employed a particle swarm optimization algorithm to choose optimal features from traffic data. They utilized a decision tree (DT) classification algorithm to construct an effective model for detecting attacks. Gumaei *et al.* [9] suggested a feature selection approach, correlation-based feature selection (CFS), to eliminate irrelevant features from traffic data. They employed an instance-based learning (IBL) algorithm to identify optimal features and conduct supervised learning for classifying normal and attack traffic. In their work, Javaid *et al.* [10] introduced an intrusion detection approach utilizing sparse autoencoders (AEs) and softmax regression. This method involved unsupervised feature extraction through sparse AEs, with softmax regression employed as a classifier for detecting network traffic anomalies.

Recent unsupervised learning-based network traffic detection emphasizes anomaly detection by correlating data features with reconstruction. Abnormal data is identified through subspace formation, leveraging significant reconstruction errors [11]. Yang *et al.* [12] introduced a fuzzy aggregation method employing the modified density peak clustering algorithm (MDPCA) to reduce and balance training data size for deep belief networks (DBNs). The MDPCA improves accuracy but incurs higher execution time, and further exploration of data purification is suggested. Jian *et al.* [13] introduced a convolutional neural network (CNN) for feature selection and data balancing by adjusting class-specific cost function weights based on class sizes. Utilizing the NSL-KDD dataset, the model is computationally intensive and demonstrated reduced detection accuracy. Karatas *et al.* [14] presented a distinctive approach to reduce imbalance using synthetic minority oversampling (SMOTE). By applying SMOTE, they achieved 99.69% precision, 99.34% recall, 99.35% F1-score, with a 0.65% error rate on the CSECIC-IDS2018 dataset. The AdaBoost approach improves accuracy but at a higher execution time cost. Further exploration of data purification is suggested. Onah *et al.* [15] implemented an IDS in fog computing, presenting the genetic algorithm wrapper-based feature selection and naive bayes for anomaly detection model (GANBADM). GANBADM effectively removes irrelevant features, maintaining high accuracy, and was evaluated on the NSL-KDD dataset.

The migration of critical data to the cloud presents an opportunity to bolster overall security. Despite the prevalence of routing-based attacks in network security, research often prioritizes identifying known attacks using imbalanced and labeled learning data. Accuracy hinges on balanced data; skewed or labeled data may distort results. Consequently, current efforts concentrate on fortifying defensive measures, notably by detecting zero-day attacks and addressing imbalanced learning through leveraging unsupervised learning capabilities, particularly in targeting routing-based attacks. This integration of an intrusion detection mechanism aims to comprehensively enhance network security. The methodology commences with dimensionality reduction using a basic auto-encoder (bAE) to extract encoded features from single-category data. The deep auto-encoder (dAE) neural network captures essential features for model training with normal network traffic [16]. The detection threshold is established by calculating the difference between input and output, ensuring accurate differentiation between normal and abnormal traffic. The abnormal traffic, processed with density-based spatial clustering of applications with noise (DBSCAN) after dAE, forms clusters based on feature density [17].

The dAE+DBSCAN approach focuses on outlier identification, involving dimension reduction, model training, anomaly detection, and clustering anomalies. In dimensionality reduction, bAE extracts feature and capture normal traffic data in an unsupervised manner [16], yielding distinctive patterns in latent space features. The dAE neural network reconstructs input data through encoding-decoding, producing an average reconstruction error used as the detection threshold. Anomaly detection involves inputting test data, computing the mean squared error for reconstruction, and labeling traffic as abnormal if the error exceeds the threshold [18], [19]. Subsequently, abnormal traffic from dAE undergoes DBSCAN processing, identifying clusters based on packet feature density. Crucial DBSCAN parameters, Epsilon (ε) and MinPts, influence cluster shape and density, covering density identification, noise management, shape adaptation, automatic cluster count, scalability, and false positive reduction [20].

This study introduces a hybrid algorithm based on deep learning principles. Utilizing auto-encoder for dimensionality reduction and a deep neural network for attack detection, it employs DBSCAN to form precise attack clusters. Validated on diverse datasets, it includes enhanced metrics.

2. THE COMPREHENSIVE THEORETICAL BASIS

In network-based environments, IDS are pivotal for cybersecurity. In the realm of networking, where extensive data is stored and processed on distributed servers, the risk of security breaches is elevated [21]. IDS monitor network and system activities, discerning and responding to unusual behavior signaling potential security threats. The dynamic and scalable nature of network infrastructure poses distinct challenges for intrusion detection [22], with attacks taking various forms like DDoS, unauthorized access attempts, and malware injection. In networking, IDS rely on advanced algorithms and machine learning to analyze extensive datasets, detecting patterns indicative of malicious activity [21], [22]. Timely and accurate intrusion detection is crucial for ensuring the confidentiality, integrity, and availability (CIA) of data in network environments, contributing to the overall security posture. It is regarded as the second security gateway [2], as illustrated in Figure 1, following the firewall, it plays a crucial role in network security. The elements constituting the IDS, illustrated in Figure 2, collaborate to assist organizations in identifying and addressing security incidents, thus strengthening their cybersecurity stance.

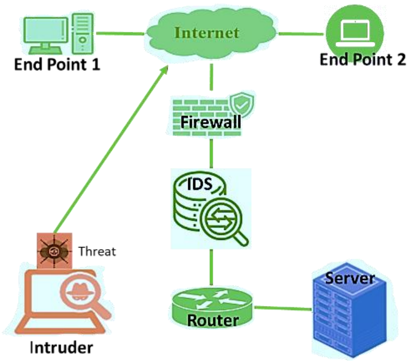

Figure 1. IDS

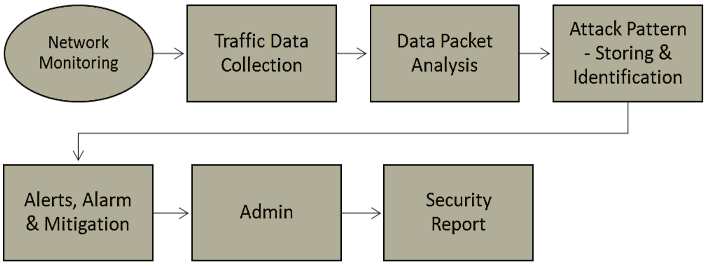

Figure 2. Elements of IDS

Anomaly-based IDS, or misuse data-based IDS, relies on predefined norms for normal network or host behavior, identifying deviations as potential intrusions. Signature-based IDS, also rule-based IDS, uses predefined attack signatures, scrutinizing network activity against them [21]. Integrated approaches combining these methods are the focus of most research efforts.

2.1. Auto-encoders

The AE, a neural network for unsupervised learning and dimensionality reduction, processes input normal traffic data through its encoder network. Each layer applies weights and biases to transform the data into a lower-dimensional representation, culminating in the latent layer producing an encoded representation [23]. The entire network, comprising the encoder and decoder, undergoes end-to-end training using backpropagation and gradient descent. This unsupervised learning process yields a compressed representation of the input data, capturing its essential features [24]. Trained on unlabeled data, the AEs encoded representation serves as a lower-dimensional, pre-clustered version of the original normal data, facilitating efficient analysis and anomaly detection. Training the AE minimizes reconstruction error, facilitating accurate differentiation between normal and malicious traffic [25]. The illustrated architecture in Figure 3 demonstrates the AE model's effectiveness in extracting essential information for robust classification outcomes.

2.2. DBSCAN

Clustering algorithms categorize data into meaningful groups. K-Means employs iterative centroid-based partitioning, fuzzy C-Means accommodates varied memberships, and hierarchical clustering reveals hierarchy. DBSCAN, a non-parametric, density-based unsupervised algorithm, excels in clustering closely packed data points of varying shapes without requiring a predetermined cluster count [26]. Unlike partitioning-based methods, it identifies clusters based on data density distribution Figure 4.

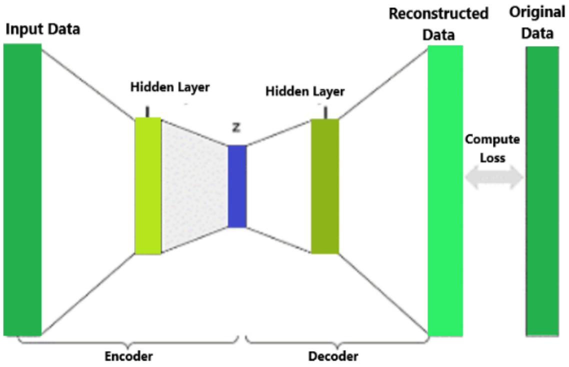

Figure 3. AE architecture

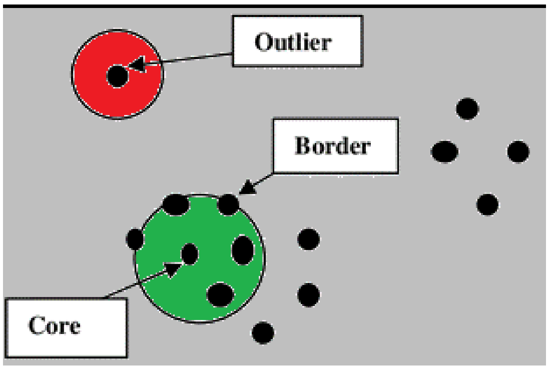

Figure 4. DBSCAN clustering

This DBSCAN partitions data into high-density clusters and merges those sharing common neighboring points, yielding flexible clusters. DBSCAN condenses clusters and ensures a balanced training dataset by creating sub-datasets within each cluster. These sub-datasets become subsets of training data, each with its trained classifier. During test data prediction, the algorithm assigns labels based on the classifier from the most relevant cluster, offering an effective and adaptable clustering approach. Clustering algorithms categorize data into meaningful groups. K-Means employs iterative centroid-based partitioning, fuzzy C-Means accommodates varied memberships, and hierarchical clustering reveals hierarchy.

3. METHOD

This proposed detection model primarily comprises phases such as traffic data preprocessing, Pre-Clustering, training the detection model, anomaly detection, and attack clustering. The design of the attack detection and clustering method is illustrated through the architecture depicted in Figure 5. In the preprocessing of traffic data, normalization and standardization are initial steps to prepare normal network traffic datasets. Unsupervised pre-clustering with bAE yields distinct subsets in the latent layer [19], each showcasing unique patterns. These subsets are then individually processed by dAE during the detection model's training phase. The "encoding-decoding" process assesses the average reconstruction error, establishing the detection threshold. In the anomaly detection phase, the trained model uses test data to generate reconstructed outputs, and the mean squared error calculates the reconstruction error. If this error exceeds the predefined threshold, the traffic is flagged as abnormal; otherwise, it's classified as normal [19]. Abnormal traffic undergoes further classification through the DBSCAN algorithm, aimed at identifying various attack types, including novel attacks. This multi-step approach enhances the model's ability to discern and categorize network anomalies effectively.

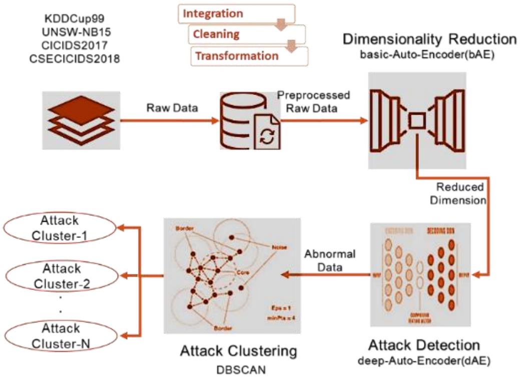

Figure 5. Architecture of proposed model

3.1. Data preprocessing and dimensionality reduction

This research utilizes benchmark datasets for intrusion detection, including KDDCUP99 [27], CICIDS2017 [28], CSECICIDS2018 [28], and UNSWNB15 [29]. Standard Scaler techniques are applied to standardize numerical data, transforming it linearly. The normalized data is then mapped to a mean (μ) of 0 and a standard deviation (σ) of 1 range using Z-score normalization. The AE, a neural network for unsupervised learning and dimensionality reduction, processes input normal traffic data through its encoder network. Each layer applies weights and biases to transform the data into a lower-dimensional representation, culminating in the latent layer producing an encoded representation. The entire network, comprising the encoder and decoder, undergoes end-to-end training using backpropagation and gradient descent. This unsupervised learning process yields a compressed representation of the input data, capturing its essential features. Trained on unlabeled data, the AEs encoded representation serves as a lower-dimensional, pre-clustered version of the original normal data, facilitating efficient analysis and anomaly detection.

3.2. Anomaly detection using deep auto-encoder

The AEs, an unsupervised neural network, extracts hierarchical features for improved classification. Utilizing the bAE, this research builds a dAE model to proficiently capture data features within clustering subsets. Training the dAE minimizes reconstruction error, facilitating accurate differentiation between normal and malicious traffic. The illustrated architecture in Figure 3 demonstrates the dAE model's effectiveness in extracting essential information for robust classification outcomes. The training phase involves the deep AE model processing batches of normal traffic data with the aim of minimizing the reconstruction error. The detection threshold is dynamically set based on the accumulated loss. The process continues for a specified number of epochs and the trained deep AE model along with the detection threshold is returned as the training model output. The detection phase evaluates the trained AE model using the test dataset to identify anomalies. It calculates the reconstruction error for each data point by comparing the original and reconstructed data. If the reconstruction error surpasses the detection threshold, the data point is categorized as anomaly traffic; otherwise, it is labeled as normal traffic. The resulting output provides a detection result for each data point in the test dataset, indicating its classification as either normal or abnormal traffic. This testing procedure plays a pivotal role in assessing the trained detection model's efficacy on new data, aiding in the detection of potential anomalies.

3.3. Clustering anomalies using DBSCAN

The AE dAE produces outputs containing attack pattern information, representing points in the AEs output space. DBSCAN requires ε (epsilon-0.14) and MinPts parameters, starting from an unvisited point. ε sets the maximum distance for neighborhood consideration, defining a point's radius. It retrieves ε-neighborhoods, initiating clusters if sufficiently populated or classifying as noise. MinPts-5 denotes the minimum samples needed for core point qualification, representing central elements in dense regions. The fit method assigns clusters based on density, extending to densely populated ε-neighborhoods. This process iterates until the entire density-connected cluster is identified. New unvisited points are processed, revealing additional clusters or noise. DBSCAN optimizes the set of all possible clusterings from the entire set, minimizing the number of clusters. It ensures each pair of points within a cluster is density-reachable, preserving the original properties of "maximality" and "connectivity" of clusters.

4. RESULTS AND DISCUSSION

The proposed approach is assessed on the top four attacks in each dataset using benchmark intrusion detection datasets. The evaluation includes comparing the proposed model with standard clustering models (K-Means, fuzzy C-means, hierarchical, and OPTICS) on un-clustered data, followed by performance comparison.

4.1. Experimental dataset usage

The dataset used and the number of selected samples for this research are presented in Table 1. This table provides a comprehensive overview of the experimental samples, highlighting the different datasets and their respective attack categories. These datasets include KDDCup99, UNSW-NB15, CICIDS2017, and CSECICIDS2018, each featuring a variety of attack types such as DoS, Probe, R2L, U2R, worms, backdoors, fuzzers, DDoS, botnet, and brute force.

Table 1. Experimental samples

	KDDCup99	UNSW-NB15	CICIDS2017	CSECICIDS2018
Class	Dos, Probe, R2L and U2R	DoS, worms, Backdoors, and Fuzzers	DoS, DDoS, Botnet and Bruteforce	DoS, DDoS, Botnet and Bruteforce
Features	41	49	83	80
Benign	972,781	2,218,761	2,359,087	2,374,871
Attack	109,291	321,465	224,893	239,842

4.2. Performance metrics

Assessing the intrusion model's effectiveness with a confusion matrix [21], includes evaluating accuracy, precision, recall, and F-score. Precision in (1) gauges optimistic prediction accuracy, recall in (2) captures correctly classified positive instances, F-score in (3) provides a comprehensive metric for evaluation and accuracy in (4) measures correct classifications.

$$Precision = \frac{True\ Positive}{Predicted\ Positives} \tag{1}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$F - Measure = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{Predicted\ Positive + Predicted\ Negative} \tag{4}$$

The model's effective performance relies on crucial factors such as the appropriate structure and hyperparameter settings for the AE detection model. Extensive experiments were conducted to determine the optimal hyperparameter settings. The proposed method's bAE consists of a single input layer and a single output layer, while the dAE incorporates three hidden layers, each with unit sizes tailored to the loss function, thereby ensuring an optimal fit for the varying dimensions of features present in the training data.

4.3. Training and testing of detection model

This approach involves splitting the dataset into training (70%) and testing (30%) sets. The model is trained with early stopping criteria to determine the optimal number of epochs. ReLU serves as the activation function, and Adam optimization with a learning rate of 0.0001 is used.

Detection accuracy is assessed using mean square error (MSE) as the loss function. Figure 6 displays the testing accuracy of the proposed approach on various datasets, highlighting the model's effectiveness. Detection models, built using normal and attack data from KDDCUP99, UNSWNB15, CICIDS2017, and CSECICIDS2018 datasets, exhibit notable performance across all dataset. Averaging outcomes from ten experiments, the AE Model's findings demonstrate particularly superior results on CSECICIDS2018 and CICIDS2017. Averaging outcomes from ten experiments, the AE Model's experimental findings, showcase notable performance across all datasets, with particularly superior results on CSECICIDS2018 and CICIDS2017.

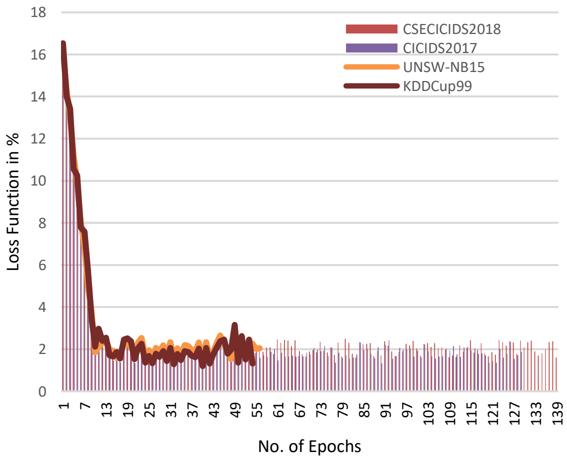

Figure 6. Detection model loss function

4.4. Performance evaluation-clustering model

The clustering model is compared with prevalent clustering algorithms such as K-means, fuzzy C Means, hierarchical clustering divisive hierarchical clustering (DHC), and OPTICS [3] using network traffic datasets. The experiment results, including confusion matrices for CSECICIDS2018 Figure 7 and CICIDS2017 are presented Figure 8.

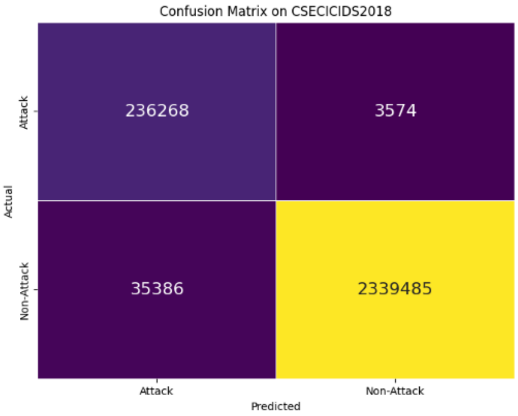

Figure 7. Confusion matrix on CSECICIDS2018

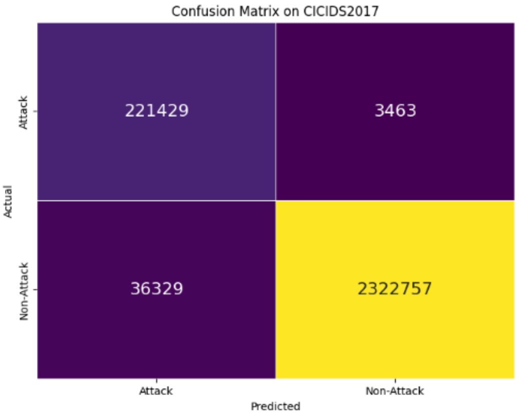

Figure 8. Confusion matrix on CICIDS2017

4.5. Evaluation on CICIDS2017 dataset

Table 2 represents, the suggested approach consistently surpasses alternative clustering algorithms across diverse metrics, showcasing its efficacy in clustering, notably in precision, recall, specificity, and overall accuracy. The proposed approach demonstrates the highest precision (0.9854), indicating a low false positive rate.

Table 2. Performance comparison of proposed method on CICIDS2017

Methods	Precision	Recall	Specificity	Accuracy	F-Measure
K-Means	0.9530	0.9543	0.9497	0.9522	0.9537
FC-Means	0.9441	0.9454	0.9408	0.9433	0.9448
Hierarchical	0.9499	0.9512	0.9466	0.9491	0.9506
OPTICS	0.9321	0.9335	0.9288	0.9314	0.9328
Proposed	0.9854	0.9867	0.9820	0.9846	0.9860

4.6. Evaluation on KDDCUP99 dataset and UNSW-NB15 dataset

Figure 9 demonstrates that the proposed approach achieves superior results in the KDDCup99 dataset, encompassing higher f1-score, recall, precision, and accuracy. The overall correctness of the clustering is highest for the proposed approach (0.983609). The proposed approach achieves the highest F1-Measure (0.9859), indicating a good balance between precision and recall. Figure 10 demonstrates that the proposed approach achieves superior results in the UNSW-NB15 dataset, encompassing higher f1-score, recall, precision, and accuracy. The ability to correctly identify negative instances is highest for the proposed approach (0.9807), indicating a low false negative rate.

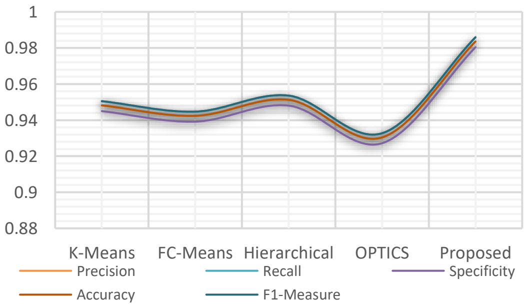

Figure 9. Performance comparison on KDDCUP99

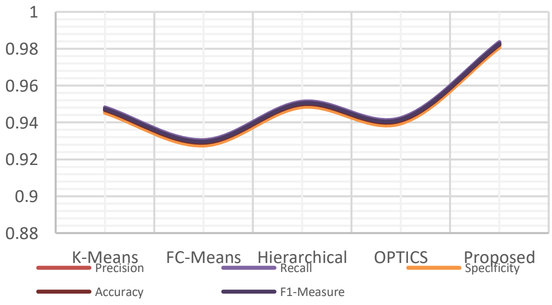

Figure 10. Performance comparison on CSECICIDS2018

4.7. Evaluation on CSECICIDS2018 dataset

Table 3 represents the proposed approach consistently outperforms other clustering algorithms across various metrics, demonstrating its effectiveness in clustering. The proposed approach achieves the highest recall (0.9870), suggesting a strong ability to capture positive instances.

Table 3. Performance comparison of proposed method on CSECICIDS2018

Methods	Precision	Recall	Specificity	Accuracy	F-Measure
K-Means	0.9529	0.9516	0.9469	0.9497	0.9522
FC-Means	0.9471	0.9458	0.9411	0.9439	0.9464
Hierarchical	0.9351	0.9338	0.9291	0.9319	0.9344
OPTICS	0.9560	0.9547	0.9500	0.9528	0.9553
Proposed	0.9883	0.9870	0.9823	0.9851	0.9877

4.8. Accuracy Comparison across various methods and datasets

The Figure 11 presents the performance metrics (Accuracy) of different clustering algorithms across four datasets KDDCUP99, UNSW-NB15, CSE-CIC-IDS2018, and CIC-IDS2017. For KDDCUP99, the proposed approach achieves the highest accuracy at 0.9836. In UNSWNB15, the proposed approach also outperforms others with an accuracy of 0.9822. CSECICIDS2018 shows the proposed approach leading in accuracy at 0.9851. For CICIDS2017, the proposed approach maintains a high accuracy of 0.9845. Overall, the proposed approach consistently demonstrates superior accuracy compared to other clustering algorithms across the datasets.

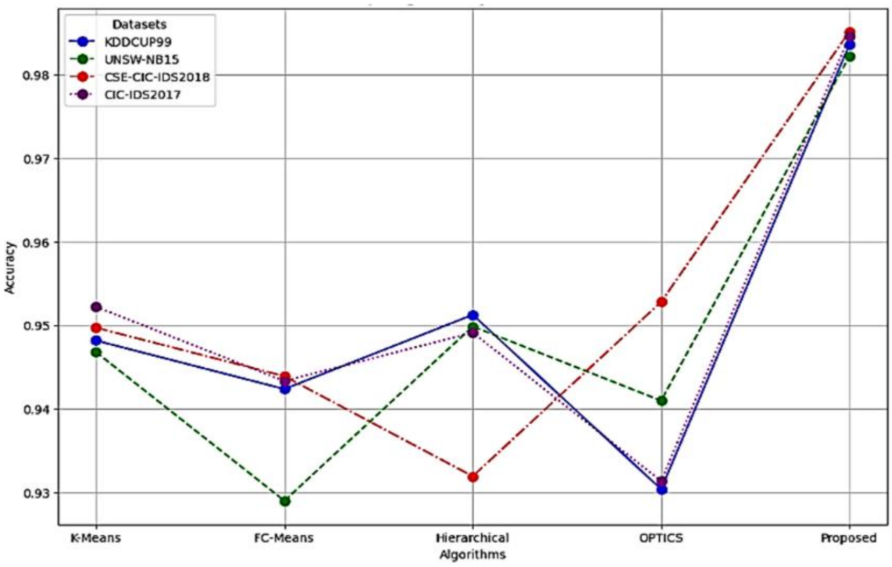

Figure 11. Accuracy comparison across various methods and datasets

The insights gleaned from the study, which emphasize the effectiveness of the proposed intrusion detection technique and stress the importance of employing a larger training dataset comprising solely benign data within the AE, are critical for advancing cybersecurity. These findings can inform researchers in the development of more effective IDS, addressing challenges posed by zero-day attack detection, imbalanced data, and bolstering overall network security. The heightened accuracy observed in the proposed method, as identified in the study, sets a benchmark for innovative methodologies in addressing cybersecurity challenges.

5. CONCLUSION

This research proposes an innovative unsupervised learning approach, combining AE and DBSCAN, for enhanced network intrusion detection. By training on preprocessed, unlabeled normal network traffic, the system identifies unknown attacks and mitigates imbalanced training data effects. The dAE+DBSCAN model achieves over 98.39% accuracy across datasets, minimizing false positives, and excelling in diverse attack pattern identification and clustering. The approach consistently outperforms other clustering algorithms, offering a reliable means to predict potential attacks post-training. The model, trained on prepared data, analyzes real-time network traffic, improving intrusion detection efficiency. The study emphasizes the success of the proposed intrusion detection technique, advocating for a larger training dataset comprising solely benign data within the AE. These findings guide the development of more effective IDS, addressing zero-day attack detection, imbalanced data challenges, and enhancing overall network security, setting a benchmark for innovative methodologies. Future research targets scalability for larger networks, adaptation to emerging attack vectors, integration of advanced techniques, and exploration of varied environments for optimization.

REFERENCES

[1] T. F. Schindler, S. Schlicht, and K.-D. Thoben, "Towards benchmarking for evaluating machine learning methods in detecting outliers in process datasets," *Computers*, vol. 12, no. 12, Dec. 2023, doi: 10.3390/computers12120253.

[2] F. Alrowais *et al.*, "Intelligent intrusion detection using arithmetic optimization enabled density based clustering with deep learning," *Electronics*, vol. 11, no. 21, Oct. 2022, doi: 10.3390/electronics11213541.

[3] S. B. Mallampati and H. Seetha, "A review on recent approaches of machine learning, deep learning, and explainable artificial intelligence in intrusion detection systems," *Majlesi Journal of Electrical Engineering*, vol. 17, no. 1, pp. 29–54, 2023, doi: 10.30486/mjee.2023.1976657.1046.

[4] B. Cao, C. Li, Y. Song, Y. Qin, and C. Chen, "Network intrusion detection model based on CNN and GRU," *Applied Sciences*, vol. 12, no. 9, Apr. 2022, doi: 10.3390/app12094184.

[5] M. Ramasamy and P. Vinitha Eric, "A novel classification and clustering algorithms for intrusion detection system on convolutional neural network," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2845–2855, Oct. 2022, doi: 10.11591/eei.v11i5.4145.

[6] S. Guarino, F. Vitale, F. Flammini, L. Faramondi, N. Mazzocca, and R. Setola, "A two-level fusion framework for cyber-physical anomaly detection," *IEEE Transactions on Industrial Cyber-Physical Systems*, vol. 2, pp. 1–13, 2024, doi: 10.1109/TICPS.2023.3336608.

[7] J. F. J. Fan, G. Y. J. Fan, and J. G. G. Yang, "DDoS attack detection system based on RF-SVM-IL model under SDN," *Journal of Computers*, vol. 32, no. 5, pp. 031–043, Oct. 2021, doi: 10.53106/199115992021103205003.

[8] A. A. Saeed and N. G. M. Jameel, "Intelligent feature selection using particle swarm optimization algorithm with a decision tree for DDoS attack detection," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 1, pp. 37–48, Mar. 2021, doi: 10.26555/ijain.v7i1.553.

[9] A. Gumaei *et al.*, "A robust cyberattack detection approach using optimal features of SCADA power systems in smart grids," *Applied Soft Computing*, vol. 96, Nov. 2020, doi: 10.1016/j.asoc.2020.106658.

[10] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, 2016, doi: 10.4108/eai.3-12-2015.2262516.

[11] P. R. Chandre, P. Mahalle, and G. Shinde, "Intrusion prevention system using convolutional neural network for wireless sensor network," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp504-515.

[12] Y. Yang, K. Zheng, C. Wu, X. Niu, and Y. Yang, "Building an effective intrusion detection system using the modified density peak clustering algorithm and deep belief networks," *Applied Sciences*, vol. 9, no. 2, Jan. 2019, doi: 10.3390/app9020238.

[13] K. Jiang, W. Wang, A. Wang, and H. Wu, "Network intrusion detection combined hybrid sampling with deep hierarchical network," *IEEE Access*, vol. 8, pp. 32464–32476, 2020, doi: 10.1109/ACCESS.2020.2973730.

[14] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based idss on an imbalanced and up-to-date dataset," *IEEE Access*, vol. 8, pp. 32150–32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

[15] J. O. Onah, S. M. Abdulhamid, M. Abdullahi, I. H. Hassan, and A. Al-Ghusham, "Genetic algorithm based feature selection and naïve Bayes for anomaly detection in fog computing environment," *Machine Learning with Applications*, vol. 6, Dec. 2021, doi: 10.1016/j.mlwa.2021.100156.

[16] C. Brunner, A. Kő, and S. Fodor, "An autoencoder-enhanced stacking neural network model for increasing the performance of intrusion detection," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 12, no. 2, pp. 149–163, Apr. 2021, doi: 10.2478/jaiscr-2022-0010.

[17] P. Jain, M. S. Bajpai, and R. Pamula, "A modified DBSCAN algorithm for anomaly detection in time-series data with seasonality," *The International Arab Journal of Information Technology*, vol. 19, no. 1, Jan. 2022, doi: 10.34028/iajit/19/1/3.

[18] H. G. Mohamed *et al.*, "Feature selection with stacked autoencoder based intrusion detection in drones environment," *Computers, Materials & Continua*, vol. 73, no. 3, pp. 5441–5458, 2022, doi: 10.32604/cmc.2022.031887.

[19] M. Leon, T. Markovic, and S. Punnekkat, "Feature encoding with autoencoder and differential evolution for network intrusion detection using machine learning," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Jul. 2022, pp. 2152–2159, doi: 10.1145/3520304.3534009.

[20] M. Monshizadeh, V. Khatri, R. Kantola, and Z. Yan, "A deep density based and self-determining clustering approach to label unknown traffic," *Journal of Network and Computer Applications*, vol. 207, Nov. 2022, doi: 10.1016/j.jnca.2022.103513.

[21] 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," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, Jan. 2021, doi: 10.1002/ett.4150.

[22] T. A. Jasim Ali and M. M. T. Jawhar, "Detecting network attacks model based on a convolutional neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 3, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3072-3078.

[23] X. Li, W. Chen, Q. Zhang, and L. Wu, "Building auto-encoder intrusion detection system based on random forest feature selection," *Computers & Security*, vol. 95, Aug. 2020, doi: 10.1016/j.cose.2020.101851.

[24] G. Andresini, A. Appice, and D. Malerba, "Autoencoder-based deep metric learning for network intrusion detection," *Information Sciences*, vol. 569, pp. 706–727, Aug. 2021, doi: 10.1016/j.ins.2021.05.016.

[25] A. Binbusayyis and T. Vaiyapuri, “Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM,” *Applied Intelligence*, vol. 51, no. 10, pp. 7094–7108, Oct. 2021, doi: 10.1007/s10489-021-02205-9.

[26] A. S. Alfoudi *et al.*, “Hyper clustering model for dynamic network intrusion detection,” *IET Communications*, Oct. 2022, doi: 10.1049/cmu2.12523.

[27] U. of C. (Irvine), “KDD’99 Dataset,” *University of California, Irvine*, 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed Dec. 30, 2023).

[28] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 2018, pp. 108–116, doi: 10.5220/0006639801080116.

[29] N. Moustafa and J. Slay, “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” in *2015 Military Communications and Information Systems Conference (MilCIS)*, Nov. 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.

BIOGRAPHIES OF AUTHORS

Prabu Kaliyaperumal Assistant Professor, SCSE at Galgotias University, has 16 years of teaching experience. Currently pursuing a Ph.D., he holds an M.Tech. in CSE from SRM University. He has published 4 patents and 9 research papers in international journals and conferences. His expertise includes cyber security, networks, cloud computing, and machine learning. He can be contacted at email: k.prabu@galgotiasuniversity.edu.in.

Prof. Sudhakar Periyasamy SCSE at Galgotias University. With 18 years of teaching experience, he holds a Ph.D. from Anna University. He has published 7 patents, 5 book chapters, and 18 research papers published in reputable international journals and conferences. His expertise includes networks, cyber security, cloud computing, and machine learning. He can be contacted at email: p.sudhakar@galgotiasuniversity.edu.in.

Prof. Muthusamy Periyasamy Department of Cyber Security at Paavai Engineering College, has 20 years of teaching experience. he holds a Ph.D. from Anna University. He has published 17 patents, 8 book chapters, and 30 research papers published in reputable international journals and conferences. His expertise includes cloud computing, cyber security, artificial intelligence, and machine learning. He can be contacted at email: muthu.namakkal@gmail.com.

Abinaya Alagarsamy Assitant Professor, Department of Information Technology, Vel Tech High Tech Dr. Rangarajan, Dr. Sakunthala Engineering College. Currently pursuing a Ph.D., she holds an M.E in CSE from Anna University. She has published 2 patents and 6 research papers in international journals and conferences. Her expertise includes machine learning, cyber security, networks, and cloud computing. She can be contacted at abinayaalagar1992@gmail.com.