Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFECTIVE IDS FRAMEWORK FOR IOT USING FEATURE SELECTION AND CLASSIFICATION MODEL

L. SARALADEVE[1], A. CHANDRASEKAR[2]

[1]Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India.

[2]Research Supervisor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India.

E-mail: [1]saraladeve@gmail.com

ABSTRACT

Devices and services that are part of the Internet of Things (IoT) bring convenience but are loaded with significant security risks. When protecting the IoT environment, the Intrusion Detection System (IDS) for the network is really important. Thus, a new hybrid IDS model is developed in this research work to classify and detect attacks present in IoT networks. The proposed model integrates feature selection and classification processes implemented using machine learning (ML) models. The Binary-Black Widow Optimization (BBWO) algorithm selects the optimal feature sets from the given datasets, and the Logistic Regression (LR) algorithm performs the binary classification. The proposed model initially performed data preprocessing using min-max scaling normalization for dataset standardization. After preprocessing, the datasets are split into training and test sets for evaluation. Using the datasets, the features are selected optimally using the BBWO algorithm. The classification is performed using the LR algorithm based on the selected optimal feature sets. The performance of this research model is evaluated based on accuracy, detection rate, FPR, f1-score and precision. The results are evaluated individually for both datasets and correlated. The BBWO-LR model obtained 98.83% accuracy, 98.32% detection rate, 99.58% precision, and 98.95% f1-score for the CICIDS-2017 data set. Using the CICIDS-2018 data set, the BBWO-LR model obtained 98.92% accuracy, 98.17% detection rate, 99.76% precision, and 98.97% f1-score.

Keywords: *Internet of Things, IDS, Feature Selection, BBWO, LR, CICIDS-2017, CICIDS-2018*.

1. INTRODUCTION

The conventional methods of perceiving one's physical surroundings have been largely rendered obsolete by recent developments in communication technologies, such as the IoT. IoT technologies have the potential to provide modernizations that improve the quality of life and have the ability to gather, quantify, and analyze the environments around them [1]. IoT is a rapidly developing domain in the computing history. It plays an essential part in the improvement of real-world smart applications, like smart homes, smart transportation, smart education, and smart healthcare. In contrast, the interconnected and expansive manner of IoT systems, together with the myriad of elements that consider the deployment of such systems, has resulted in the emergence of new security challenges [2].

The architecture of the IoT can be readily understood as an abstraction of numerous different hierarchical layers. In figure 1, the architecture of the IoT is portrayed as having four layers. Physical attacks, network attacks, software attacks, and breaches of privacy are all threats that could potentially affect IoT systems, which are comprised of items, services, and networks [3].

The foundation of security systems is the application of AI to the creation of a variety of models that are able to rapidly evaluate network data and forecast the nature of the data. New challenges in machine learning have emerged in the field of data science as a result of the high velocity and enormous amount of data generated by the IoT. These issues have been broken down into a variety of categories, including clustering, classification, forecasting, and regression, amongst others. Nearly every year, at least one or many

Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

attacks are carried out, resulting in the malfunction of a variety of cloud-based platforms and applications or in the potential for data to be compromised [4].

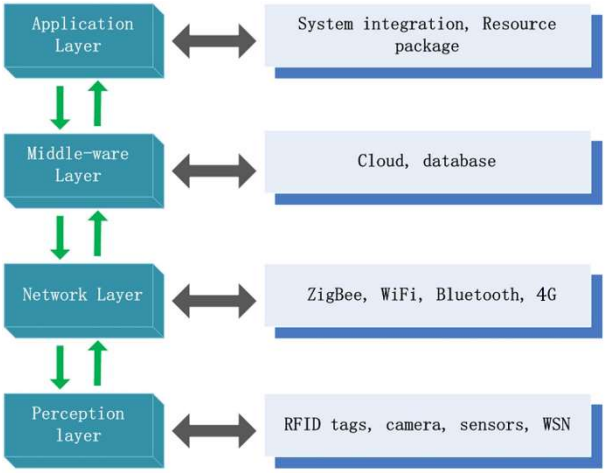

Figure 1: IoT Architecture

Operating attacks are an excellent illustration of an application that takes use of stream data because of the nature of its streaming. In addition, every attack identification or detection system needs to evaluate the information, extricate its features, and use machine learning to either interpret the data directly or implicitly depending on the situation. In this part of the process, systems are trained on past experiences using labelled information from well-known as well as less common threats. This suggests that attack detection systems learn from the stream of data that they receive. In addition, ML and DL approaches are essential when it comes to the prediction of new attacks, which are variants of earlier attacks. This is because these methods are able to intelligently predict future unknown attacks by learning from prior examples. IDS is absolutely necessary for detecting these kinds of attacks on computers and networks and alerting the users. An IDS can be installed on individual hosts inside a network, at a central location with dedicated resources, or spread across a network. IDSs that are designed to detect attacks on networks of computers, as opposed to attacks on a single host, are referred to as Network IDS [5].

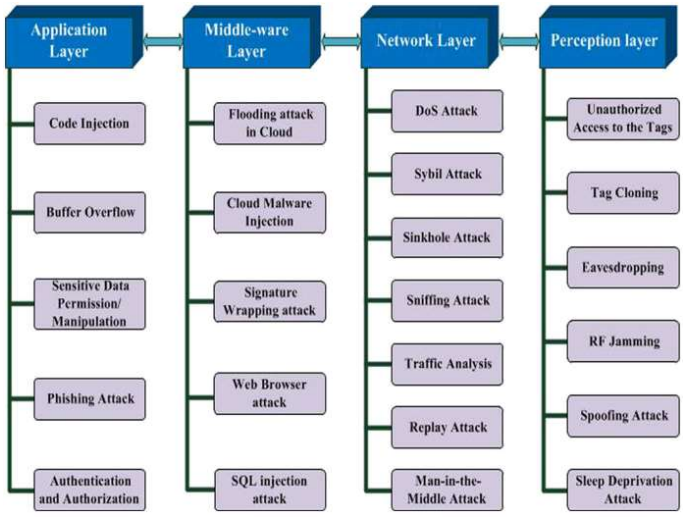

Figure 2: Attack Categories in Different IoT Layers

The IoT has various different traffic aspects that make up its network traffic. Because there are so many characteristics present in IoT network traffics, the construction of the ML models takes significantly more time and has a negative effect on the overall performances of the IDS. As a result, feature selection is necessary for intrusion detection in IoT in order to construct models in the shortest amount of time possible while yet achieving a higher level of performance in attack classification.

A procedure that is utilized to improve the overall performances of a ML model is known as feature selection. This approach involves selecting a small subset of relevant features from a larger collection of available features. The process of selecting features can be thought of as an optimization problem, the aim of which is to discover the subset of characteristics that minimizes a specified objective function while adhering to a set of predetermined limitations. The objective function that is applied during feature selection can be different for each problem that is being solved. For instance, it could quantify the accuracy of classification, the error in regression, or the complexity of the model [6].

A wrapper methodology is a typical method that examines the performances of a ML model by using a subset of features, and then selects the subset that maximizes the performance of the model. Overall, the feature selection formulation is an essential stage in the process of creating efficient machine learning models. This is because it can assist in lowering the likelihood of overfitting, enhancing generalization performance, and enhancing interpretability.

The contribution of the research includes:

- A feature selection method using a modified BWO algorithm called Binary-BWO is proposed to select optimal features for the developed IDS model.
- The proposed feature selection model was trained and tested on CIC-IDS 2017 & 2018 datasets.
- The BBWO method performs better using selected features with the Logistic Regression classifier on CIC-IDS-2018 and 2017 datasets.
- The proposed system is also compared to the existing models for validation based on accuracy, detection rates, f-measure, false positive rates (FPR), and precision.

A feature selection method using a modified BWO algorithm called Binary-BWO is proposed to select optimal features for the developed IDS model:

Feature selection is an important step in building an IDS. The BBWO algorithm is a modified version of the BWO. The BBWO algorithm is designed to identify and select the most relevant features from a given dataset. By applying BBWO, the proposed method aims to optimize the feature set used in the IDS model, improving its performance. The proposed feature selection model was trained and tested on CIC-IDS 2017 & 2018 datasets: To evaluate the effectiveness of the proposed feature selection model, it was trained and tested on the CIC-IDS 2017 and 2018 datasets. The CIC-IDS datasets are commonly used in research for evaluating IDS models. By using these datasets, the researchers could assess the performance of their feature selection model in the context of real-world network traffic data. The BBWO method performs better using selected features with the Logistic Regression classifier on CIC-IDS-2018 and 2017 datasets: The results of the experiments conducted with the BBWO method and the selected features were compared to other feature selection algorithms. The performance of the BBWO method, in combination with the Logistic Regression classifier, was found to be superior when applied to the CIC-IDS-2018 and 2017 datasets. This suggests that the BBWO algorithm effectively identifies relevant features that contribute to the accuracy and effectiveness of the IDS model. The proposed system is also compared to the existing models for validation based on accuracy, detection rates, f-measure, FPR, and precision: To validate the proposed system, it was compared to existing models using various evaluation metrics such as accuracy, detection rates, f-measure, and precision. These metrics provide insights into different aspects of the system's performance. By comparing the proposed system to existing models, the researchers can determine if their approach offers improvements in terms of these evaluation metrics, indicating the effectiveness and superiority of their system.

In summary, the statements describe a research proposal that introduces a hybrid method (BBWO-LR) for an IDS model. The proposed method is evaluated using real-world datasets, and the results indicate improved performance compared to existing models. The validation is based on several evaluation metrics to assess the effectiveness of the proposed system.

The paper is structured into several sections. Section II presents an analysis of the relevant literature. Section III discusses the implementation of the research model, including BBWO-based feature selection and LR-based classification. Section IV presents the experimental results of the BBWO-LR model and comparison with existing IDS models from the literature review. Section V concludes the work with the summary of the findings and recommendations for future research.

2. LITERATURE REVIEW

In recent developments, one new trend that offers an increased discovery rate is the detection of intruders utilizing metaheuristics and machine learning approaches. Thus, an enhanced Binary Manta Ray Foraging Optimizations based on an adaptive S-shaped function and a Random Forest (RF) algorithm was proposed to be used in an intrusion detection model in [6]. The BMFR was conceived with the intention of locating the features that were the most important for intrusion detection datasets and removing those that were unnecessary or redundant. In addition, the RF was applied in both the evaluation of the features and the construction of the intrusion detection model. Using the CICIDS-2017 data set, this model demonstrated significantly improved performance. When compared to other methodologies, this model for the CIC-IDS2017 dataset displayed a substantial departure from the others. In [7], a novel process flow for filter-based features selection using a transformation technique was presented. In most cases, the process of normalization or transformation comes first, followed by classification. Before moving on to the feature selection process, the impacts of

normalization were first introduced and then evaluated.

The network IDS is an essential component in ensuring the safety of the IoT network. Therefore, a dual-phase IDS model combined with ML and deep learning was developed in [8] to manage the network traffic data class imbalance. This model achieved fine-grained attack classification on large scale data based on the CICIDS-2018 data set. At the first phase, the LightGBM model was used to differentiate between typical and unusual patterns of network traffic. The samples that were anticipated to be abnormal in stage one was used in stage two of the Convolutional Neural Networks (CNN) process to perform fine-grained detection of attack class on those data. The model's performances could have been enhanced even more by using methodical approaches to feature selection.

The vast majority of the currently available IoT IDS systems are static, meaning that they are not able to learn from the experience of a past attack. Artificial intelligence (AI) is an effective technology that could learn from prior attacks eventually, recognize attack from the normal traffics, and inform the appropriate systems. AI techniques like ML and deep learning have the potential to give tremendous capabilities to meet the demands of the IoT. Using the CICIDS-2017 data set, a comparison study of the performances of deep neural networks, CNN, and long short-term memories was reported in [9]. According to the findings, DNN was able to obtain an accuracy of 94.61%, while LSTM and CNN were able to achieve 97.67% and 98.61%, respectively.

A hybrid framework that used a deep learning model referred to as "ImmuneNet" was developed in [10] as a way to detect intrusion threats. In order to achieve higher accuracy and performances, this model made use of several processes of feature engineering, oversampling approaches for improving class balance, and hyper-parameters optimization strategies. The design has approximately one million parameters, which allows it to be lightweight, quick, and IoT-friendly. As a result, it is appropriate for implementing the IDS on clinical devices and health care devices. ImmuneNet obtained higher true positive rates compared to a low false positive rate, which indicates that it does not have a bias towards false positives or negatives when using the CICIDS-2018 and 2017 datasets.

A plausible IDS that is capable of thwarting the vast majority of attacks used in current times. A hybrid feature selection strategy was developed in [11] for reducing the latency of prediction with not impacting the attack identification performances by reducing the complexity of the model. This was done in order to lower the complexity of the model. For the most recent CICIDS-2018 data set, the model was constructed with the help of a tool called Light Gradient Boosting Machine (GBM), which was a rapid GBM. Random Forest was utilized to select the essential features, and principal components analysis (PCA) was then implemented to those features as a method for reducing the dimensions of the data. In conclusion, the LightGBM was applied for attack classification, where it achieved an accuracy of 97.73% while also having a low prediction latency.

The research that was presented in [12] made use of the idea of a one class classifier for solving the issue of finding anomalies in the communication network. The algorithm was constructed using a polynomial interpolation method and statistical analysis. This methodology was utilized in the benchmarking of datasets such as UNSW-NB15, KDD99, CICIDS-2018, and EDGE-IIOTSET 2022. In order to detect anomalies, the one-class classifier was more important than being able to specify the type of anomaly. Because the attack scenarios were always changing, taking this approach was very necessary for safety-critical applications like defense.

The effectiveness of machine learning ensemble approaches in terms of learning has been thoroughly demonstrated. The work in [13] presented an original IDS model that makes use of ensemble machine learning techniques. The CICIDS-2017 dataset was combed through for attributes that may be used to enhance classification accuracy and reduce the number of false positives. ML methodologies like decision trees, RF, and support vector machine were utilized in the creation of an IDS model. An ensemble approach voting classifier was implemented after the training of these models, and it attained an accuracy of 96.25%.

Specifically for the features selection and analysis of intrusions, the most recent variants of AI models are required. Hence, the research published in [14] highlighted the most up-to-date AI models for IDS in the IoT network in order to create a safe network. Moreover, the research presented the security challenges that exist within IoT-based environments. CNN was utilized for classifying the attacks present in the IoT network.

Additionally, the work provided a particles swarm optimization methodology for extracting significant characteristics from the information. This method was able to choose relevant characteristics automatically, which were then utilized to classify the datasets. This model performed better on the CIC-IDS 2017 data set, which was intended to evaluate how well it could detect intrusions.

IDS models provide the ability to automatically detect dangerous attacks. On the other hand, hostile threats are always emerging and evolving, which means that the network demands an advanced security solution. The work presented in [15] described the development of a hybrid IDS based on deep learning that used convolutional recurrent neural network (CRNN) to analyze, forecast, and classify harmful intrusions that occur within a network. To improve the IDS performance and prediction based on the CSE-CIC-DS2018 dataset, this model utilized a CNN for capturing local characteristics and a RNN to capture temporal features. Both networks were trained utilizing the same data set. This model improved both the accuracy and detection rate of the IDS while simultaneously lowering the computational complexity.

By utilizing BBWO, the proposed model can enhance the feature selection process compared to other methodologies. BBWO has the potential to handle class imbalance effectively, thus improving the classification accuracy for IoT intrusion detection. By learning from past attacks, our model can adapt and improve its intrusion detection capabilities over time. The Logistic Regression algorithm is known for its efficiency and low computational complexity, further contributing to reducing latency and complexity in the model. The proposed BBWO-LR model has the potential to address the limitations mentioned in the paragraphs by offering an efficient feature selection process, handling class imbalance, learning from past attacks, improving accuracy and performance, and reducing prediction latency and model complexity. However, it is important to conduct thorough experimentation and comparative analysis to validate the effectiveness of our model in comparison to existing approaches. BBWO-LR model, with its novel combination of the BBWO algorithm and Logistic Regression, can contribute to fulfilling this need by providing an advanced AI-based solution for feature selection and classification in IoT IDS. BBWO-LR model can overcome the limitations of the existing hybrid model by incorporating the Binary-Black Widow Optimization algorithm for feature selection, which can enhance the input representation for deep learning models like CRNN. This can potentially improve the accuracy, detection rate, and computational complexity of the IDS. by incorporating the BBWO algorithm for feature selection and Logistic Regression for classification, can potentially address this limitation by not only detecting anomalies but also providing fine-grained classification of different types of intrusions in IoT networks.

3. PROPOSED IDS MODEL

This work proposes a new hybrid IDS model for detecting and classifying attacks. This proposed hybrid model combines feature selection and classification process, developed using the BBWO and logistic regression algorithms. The BBWO algorithm is implemented for features selection, and based on the selected features, the logistic regression algorithm is implemented for classification. This research used CIC-IDS-2017 and 2018 datasets to evaluate the research model. The proposed research model is a binary classification model, which includes data preprocessing, feature selection, classification, and performance analysis stages. Figure 3 represents the pipeline of the proposed IDS model.

3.1. Datasets

CIC-IDS-2017: The Canadian Institute for Cybersecurity's Intrusion Detection System-2017 (CICIDS-2017) data set is an accumulation of network traffics data that was generated specifically for the purpose of being used in research regarding network security. The dataset contains network traffic captures of attacks, such as Botnet attacks, DDoS attacks, and Brute-force attacks, that were carried out on an emulated network. It is comprised of 3.2 million packets and has been compiled from nine distinct attack scenarios. In the field of research on network intrusion detection and security, the CIC-IDS-2017 dataset is utilized extensively. It has been applied to the process of determining the efficacy of various ML algorithms in the detection of network intrusions as well as the creation of novel intrusion detection strategies. The dataset known as CIC-IDS-2017 has 82 attributes that are connected to network traffic. The dataset can be obtained for research purposes by downloading it from the CIC website [16].

Table 1: CIC-IDS-2017 Dataset Description

Type of Attack	For Training	For Testing
Web_Attack_SQL_Injection	19	4
Web_Attack_XSS	575	129
Web_Attack_Brute_Force	1329	299
SSH_Patator	5201	1169
FTP_Patator	6997	1574
DDoS	112901	25388
PortScan	140043	31492
Benign	727397	163572
Total	904056	223627

CIC-IDS-2018: This is an evaluation dataset that is made available to the public and is used for intrusion detection system testing. It was developed at the University of New Brunswick in Canada, which is located in Canada, by the CIC. The dataset includes statistics on both good and harmful network traffic. These malicious attacks include Brute Force Password Guessing, Distributed Denial of Services (DDoS), and Web Application Attacks, among others. A simulated network environment was used to collect real-world data traffic, which was then used to construct the dataset. Raw packet capture data and preprocessed features are both included in the CICIDS-2018 data set, which was created with the intention of accurately representing the network traffic that may be seen in real-world scenarios. The flow duration, the number of packets and bytes, as well as a variety of statistics generated from the payloads of the packets are all included in the preprocessed features. This dataset includes 89 different types of traffic features, ranging from the most fundamental to the most advanced, including packet lengths, protocol types, flow, timestamps, and content-related aspects. Researchers working in the fields of intrusion detection and cybersecurity have made substantial use of this dataset, which can be downloaded from the CIC website [16].

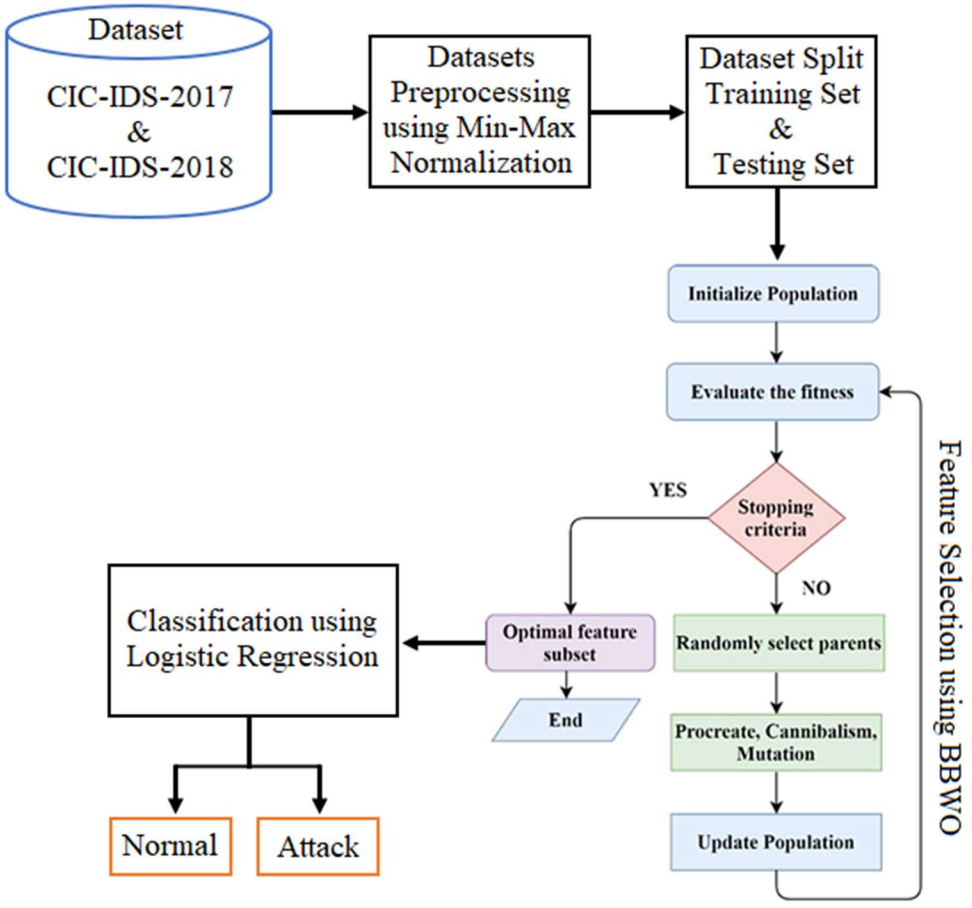

Figure 3: Proposed IDS Model using BBWO-LR

Table 2: CIC-IDS-2018 Dataset Description

Classes	Samples to Test	Samples to Train	Total
Benign	2696942	10787766	13484708
Bot	57238	225983	286191
Brute force-XSS	46	184	230
DDoS attacks LOIC UDP	346	1384	1730
Brute-force Web	122	489	611
DDoS attacks HOIC	137203	548809	686012
DoS attacks SlowHTTPtest	27978	111912	139890
DoS attacks slowloris	2198	8792	10990
FTP Brute Force	38672	154688	193360
Dos attacks goldeneye	8302	33206	41508
SSH Brute Force	37518	150071	187589
SQL injection	17	70	87
DDoS attacks LOIC	115238	460953	576191
Infiltration	32387	129547	161934
DoS attacksHulk	92382	369530	461192

3.2. Data Preprocessing

The phase of data preprocessing is a key one in the process of developing machine learning models. It is of particular significance in the context of IoT attack datasets, which can be rather vast and complicated. The actions listed below are those that were taken in order to pre-process the data for the used IoT attack datasets:

Data Cleaning: During this stage, the data with any missing values, duplicated data, or outliers that have been found and eliminated. In the context of a dataset pertaining to an IoT attack, this may involve erasing data that is either insufficient or corrupted, as well as data that falls outside the typical range of values that are anticipated.

Scaling the data: This method was utilized in order to standardize the range of values present in numerical characteristics to a single scale. In the process of preparing the data for machine learning models, this step is significant since it helps to ensure that each feature contributes equally to the model, regardless of the scale on which it was originally measured. The min-max scaling normalization method was utilized in the performance of this process. A technique for the preprocessing of data known as normalization scales the values of numerical features to fall into a similar range, which is most commonly between 0 and 1. The purpose of the normalization process is to ensure that each feature contributes the same amount to the machine learning model, despite the scale at which it was first measured. The min-max scaling algorithm converts the value of each feature, x, into a new value, x_{norm}, by utilizing the following formula:

$$x_{norm} = \frac{(x - m\ \ (x))}{max(x) - min(x)} \quad (1)$$

In this context, the values of minimum and maximum of the feature x in the data set are denoted by $max(x)$ and $min(x)$, respectively. The final normalized value, x_{norm}, will fall in the range of 0 to 1 [17].

Data Splitting: This process refers to the process of dividing the dataset into the training and test sets. In the context of a data set representing an IoT attack, it is essential to check that the machine learning model does not inappropriately match the data and has the ability to generalize well to data that has not been seen before. In this case, the datasets were divided up so that 80% would be utilized for training the model and 20% would be utilized to test it.

3.3. Feature Selection using BBWO Algorithm

The hunting behaviour of black widow spiders served as the inspiration for the BWO algorithm, which is a populations-based metaheuristic algorithm. The technique has been used to find solutions to a number of different optimization issues, one of which is feature selection. The BWO method begins by randomly initializing a population of black widows, and each black widow in this population indicates a possible solution (or a subset of features). The fitness levels of the black widows are then taken into consideration. BWO is an excellent method for handling many different optimization issues, one of which is feature selection. The Particles Swarm Optimization, Ant Colony Optimization, and Genetic Algorithm are just a few examples of popular feature selection algorithms that are outperformed by this technique when applied to several datasets. The BWO is employed for the purpose of improving the efficiency of solution and dependability while identifying the most considerable solutions to difficulties involving

feature selection. The Binary-BWO (BBWO) is a modified version of the BWO that is offered in this research for improved feature selection. The goal of this research is to further increase the effectiveness of the BWO. In BBWO, a potential solution is denoted by a "black widow," which is a representation of the solution in the form of a binary string (that is, a subset of characteristics represented by 0s and 1s). The algorithm begins by first seeding a population of black widows with random information, and then evaluates the fitness of each individual based on how well they solve the problem that is presented [18].

In the BWO algorithm, every conceivable issue is recast in relation to the characteristics of the black widow spiders, which acts as a model for the various solutions. Consider the structure as an array while attempting to solve an optimization issue with N_v dimensions. This will allow to find a solution to the optimization issue. A widow was the array that describes the solutions to the issue, and this array could be described as the following:

$$w = [x_1, x_2, \dots, x_{N_v}] \quad (2)$$

The binary format is utilized by the proposed BBWO algorithm in order to present the population of solutions denoted by N_p. One widow is represented by each individual solution. In the binary model, the solutions are illustrated by the vector that only has one dimension. The vector's length shifts according to the amount of features present in the primary data set. With this representation, a solution is shown in the form of a vector that only has one dimension. For instance, if there are *S* features included in the data set, then the length of the solution was also *S*. The cell value in the vectors are denoted as either "1" or "0." If the value is 1, it indicates that the associated feature has been selected, while if the value is 0, it denotes the features are not selected. During the creation of the first solution, random results are generated; more specifically, a value of one or zero was arbitrarily allotted to all the cells in the vectors. Because BBWO acts on the populations of solutions, the population was described by the array, with all the rows indicating a solution of candidate. Consider the number of features was F_n and the size of the population was $|N_p|$, then the size of the array will be $F_n \times |N_p|$.

Initialization: The solution populations that BBWO provides for the problems of feature selection was initially produced in a random way by assigning a value of either "0" or "1" to each cell of the solution. The procedure starts out by setting the size of population and the count of features to their default values. The method will then iterate through each solution in the population, at which point it will arbitrarily assign either a "0" or a "1." This method is performed as many times as necessary until all the possible solutions in the population have been started.

Fitness Functions & Assessment: Since the solutions of population was initialized, a fitness value was assigned to each solution (widow) to represent the solution's quality. All the solutions fitness values are determined by calculating it with the help of the fitness functions (FF) of the wrapper approach, as shown in equation (3). The wrapper approach is favoured as the evaluation strategy over the filter approach and the embedded approach due to the fact that it has a greater level of accuracy and because a large number of researchers are in favour of using it. As a result, the wrapper approach was utilized to ascertain the value of fitness and to facilitate the balance among the amount of chosen features in all the solutions (minimum) and the accuracy of classification (maximum). The classification performance of a classifier is taken into account when doing an evaluation of the solutions by the feature selection wrapper approach (accuracy). In particular, the k-NN classifier was utilized as the learning algorithm phase for the purpose of determining how accurate the solutions that were produced by the BBWO were.

The problem of feature selection can be thought of as an example of a multi-objective optimization problem, in which the solution must simultaneously satisfy two goals that are in direct opposition to one another: reducing the total number of features chosen while simultaneously improving classification accuracy. If there is a lesser count of characteristics in the solutions, then the accuracy of classification will be higher, and the solution will be of higher quality. All the solutions are assessed based on the FF proposed, which are dependent on the classification model to obtain the solution's accuracy of classification and the features selected in the solutions produced by the search algorithm. The FF also takes into account the total count of features in the solutions. The following equation illustrates the FF that is determined by the feature selection method:

$$FF = \alpha \gamma_C(D) + \beta \frac{|C|}{|T|} \quad (3)$$

Here, $\gamma_C(D)$ represents the classification error rate of the proposed classifier, $|C|$ was the selected subset's cardinality, $|T|$ was the original feature's total count in the data set, and, α, β were

the parameters of weight that correspond to the significance of classification efficiency and length of subset, $\alpha \in [0,1]$ and $\beta = (1-\alpha)$.

Transformation Function: The search agent's positions that are created from the conventional BWO were constant values. Because it goes against the binary type of the feature selections on choosing or not choosing, this cannot be directly applied to the issue (zero or one). The sigmoid functions in equations (4) and (5), which are regarded as the type of the transformation functions and utilized by the BBWO technique as the reproduction process part to transform any constant values to the binary related. This can be done in order to produce a binary representation of the original continuous value.

$$z_{S_w} = \frac{1}{1+e^{-z_w}} \quad (4)$$

$$z_{binary} = \begin{cases} 0, & if\ rand < z_{S_w} \\ 1, & if\ rand \geq z_{S_w} \end{cases} \quad (5)$$

Every z_{S_w} was a constant value (features) in the search agents for the *S*-shaped functions, more precisely in the solutions w at dimensions $d(w = 1, \dots, d)$, where $rand$ was the random value chosen from the uniform distributions $\in [0,1]$. The values of z_{binary} could either be zero or one depending on what the values of the $rand$ was in relation to the z_{S_w} value, where '*e*' was the numerical constant also called as Euler's value.

Process of Reproduction: In order to bring out a newer generation, the process of procreation must first begin. The parents were chosen at random to conduct the steps of the procreation process by mating in order to generate the newer generations. In order to finish the reproduction process, an array that will be referred to as Alpha should also be constructed. By combining the α with the below equation, in that p_1 and p_2 play the roles of parents, the offspring o_1 and o_2 will be generated.

$$\begin{cases} o_1 = \alpha \times p_1 + (1-\alpha) \times p_2 \\ o_2 = \alpha \times p_2 + (1-\alpha) \times p_1 \end{cases} \quad (6)$$

This process was carried out for each and every couple, despite the fact that the parents should not be the same each time. In the final step, the offspring together with their maternal parents were integrated to the array and then arranged according to their respective fitness values.

Cannibalism Process: The act of cannibalism are broken down into three distinct subtypes: sexual cannibalism, in which the mate is consumed by its black widow either after or during mating; sibling-cannibalism, in which the less powerful siblings are consumed by their more powerful sibling; and finally, parental cannibalism, in which the mother is consumed by her most powerful offspring. The BBWO model computes and assesses the relative levels of fitness possessed by weak or powerful spiders. In light of this, the most effective solutions (surviving spiders) will be chosen from the process of reproduction and saved in the variable known as $pop2$.

Process of Mutations: The process of mutations begins with the random selection of the count of solutions from population $pop1$. These solutions would each be modified in their own unique way. There will be a random swap of two cells from all the chosen solutions (widows), and the newer mutation solution would be stored in $pop3$.

Generation of Newer Population: The newer population could be formed as the composite of $pop2$ and $pop3$, which would be assessed to return the values of optimal solutions (W^*) carrying the *N* dimensions. The BWO method has various parameters, and with the right combination of those parameters, one can attain amazing results. These aspects include the rate of cannibalism, the rate of procreation (P_r), and the rate of mutation (M_r). The cannibalism rate is calculated using the fitness values equation (3) by the BBWO technique. However, the standard BWO's Pr and Mr parameter rates are applied to use in the BBWO process.

Step-1: Choose the initial values for the BBWO parameters: N_p, maxIteration, F_n, M_r, P_r.

Step-2: Generate the initial population of solutions in a completely random manner using $F_n \times |N_p|$. All the solutions comprise a single widow, which was reflected in the 1D vector representation denoted by the value $1 \times F_n$.

Step-3: Compute the FF for the starting populations utilising equation (3), store the outputs in the $pop1$ file, and designate the best solutions as W^*.

Step-4: Reproduce a newer generation using the cannibalism and procreate process through choosing two parents from $pop1$ for generating *C* children utilizing equation (6), transforming *C* to a binary type using equations (4) and (5), calculate the values of FF for *C*, removing few children and father, and thus saving the remaining solutions as remaining spiders in $pop2$.

Step-5: Perform the process of mutation through picking the certain solutions from $pop1$,

after selecting two spots at random in all the solutions and exchanging them. Finally, save the newer solutions in $pop3$.

Step-6: Updating the populations that are equal to $(pop2 + pop3)$.

Step-7: Evaluate the population utilising equation (1), and update the W^* if there was more optimal solutions.

Step-8: Check if the convergence criteria have been met before continuing on to maxIteration. The algorithm will then come to an end and return the optimal solution W^*; if this is not the case, it will go on to Step 4 [19].

By applying the proposed datasets, the BBWO algorithm selects optimal features individually. From the CIC-IDS-2017 dataset, the BBWO algorithm selects 14 features (feature # 3, 4, 6, 8, 9, 10, 12, 13, 16, 17, 39, 40, 67, 68) and 22 features (feature # 7, 9, 11, 13, 19, 20, 21, 22, 25, 26, 27, 30, 39, 50, 51, 55, 56, 68, 69, 70, 76, 78) from CIC-IDS-2018 was selected and applied for classification.

3.4. Classification using Logistic Regression

For the purpose of classifying intrusions in IoT network, the machine-learning strategy of logistic regression is an appropriate option. The reason for this is that LR is a technique for binary classification that can represent the probability of an event taking place. This makes it ideal for recognizing attacks in IoT networks that are binary in nature, meaning that they can either be present or not present. This research's key objective was to discover a subset of characteristics that could enhance the classification performance of a logistic regression classifier by the use of BBWO to the search process. The classification accuracy of the LR classifier based on the selected feature subset would be the fitness function for BBWO if it were to be defined at all. During each iteration of BBWO, the chosen subset of features would be put to use in the process of training a LR classifier using the training set, and the accuracy of the classifier's classification would be measured using the testing set. The binary string's "fitness value" would be equal to the classification accuracy that was achieved on average over the entire cross-validation process.

The hybridization of BBWO with logistic regression can also involve putting the logistic regression classifier into the attraction force of BBWO. This may be accomplished by applying the gradient of the logistic regression loss function to the weights of the logistic regression classifier and utilizing it to update the placements of the black widows throughout the population. In general, the combination of BBWO and logistic regression has the potential to increase performance when it comes to the classification of attacks.

The logistic functions, often called as the Sigmoid functionality was utilized in the supervised learning technique known as LR. LR is almost same as linear regression, with the key differences being that rather than making predictions based on continuous data, LR is utilized to classify data as either true or false. Comparatively, logistic regression only accepts values between 0 and 1, whereas linear regression can take on any value. Logistic regression is one of the models used in intrusion detection, but it is not as commonly used as other models. Nonetheless, a logistic regression-based intrusion detection model has been investigated in some previous research. When compared to the other models through the use of multi-class classification, this model performed significantly better [20-23].

The logistic regression is as a form of linear regression that is used for classification problems. Because linear regression allows for the possibility that the hypothesis $h_o(x)$ could either be higher than 1 or less than 0, logistic regression was chosen for this research work. In the case of logistic regression, the hypothesis falls in the range of 0 and 1, i.e., $0 \leq h_o(x) \leq 1$. A single hypothesis, denoted by h_o in this case, can be used to map input to output, and it could be assessed and applied in order to create classifications. The Sigmoid function, which is represented in the following way, is applied so as to obtain a value that falls between 0 and 1.

$$S(x) = \frac{1}{1+e^{-x}} \quad (7)$$

The above function will return a number among 1 and 0 that could be translated to the certain category of data by applying the decision boundary for determining the probability that the data belong to the particular category, which could be represented as:

$$p \geq 0.5 \; class = 1$$
$$p < 0.5 \; class = 0 \quad (8)$$

After the threshold has been established, it is possible to make predictions with the help of the Sigmoid function to determine the probability that the data are part of class 1 as follows:

$$S(class = 1) = \frac{1}{1+e^{-x}}$$
(9)

The function above gives back a number that represents the probability that the data should be classified as belonging to Class 1 (attack) or Class 0 (normal). According to the criteria that had been established in the beginning, the data will be classified as Class 1 if the number is 0.5 or above, while everything that is less than 0.5 will be classified as Class 0.

4. EXPERIMENTAL ANALYSIS

This section presents a detailed analysis of the experimental results of the proposed BBWO-LR model. The MATLAB 2019b simulation tool equipped in the laptop with a Core i7 processor operating at 3.20 GHz and 12GB of RAM with Windows 11 OS is used to experiment and evaluate the BBWO-LR model. The performance metrics like detection rate, accuracy, FPR, precision, and F1-score are calculated based on the classified attacks. The experimental evaluations of the BBWO-LR model were compared with various models for validation.

4.1. Performance Evaluation

To determine the performance of the research model in detecting intrusions in IoT environments, an assessment of its performance was conducted using several metrics, including accuracy, detection rate, FPR, f1-score, and precision. These performance indicators were evaluated utilizing the confusion matrix, which takes into account the values of true positive (TP), false positive (FP), true negative (TN) and false negative (FN), as represented in the equations (10) to (14).

$$Detection\ rate = \frac{TP}{TP+FN}$$
(10)

$$Accuracy = \frac{TP+TN}{FN+TP+FP+TN}$$
(11)

$$FPR = \frac{FP}{FP+TN}$$
(12)

$$Precision = \frac{TP}{TP+F}$$
(13)

$$F1score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$
(14)

Table 3: Results of the BBWO-LR model on Training Set

Metric	CICIDS-2017	CICIDS-2018
Accuracy	99.11	99.23
Detection Rate	98.64	99.40
Precision	99.64	99.10
F1-score	99.14	99.25
FPR	0.038	0.096

Table III represents the results evaluated using the training sets of datasets according to the performance parameters. As shown in the table, the accuracy attained by the BBWO-LR model for the CICIDS-2017 training set was 99.11%, and 99.23% for CICIDS-2018 set. The detection rate was 98.64% for dataset-1 and 99.40% for dataset-2. The precision rate for dataset-1 was 99.64% and 99.10% for dataset-2. The f1-score for dataset-1 was 99.14% and 99.25% for dataset-2. According to these results from the training sets of datasets, the accuracy, detection rate and f1-scores of the research model evaluated with the CICIDS-2018 were higher than those obtained using the CICIDS-2017 dataset. Figure 4 represents the graphical plot for these obtained results as a comparison.

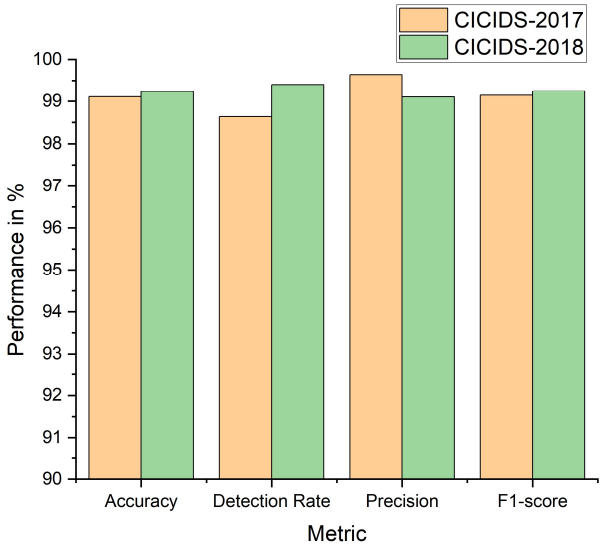

Figure 4: Performance Analysis of BBWO-LR Model using Training Sets

Table 4: Results of the BBWO-LR model on Testing Set

Metric	CICIDS-2017	CICIDS-2018
Accuracy	98.83	98.92
Detection Rate	98.32	98.17
Precision	99.58	99.76
F1-score	98.95	98.97
FPR	0.052	0.074

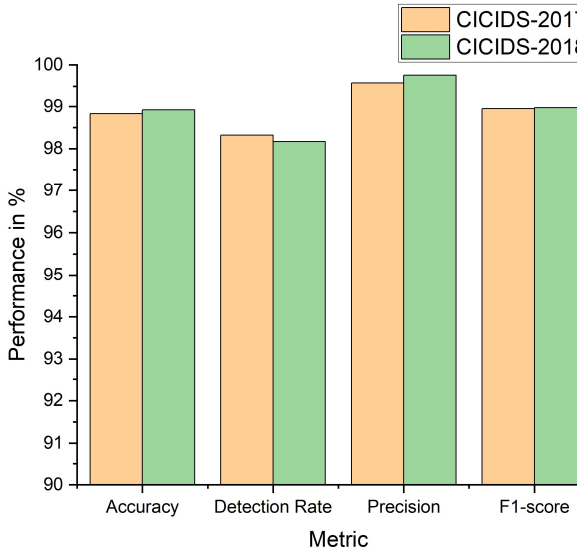

Figure 5: Performance Analysis of BBWO-LR Model using Testing Sets

The results evaluated using the testing sets of datasets according to the performance metrics are represented in table IV. The accuracy attained by the BBWO-LR model for the CICIDS-2017 test set was 98.83%, and 98.92% for the CICIDS-2018 test set. The detection rate was 98.32% for dataset-1 and 98.17% for dataset-2. The precision rate for dataset-1 was 99.58% and 99.76% for dataset-2. The f1-score for dataset-1 was 98.95% and 98.97% for dataset-2. As shown in the table, the accuracy, precision, and f1-scores of the research model evaluated with the test set of CICIDS-2018 were higher than the results obtained using the CICIDS-2017 dataset. Figure 5 represents the graphical plot for results obtained using the test sets as a comparison.

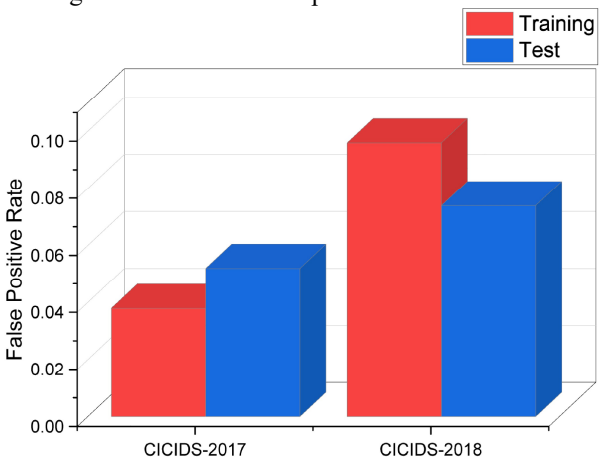

Figure 6: FPR Comparison with Training and Testing Sets

The FPR or false alarm rate for the training set of CICIDS-2017 was 0.038 and 0.096 for the training set of CICIDS-2018. The FPR rate for the test set of CICIDS-2017 was 0.052, and 0.074 for the test set of CICIDS-2018. As the FPR score is minimum, the research model is efficient for the proposed application. Based on the comparison between the training set and test set results, the FPR score was better evaluated using CICIDS-2017 training set and better evaluated using the CICIDS-2018 test set. Figure 6 depicts the comparison of FPR scores of the research model based on training and test sets of datasets.

Table 5: Performances Comparison of the BBWO-LR Model

Models	Accuracy	Detection Rate	Precision	F1-score
LR [10]	92.96	90.96	90.80	90.87
RF+PCA [11]	97.73	96.06	99.33	97.57
BMRF+XGBoost [6]	95.80	89.20	94.80	91.90
CNN [9]	98.61	95.00	97.05	93.09
Poly-BR [12]	94.50	93.20	94.39	NA
Voting Classifier [13]	96.25	89.00	89.00	89.00
HCRNN [15]	97.75	97.12	96.33	97.60
BBWO-LR (CICIDS-2017)	98.83	98.32	99.58	98.95
BBWO-LR (CICIDS-2018)	98.92	98.17	99.76	98.97

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

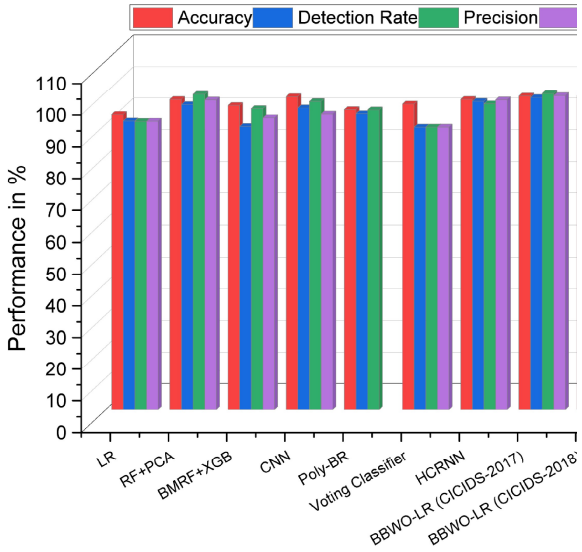

Figure 7: FPR Comparison with Training and Testing Sets

As shown in table V, a comparison between the proposed BBWO-LR model with the existing IDS models was presented. The compared models were discussed in the literature review above. Based on the comparison, the research model BBWO-LR obtained 98.92% higher accuracy, which is better than the compared models with a difference of 0.31% to 5.96%. The CNN model in the comparison has a close value of accuracy with 98.61%. The least accuracy score was obtained by LR, with 92.96%. The highest detection rate of the BBWO-LR model was 98.32%, which was 1.2% to 9.32% better than the other approaches. The HCRNN model has obtained a 97.12% detection rate, which is close to the proposed model's score, and the poor detection rate was obtained by the voting classifier with 89%. The precision score of the research model was 99.76%, which is 0.43% to 10.76% improved than the other models. The RF+PCA model has obtained a close precision value of 99.33%, and the least performed model was the voting classifier with 89%. The f1-score of the BBWO-LR model was 98.97%, which is 1.3% to 9.9% higher than the other models. The models like RF+PCA and HCRNN have obtained close performance with 97.57% and 97.60%. The voting classifier has obtained the least performance with an 89% f1-score. Based on this comparison, the research model has outperformed the other models regarding all the parameters with improved performances.

5. CONCLUSION

A new hybrid intrusion detection model was proposed for classifying the attacks in the IoT network based on CICIDS-2017 and CICIDS-2018. The proposed hybrid model integrated feature selection and classification process implemented using machine learning models. The BBWO algorithm was used for selecting features with the optimal feature sets from the given data sets. For classification, the logistic regression model was used to perform the binary classification. The proposed model initially performed data preprocessing using min-max scaling normalization for dataset standardization. After preprocessing, the datasets were split into training and test sets for evaluation. Using these datasets, the features were chosen optimally with the BBWO approach. The classification was performed using the LR algorithm based on the selected optimal feature sets. The performance of this research model was evaluated based on accuracy, detection rates, FPR, precision, and f1-scores. The results were evaluated individually for both datasets and correlated. The BBWO-LR model obtained 98.83% accuracy, 98.32% detection rate, 99.58% precision, and 98.95% f1-score for the CICIDS-2017 data set. Using the CICIDS-2018 data set, the BBWO-LR model obtained 98.92% accuracy, 98.17% detection rate, 99.76% precision, and 98.97% f1-score. These results were compared with the existing models for validation, representing that the proposed BBWO-LR model clearly outperformed the other models with better performances.

Pros:

Hybrid approach: The model combines feature selection and classification processes, leveraging the strengths of both techniques. This integration may lead to improved accuracy and performance in detecting and classifying attacks in IoT networks.

Feature selection with BBWO algorithm: The use of the BBWO algorithm for feature selection helps identify the most relevant features from the datasets. This can enhance the efficiency of the model by reducing the dimensionality of the data and focusing on the most informative features.

Logistic regression classification: Logistic regression is a widely used algorithm for binary classification tasks. It is computationally efficient and can provide interpretable results, making it a suitable choice for this intrusion detection model.

High-performance metrics: According to the results reported in the statement, the proposed BBWO-LR model achieved high accuracy, detection rates, precision, and f1-scores for both the CICIDS-2017 and CICIDS-2018 datasets. These metrics indicate that the model performs well in accurately classifying attacks.

Cons:

Dataset limitations: The statement mentions the use of the CICIDS-2017 and CICIDS-2018 datasets. It is important to consider the representativeness and diversity of these datasets. If the datasets do not sufficiently cover all possible attack scenarios or if they are biased in any way, the model's performance may not generalize well to real-world scenarios.

Generalization to other datasets: The statement does not provide information on how well the proposed model performs on other datasets or in different IoT network environments. The model's effectiveness and robustness in handling various datasets and network conditions are important factors to consider.

In future, this research model can experiment with focussing on Cons discussed above as new datasets can be implement to the proposed model containing only IoT attacks. Further, this research can be improved by implementing a deep learning classifier model for multi-class classification with better performance.

REFERENCES:

[1] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security," *IEEE Communications Surveys & Tutorials*, Vol. 22, No. 3, 2020, pp. 1646-1685.

[2] E. Leloglu, "A Review of Security Concerns in Internet of Things," *J. Comput. Commun.*, Vol. 5, 2017, pp. 121-136.

[3] K. Chen et al., "Internet-of-Things Security and Vulnerabilities: Taxonomy, Challenges, and Practice," *J. Hardware Syst. Secur.*, Vol. 2, 2018, pp. 97–110.

[4] A. Adnan, A. Muhammed, A. A. Abd Ghani, A. Abdullah, and F. Hakim, "An Intrusion Detection System for the Internet of Things Based on Machine Learning: Review and Challenges," *Symmetry*, Vol. 13, No. 1011, 2021, pp. 1-13.

[5] S. Gamage and J. Samarabandu, "Deep learning methods in network intrusion detection: A survey and an objective comparison," *J. Netw. Comput. Appl.*, Vol. 169, No. 102767, 2020, pp. 1-21.

[6] I. H. Hassan, M. Abdullahi, M. M. Aliyu, S. A. Yusuf, and A. Abdulrahim, "An improved binary manta ray foraging optimization algorithm-based feature selection and random forest classifier for network intrusion detection," *Intell. Syst. Appl.*, Vol. 16, 2022, pp. 1-14.

[7] M. A. Siddiqi and W. Pak, "Optimizing Filter-Based Feature Selection Method Flow for Intrusion Detection System," *Electronics*, Vol. 9, No. 2114, 2020, pp. 1-18.

[8] 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. 77, 2023, pp. 1-17.

[9] J. Jose and D. V. Jose, "Deep learning algorithms for intrusion detection systems in internet of things using CIC-IDS 2017 dataset," *Int. J. Electr. Comput. Eng.*, Vol. 13, No. 1, 2023, pp. 1134-1141.

[10] M. A. Kumaar, D. Samiayya, P. M. D. R. Vincent, K. Srinivasan, C. Y. Chang, and H. Ganesh, "A Hybrid Framework for Intrusion Detection in Healthcare Systems Using Deep Learning," *Front. Public Health*, Vol. 9, 2022, pp. 1-18.

[11] S. Seth, G. Singh, and K. K. Chahal, "A novel time efficient learning-based approach for smart intrusion detection system," *J. Big Data*, Vol. 8, No. 111, 2021, pp. 1-28.

[12] P. Dini et al., "Design and Testing Novel One-Class Classifier Based on Polynomial Interpolation with Application to Networking Security," *IEEE Access*, Vol. 10, 2022, pp. 67910-67924.

[13] S. Patil et al., "Explainable Artificial Intelligence for Intrusion Detection System," *Electronics*, Vol. 11, No. 3079, 2022, pp. 1-23.

[14] J. B. Awotunde and S. Misra, "Feature Extraction and Artificial Intelligence-Based Intrusion Detection Model for a Secure Internet of Things Networks," *Illumination of Artificial Intelligence in Cybersecurity and Forensics*, Cham, Springer, 2022, pp. 21-44.

[15] M. A. Khan, "HCRNNIDS: Hybrid Convolutional Recurrent Neural Network-Based Network Intrusion Detection System," *Processes*, Vol. 9, No. 834, 2021, pp. 1-14.

[16] L. Liu, G. Engelen, T. Lynar, D. Essam and W. Joosen, "Error Prevalence in NIDS datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018," *2022 IEEE Conference on Communications and Network Security (CNS)*, Austin, TX, USA, 2022, pp. 254-262.

[17] D. Y. Mahmood, "Classification Trees with Logistic Regression Functions for Network Based Intrusion Detection System," *IOSR J. Comput. Eng*., Vol. 19, No. 3, 2017, pp. 48-52.

[18] A. J. Daniel and M. J. Meena, "A hybrid sentiment analysis approach using black widow optimization-based feature selection," *J. Eng. Res*., 2021, pp. 1-15.

[19] M. Kaya Keleş and Ü. Kiliç, "Binary Black Widow Optimization Approach for Feature Selection," *IEEE Access*, Vol. 10, 2022, pp. 95936-95948.

[20] A. Churcher et al., "An Experimental Analysis of Attack Classification Using Machine Learning in IoT Networks," *Sensors*, Vol. 21, No. 446, 2021, pp. 1-32.

[21] D. Sathiya and S. Sheeja, "Data Delivery and Node Positioned Learning Automaton in Mobile Ad Hoc Networks," *J. Comput. Sci. Intell. Technol.,* Vol. 3, No. 2, 2022, pp. 1-14. https://doi.org/10.53409/MNAA/JCSIT/e202203020114

[22] S. S. Lutfi and M. L. Ahmed, "A Novel Intrusion Detection System in WSN using Hybrid Neuro-Fuzzy Filter with Ant Colony Algorithm," *J. Comput. Sci. Intell. Technol.,* Vol. 1, No. 1, 2020, pp. 1-8. https://doi.org/10.53409/mnaa.jcsit1101

[23] R. Khilar et al., "Artificial Intelligence-Based Security Protocols to Resist Attacks in Internet of Things," *Wireless Commun. Mobile Comput.*, Vol. 2022, No. 1440538, 2022, pp. 1-10. https://doi.org/10.1155/2022/1440538