Healthcare Intrusion Detection using Hybrid Correlation-based Feature Selection-Bat Optimization Algorithm with Convolutional Neural Network

A Hybrid Correlation-based Feature Selection for Intrusion Detection Systems

H. Kanakadurga Bella[1], S. Vasundra[2]
Department of CSE, Jawaharlal Nehru Technological University Anantapur, Ananthapuramu, India[1]
Department of CSE, JNTUA College of Engineering, Ananthapuramu,
Constituent College of Jawaharlal Nehru Technological University Anantapur, Ananthapuramu, India[2]

***Abstract*—Cloud computing is popular among users in various areas such as healthcare, banking, and education due to its low-cost services alongside increased reliability and efficiency. But, security is a significant problem in cloud-based systems due to the cloud services being accessed via the Internet by a variety of users. Therefore, the patient's health information needs to be kept confidential, secure, and accurate. Moreover, any change in actual patient data potentially results in errors during the diagnosis and treatment. In this research, the hybrid Correlation-based Feature Selection-Bat Optimization Algorithm (HCFS-BOA) based on the Convolutional Neural Network (CNN) model is proposed for intrusion detection to secure the entire network in the healthcare system. Initially, the data is obtained from the CIC-IDS2017, NSL-KDD datasets, after which min-max normalization is performed to normalize the acquired data. HCFS-BOA is employed in feature selection to examine the appropriate features that not only have significant correlations with the target variable, but also contribute to the optimal performance of intrusion detection in the healthcare system. Finally, CNN classification is performed to identify and classify intrusion detection accurately and effectively in the healthcare system. The existing methods namely, SafetyMed, Hybrid Intrusion Detection System (HIDS), and Blockchain-orchestrated Deep learning method for Secure Data Transmission in IoT-enabled healthcare systems (BDSDT) are employed to evaluate the efficacy of HCFS-BOA-based CNN. The proposed HCFS-BOA-based CNN achieves a better accuracy of 99.45% when compared with the existing methods: SafetyMed, HIDS, and BDSDT.**

Keywords*—Convolutional neural network; deep learning; intrusion detection system; healthcare; security

I. INTRODUCTION

Network Intrusion Detection Systems (NIDSs) identify malicious activities and safeguard the vulnerable services by monitoring network traffic and providing alerts when anomalous events are recognized. Some organizations that are primarily focused on obtaining private user data, establishing the foundation for modern-day detection and protection are attacked by cyber-attackers. Furthermore, the healthcare sector keeps growing, and most hospitals are integrating e-healthcare systems as quickly as feasible to fulfill the needs of their patients. IDS based on cloud networks employ anomaly-based techniques to protect the cloud-based applications [1]. In network security, there are two common detection techniques for NIDS, anomaly-based detection, and signature-based detection [2]. An anomaly-based IDS analyzes the network traffic and correlates it to a created baseline for unknown or known attacks, where a signature-based IDS is allowed to be employed while the attack patterns are established and pre-determined [3, 4]. To address numerous security issues, the cloud utilizes numerous cybersecurity techniques like IDS, Intrusion Prevention Systems (IPS), and firewalls [5]. The centralized processing technique used by cloud computing involves uploading every transaction and processing the end-user service requests based on the transmission bandwidth, capacity of storage, and computer resources [6]. Proactive network security defenses are required to protect essential assets and data because the cloud attack vector has the potential to result in successful security breaches [7].

Network security has always placed a high priority on intrusion detection since it is crucial for identifying anomalous activity on secured internal networks [8]. The network of intermediate, source, and endpoint are used to identify the Distributed Denial-of-Services (DDoS) attacks. The attack's endpoint is easily detected because of the massive volume of network traffic that is generated [9]. A significant number of traditional intrusion detection systems use either a port-based or Deep Packet Inspection (DPI) technique. The port-based technique identifies traffic by using the ports established by the Internet Assigned Numbers Authority (IANA) [10]. Software Defined Network (SDN) is an emerging design that is cost-effective, flexible, adaptable, and controlled, thereby making it more suitable for presently employed complicated applications and bandwidth [11, 12]. SDN's goal is to create a logically centralized hub for internet and networking architects so that they quickly respond to the evolving client demands [13]. Deep learning techniques, especially CNN represent remarkable capacity in automatically extracting features and intricate patterns from complex data, including network traffic [14]. By employing Deep CNN, the IDS efficiently recognizes anomalous behavior and emerging threats in real time [15]. Since cloud services are accessed via the internet by a variety of users, security is a significant concern in cloud-based systems because the health information of patients must be

kept confidential, secure, and accurate. Moreover, any change in actual patient data results in errors during the diagnosis and treatment [31-34]. Therefore, the HCFS-BOA based on CNN is proposed in this research, for intrusion detection to secure the entire network in the healthcare system. The main contributions of this research are as follows:

- The proposed HCFS-BOA approach is evaluated on the CIC-IDS2017, NSL-KDD benchmark datasets, and the Min-max normalization technique is employed to normalize the raw data.
- For feature selection, HCFS-BOA is employed to examine the appropriate features that not only have significant correlations with the target variable, but also contribute to the optimal performance of intrusion detection in the healthcare system.
- Finally, CNN is employed for classification to identify and classify intrusion detection accurately and effectively. The efficacy of HCFS-BOA is analyzed based on the performance measures of accuracy, precision, recall, and f1-score.

The rest of the paper is organized as follows: Section II presents the literature survey. The block diagram of the proposed method is discussed in Section III. The results are illustrated in Section IV, while Section V discusses the conclusion of this paper.

II. Literature Survey

Faruqui et al. [16] presented a SafetyMed for Internet of Medical Things (IoMT) IDS by employing hybrid CNN-Long Short-Term Memory (CNN-LSTM). The SafetyMed was the first IDS that included an optimization approach based on the trade-off between Detection Rate (DR) and False Positive Rate (FPR). The SafetyMed enhanced the safety and security of medical devices and patient information. However, the presented SafetyMed method had no defense mechanism against an attack of Adversarial Machine Learning (AML).

Vashishtha et al. [17] implemented a HIDM for cloud-based healthcare systems to detect all kinds of attacks. The hybrid approach was a mixture of a Signature-based Detection Model (SDM) and an Anomaly-based Detection Model (ADM). The datasets of NSL-KDD, CICIDS2017, and UNSW-NB15 were employed to evaluate the efficacy of the HIDM approach. The implemented method had a higher detection rate with the error of Type-I and Type-II for both ADM and SDM. However, combining various detection systems increased the risk of false negatives and false positives.

Kumar et al. [18] introduced a BDSDT for the transmission of secure data in IoT-based healthcare systems. Initially, the architecture of blockchain was created in all IoT devices that were identified and established using a zero-knowledge proof, and then connected to the blockchain network using a smart contract-based ePOW consensus. Then, a bidirectional LSTM was employed using a DL to recognize IDS in the healthcare system. The BDSDT enhanced the privacy and security by combining both DL and blockchain methods. However, BDSDT wasn't effective against web and Bot threat attacks as there were fewer instances of these two classes which led to changes in actual patient data resulting in errors during the diagnosis and treatment.

Halbouni et al. [19] presented a CNN-Long Short-Term Memory (CNN-LSTM) for IDS system. The ability of CNN to extract the spatial features alongside the ability of LSTM to extract the temporal features were the highlights of this model. In order to improve performance, batch normalization and the layers of dropout were created to the presented method. The presented method decreased the false alarm rate and improved the rate of detection. However, CNN-LSTM failed to provide a high detection rate for specific kinds of attacks like web attacks and worms which led to changes in actual patient data resulting in errors during the diagnosis and treatment.

Han et al. [20] presented an Intrusion Detection Hyperparameter Control System (IDHCS) to regulate and train a Deep Neural Network (DNN) extracted feature and the module of k-means clustering in terms of Proximal Policy Optimization (PPO). The most valuable network features were extracted by the DNN under the control of an IDHCS, which also used K-means clustering to detect intrusion. The IDHCS performed effectively for each dataset, as well as the combined dataset. However, to represent a more realistic network environment, a diverse dataset needed to be examined.

Bakro et al. [21] introduced a hybrid feature selection approach that combined filter techniques such as Particle Swarm Optimization (PSO), Chi-Square (CS), and Information Gain (IG). Combining each of these three techniques was a novel method that generated a more reliable process of feature selection by using every technique's strength to increase the possibilities of selecting the most associated features. The introduced method had the benefits of flexibility, time complexity, interpretability, and scalability. But, the feature selection approach was not done properly which resulted in overfitting.

Sudar et al. [22] implemented a Machine Learning (ML) approach based on Decision Tree (DT) and Support Vector Machine (SVM) to detect Distributed Denial of Service (DDoS) attacks. The classification approach was established in the environment of Software Defined Network (SDN). The DT and SVM approaches were deployed to distinguish among malicious and normal traffic data. This approach provided better accuracy and detection rate. Nonetheless, this implemented approach struggled to adapt to evolving attack strategies.

Praveena et al. [23] developed a Deep Reinforcement Learning approach that was optimized by Black Widow Optimization (DRL-BWO) for intrusion detection in Unmannered Aerial Vehicles (UAV). The BWO approach was deployed for parameter optimization of the DRL method which assisted in enhancing the performance of intrusion detection in UAV networks. This approach was fit for the tasks of information extraction in high dimensional space. Nonetheless, the intricate nature of the DRL-BWO approach resulted in minimized interpretability.

Chinnasamy et al. [24] presented a Blockchain DDoS flooding attack with dynamic path detectors. The ML approach

was established to identify the attacks which focused on the DDoS assault. The primary essential traits were employed to predict the accurate DDoS attacks by utilizing a different attribute selection approach. Nevertheless, this presented approach led to severe network congestion which hindered the processing of transactions and slowed down the overall system's performance.

Chinnasamy et al. [25] developed an ML approach for effective phishing attack detection. Based on the input features such as Uniform Resource Locator (URL) and Web Traffic, the link was classified as phishing or non-phishing. This approach was determined by retrieving datasets from ML and phishing cases by employing SVM, Random Forest (RF), and Genetic. Nevertheless, ML approaches in phishing detection struggled to maintain pace with constantly evolving phishing tactics which led to potential delays in identifying the new attacks.

Anupriya et al. [26] implemented an ML approach for fraud account detection. To compute buddy similarity criteria, the adjacency network matrix graph was employed and then new features were acquired by utilizing the Principle Component Analysis (PCA). This was employed to equalize the data and transform it into the classifier in the next phase of cross-validation for training and testing the classifier. Nevertheless, due to imbalanced datasets, this approach struggled with evolving the fraud pattern and generated false positives or negatives.

There are some limitations with the existing methods that are mentioned above such as the methods not being effective in detecting attacks which led to changes in actual patient data resulting in errors during the diagnosis and treatment. In order to overcome these issues, the HCFS-BOA-based CNN is proposed for intrusion detection to secure the entire network in the healthcare system.

III. Proposed Methodology

In this research, a hybrid CFS-BOA-based CNN approach is proposed for intrusion detection in healthcare systems using deep learning. It includes datasets, min-max normalization, feature selection using HCFS-BOA, classification using CNN, and performance evaluation. The overview of the proposed method is represented in Fig. 1.

A. Datasets

The proposed HCFS-BOA approach is evaluated on CIC-IDS2017 [27] and NSL-KDD benchmark datasets. The CIC-IDS-2017 dataset includes malicious and normal traffic data that is considered new and does not include an enormous amount of redundant data. It includes eleven new attacks namely, PortScan, Brute Force, DoS, web attacks like SSH, Patator, FTP-Patator, SQL injection, and XSS. It is created by the Canadian Institute for Cybersecurity in 2017, and its 80 features are employed to monitor malicious and benign traffic. The NSL-KDD is an extension of the KDD cup 99 database and contains 41-dimensional vectors with numerical and categorical values. The intrusion attacks in the NSL-KDD database are probe attacks, Remote to a user (R2L), Denial of Service (DoS), and the User to Root attack (U2R). NSL-KDD is an IoT dataset used for model training purposes in healthcare applications.

B. Pre-processing

After data collection, the normalizing process is established by rescaling the attributes with a uniform contribution. Typically, the data normalization technique addresses two key problems: the presence of outliers and the presence of dominant features. The various methods for normalizing data based on the measures of statistics are examined. Consider the data with z records and N instances, as expressed numerically in Eq. (1).

$$Data = \{p_{i,n}, q_n | i \in z \; and \; n \in N\} \quad (1)$$

where, q indicates the label of class and p represents the data to be learned via a learning process. The Min-max normalization technique [28] is employed to normalize the raw data, which is one of the various normalization techniques. This approach greatly minimizes the outlier's impact on the data. It scales the obtained data within the range of 0 to 1 which is numerically expressed in Eq. (2).

$$p_{i,n} = \frac{p_{i,n} - min(p_i)}{max(p_i) - min(p_i)}(nMax - nMin) + nMin \quad (2)$$

where, max and min represent the ith attribute's maximum and minimum values. By employing $nMax$ and $nMin$, the acquired data are rescaled by the upper and lower boundaries. This acquired data is then passed as input to the feature selection.

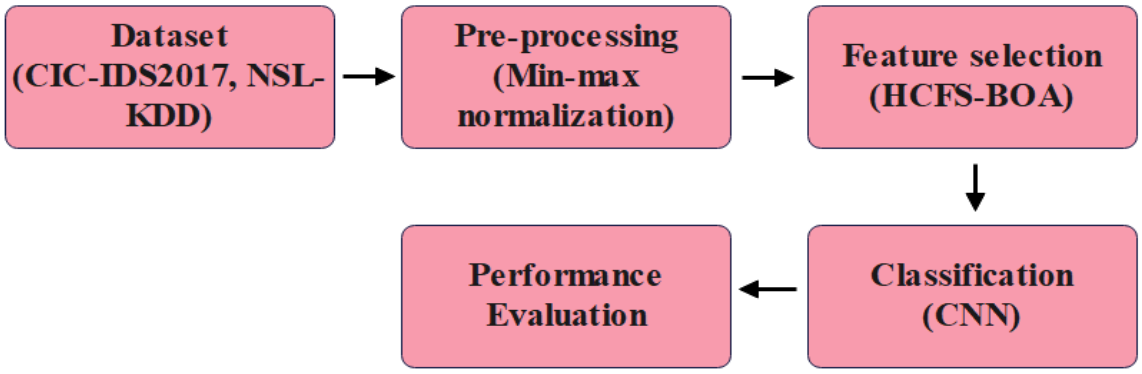

Fig. 1. Block diagram for the proposed method

C. Feature Selection

After normalizing the acquired data, the hybrid CFS-BOA approach is implemented for feature selection. In CFS-BOA, the features are selected by using a nature-inspired optimization technique to enhance the optimization process. The CFS-BOA's goal is to choose the most useful feature subset for detecting and avoiding security vulnerabilities while minimizing the redundancy and computational complexity. When compared to other optimization algorithms like Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO), the BOA tunes the optimization process for maximum efficiency for combining with CFS. The HCFS-BOA examines appropriate features that not only have significant correlations with the target variable, but also contribute to its optimal performance of intrusion detection in the healthcare system. This hybrid method has the potential to result in a more efficient and effective IDS that is specific to the unique characteristics of healthcare data and security requirements.

*1) Correlation-based Feature Selection (CFS):*One of the most known filter algorithms is CFS which selects features based on the output of a heuristic (correlation-based) evaluation function. It seeks to choose subsets whose attributes are highly correlated with the class but unassociated with one another. Repetitive features are selected based on their high correlation with at least one other feature, while low-association features are ignored. The function of the CFS feature subset assessment is mathematically expressed in Eq. (3).

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)+\overline{r_{ff}}}} \quad (3)$$

M_s – feature subset's heuristic evaluation for a feature set that includes k features

$\overline{r_{cf}}$ – average degree of connection between the category label and the features

$\overline{r_{ff}}$ – average degree of inter-connection between features

A correlation technique based on the feature subsets is used for the evaluation of CFS. During the procedure, the feature set with the greatest value is determined to decrease the training and testing set size. A larger $\overline{r_{cf}}$ or smaller $\overline{r_{ff}}$ out of the obtained subsets by the approach provides a greater evaluation value.

2) Bat Optimization Algorithm (BOA): BOA is the first algorithm for optimization and computational intelligence, influenced by microbat echolocation behavior. In a d-dimensional search, every bat flies at random with v_i^t velocity, x_i^t location and f_i frequency at t iteration. The current best solution x_* is archived for n bats in a population through an iterative search process.

The procedures for updating the location x_i^t and velocity v_i^t at each time step t are mathematically presented in Eq. (4), Eq. (5) and Eq. (6).

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad (4)$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_*)f_i \quad (5)$$

$$x_i^t = x_i^{t-1} + v_i^t \quad (6)$$

where, $\beta \in [0,1]$ is a vector selected at random from a uniform distribution.

Once a solution is chosen from the existing ideal solutions, a new solution for every bat is produced via a local random walk which is numerically expressed in Eq. (7).

$$x_{new} = x_{old} + \varepsilon A' \quad (7)$$

where, ε is a random vector generated from uniform or Gaussian distribution in the range [-1,1].

A' is the average loudness of all bats at a time step.

Furthermore, the rate of pulse emission and loudness are modified as the iterations progress. They are updated using the following Eq. (8) and Eq. (9).

$$A_i^{t+1} = \alpha A_i^t \quad (8)$$

$$r_i^{t+1} = r_i^0(1 - e^{-\gamma t} \quad (9)$$

where, $0 < \alpha < 1$ and $\gamma < 0$ are constant.

3) HCFS-BOA method for feature selection: The significance and correlation of the chosen feature subset are evaluated using the HCFS-BOA-based feature selection method. Correlation-based feature method is used in the HCFS-BOA to create a fitness function and assess the reliability of the reduced feature subset. CFS evaluates the correlation of mean feature class and the average inter-correlation between features for S feature subset with k features, where $S = (s_1, s_2, \ldots s_k)$ using (3). CFS is a classical filter method that selects relevant features based on correlation-based evaluation due to feature redundancy. By storing solutions in a bat's vector, BA is inspired by the echolocation activity of microbats, eliminating redundant features and reducing dimensionality. When a bat moves, it archives the best solution at the time. During the process of iterative search, the population scans for the optimum arrangement by updating and refreshing the position of each bat based on Eq. (4), Eq. (5), and Eq. (6). An ideal intrusion-detection approach has a higher detection rate and a lower false positive rate. Hence, a weighted fitness function is shown in Eq. (10).

$$Maximize\ Fit = w_1 \times DR \times w_2 \times (1 - FPR) \quad (10)$$

where, w_1 and w_2 are the weights for the Detection Rate and False Positive Rate, respectively. A higher fitness Fit means higher intrusion detection performance. In one iteration of the HCFS-BOA, the algorithm chooses a feature subset that depends on its correlation coefficients with the target variable. The bat optimization process involves updating the virtual bat's positions in the search space with each bat representing a potential feature subset. The technique iteratively refines feature selection by adjusting the position of bats and evaluates their performances via correlation-based metrics during both the testing and training phases. Thus, the rescaling acquired

data is passed into the feature selection phase which is sufficient for the classification of intrusion detection.

D. Classification

The selected features are classified using the CNN model which produces enormous results in domains such as Natural Language Processing (NLP), image processing, and healthcare diagnosis systems. For recognizing patterns and anomalies in network traffic or system logs, CNN classification is employed to improve intrusion detection in healthcare systems. Using CNN classification for IDS in healthcare helps to protect sensitive patient data, ensure the integrity of healthcare information systems, and avoid security breaches. It is an essential component of healthcare cybersecurity measures to protect electronic health records and vital healthcare infrastructure.

In contrast to Multi-Layer Perceptron (MLP), CNN reduces the number of neurons and parameters, resulting in rapid adaptability and minimal complexity. The CNN model offers an extensive number of clinical classification applications. CNN models are a subset of Feed-Forward Neural Network (FFNN) [29, 30] and Deep Learning models. The convolution operations convention is constant which implies that the filter is independent in function, thereby reducing the amount of parameters. Pooling, convolution, and fully connected layers are the three types of layers used in the CNN method. These layers are required for performing feature extraction, dimensionality reduction, and classification. The filter is slid on the computers through the forward pass of convolution operation, and the input capacity of the activation map that assesses the point-wise result of every score is added to obtain the activation. The sliding filter is employed by linear and convolution operators, being stated as a quick distribution of dot product. Consider w is the kernel function, x is the input, $(x \times w)(a)$ at time t is formulated as in Eq. (11).

$$(x \times w)(a) = \int x(t)w(a-t)da \tag{11}$$

Where, a is R^n for each $n \geq 1$. The parameter t is the discrete which is presented in Eq. (12).

$$(x \times w)(a) = \sum_a x(t)w(t-a) \tag{12}$$

The 2D image I is given as input, K is a 2D kernel, and the convolution is formulated as in Eq. (13).

$$(I \times K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{13}$$

In order to improve the non-linearity, two activation functions, ReLU and softmax are utilized. The ReLU is mathematically represented as in Eq. (14).

$$ReLU(x) = max(0,x)\ x \in R \tag{14}$$

The gradient $ReLU(x) = 1$ for $x > 0$ and $ReLU - (x) = 0$ for $x < 0$. The ReLU convergence ability is better than the sigmoid non-linearities. The next layer is softmax, preferable when the result requires including two or more classes which is mathematically formulated as in Eq. (15).

$$softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_i)} \tag{15}$$

The pooling layers are applied to the result in a statistic of input, and the structure of output is rescaled without losing the essential information. There are various types of pooling layers, this paper utilizes the highest pooling that individually produces large values in the rectangular neighborhood of individual points (i,j) in 2D information for every input feature correspondingly. The fully connected (FC) layer, which is the last layer with m and n output and input are illustrated further. The parameter of the output layer is stated as a weight matrix $\in M_{m,n}$. Where, m and n are rows and columns, and the bias vector $b \in R^m$. Consider the input vector $x \in R^n$, the fully connected layer output with an activation function f is formulated as in Eq. (16).

$$FC(x) \coloneqq f(Wx = b) \in R^m \tag{16}$$

where, Wx is the matrix product where function f is employed as a component. This fully connected layer is applied for classification difficulties. The FC layer of CNN is commonly involved at the topmost level. The CNN production is compressed and displayed as a single vector.

Table I shows the notation description.

TABLE I. NOTATION DESCRIPTION

Symbol	Description
$\boldsymbol{q}$	label of class
$\boldsymbol{p}$	data to be learned via a learning process
$\boldsymbol{min}$ and $\boldsymbol{max}$	$\boldsymbol{ith}$ attribute minimum and maximum values
$\boldsymbol{M_s}$	feature subset's heuristic evaluation for a feature set that includes $\boldsymbol{k}$ features
$\overline{\boldsymbol{r_{cf}}}$	the average degree of connection between the category label and the features
$\overline{\boldsymbol{r_{ff}}}$	average degree of inter-connection between features
$\boldsymbol{v_i^t}$	velocity
$\boldsymbol{x_i^t}$	location
$\boldsymbol{f_i}$	frequency at $\boldsymbol{t}$ iteration
$\boldsymbol{\varepsilon}$	random vector generated from a uniform or Gaussian distribution in the range [-1,1]
$\boldsymbol{A'}$	average loudness of all bats at time step
$\boldsymbol{w_1}$ and $\boldsymbol{w_2}$	weights for the Detection Rate and False Positive Rate
$\boldsymbol{Wx}$	Matrix product

IV. EXPERIMENTAL RESULTS

In this research, the HCFS-BOA based CNN is simulated using a Python environment with the system configuration of 16GB RAM, Intel core i7 processor, and Windows 10 operating system. The parameters like accuracy, precision, recall, and f1-score are utilized to estimate the performance of the model. The mathematical representation of these parameters is shown Eq. (17) to Eq. (20).

- Accuracy – Accuracy is the proportion of accurate predictions to all input samples and it is calculated using Eq. (10).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{17}$$

- Precision - The precision measures the percentage of actual data records versus expected data records. The performance of the classification model is greater if the precision is higher.

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

- Recall – Recall is calculated as the sum of the true positives and the positive class images.

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

- F1-Score – It is also known as the harmonic mean which seeks a balance between recall and precision.

$$F1-Score = \frac{2TP}{2TP+FP+FN} \quad (20)$$

A. Quantitative and Qualitative Analysis

This section shows the quantitative and qualitative analysis of the proposed CSF-BOA-based CNN model in terms of precision, accuracy, f1-score, and recall, as presented in Tables II, III and IV. Table II illustrates the performance of feature selection on the CIC-IDS2017 dataset. The performances of ACO, PSO, CFS, and BOA are measured and matched with the proposed HCFS-BOA. Fig. 2 represents a graphical illustration of the feature selection methods. The obtained result shows that the proposed HCFS-BOA algorithm attains an accuracy of 95.98%, precision of 94.23%, recall of 93.62%, and f1-score of 94.96% which is better when compared to the existing optimization algorithms.

Table III illustrates the performance of classification with default features using CIC-IDS2017 dataset. The performance of Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), and Recurrent Neural Network (RNN) are measured and matched with the proposed HCFS-BOA. Fig. 3 represents the graphical illustration of classification performances. The obtained result shows that the proposed HCFS-BOA algorithm attains an accuracy of 93.68%, precision of 92.92%, recall of 91.69%, and f1-score of 92.73% which is superior when compared to the existing optimization algorithms.

TABLE II. PERFORMANCE OF FEATURE SELECTION USING CIC-IDS2017 DATASET

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ACO	89.45	85.61	90.12	89.23
PSO	91.82	81.20	88.65	90.14
CFS	94.37	90.47	91.52	92.74
BOA	93.26	92.82	92.76	93.85
HCFS-BOA	95.98	94.23	93.62	94.96

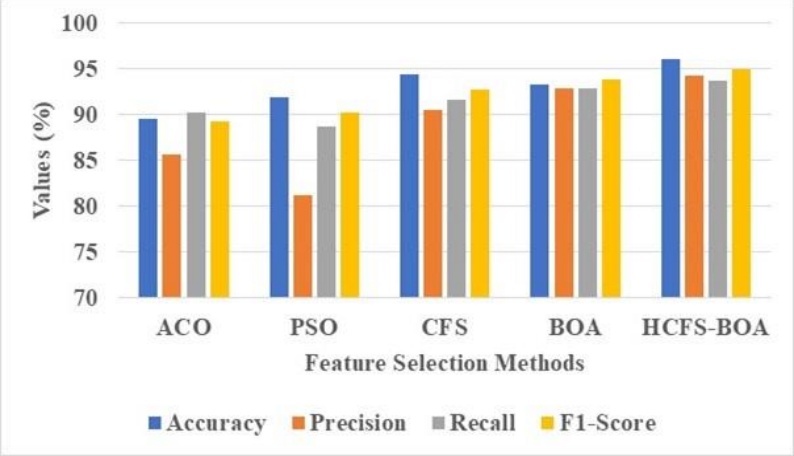

Fig. 2. Graphical representation of feature selection performances.

TABLE III. PERFORMANCE OF CLASSIFICATION WITH DEFAULT FEATURES USING CIC-IDS2017 DATASET

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	88.21	89.65	89.33	88.37
ANN	86.45	85.37	90.27	89.91
KNN	89.98	88.83	89.24	90.28
RNN	92.47	90.61	90.49	91.96
CNN	93.68	92.92	91.69	92.73

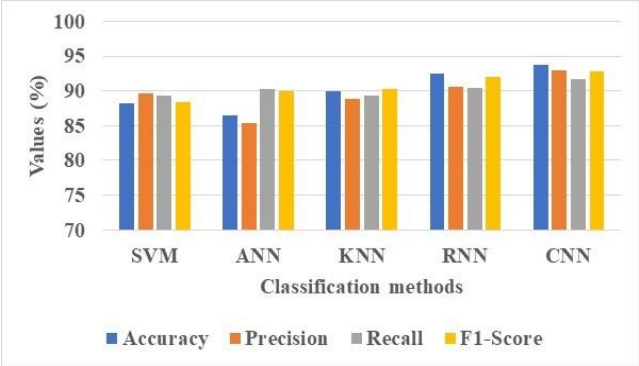

Fig. 3. Graphical representation of classification performances.

Table IV illustrates the classification outcomes with optimized features using CIC-IDS2017 dataset. The performance of SVM, ANN, KNN, and RNN are measured and matched with the optimized feature CNN. Fig. 4 illustrates the graphical representation of classification performances with optimized features. The obtained outcomes prove that the CNN algorithm accomplishes an accuracy of 99.45%, precision of 98.89%, recall of 98.67%, and f1-score of 97.98%, therefore being superior in contrast to the existing optimization algorithms. The ACO, PSO, CFS, and BOA consume 25 seconds, 29 seconds, 31 seconds, and 35 seconds of time, respectively. The time analysis of HCFS-BOA with CNN demands a training time of 20 seconds, being more robust in comparison with other optimization techniques like ACO, PSO, CFS, and BOA on the CIC-IDS2017 dataset. Table V shows the performance of classification with optimized features on the NSL-KDD dataset. Fig. 5 shows that the obtained outcomes of optimized results of the CNN algorithm accomplishes an accuracy of 98.13%, precision of 97.36%, recall of 97.07%, and f1-score of 95.34%, in that way, proving more robust in contrast to the previous optimization algorithms. The ACO, PSO, CFS, and BOA require 22 seconds, 25 seconds, 28 seconds, and 34 seconds of time, respectively. The time analysis of HCFS-BOA with CNN needs a training time of 15 seconds which is lesser than that of the previous optimization techniques like ACO, PSO, CFS, and BOA on the NSL-KDD dataset.

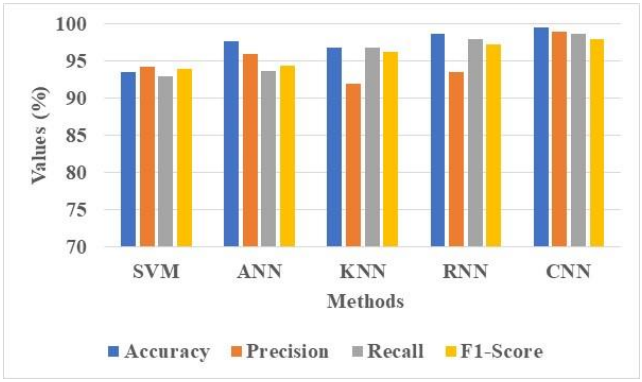

Fig. 4. Graphical representation of optimized feature performances using CIC-IDS 2017.

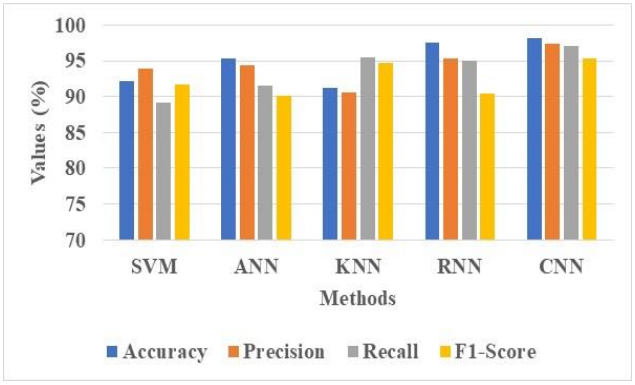

Fig. 5. Graphical representation of optimized feature performances using NSL-KDD.

TABLE IV. PERFORMANCE OF CLASSIFICATION WITH OPTIMIZED FEATURES USING CIC-IDS2017 DATASET

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	93.54	94.21	92.89	93.96
ANN	97.68	95.86	93.65	94.37
KNN	96.73	91.85	96.73	96.18
RNN	98.66	93.47	97.87	97.25
CNN	99.45	98.89	98.67	97.98

TABLE V. PERFORMANCE OF CLASSIFICATION WITH OPTIMIZED FEATURES USING NSL-KDD DATASET

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	92.17	93.95	89.20	91.66
ANN	95.38	94.36	91.51	90.17
KNN	91.20	90.51	95.47	94.68
RNN	97.54	95.34	94.93	90.50
CNN	98.13	97.36	97.07	95.34

B. Comparative Analysis

This section provides the comparative analysis of proposed HCFS-BOA based CNN model on the evaluation metrics: precision, accuracy, f1-score, and recall as shown in Table VI. The previous methods namely, SafetyMed, HIDM, and BDSDT are employed to assess the HCFS-BOA based CNN performance. SafetyMed [16] achieves 97.63% accuracy, 98.47% precision, 97% recall, and 97.73% f1-score. HIDM [17] achieves 85% accuracy and BDSDT [18] achieves 99.04% accuracy, 87.31% precision, 82.89% recall, and 83.2% f1-score. When compared with the existing methods, the proposed HCFS-BOA based CNN achieves higher accuracy of 99.45%, precision of 98.89%, 98.67% of recall, and 97.98% of f1-score.

TABLE VI. COMPARATIVE ANALYSIS WITH EXISTING METHODS

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SafetyMed [16]	CIC-IDS2017	97.63	98.47	97	97.73
HIDM [17]	CIC-IDS2017	85	N/A	N/A	N/A
BDSDT [18]	CIC-IDS2017	99.04	87.31	82.89	83.2
Proposed HCFS-BOA based CNN	CIC-IDS2017	99.45	98.89	98.67	97.98

C. Validation of Real-Time Applications

The NSL-KDD dataset is commonly deployed for intrusion detection in IoT to ensure reliability and security for healthcare systems. This research uses the NSL-KDD dataset for training and validation purposes on real-time applications in the cloud. The NSL-KDD dataset is split into training, testing, and validation in the ratio of 70:15:15. IDS is created to detect the different types of attacks by evaluating system logs, network traffic, and behavioral patterns. Malware attacks, DoS attacks, Cross-Site Scripting (XSS), etc., are different attacks. These types of attacks are performed when the patient information is blocked or stolen by attackers. Therefore, the NSL-KDD dataset is employed for model training purposes to reduce the attacks in real-time healthcare applications.

D. Discussion

The CIC-IDS-2017 dataset is beneficial for intrusion detection because of its comprehensive representation of realistic traffic network scenarios with different types of attacks and normal activities. It provides a labelled and large-scale dataset that assists the evaluation and enhancement of intrusion detection with enhanced robustness and accuracy. The NSL-KDD dataset is beneficial for intrusion detection as it solves limitations in the original KDD Cup dataset by minimizing redundancy and managing a more balanced distribution of classes. It generates the representation of a more realistic modern traffic network that contains normal behavior and different wider attacks that maximize intrusion detection robustness. By using these two datasets, the proposed approach is analyzed by generic type. Moreover, the advantages of the proposed method and the limitations of existing methods are discussed. The existing methods have some limitations such as the SafetyMed method [16] has no defense mechanism against an attack of AML. Combining various detection systems increases the risk of false negatives and false positives in HIDM [17]. BDSDT [18] isn't effective against web and Bot threats since there are fewer instances of these two attack classes. The proposed HCFS-BOA-based CNN model overcomes the existing models' limitations.

To overcome the problem of AML attack, CFS is used to identify highly informative features for minimizing the risk of adversarial manipulations compared to other algorithms. BOA assists in identifying an optimal subset of features that maximizes detection accuracy and reduces the risk of false positives and false negatives. This is done by focusing on informative features in CFS that assist in enhancing the model's ability to discriminate between various attack classes like web and Bot threat. Combining CFS with BOA enables appropriate features that not only have significant correlations with the target variable but also contribute to the optimal performance of intrusion detection in the healthcare system, in contrast to the other methods. The CNN is deployed to identify and classify intrusion accurately and effectively. New attacks such as web and Bot threat attacks are classified effectively by using CNN. The proposed HCFS-BOA-based CNN achieves a superior accuracy of 99.45% when compared with the existing methods namely, SafetyMed, HIDS, and BDSDT.

V. CONCLUSION

In this research, the HCFS-BOA based on the CNN model is proposed for intrusion detection to secure the entire network in the healthcare system. The proposed method mainly comprises four stages: dataset, min-max normalization, feature selection, and classification. Initially, the data is obtained from the CIC-IDS2017 and NSL-KDD datasets, after which the min-max normalization is performed to normalize the acquired data. For feature selection, HCFS-BOA is employed for optimal performance of intrusion detection in healthcare systems. Finally, the CNN is deployed to identify and classify intrusion accurately and effectively. The proposed HCFS-BOA-based CNN achieves a better accuracy of 99.45% when compared with the existing methods like SafetyMed, HIDS, and BDSDT. In the future, hyperparameter tuning can be applied in feature selection for improving the model's performance.

REFERENCES

[1] K. Samunnisa, G. S. V. Kumar, and K. Madhavi, "Intrusion detection system in distributed cloud computing: Hybrid clustering and classification methods," Meas.: Sens., vol. 25, p. 100612, Feb. 2023.

[2] A. H. Janabi, T. Kanakis, and M. Johnson, "Overhead Reduction Technique for Software-Defined Network Based Intrusion Detection Systems," IEEE Access, vol. 10, pp. 66481-66491, Jun. 2022.

[3] P. B. Udas, M. E. Karim, and K. S. Roy, "SPIDER: A shallow PCA based network intrusion detection system with enhanced recurrent neural networks," J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 10B, pp. 10246–10272, Nov. 2022.

[4] O. A. Alzubi, J. A. Alzubi, M. Alazab, A. Alrabea, A. Awajan, and I. Qiqieh, "Optimized Machine Learning-Based Intrusion Detection System for Fog and Edge Computing Environment," Electronics, vol. 11, no. 19, p. 3007, Sep. 2022.

[5] M. Bakro, R. R. Kumar, A. A. Alabrah, Z. Ashraf, S. K. Bisoy, N. Parveen, S. Khawatmi, and A. Abdelsalam, "Efficient Intrusion Detection System in the Cloud Using Fusion Feature Selection Approaches and an Ensemble Classifier," Electronics, vol. 12, no. 11, p. 2427, May 2023.

[6] H. Lin, Q. Xue, J. Feng, and D. Bai, "Internet of things intrusion detection model and algorithm based on cloud computing and multi-feature extraction extreme learning machine," Digital Commun. Networks, vol. 9, no. 1, pp. 111–124, Feb. 2023.

[7] A. K. Samha, N. Malik, D. Sharma, S. Kavitha, and P. Dutta, "Intrusion Detection System Using Hybrid Convolutional Neural Network," Mobile Networks Appl., Aug. 2023.

[8] V. Hnamte, H. Nhung-Nguyen, J. Hussain, and Y. Hwa-Kim, "A Novel Two-Stage Deep Learning Model for Network Intrusion Detection: LSTM-AE," IEEE Access, vol. 11, pp. 37131-37148, Apr. 2023.

[9] R. A. Bakar, X. Huang, M. S. Javed, S. Hussain, and M. F. Majeed, "An Intelligent Agent-Based Detection System for DDoS Attacks Using Automatic Feature Extraction and Selection," Sensors, vol. 23, no. 6, p. 3333, Mar. 2023.

[10] R. Zhao, Y. Mu, L. Zou, and X. Wen, "A Hybrid Intrusion Detection System Based on Feature Selection and Weighted Stacking Classifier," IEEE Access, vol. 10, pp. 71414-71426, Jun. 2022.

[11] R. Shrestha, A. Omidkar, S. A. Roudi, R. Abbas, and S. Kim, "Machine-Learning-Enabled Intrusion Detection System for Cellular Connected UAV Networks," Electronics, vol. 10, no. 13, p. 1549, Jun. 2021.

[12] D. Javeed, M. S. Saeed, I. Ahmad, P. Kumar, A. Jolfaei, and M. Tahir, "An Intelligent Intrusion Detection System for Smart Consumer Electronics Network," IEEE Trans. Consum. Electron., May 2023.

[13] A. Bhardwaj, R. Tyagi, N. Sharma, A. Khare, M. S. Punia, and V. K. Garg, "Network intrusion detection in software defined networking with self-organized constraint-based intelligent learning framework," Meas.: Sens., vol. 24, p. 100580, Dec. 2022.

[14] V. Hnamte and J. Hussain, "Dependable intrusion detection system using deep convolutional neural network: A Novel framework and performance evaluation approach," Telematics and Informatics Reports, vol. 11, p. 100077, Sep. 2023.

[15] S. Shitharth, P. R. Kshirsagar, P. K. Balachandran, K. H. Alyoubi, and A. O. Khadidos, "An Innovative Perceptual Pigeon Galvanized Optimization (PPGO) Based Likelihood Naïve Bayes (LNB) Classification Approach for Network Intrusion Detection System," IEEE Access, vol. 10, pp. 46424-46441, May 2022.

[16] N. Faruqui, M. A. Yousuf, M. Whaiduzzaman, A. K. M. Azad, S. A. Alyami, P. Liò, M. A. Kabir, and M. A. Moni, "SafetyMed: A Novel IoMT Intrusion Detection System Using CNN-LSTM Hybridization," Electronics, vol. 12, no. 17, p. 3541, Aug. 2023.

[17] L. K. Vashishtha, A. P. Singh, and K. Chatterjee, "HIDM: A hybrid intrusion detection model for cloud based systems," Wireless Pers. Commun., vol. 128, no. 4, pp. 2637–2666, Feb. 2023.

[18] P. Kumar, R. Kumar, G. P. Gupta, R. Tripathi, A. Jolfaei, and A. K. M. N. Islam, "A blockchain-orchestrated deep learning approach for secure data transmission in IoT-enabled healthcare system," J. Parallel Distrib. Comput., vol. 172, pp. 69–83, Feb. 2023.

[19] A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, "CNN-LSTM: Hybrid Deep Neural Network for Network Intrusion Detection System," IEEE Access, vol. 10, pp. 99837-99849, Sep. 2022.

[20] H. Han, H. Kim, and Y. Kim, "An Efficient Hyperparameter Control Method for a Network Intrusion Detection System Based on Proximal Policy Optimization," Symmetry, vol. 14, no. 1, p. 161, Jan. 2022.

[21] M. Bakro, R. R. Kumar, A. Alabrah, Z. Ashraf, M. N. Ahmed, M. Shameem, and A. Abdelsalam, "An Improved Design for a Cloud Intrusion Detection System Using Hybrid Features Selection Approach With ML Classifier," IEEE Access, vol. 11, pp. 64228-64247, Jun. 2023.

[22] K. M. Sudar, M. Beulah, P. Deepalakshmi, P. Nagaraj, and P. Chinnasamy, "Detection of Distributed Denial of Service Attacks in SDN using Machine learning techniques," in 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, IEEE, 2021, pp. 1-5.

[23] V. Praveena, A. Vijayaraj, P. Chinnasamy, I. Ali, R. Alroobaea, S. Y. Alyahyan, and M. A. Raza, "Optimal deep reinforcement learning for intrusion detection in UAVs," Computers, Materials & Continua, vol. 70, no. 2, pp. 2639-2653, 2022.

[24] P. Chinnasamy, S. Devika, V. Balaji, S. Dhanasekaran, B. J. A. Jebamani, and A. Kiran, "BDDoS: Blocking Distributed Denial of Service Flooding Attacks With Dynamic Path Detectors," in 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, IEEE, 2023, pp. 1-5.

[25] P. Chinnasamy, N. Kumaresan, R. Selvaraj, S. Dhanasekaran, K. Ramprathap, and S. Boddu, "An Efficient Phishing Attack Detection using Machine Learning Algorithms," in 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, IEEE, 2022, pp. 1-6.

[26] E. Anupriya, N. Kumaresan, V. Suresh, S. Dhanasekaran, K. Ramprathap, and P. Chinnasamy, " Fraud Account Detection on Social Network using Machine Learning Techniques," in 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), IEEE, 2022, pp. 1-4.

[27] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "RTIDS: A Robust Transformer-Based Approach for Intrusion Detection System," IEEE Access, vol. 10, pp. 64375–64387, Jun. 2022.

[28] S. M. Kasongo, "A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework," Comput. Commun., vol. 199, pp. 113–125, Feb. 2023.

[29] H. Zhang, B. Zhang, L. Huang, Z. Zhang, and H. Huang, "An Efficient Two-Stage Network Intrusion Detection System in the Internet of Things," Information, vol. 14, no. 2, p. 77, Jan. 2023.

[30] M. A. Siddiqi and W. Pak, "Tier-Based Optimization for Synthesized Network Intrusion Detection System," IEEE Access, vol. 10, pp. 108530–108544, Oct. 2022.

[31] V. Ravuri and S. Vasundra, "Moth-flame optimization-bat optimization: Map-reduce framework for big data clustering using the Moth-flame bat optimization and sparse Fuzzy C-means," Big Data, vol. 8, no. 3, pp. 203-217, Jun. 2020.

[32] S. Vasundra and G. Rajeswarappa, "Hybrid Grasshopper and Improved Bat Optimization Algorithms-based clustering scheme for maximizing lifetime of Wireless Sensor Networks (WSNs)," International Journal of Intelligent Engineering and Systems, vol. 15, pp. 536-546, May. 2022.

[33] S. Vasundra and A. Balaram, "A Hybrid Soft Computing Technique for Software Fault Prediction based on Optimal Feature Extraction and Classification," International Journal of Computer Science and Network Security, vol. 22, no. 5, pp. 348-358, May. 2022.

[34] S. Vasundra and Vasavi Ravuri, "An effective weather forecasting method using a deep long-short-term memory network based on time-series data with sparse fuzzy c-means clustering," Engineering Optimization, vol. 55, no. 9, pp. 1437-1455, Sep. 2022.