ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20231011001016

Original scientific paper

Improving Spam Intrusion Detection with the Machine Learning-Enhanced Chaotic Horse Ride Optimization Algorithm

Amutha THANGARAJ*, Geetha SADAYAN

Abstract: Automated spam detection, utilizing feature selection (FS) and machine learning (ML), categorizes and identifies unsolicited messages, like spam emails. The goal is to accurately differentiate and filter out spam, enhancing overall email security. This research presents the Chaotic Horse Ride Optimization Algorithm with Machine Learning-Driven Spam Detection and Classification (CHROA-MLSDC). CHROA-MLSDC efficiently classifies spam and non-spam through preprocessing and CHROA-based feature selection. It incorporates the Variation Auto Encoder (VAE) model and the Bat Algorithm (BA) for potential performance improvements. Simulations on Ling spam, Enron, Spam Assassin, and CSDM C2010 datasets demonstrate significant enhancements in precision, recall, accuracy, and execution speed compared to existing systems. CHROA-MLSDC achieves notable accuracy: 99.35% on Ling spam, 98.80% on Enron spam, 99.92% on Spam Assassin spam, and 98.02% on CSDM C2010 spam. Recall rates range from 97.82% to 98.05%. CHROA-MLSDC consistently outperforms similar methods, exhibiting accuracy from 96.91% to 97.55%. Execution time analysis reveals CHROA-MLSDC's consistently faster performance across all datasets. In summary, CHROA-MLSDC excels in spam detection, surpassing other methods across various evaluation metrics.

Keywords: feature selection; horse ride optimization; machine learning; metaheuristics; spam detection

1 INTRODUCTION

In today's age of information technology, the sharing of data has become incredibly efficient and rapid. Various platforms for global data sharing are available, with email being a cost-effective and straightforward method for worldwide data exchange [1, 2]. However, the simplicity of email also makes it susceptible to various attacks, with spam being one of the most common and dangerous threats. Unwanted and irrelevant emails are a significant annoyance to recipients, consuming their time and resources [3]. Moreover, emails may contain malicious elements such as URLs or attachments, which can lead to security breaches in the recipient's system. In essence, spam comprises unsolicited and irrelevant messages sent by hackers to multiple recipients via emails or other means. As a result, there is a substantial demand for securing the email system [4]. Spam emails often carry Trojans, viruses, and more, as hackers use this method to lure users into various online scams. They send spam messages containing attachments with various file extensions and URLs that direct users to spam websites, often resulting in identity theft or financial fraud [5]. Some email providers allow users to create rules based on keywords for automatically filtering emails. However, this approach is not very effective, as it is intricate, and users are generally reluctant to customize their email settings, leaving their email accounts vulnerable to attacks [6]. A more effective solution is to develop a model that enables email recipients to automatically detect spam [7]. This model is commonly referred to as a spam detection technique and is primarily categorized into two types: statistical and non-statistical approaches [8]. Generally, non-statistical approaches are more robust than statistical ones. Statistical detection models search for specific keyword patterns in email messages. Several spam detection techniques using machine learning (ML) models have been explored. It is crucial to minimize the computational resources required for spam detection, as it must keep pace with the massive volume of emails from bulk mailers [9]. Feature selection (FS) and parameter optimization techniques are employed to reduce resource consumption while maintaining a high detection rate. FS involves selecting only relevant features or feature sets from a pool of all available features [10]. FS helps eliminate unnecessary features to avoid computational overhead. Parameter optimization aims to fine-tune the parameters of the spam detection model to discover the optimal configuration for the detection system [11]. While previous techniques have considered parameter optimization for spam detection models, they often lack detailed explanations of the process. In this study, a novel approach called the Chaotic Horse Ride Optimization Algorithm with Machine Learning-Driven Spam Detection and Classification (CHROA-MLSDC) is introduced. The primary goal of the CHROA-MLSDC method is to effectively categorize spam and non-spam content. To achieve this, the CHROA-MLSDC technique initially preprocesses the input data. For the feature selection process, the CHROA-MLSDC approach employs the CHROA technique to select relevant subsets of features.

2 RELATED WORKS

Srinivasarao and Sharaff [12] present a hybrid classifier that depends on sentiment analysis (SA) and SMS spam classification. Word2vec data augmentation was utilized after the data pre-processing to extract the features. Next, the features were given to equilibrium optimization (EO) and six different feature selection approaches. Then, optimal elements are given into a hybrid SVM and KNN method to categorize SMS messages. Bhardwaj and Sharma [13] attempt to find email spam by building an ensemble mechanism with the use of bagging and boosting ML approaches. The data utilized in the study is Ling-Spam Corpus. The system identifies spam emails by bagging the J48 decision tree and ML-related multinomial Naïve Bayes (MNB) classifiers and boosting the method of transforming weak classifiers into strong ones by applying the Adaboost method. Samarthrao and Rohokale [14] present the email spam recognition approach for text and image datasets. In this study, the main contribution is the growth of an adaptive capsule network and multi-objective feature selection for email spam recognition. While utilizing the textual data, two feature extraction methodologies like Term Frequency-Inverse Document Frequency (TF-IDF), and Term Variance (TV) utilized, where the colour correlogram and Walsh-Hadamard Transform (WHT) were utilized as the feature extraction algorithms to manage the image dataset. Rayan et al. [15] introduce a new ML-related hybrid bagging approach for e-mail spam detection by integrating dual ML approaches: J48 (decision tree) and RF. The presented structure classifies the e-mail into spam and ham. Likewise, stop word removal, tokenization, and stemming were effectuated in the pre-processing stage. Additionally, for choosing the essential attributes from the pre-processed data correlation feature selection (CFS) was utilized in this study. Ramanujam et al. [16] introduce multilingual SMS spam data and devise a hybrid DL approach that integrates the LSTM and CNN methods for categorizing the message datasets. Poonkodi et al. [17] developed the Enriched Firefly Optimization Algorithm (EFOA) approach efficiently choosing appropriate features from upper dimensional spaces utilizing the fitness function. The spam classification was accomplished through ANN, once the optimal feature space was recognized using EFOA. The E-mail spam dataset was pre-processed initially, and then the abstracted textual attributes were Semantic-related reduction and Features weights upgraded with the use of optimized semantic WordNet. Rajakani et al. [18] develop a method to perform intrution recognition. Different to classical ML methods namely SVM, and NB, a DNN is immune to different fluctuating surroundings. This study devises the application of CountVectorizer for performing feature extraction on text-related data. Bosaeed et al. [19] present a tool to identify spam from outgoing SMS messages, though the work can be implemented for outgoing and incoming SMS messages. To be specific, the author developed a system that has several ML-related classifiers built through three classification approaches NB, SVM, and Naïve Bayes Multinomial (NBM) and five pre-processing and feature extraction approaches. Isa Avci and M. Koca [20] BMS DDoS assaults are the emphasis of the article on smart building cybersecurity issues caused by IoT expansion. A technique integrating Slime Mould Optimization, Artificial Neural Network, and Support Vector Machine was proposed after conventional methods failed. The program estimates DDoS risk (97.44%) and assaults (99.19%) with great accuracy. The CIC IoT Dataset 2022 validates the model's performance in various IoT devices and protocols. The research stresses the need of BMS cyber security for smart building security and safety. İsa Avc and Mehmet Y. [21] combinatorial optimization problem of weapon target assignment (WTA). It uses the Salp Swarm Algorithm (SSA) to approximate solutions and the Salp Hybrid Algorithm (SHA) to forecast target probabilities. Isa Avc and M. Koca [22] proposes an ML-based network intrusion detection solution to improve cybersecurity for millions of online users. Traditional systems are ineffective, thus AI is used. The study uses feature selection approaches to assess Random Forest, K-Nearest Neighbors, Support Vector Machine, and Decision Tree ML algorithms. Random Forest trumps others with 99.72% accuracy. The research emphasizes detecting dangerous and benign cyber-attacks, improving intrusion detection accuracy.

3 THE PROPOSED MODEL

In this study, we have presented an automated spam detection and classification technique, called the CHROA-MLSDC model. The intention of the CHROA-MLSDC method lies in the effectual classification of spam and non-spam content. It comprises several sub-processes such as data pre-processing, feature selection using CHROA, VAE-based classification, and BA-based parameter tuning. Fig. 1 depicts the workflow of the CHROA-MLSDC algorithm.

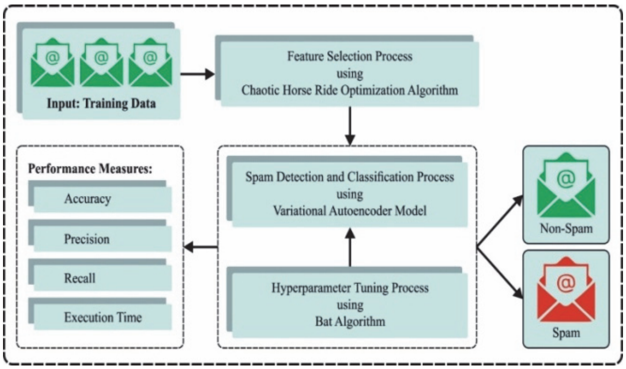
Input: Training Data
Feature Selection Process using Chaotic Horse Ride Optimization Algorithm
Performance Measures:
Accuracy
Precision
Recall
Execution Time
Spam Detection and Classification Process using Variational Autoencoder Model
Hyperparameter Tuning Process using Bat Algorithm
Non-Spam
Spam

Figure 1 Workflow of CHROA-MLSDC system

3.1 Dataset

1. Ling Spam Dataset is most likely associated with the task of spam detection and may consist of a collection of emails or messages that have been classified as either spam or non-spam. It is possible that this dataset involves the identification of spam. A wide variety of attributes that are obtained from these conversations may be included in the dataset. These qualities may include the content, information about the sender, and other relevant factors.

2. The Enron dataset is a dataset that is employed extensively in the domains of study pertaining to machine learning and email correspondence analysis. The Enron Corporation, an energy company that had a crisis in the early 2000s, is the source of both of these terms. The dataset contains a wide variety of emails that were sent by employees of Enron. As a result, it is extremely useful for evaluating patterns of email communication, identifying spam, and carrying out other operations linked with these activities.

3. The Spam Assassin Dataset is an open-source system that is in widespread usage for filtering spam emails. The dataset that is thought to be associated with Spam Assassin is most likely made up of a collection of emails that have been classified as either spam or non-spam. For the aim of training and testing spam detection algorithms, the data may comprise aspects such as the content of emails, information about the sender, and other relevant features.

4. The abstract of the CSDM C2010 Dataset does not contain any information that is easily accessed with ease. It is possible that CSDM C2010 refers to a dataset that is directly related with the detection or classification of spam,

which may have been compiled for the express aim of research. In order to get knowledge on the properties and characteristics of this dataset, it is necessary to consult the literature or material that is associated with CSDM C10000.

3.2 BC Technology Data Pre-processing

The message database in consideration has been unprocessed. Consequently, it must be pre-processed still being regarded any more [23]. The pre-processing step has 3 stages. First, the tokenization of text data can be measured. Tokens can be words which can be divided into the rest of the phrases.

3.3 Feature Selection using CHROA

To elect the optimal set of features, the CHROA was used. The CHROA algorithm is based primarily on chaos theory and HROA concepts [24]. Chaos is in an unbalanced state by focusing on earlier conditions. It improves the quality of solutions and prevents optimum local difficulties in different optimization techniques. CHROA is represented as "dominant horse" with dysfunctional social abilities. Once the grain was rare, high status horse might ban low status horse from blocking animal of low social rank from eating at all. Access to resource can be ranked in the horse herd directed by the dominant stallion and mare, based on hierarchical position of horses within the herd. At first, the hierarchy of horse in a herd can be described by the fitness level. Assume a herd of *k* horses, where *P* indicates the function.

$$Herd = \{H_1, ..., H_k\} \tag{1}$$

$$P = Herd \to 1, ..., \tag{2}$$

If fitness (H_x) < fitness (H_y), where $x \neq y$ and, $y \in \{1 \dots K\}$, then

$$P(H_x) > P(H_y) \tag{3}$$

If fitness (H_x) = fitness (H_y), where $x \neq y$ and, $y \in \{1 \dots K\}$, then

$$\left[P(H_x) - P(H_y)\right](x - y) > 0 \tag{4}$$

The rank of horse H_x can be determined by:

$$H_x - Rank\ of\ each\ horse = \frac{P(H_x)}{K} \tag{5}$$

All the herds have centres that are corresponding to weighted average of position of horse from the herds; therefore, the weights represent a rank of the horse:

$$Herd_{Center} = \frac{\sum_{x=1}^{k} Z_x H_{x.rank}}{\sum_{x=1}^{k} H_{x.rank}} \tag{6}$$

All the herds have centers that are corresponding to the weighted average of position of the horses; hence, the weight was representing the horse's location in the herd:

$$Dim(Stallion, herd) = \sqrt{\sum_{y=1}^{Dim} \left(Stallion_y - Herd_{Center}\right)^2} \tag{7}$$

In Eq. (7), *Dim* represents the amount of dimension of searching space. Once the horses belong to the herd group of horses then it upgrades the velocity using the following expression:

$$Vel_{x,y}^{T+1} = Vel_{x,y}^{T} + H_{x.rank} \cdot \left(Herd_{center.y}^{T} - Z_{x,y}^{t}\right)$$

(8)

$$Vel_{x,y}^{T+1} = Vel_{x,y}^{T} + Rand \cdot \left(Herd_{center.y}^{T} - Z_{x,y}^{t}\right) \tag{9}$$

where *T* shows the existing iteration, *Rand* signifies a randomly generated value within [0, 1] and *T* + 1 denotes the new iteration. The memory of horse (*Mem)* refers to a matrix that has several rows corresponding to values of Horse Memory Pool (*HMP*) and *D* columns.

$$Mem_x^{T+1} = \begin{bmatrix} Mem_{1,x,1}^{T+1} & \cdots Mem_{1,x,D}^{T+1} \\ \vdots & \ddots \quad \vdots \\ Mem_{HMP,x,1}^{T+1} & \cdots Mem_{HMP,x,D}^{T+1} \end{bmatrix} \tag{10}$$

The equation to update the cell of memory matrix is:

$$Mem_{K,x,y}^{T+1} = Z_{x,y}^{T+1} \cdot N(0, SD) \tag{11}$$

In Eq. (11), *N* signifies a standard distribution, 0 acts as a mean and *SD* serves as standard deviation. The chaos concept was integrated to improve global optimization ability of the HROA technique. It can be utilized in different optimization algorithms to improve solution quality and prevent local optimum issues. The chaos concept is introduced to maintain effective trade-offs between exploitation and exploration since the metaheuristic method depends on exploration and exploitation stages, thus obtaining the best solution effectively. In HROA, variable H_x has considerable impact on convergence speed of Artificial Fish Optimization (AFO) technique. The efficiency of HROA approach relies on its parameters. It is noted by a larger momentum to start using the possible search spaces, and it could not be potentially exposed. The chaos was used to attain better search features for the exploration and exploitation region, thereby enhancing the outcome to find optimum global results. The chaotic map was used to determine the location x_i^k, where the variable θ is replaced with the value obtained by the chaotic map as follows

$$x_i^{k+1} = x_i^k + C_{nap} \cdot \left(x_{BH} - x_i^k\right), i = 1, 2, ..., N_\nabla \tag{12}$$

In Eq. (12), x_{BH} indicates the position of *BH* from space, C_{ap} shows the chaotic map, and N_∇ represents the overall amount of individuals. Briefly, ten chaotic maps are used for manipulating the value of a random variable from the HROA algorithm. The fitness function (FF) derived from Eq. (10) signifies the FF to measure solutions.

$$Fitness = \alpha \gamma_R \left(D\right) + \beta \frac{|R|}{|C|} \tag{13}$$

where $\gamma_R(D)$ denotes the classifier rate of error of the provided classifier. $|R|$ stands for the cardinality of chosen subset and $|C|$ means the entire count of features from the database, α and β are 2 parameters equal to the significance of classification quality and subset length. $\in$ [1, 0] and $\beta = 1 - \alpha$.

3.4 VAE-based Classification

For the automated detection of spam data, the VAE model is exploited in this study. AE and VAE are types of neural network that uses an encoder-decoder approach [25]. The encoder turns the higher dimension into lower dimension data, while the decoder converts vice versa. C_{AE} signifies layer values that are compressed and require lower dimensional data. Every unit's weights and biases are adapted, to enhance the network parameter, and the network learns $x = x_{out}$ out identity function. As the loss function, the AE defines the difference between x and x_{out}. The more commonly applied loss function in AE is Mean Squared Error (MSE). MSE signifies the value of mean position data. As the loss function, the AE defines the difference between x and x_{out}. The function of auto-loss encoders was explained as follows.

$$f_{loss} = \left(W^T \left(W\left(x\right) + b\right) + b', x\right) \tag{14}$$

where b shows the encoder or decoder bias. $W(\cdot)$ denotes the encoder or decoder weight. The encoder output was evaluated depending on the data given

$$C_{AE} = W_{AE} \cdot I + b \tag{15}$$

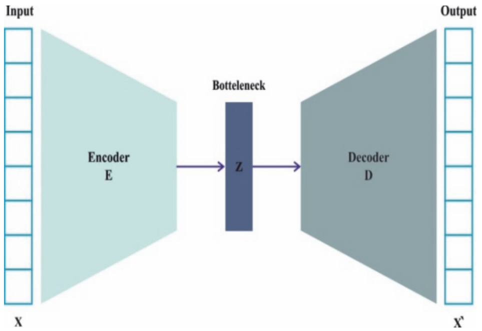

Figure 2 Workflow of CHROA-MLSDC system

It is critical to the AE for locating lower dimensional data to initialize the weight AE that is performed through RBM or random distribution. Resultant values are often off as soon as the weight is initialized at random. RBM generates weight and bias depending on input data of hidden data structure, which enable backpropagation to avoid bad local minima for a certain range. The RBM-initialized automatic encoder accomplishes the outcome in a superior way. Fig. 2 represents the structure of VAE.

3.5 Parameter Tuning using BA

At the final stage, the parameter tuning of the VAE method takes place using the BA. Based on the fundamental characteristics of bats, namely wavelength, velocity, pulse emission rate and position, Yang [26] analyzed these behaviours of bats and the food searching technique. Echo location is used for the prediction and discovery of prey. The main characteristic of bats is avoiding the surrounding using ultrasonic pulse via different pulses.

$$f_{j^\tau} = f_{\min} + \psi\left(f_{\max} - f_{\min}\right) \tag{16}$$

where Ψ denoted the arbitrary number ranges from 0 to 1. Furthermore, the velocity V_{j^τ} and position X_{j^τ} are expressed as follows:

$$V_{j^\tau} = V_j^{\tau-1} + f_j \left(X_{j^\tau} - X^*\right) \tag{17}$$

$$X_{j^\tau} = X_j^{\tau-1} + V_j^\tau \tag{18}$$

The updated velocity value depends on the global best, present frequency, prior velocity and position. Then, the existing location was also updated. Next, the algorithm chooses a better parameter set from the population. Similar to bats, later, it initiates a random walking to evaluate the personal best with the given equation:

$$X_{new} = X_{old} + \theta L^\tau,\ \theta \in \left[-1, 1\right] \tag{19}$$

Bat randomly moves in the vicinity of the previously defined global optimum place. This computation depends on the average loudness L^τ and the global optimal X^*, multiplied by the arbitrary factor ε. The pulse rate and loudness owned by all the bats in the search space are evaluated by the random number [24]. During the cooling temperature, if the bat methods its target, its loudness reduces, and its pulse rate φ_j^τ raises till it attains φ_j^0. At $\tau + 1$ time, the equations to determine the loudness L_j and updated pulse rate φ_j are shown as follows, where ε and η are constant.

$$L_j^{\tau+1} = \varepsilon L_j^\tau,\ \varepsilon \in \left[0, 1\right] \tag{20}$$

$$\varphi_j^{\tau+1} = \varphi_j^0 \left(1 - e^{-\eta\tau}\right),\ \eta \tag{21}$$

Population fitness can be evaluated by the cost function, and if a better cost is attained, both values are updated. The fittest set of parameters is evaluated by the initial metaheuristic approach and is transferred to the contrast modification function that produces enhanced contrast. The BA method grows a FF to bring about a greater classifier solution. It solved a positive integer that demonstrates the good solution of candidate performances. Here, the minimized classifier error rate was assumed that FF, as depicted in Eq. (22).

$$fitness\left(x_i\right) = Classifier\ Error\ Rate\left(x_i\right) = \\ = \frac{no.\ of\ misclassified\ ins\tan ces}{Total\ no.\ of\ ins\tan ces} \cdot 100 \qquad (22)$$

4 RESULTS AND DISCUSSION

The performance analysis of the CHROA-MLSDC method takes place using four datasets, such as Ling spam, Enron, Spam Assassin, and CSDM C2010. In Tab. 1 and Fig. 3, a brief comparative $prec_n$ assessment of the CHROA-MLSDC technique takes place. The experimental results indicate that the CHROA-MLSDC technique obtains improved $prec_n$ values. Based on ling spam, the CHROA-MLSDC technique offers enhanced $prec_n$ of 99.35% while the KNN, NB, IRBNN-LK, IRBNN-PK, and IRBNN-QK techniques obtain decreased $prec_n$ of 83.12%, 88.25%, 98.72%, 96.57%, and 95.66% respectively. Meanwhile, based on spam assassin, the CHROA-MLSDC algorithm offers enhanced $prec_n$ of 99.92% while the KNN, NB, IRBNN-LK, IRBNN-PK, and IRBNN-QK approaches obtain decreased $prec_n$ of 81.64%, 86.23%, 99.31%, 96.87%, and 97.69% correspondingly.

Table 1 Precision analysis of the CHROA-MLSDC approach with other methods on four datasets

Precision / %						
Dataset	KNN	Naive Bayes	IRBNN-LK	IRBNN-PK	IRBNN-QK	CHROA-MLSDC
Ling spam	83.12	88.25	98.72	96.57	95.66	99.35
Enron	85.89	90.01	96.54	93.12	94.78	98.80
Spam Assassin	81.64	86.23	99.31	96.87	97.69	99.92
CSDM C2010	88.58	91.98	94.22	93.58	91.27	98.02

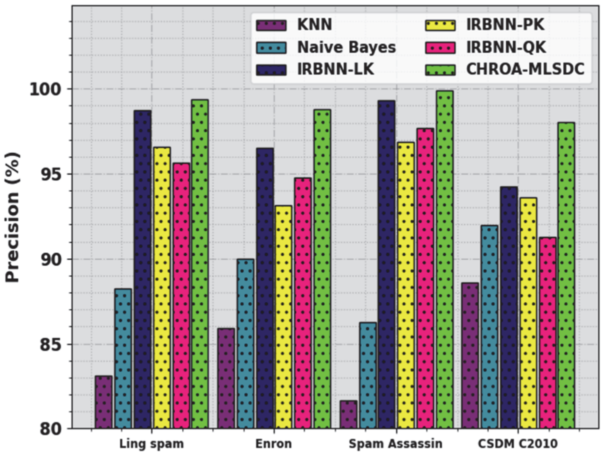

Figure 3 Precision analysis of CHROA-MLSDC approach on four datasets

In Tab. 2 and Fig. 4, a comparative $recal_l$ assessment of the CHROA-MLSDC approach takes place. The figure indicates that the CHROA-MLSDC algorithm gains improved $recal_l$ values. Based on ling spam, the CHROA-MLSDC approach offers enhanced $prec_n$ of 97.95% while the KNN, NB, IRBNN-LK, IRBNN-PK, and IRBNN-QK methods obtain decreased $recal_l$ of 82.38%, 89.82%, 91.94%, 95.83%, and 96.08% correspondingly.

Table 2 Recall analysis of the CHROA-MLSDC approach with other methods on four datasets

Recall / %						
Dataset	KNN	Naive Bayes	IRBNN-LK	IRBNN-PK	IRBNN-QK	CHROA-MLSDC
Ling spam	82.38	89.82	91.94	95.83	96.08	97.95
Enron	85.43	89.69	85.86	92.41	94.48	98.05
Spam Assassin	80.89	85.82	90.41	90.28	96.05	97.82
CSDM C2010	88.36	90.87	83.15	92.38	90.52	97.90

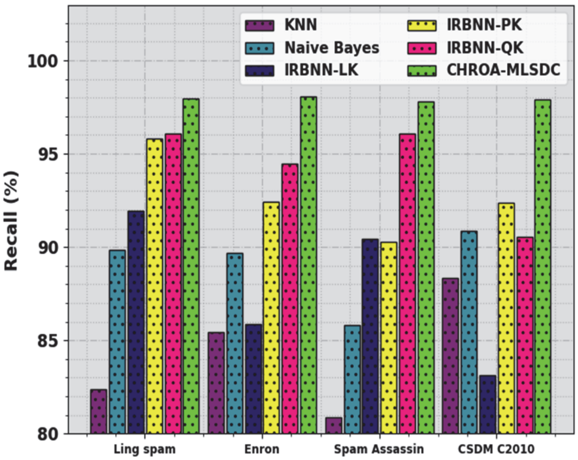

Figure 4 Recall analysis of CHROA-MLSDC approach on four datasets

In Tab. 3 and Fig. 5, a brief $accu_y$ assessment of the CHROA-MLSDC technique takes place. The experimental results indicate that the CHROA-MLSDC method obtains improved $accu_y$ values. Based on ling spam, the CHROA-MLSDC algorithm offers enhanced $accu_y$ of 97.55% while the KNN, NB, IRBNN-LK, IRBNN-PK, and IRBNN-QK techniques obtain decreased $accu_y$ of 79.35%, 85.21%, 95.74%, 93.87%, and 93.78% respectively. In the meantime, based on spam assassin, the CHROA-MLSDC technique offers enhanced $accu_y$ of 96.91% while the KNN, NB, IRBNN-LK, IRBNN-PK, and IRBNN-QK techniques obtain decreased $accu_y$ of 77.63%, 86.54%, 94.79%, 92.66%, and 92.19% correspondingly.

Table 3 Accuracy analysis of CHROA-MLSDC approach with other methods on four datasets

Accuracy / %						
Dataset	KNN	Naive Bayes	IRBNN-LK	IRBNN-PK	IRBNN-QK	CHROA-MLSDC
Ling spam	79.35	85.21	95.74	93.87	93.78	97.55
Enron	78.22	84.98	94.65	93.06	92.95	97.07
Spam Assassin	77.63	86.54	94.79	92.66	92.19	96.91
CSDM C2010	78.51	81.65	92.97	91.98	91.90	96.82

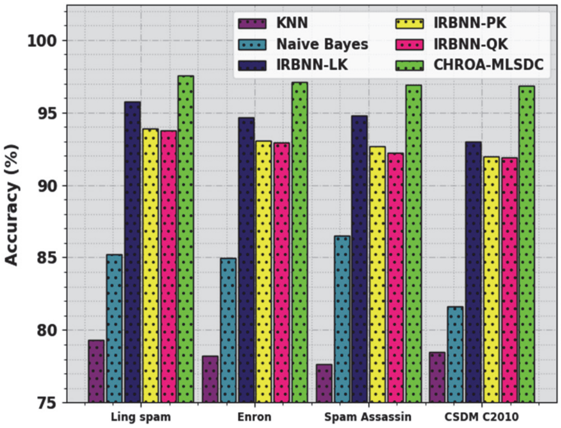

Figure 5 Accuracy analysis of CHROA-MLSDC approach on four datasets

In Tab. 4 and Fig. 6, a detailed execution time (ET) assessment of the CHROA-MLSDC technique with other approaches takes place. The resultant values demonstrated that the CHROA-MLSDC technique gains improved performance with the least ET values. On the ling spam dataset, the CHROA-MLSDC technique accomplishes minimal CT of 0.03 s while the KNN, NB, IRBNN-LK, IRBNN-PK, and IRBNN-QK techniques resulted in increased CT of 0.35 s, 0.25 s, 0.07 s, 0.09 s, and 0.08 s respectively. Eventually, on the spam assassin dataset, the CHROA-MLSDC method accomplishes minimal CT of 0.05 s while the KNN, NB, IRBNN-LK, IRBNN-PK, and IRBNN-QK algorithms resulted in increased CT of 0.37 s, 0.29 s, 0.10 s, 0.12 s, and 0.13 s respectively.

Table 4 Execution Time analysis of the CHROA-MLSDC approach with other methods on four datasets

Execution Time (sec)						
Data set	KNN	Naive Bayes	IRBNN -LK	IRBNN -PK	IRBNN -QK	CHROA-MLSDC
Ling spam	0.35	0.25	0.07	0.09	0.08	0.03
Enron	0.31	0.23	0.05	0.07	0.06	0.02
Spam Assassin	0.37	0.29	0.10	0.12	0.13	0.05
CSDM C2010	0.32	0.21	0.07	0.11	0.09	0.04

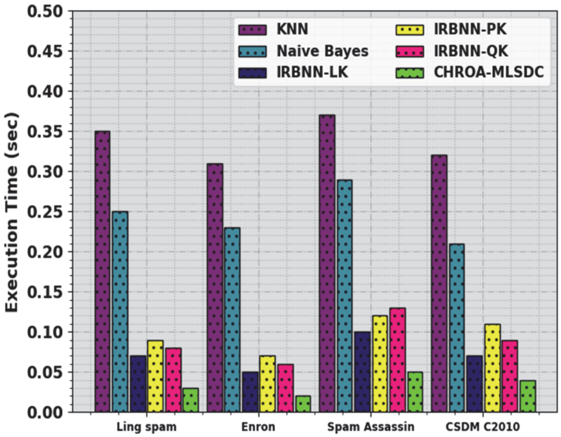

Figure 6 Execution Time analysis of CHROA-MLSDC approach on four datasets

5 CONCLUSIONS

This research presents an automated approach for spam detection and classification called the CHROA-MLSDC model. The primary goal of this method is to effectively categorize content as either spam or non-spam. It comprises several key steps, including data preprocessing, the utilization of CHROA for feature selection, VAE-based classification, and parameter tuning with the assistance of BA. To demonstrate the enhanced performance of the CHROA-MLSDC approach, an extensive series of simulations was carried out. The experimental results consistently showcased the superior performance of CHROA-MLSDC across various evaluation metrics, highlighting its effectiveness compared to existing methods. The incorporation of CHROA and BA-based parameter tuning techniques brings a novel dimension to this work. Through rigorous simulations on many datasets, including Ling spam, Enron, Spam Assassin, and CSDM C2010, the CHROA-MLSDC model consistently exhibits greater performance when compared to alternative techniques. CHROA-MLSDC showcases its exceptional abilities in accurately classifying spam while maintaining computational efficiency, as seen by its utilization of metrics such as precision, recall, accuracy, and execution time. The model effectively resolves the persistent problems posed by spam emails by employing algorithms like CHROA, VAE, and BA. This accomplishment is the result of the collective utilization of these algorithms. The recommended method not only exhibits improved numerical results but also offers a potential pathway for enhancing automated spam detection systems. Subsequent research might explore additional adjustments, such as including outlier identification approaches, to enhance the model's capabilities and robustness.

6 REFERENCES

[1] Salama, W. M., Aly, M. H., & Abouelseoud, Y. (2023). Deep learning-based spam image filtering. *Alexandria Engineering Journal, 68*, 461-468. https://doi.org/10.1016/j.aej.2023.01.048

[2] Acko, B., Weber, H., Hutzschenreuter, D., & Smith, I. (2020). Communication and validation of metrological smart data in IoT-networks. *Advances in Production Engineering & Management, 15*(1), 107-117. https://doi.org/10.14743/apem2020.1.353

[3] Magdy, S., Abouelseoud, Y., & Mikhail, M. (2022). Efficient spam and phishing emails filtering based on deep learning. *Computer Networks, 206*, 108826. https://doi.org/10.1016/j.comnet.2022.108826

[4] Kavitha, S., Umamaheswari, A., & Venkatesh, R. (2021). Intelligent Intrusion Detection System using Enhanced Arithmetic Optimization Algorithm with Deep Learning Model. *Technical Gazette, 30*(4), 1217-1224. https://doi.org/10.17559/TV-20221128071759

[5] Amin, I. & Dubey, M. K. (2022). Hybrid ensemble and soft computing approaches for review spam detection on different spam datasets. *Materials Today: Proceedings,* 62, 4779-4787. https://doi.org/10.1016/j.matpr.2022.03.342

[6] Rapacz, S., Chołda, P., & Natkaniec, M. (2021). A method for fast selection of machine-learning classifiers for spam filtering. *Electronics, 10*(17), 2083. https://doi.org/10.3390/electronics10172083

[7] Sisodia, D. S., Mahapatra, S., & Sharma, A. (2020). Automated sms classification and spam analysis using topic modeling. *2nd International Conference on Data, Engineering and Applications (IDEA),* 1-6. https://doi.org/10.1109/IDEA49133.2020.9170710

[8] Ahmad, S. B. S., Rafie, M., & Ghorabie, S. M. (2021). Spam detection on Twitter using a support vector machine and users' features by identifying their interactions. *Multimedia Tools and Applications, 80*(8), 11583-11605. https://doi.org/10.1007/s11042-020-10405-7

[9] Elakkiya, E., Selvakumar, S., & Velusamy, R. L. (2020). CIFAS: Community Inspired Firefly Algorithm with fuzzy cross-entropy for feature selection in Twitter Spam detection. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-7. https://doi.org/10.1109/ICCCNT49239.2020.9225321

[10] Li, Y. & Zhanyong, W. (2022). A Cloud Based Network Intrusion Detection System. *Technical Gazette, 29*(3), 987-992. https://doi.org/10.17559/TV-20211130024245

[11] Gaurav, D., Tiwari, S. M., Goyal, A., Gandhi, N., & Abraham, A. (2020). Machine intelligence-based algorithms for spam filtering on document labeling. *Soft Computing, 24,* 9625-9638. https://doi.org/10.1007/s00500-019-04473-7

[12] Srinivasarao, U. & Sharaff, A. (2023). Machine intelligence based hybrid classifier for spam detection and sentiment analysis of SMS messages. *Multimedia Tools and Applications,* 1-31. https://doi.org/10.1007/s11042-023-14641-5

[13] Bhardwaj, U. & Sharma, P. (2023). Email spam detection using bagging and boosting of machine learning classifiers. *International Journal of Advanced Intelligence Paradigms, 24*(1-2), 229-253. https://doi.org/10.35940/ijitee.K1365.0981119

[14] Samarthrao, K. V. & Rohokale, V. M. (2022). A hybrid meta-heuristic-based multi-objective feature selection with adaptive capsule network for automated email spam detection. *International Journal of Intelligent Robotics and Applications, 6*(3), 497-521. https://doi.org/10.1007/s41315-021-00217-9

[15] Rayan, A. (2022). Analysis of e-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique. *Computational Intelligence and Neuroscience.* https://doi.org/10.1155/2022/2500772

[16] Arun Prasad, P. B., Mohan, V., & Vinoth Kumar, K. (2024). Hybrid Metaheuristics with Deep Learning Enabled Cyberattack Prevention in Software Defined Networks. *Technical Gazette, 31*(1), 208-214. https://doi.org/10.17559/TV-20230621000752

[17] Poonkodi, T. (2021). E-Mail Spam Filtering Through Feature Selection Using Enriched Firefly Optimization Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(5), 1248-1255. https://doi.org/10.17762/turcomat.v12i5.1791

[18] Rajakani, V. & Vinoth Kumar, K. (2023). Barnacles Mating Optimizer with Hopfield Neural Network Based Intrusion Detection in Internet of Things Environment. *Tehnički vjesnik, 30*(6), 1821-1828. https://doi.org/10.17559/TV-20230414000533

[19] Busaeed, S., Katib, I., & Mehmood, R. (2020). A fog-augmented machine learning based SMS spam detection and classification system. *2020 fifth international conference on fog and mobile edge computing (FMEC),* 325-330. https://doi.org/10.1109/FMEC49853.2020.9144833

[20] Avcı, İ. & Koca, M. (2023). Predicting DDoS Attacks Using Machine Learning Algorithms in Building Management Systems. *Electronics,* 12(19), 4142. https://doi.org/10.3390/electronics12194142

[21] Avci, I. & Yildirim, M. (2023). Solving Weapon-Target Assignment Problem with Salp Swarm Algorithm. *Tehnički vjesnik, 30*(1), 17-23. https://doi.org/10.17559/TV-20220113192727

[22] Avcı, İ. & Koca, M. (2023). Cybersecurity Attack Detection Model, Using Machine Learning Techniques. *Acta Polytechnica Hungarica, 20*(7), 29-44. https://doi.org/10.12700/APH.20.7.2023.7.2

[23] Aqeel, I., Khormi, I. M., Khan, S. B., Shuaib, M., Almusharraf, A., Alam, S., &Alkhaldi, N. A. (2023). Load Balancing Using Artificial Intelligence for Cloud-Enabled Internet of Everything in Healthcare Domain. *Sensors, 23*(11), 5349. https://doi.org/10.3390/s23115349

[24] Putra, C. D., Al Isyrofie, A. I. F., Astuti, S. D., Putri, B. D., Ummah, D. R., Khasanah, M., Permatasari, P. A. D., & Syahrom, A. (2023). Variationalautoencoder analysis gas sensor array on the preservation process of contaminated mussel shells (Mytilusedulis). *Sensing and Bio-Sensing Research, 40*, 100564. https://doi.org/10.1016/j.sbsr.2023.100564

[25] Thiruppathi, M. & Vinoth Kumar, K. (2023). Seagull Optimization-based Feature Selection with Optimal Extreme Learning Machine for Intrusion Detection in Fog Assisted WSN. *Technical Gazette, 30*(5), 1547-1553. https://doi.org/10.17559/TV-20230130000295

[26] Malik, S., Akram, T., Awais, M., Khan, M. A., Hadjouni, M., Elmannai, H., Alasiry, A., Marzougui, M., & Tariq, U. (2023). An Improved Skin Lesion Boundary Estimation for Enhanced-Intensity Images Using Hybrid Metaheuristics. *Diagnostics, 13*(7), 1285. https://doi.org/10.3390/diagnostics13071285

Contact information:

Amutha THANGARAJ, Assistant Professor
(Corresponding author)
Department of Artificial Intelligence and Data Science,
Care College of Engineering, Tiruchirapalli, India
E-mail: marckcocsephd@gmail.com; amudec25@gmail.com

Geetha SADAYAN, Senior Assistant Professor, PhD
Department of Computer Applications,
University College of Engineering (BIT Campus),
Anna University Trichirappalli, India
E-mail: kasagee1971@gmail.com