Citation: N Jaya Lakshmi, Sangeetha Viswanadham, M A Srinivasu, et al. E-mail classification using horse herd optimization algorithm. *Journal of Harbin Institute of Technology (New Series)*. DOI: 10.11916/j.issn.1005-9113.2025020

E-mail Classification Using Horse Herd Optimization Algorithm

N Jaya Lakshmi[1], *Sangeetha Viswanadham*[2], *M Appala Srinivasu*[3], *B Chakradhar*[4] *and B Kiran Kumar*[5]

(*1. Department of Computer Application, Gayatri Vidya Parishad College of Engineering, Visakhapatnam 530048, Andhra Pradesh, India;*
2. Department of Computer Science and Engineering, GITAM deemed to be University, Visakhapatnam 530045, Andhra Pradesh, India;
3. Department of Computer Science and Engineering (AI & ML), Anil Neerukonda Institute of Technology, Visakhapatnam 531162, Andhra Pradesh, India;
4. Department of Computer Science and Engineering, Raghu College of Engineering, Visakhapatnam 530045, Andhra Pradesh, India;
5. School of Computing, SRMIST deemed to be University, Tiruchirappalli 621105, Tamil Nadu, India)

Abstract: In recent decades, the proliferation of email communication has markedly escalated, resulting in a concomitant surge in spam emails that congest networks and presenting security risks. This study introduces an innovative spam detection method utilizing the Horse Herd Optimization Algorithm (HHOA), designed for binary classification within a multi-objective framework. The method proficiently identifies essential features, minimizing redundancy and improving classification precision. The suggested HHOA attained an impressive accuracy of 97.21% on the Kaggle email dataset, with a precision of 94.30%, a recall of 90.50%, and an F1-score of 92.80%. In comparison to conventional techniques, such as Support Vector Machine (93.89% accuracy), Random Forest (96.14% accuracy), and K-Nearest Neighbours (92.08% accuracy), HHOA exhibited enhanced performance with reduced computing complexity. The suggested method demonstrated enhanced feature selection efficiency, decreasing the number of selected features while preserving good classification accuracy. The results underscore the efficacy of HHOA in spam identification and indicate its potential for further applications in practical email filtering systems.

Keywords: E-mail classification; optimization technique; Support Vector Machine; binary classification; machine learning

CLC number: TP391.13, TP18 **Document code**: A **Article ID**: 1005-9113(2025)00-0000-12

0 Introduction

Various forms of undesirable email messages, such as spam, junk email, promotional, and business-oriented emails, are commonly unwanted by users in their inboxes. Despite their distinctions, for the purposes of this study, all such messages are categorized as spam emails[1]. Spam encompasses a broad range of inappropriate messages disseminated extensively across the Internet, lacking useful content for the recipient. It manifests in diverse formats and across multiple platforms, including social media, websites, forums, instant messaging, and particularly email. Among these, spamming of emails has become popular due to its wide usage for diverse objectives. Spam refers to unsolicited messages sent either explicitly or implicitly by individuals who lack any prior relationship with the recipient[2]. While emails are convenient and efficient for communication, they can turn into a nuisance when exploited by marketers for product promotion and scammers for deceptive purposes. The negative effects of spam emails go beyond just wasting resources, time, and effort. They also worsen communication challenges and contribute to cybercrime, ultimately affecting the global economy and causing substantial financial losses for both commercials and individual persons every year. Unwanted emails not only consume resources such as bandwidth, storage space, and time spent on removal, but they also present security vulnerabilities. Due to its cost-effectiveness and ease, email has become crucial for personal and professional correspondence as digital communication grows

Received 2025-02-10.

* Corresponding author: B Kiran Kumar, Ph.D, Assistant Professor. Email: kirankumar224@ gmail.com.

exponentially.However, spam emails have proliferated due to increased email use, providing substantial issues to users and companies[3-4]. Spam emails use network bandwidth and storage and expose recipients to phishing, malware, and inappropriate information. Thus, precise and effective spam detection technologies are essential to reduce these dangers. Many spam detection methods exist; however, getting high accuracy with few false positives is difficult. K-Nearest Neighbour (KNN)[5], Multilayer Perceptron (MLP)[6], Support Vector Machines (SVM), and Naive Bayesian classifiers have limited generalization, high computational cost, and poor performance with massive email data. Feature selection is crucial to classification accuracy, but current methods lack practical optimization algorithms. The current research gap uses standard classification methods or single-objective optimization. These strategies fail to solve the feature selection classification accuracy trade-off. Also, spam detection using binary metaheuristic algorithms has received little attention. Our project aims to create a new spam detection system using a Binary Horse Herd Optimization Algorithm (BHHOA) to maximize feature selection, use a multi-objective contrastive transformation technique and assess the proposed algorithm against classical classifiers on benchmark datasets.

The main contributions of our work are: A modified binary version of the Horse Herd Optimization Algorithm (HHOA) for spam email detection; Introduced a contrastive multi-objective framework for better feature selection; Utilized the Kaggle email dataset for comprehensive performance analysis, revealing enhanced classification accuracy and reduced computing complexity.

The remainder of this paper is structured as follows: Section 1 presents the related works, reviewing existing email classification techniques. Section 2 briefly explains horse herd optimization. Section 3 discusses the proposed methodology, including data pre-processing, feature extraction, and model design. Section 4 provides experimental results and performance evaluations. Finally, Section 5 concludes the paper with future research directions.

1 Related Works

Marketers' unsolicited spam emails, aimed at promoting their products, are often deemed bothersome due to their significant occupation of server space. Scammers attempt to obtain users' bank account information through these emails, aiming to steal funds. Additionally, attackers may utilize these spam emails through which viruses and other malicious software may distribute, often concealing them within enticing and alluring offer links[7]. Hence, it is imperative to promptly tackle the issue of spam emails and implement effective measures to mitigate this problem. Numerous researchers have dedicated their efforts to addressing the challenge of email spam detection, resulting in the proposal of several noteworthy approaches documented in the literature. This section delves into previous studies that concentrate on identifying and categorizing spam using machine learning methods and deep learning algorithms. While Naive Bayes stands out as a commonly utilized algorithm for this purpose, various techniques have been introduced for spam detection. Nonetheless, our current study primarily emphasizes the exploration of metaheuristic optimization algorithms. Bibi'[8] study offers a comprehensive comparison of previous algorithms for spam filtering, examining their accuracy and the datasets utilized. This research provides detailed insights into the straightforward Naive Bayes algorithm, recognized as the top classification algorithms for text mining. Their evaluation of classifier for detection of spam revealed that employing WEKA, the Naive Bayes algorithm delivers effective accuracy and precision. Srinivasan et al.[9] introduced a method for detecting spam utilizing embedding of words within a deep learning framework in the context of Natural Language Processing (NLP). Their study highlights that deep learning surpasses classical classifiers in terms of effectiveness for spam detection. A spam mail detection system developed by Sharma and Bhardwaj[10], by utilizing a hybrid machine learning approach that combines Naive Bayes and the J8 decision tree. The hybrid system comprises four models: preparation of data set, pre-processing of data, selection of features, and a hybrid bagged approach. The experiments were conducted, with the first two experiments focusing on Naive Bayes and J8, while one experiment evaluated the proposed method. The proposed system achieved an accuracy of 87.5%.

Carreras and Marquez[11], utilized a decision tree in their study to sift through unwanted emails.

Due to the inherent challenge of defining features specific to spam emails, this approach is not widely adopted in spam filtering. Meanwhile, Harisinghaney et al.[12] employed K-Nearest Neighbors (KNN), Naive Bayes, and Reverse DBSCAN algorithms to categorize based on image and text. They conducted a performance comparison of these algorithms based on four measurement factors. Soni[13] introduced a novel model for detecting spam called THEMIS, which simultaneously analyses emails at various levels such as header, body, character, and words. This innovative approach employs deep Convolutional Neural Network (CNN) algorithms for identifying emails which are not legitimate.

Our proposed method's performance and efficiency are to be rigorously assessed and evaluated by several profound and popular algorithms which can perform optimization as well as classification. For this purpose, this kind of algorithms was selected for simulation from the literature, and their performance and efficiency were compared to that of the proposed approach. The results after simulation demonstrate that the proposed method surpasses existing approaches, exhibiting higher accuracy and precision, also reduction in time for execution and rate of error. Thus, the superiority of the new method lies in its enhanced accuracy and speed, alongside lower error rates and complexity. As mentioned previously, in order to incorporate HHOA for feature selection, we converted it from its original continuous form to a discrete algorithm. Additionally, recognizing that feature selection entails a multi-objective challenge, we further adapted HHOA into a framework which supports multi-objective, employing it to select features which cause spam. To our understanding, this represents the inaugural research attempt in this domain, introducing both a binary and multi-objective rendition of HHOA.

Drawbacks of state-of-the-art methods and BHHOA addresses them as follows.

1) Inefficient feature selection: Classical classifiers have drawbacks, such as using static or manually picked features containing useless or redundant information. This reduces categorization accuracy. BHHOA uses a multi-objective contrastive transformation strategy to optimize feature selection. It selects key features to improve classification accuracy and reduce computational load.

2) Low generalization and high false positives: Conventional models like KNN and MLP may have significant false-positive rates due to difficulty in generalizing across various datasets. The suggested BHHOA solution uses a robust metaheuristic method to reduce false positives and negatives, enhancing the dataset's generalization.

3) Computational complexity: SVM and MLP can be costly, particularly for large datasets. BHHOA's solution reduces computational complexity by efficiently searching the solution space utilizing herd-based optimization, reducing resource consumption.

4) Existing techniques frequently rely on single-objective optimization, which might limit feature selection and classification accuracy. A multi objective framework optimizes objectives, resulting in a more balanced and effective spam detection model.

5) Inflexibility: Classical approaches are difficult to adjust to the changing nature of spam emails. BHHOA's adaptability makes it more resilient to evolving spam patterns and sophisticated attacks.

6) Limited comparative performance: Minor gains in earlier benchmarks hinder practical use.

2 Horse Herd Optimization Algorithm

Very recently, a plethora of heuristic optimization algorithms have found application in solving diverse optimization problems, owing to their capacity to mathematically model and tackle real-world challenges. This research sought to leverage a novel heuristic optimization algorithm for addressing the problem of selecting necessary features in spam email detection. Hence, HHOA was selected as the primary method for this purpose. HHOA, as introduced by MiarNaeimi et al.[14], is a robust heuristic optimization algorithm which was drawn from the horses' herding behaviours across varied age groups. With a multitude of control factors derived from the behaviours of horses at various stages, HHOA demonstrates exceptional performance in handling problems with complex and dimensions with high in number. Its efficacy has been assessed at large number of dimensions, reaching nearly 10,000, using test functions which are popular, and it has proven to be highly efficient in both exploration and exploitation. The integration of Taylor series within the horse herd optimization algorithm facilitates the clustering of subgraphs in a web page recommendation system.

With the ability to swiftly identify optimal solutions at minimal cost and complexity, HHOA outperforms many established metaheuristic optimization algorithms in terms of both accuracy and efficiency.

Horses exhibit diverse behaviours throughout their lifespans, with a typical maximum lifespan ranging from 25 to 30 years. Horses are categorized into four groups based on age: those aged 0 – 5, 5 – 10, 10 – 15, and older than 15, denoted as δ, γ, β, α respectively. HHOA models the social interactions of horses using six fundamental behaviours observed across different ages: grazing, hierarchy, sociability, imitation, defence mechanism, and roaming. Horse movement at each iteration is described by:

$$\vec{X}_{\mathrm{m}}^{\mathrm{iter,AGE}} = V_{\mathrm{m}}^{\mathrm{iter,AGE}} + X_{\mathrm{m}}^{(\mathrm{iter}-1),\mathrm{AGE}}, \mathrm{Age} = \alpha,\beta,\gamma,\delta \quad (1)$$

where, X is a position vector in the search space – a potential solution to the optimization problem, m is index of the individual/agent/solution in the population, and V is Variation vector (velocity/mutation/update). m^{th} horse position is given by $X_m^{\mathrm{iter},\eta}$, m^{th} horse velocity is given by $V_m^{\mathrm{iter},\eta}$, range of the horse is given by AGE, and current iteration is given by Iter.

To ascertain the horses age, each iteration necessitates a comprehensive responses matrix. This responses matrix is organized based on the most favorable responses, with the top 10% of horses selected as category δ. The subsequent 20%, 30%, and 40% of the remaining horses comprise categories γ, β, and α, respectively. To determine the velocities in terms of a vector, the simulation of the six aforementioned behaviors is mathematically executed.

$$\vec{V}_m^{\mathrm{iter},\alpha} = G\vec{a}_m^{\mathrm{iter},\alpha} + D\vec{e}_m^{\mathrm{iter},\alpha} \quad (2)$$

$$\vec{V}_m^{\mathrm{iter},\sigma} = G\vec{a}_m^{\mathrm{iter},\beta} + H\vec{r}_m^{\mathrm{iter},\beta} + S\vec{o}_m^{\mathrm{iter},\beta} + D\vec{e}_m^{\mathrm{iter},\beta} \quad (3)$$

$$\vec{V}_m^{\mathrm{iter},\gamma} = G\vec{a}_m^{\mathrm{iter},\gamma} + H\vec{r}_m^{\mathrm{iter},\gamma} + S\vec{o}_m^{\mathrm{iter},\gamma} + \vec{lm}_m^{\mathrm{iter},\gamma} + D\vec{e}_m^{\mathrm{iter},\gamma} + R\vec{o}_m^{\mathrm{iter},\gamma} \quad (4)$$

$$\vec{V}_m^{\mathrm{iter},\delta} = G\vec{a}_m^{\mathrm{iter},\delta} + \vec{lm}_m^{\mathrm{iter},\delta} + R\vec{o}_m^{\mathrm{iter},\delta} \quad (5)$$

The behaviors which are mentioned above are elaborated with their implementation are illustrated as follows.

2.1 Grazing

This is a kind of behaviour, where horses are known for their grazing behavior, consuming grasses, plants, and small animals, with a typical pasturing duration ranging from 16 to 20 h per day. This slow and persistent eating habit is a characteristic of their behavior. In the HHOA, the grazing behavior is represented mathematically by assigning a coefficient to denote the grazing space around each individual horse. This coefficient signifies the area in which the horse is actively grazing, simulating its foraging activity within a specific space.

$$G\vec{a}_m^{\mathrm{iter,age}} = g_m^{\mathrm{iter,age}}(u\breve{b} + \rho l\breve{b})\ [B_m^{\mathrm{iter}-1}],\ \mathrm{Age} = \alpha,\beta,\gamma,\delta \quad (6)$$

$$g_m^{\mathrm{iter,age}} = g_m^{\mathrm{iter}-1,\mathrm{age}} \times \omega_g \quad (7)$$

where, i^{th} position horse's motion parameter, $G\vec{a}_m^{\mathrm{iter,age}}$ indicating its tendency to graze, The limits of the pasturing area is expressed as $u\hat{b}$ and $l\hat{b}$, where $u\hat{b}$ is upper bound and $l\hat{b}$ is lower bound. In addition, ρ is a measure that lies in the range of $[0,1]$.

2.2 Hierarchy

Horses typically rely on a leader for guidance, which can be an adult stallion, or a mare, adhering to the hierarchy principle. In a horse herd, the most experienced and strongest individual often assumes the leadership role, with others following suit. Horses falling within the age range of 5 to 15 (categories β and γ) have been observed to adhere to the hierarchy principle and follow the lead of the dominant horse. It can be described as follows:

$$H\vec{r}_m^{\mathrm{iter,age}} = y_m^{\mathrm{iter,age}}[\beta_*^{\mathrm{iter}-1} - \beta_m^{\mathrm{iter}-1}],\ \mathrm{Age} = \alpha,\beta,\gamma \quad (8)$$

$$y_m^{\mathrm{iter,age}} = y_m^{\mathrm{iter}-1,\mathrm{age}} \times \omega_y \quad (9)$$

where $H\vec{r}_m^{\mathrm{iter,age}}$ is the influence of the leader horse's position on velocity and the place of that horse is indicated by B_*^{iter-1}.

2.3 Sociability

Sociability is a key behavioral trait observed in horses, which serves as a source of inspiration for HHOA. Horses naturally seek social engagement and often coexist harmoniously with other peer animals, enhancing their chances of survival. Some horses even demonstrate a preference for companionship, extending to species such as cattle and sheep. This sociable behavior is particularly prominent in horses aged 5 to 15 years old. In HHOA, sociability is reflected in the movement of horses towards the positions of other herd members, facilitating socialization within the group. This behavior is mathematically modelled as follows:

$$S\vec{o}_m^{\mathrm{iter,age}} = s_m^{\mathrm{iter,age}}\left[\left(\frac{1}{N}\sum_{m=1}^{N} B_m^{\mathrm{iter}-1}\right) - B_m^{\mathrm{iter}-1}\right],\ \mathrm{Age} = \beta,\gamma \quad (10)$$

$$s_m^{\mathrm{iter,age}} = s_m^{\mathrm{iter}-1,\mathrm{age}} \times \omega_s \quad (11)$$

where, N is total number of individuals in the

population social vector motion of i^{th} horse is given by $\overrightarrow{So}_m^{iter,age}$, and group cohesion is given by $s_m^{iter,age}$ at Iterth iteration.

2.4 Imitation

Horses have a tendency to learn both positive and negative habits and behaviors from one another through imitation, a behavioral trait that also influences HHOA. Young horses, in particular, are inclined to mimic the actions of their peers, and this imitation behavior remains prevalent throughout their lives. This imitation is described by following equations:

$$\overrightarrow{\mathrm{Im}}_m^{iter,age} = f_m^{iter,age}\left[\left(\frac{1}{pN}\sum_{m=1}^{pN}\widehat{B}_m^{iter-1}\right) - B^{iter-1}\right], \mathrm{Age} = \gamma \tag{12}$$

$$f_m^{iter,age} = f_m^{iter-1,age} \times \omega_f \tag{13}$$

where, $p \in (0,1]$ is a proportion of the population size N and f is behavioural learning rate. The vector indicating the movement of the m^{th} horse towards the mean position of optimal horses is represented as pN, where pN denotes the number of horses with favorable positions.

2.5 Defense

Horses rely on a "fight-or-flight" response mechanism to defend themselves, typically opting to flee when confronted with danger. However, when cornered, they may resort to bucking as a defensive tactic. Additionally, horses engage in confrontations to assert dominance over resources such as food and water, and to fend off threats from predators like wolves. This defensive behavior of horses is mirrored in HHOA, where horses avoid non-optimal responses by moving away from them. This process of defense mechanism is given by using below equation,

$$D\overrightarrow{e}_m^{iter,age} = -\xi_m^{iter,age}\left[\left(\frac{1}{\mathrm{P}N}\sum_{m=1}^{\mathrm{P}N}\breve{B}_m^{iter-1}\right) - B^{iter-1}\right], \mathrm{Age} = \alpha, \beta, \text{ and } \gamma \tag{14}$$

$$\xi_m^{iter,age} = \xi_m^{iter-1,age} \times \omega_\xi \tag{15}$$

where, the vector for escaping of m^{th} horse from the average of some horses with worse positions is indicated as $D\overrightarrow{e}_m^{iter,age}$. The number of horses count with worse positions is given as pN and the reduction parameter per cycle is shown as ω_ξ.

2.6 Roam

The last kind of horse behavior simulated by HHOA is their tendency of roaming. In their natural habitat, horses roam and graze, moving from one place to another if they are not confined to stables. They may swiftly change their grazing location and exhibit curiosity by exploring various pastures to familiarize themselves with their surroundings. In HHOA, behavior of roaming is represented as a movement which is random, of horses within the herd. This behavior can be depicted as,

$$R\overrightarrow{o}_m^{iter,age} = r_m^{iter,age}\rho\, B^{iter-1}, Age = \gamma, \beta \tag{16}$$

$$r_m^{iter,age} = r_m^{iter-1,age} \times \omega_r \tag{17}$$

here, $R\overrightarrow{o}_m^{iter,age}$ specifies the mth horse velocity when random movement made for a locally search and the parameter which is used for reduction per cycle is depicted as $r_m^{iter,age}$. For different age groups, the horses' velocity obtained is expressed as follows.

For the δ horses, the velocity (horses at the age of 0 – 5) is given a below,

$$\overrightarrow{V}_q^{iter,\beta} = [g_q^{iter-1,age} \times \omega_g(u\breve{b} + \rho l\breve{b})[B_q^{iter-1}]] + \left[f_q^{iter-1,age} \times \omega_f\left[\left(\frac{1}{\rho N}\sum_{q=1}^{\rho N}\widehat{B}_q^{iter-1}\right) - B^{iter-1}\right]\right] + [r_q^{iter-1,age} \times \omega_r \rho\, B^{iter-1}] \tag{18}$$

For the γ horses (horses at the age of 5 – 10), the velocity is given as,

$$\overrightarrow{V}_q^{iter,\gamma} = \left[g_q^{iter-1,age} \times \omega_g(u\breve{b} + \rho l\breve{b})[B_q^{iter-1}]\right] + \left[y_q^{iter-1,age} \times \omega_y\left[B_*^{iter-1} - B_q^{iter-1}\right]\right] + \left[s_q^{iter-1,age} \times \omega_s\left[\left(\frac{1}{N}\sum_{q=1}^{N}B_q^{iter-1}\right) - B_q^{iter-1}\right]\right] + \left[f_q^{iter-1,age} \times \omega_f\left[\left(\frac{1}{\rho nn}\sum_{q=1}^{\rho nn}\widehat{B}_q^{iter-1}\right) - B^{iter-1}\right]\right] - \left[\xi_q^{iter,age}\left[\left(\frac{1}{\mathrm{P}nn}\sum_{q=1}^{\mathrm{P}nn}\breve{B}_q^{iter-1}\right) - B^{iter-1}\right]\right] + \left[r_q^{iter-1,age} \times \omega_r \rho\, B^{iter-1}\right] \tag{19}$$

For the β horses, (horses at the age between 10 and 15 years), the velocity is given as,

$$\overrightarrow{V}_q^{iter,\beta} = \left[g_q^{iter-1,age} \times \omega_g(u\breve{b} + \rho l\breve{b})[B_q^{iter-1}]\right] + \left[y_q^{iter-1,age} \times \omega_y[B_*^{iter-1} - B_q^{iter-1}]\right] + \left[s_q^{iter-1,age} \times \omega_s\left[\left(\frac{1}{nn}\sum_{q=1}^{nn}B_q^{iter-1}\right) - B_q^{iter-1}\right]\right] - \left[\xi_q^{iter,age}\left[\left(\frac{1}{\mathrm{P}nn}\sum_{q=1}^{\mathrm{P}nn}\breve{B}_q^{iter-1}\right) - B^{iter-1}\right]\right] \tag{20}$$

For the α horses (horses older than 15), the velocity is given as,

$$\overrightarrow{V}_q^{iter,\alpha} = \left[g_q^{iter-1,age} \times \omega_g(u\breve{b} + \rho l\breve{b})[B_q^{iter-1}]\right] -$$

$$\left[\xi_q^{\text{iter,age}}\left[\left(\frac{1}{\text{P}nn}\sum_{q=1}^{\text{P}nn}\breve{B}_q^{\text{iter}-1}\right)-B^{\text{iter}-1}\right]\right] \quad (21)$$

Adult horses categorized as α initiate search locally around the aim of global optimum with exceptional precision. Horses categorized as β seek out neighbouring environment around the adult α horse, aiming to draw closer to them. Conversely, horses categorized as γ exhibit a little bit low interest in reaching the α horses and instead demonstrate a very strong inclination to find other places and uncover additional globally optimum places. Due to the distinct behavioural characteristics, young horses categorized as δ are well-suited for the search phase which is random.

3 The Proposed Method

Initially, the metaheuristic algorithm HHOA undergoes modification, followed by the utilization of the adapted Horse herd Optimization Algorithm version for the selection of features in spam email detection. This continuous HHOA is initially converted into a kind of binary to suit the discrete nature of the problem of selecting features. Subsequently, the inputs of the resultant algorithm are transformed into contrastive inputs. Following this, the binary contrastive HHOA is further enhanced to support multi-objective optimization, facilitating the resolution of the problems which are multi-objective. At last, this multi-objective contrastive binary HHOA is employed for spam detection purposes. Spam emails are often received by users from unknown senders with unusual email addresses. Therefore, it´s crucial to employ suitable methods for detecting and distinguishing emails which are spam, from legitimate ones containing important information. Each email received from the web server undergoes a series of steps to determine whether it is spam or not spam. The initial step after receiving an email involves extraction of features, where a set of general or specific features from the email body. Following feature extraction, the subsequent phase is selection of features, which discerns relevant features while eliminating irrelevant and duplicate ones. At last, the step entails classification, where emails are categorized as either spam or not spam. The complete framework of this method is illustrated in Fig. 1 and Fig. 2 depicting the flow of steps of the novel approach and its functionality in identifying spam emails. Subsequent sections offer comprehensive insights into each step involved in adapting the HHOA.

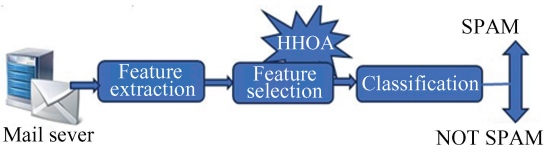

Fig. 1 Frame work of the proposed method

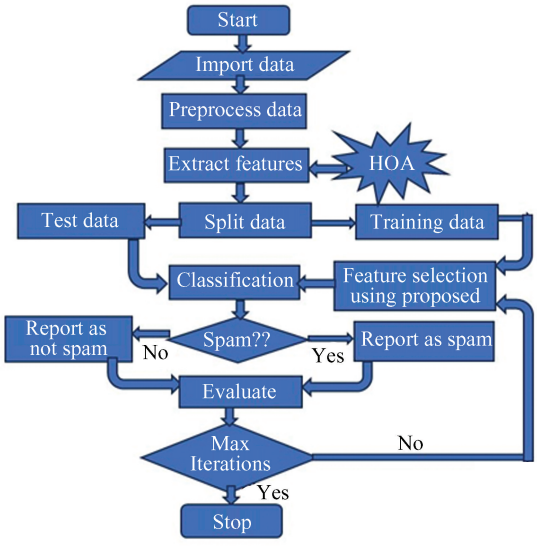

Fig. 2 Flowchart related to the proposed approach

3.1 Binary HHOA

The process of optimization varies notably between binary and continuous search spaces. In a search space of continuous, a step length is added to adjust the position vector by horse search agents. However, in a binary search space, this method is not applicable, as the position vector of search agents can only hold values of 0 or 1. Consequently, it was necessary to devise a binary adaptation of the HHOA tailored for selection of features, which inherently entails discrete problem - solving[15]. The binary version of the HHOA algorithm can be devised in simple and straightforward way. Here, simply need to set the variables´ lower and upper bound values between zero and one, then execute the algorithm. Just before inputting the values into the cost function, apply the greatest integer function to round them to a zero-one vector. While the variables maintain their continuous nature, they are treated as binary by the cost function, which only occurs before entering the cost function. Essentially, the algorithm views the problem as continuous, while the cost function treats it as discrete. Additionally, a function facilitates communication between the discrete cost function

(binary) and the continuous algorithm. This function is achieved by utilizing the greatest integer function, where x represents a real value between two consecutive integers m and n, resulting in an integer k after applying the greatest integer function to x. This approach effectively addresses the challenge of adapting a continuous algorithm for use in discrete problems[16].

3.2 Contrastive Binary HHOA

By examining contradictory solutions, contrastive learning enhances the likelihood of commencing with a superior initial population. This approach is not only applicable to initial solutions but can also be continuously applied to any solution within the current population. Typically, contrastive learning is integrated into metaheuristic approaches to enhance convergence[17]. As the temporal complexity of metaheuristic algorithms escalates, contrastive learning serves to mitigate these constraints. This strategy entails the metaheuristic method seeking solutions which are optimal, in the reverse direction of the current solution. Subsequently, it evaluates and identifies the best solution from the current and opposite directions. This methodology accelerates solution convergence, moving it nearer to the optimal solution[18].

3.3 Multi-Objective Contrastive Binary HHOA

Optimization models aimed at solving problems with only single objective function are referred to as single-objective models. In such problems, the goal is to identify the optimal solution from a set of available alternatives. However, in practical scenarios within design and engineering domains, many problems involve multiple objective functions[19]. These kinds of problems are known as multi-objective optimization. Spam detection poses a multi-objective challenge, aiming to optimize two primary objectives: reducing the number of features while maximizing classification accuracy. Achieving higher classification accuracy ensures most emails are correctly categorized, with minimal classification errors. Given that the modified HHOA metaheuristic algorithm's feature selection significantly influences classification, minimizing feature count is crucial to prevent complexity. Since multiple objective functions are involved, employing a multi-objective optimization method becomes necessary. Such methods offer engineers and system designers multiple solutions that strike a balance between various objectives[20].

The primary distinction between single-objective and multi-objective HHOA lies in their approach to updating objectives. In single-objective search spaces, selecting the best solution is straightforward. Conversely, in multi-objective HHOA, from a set of optimal solutions, the objective must be chosen. These optimal solutions are preserved, and one of them ultimately serves as the objective. The challenge here lies in enhancing the distribution of stored solutions by finding an objective. To achieve this, the number of neighboring solutions within the existing solution's vicinity is initially computed.

Selection of features in spam detection presents a multi-objective optimization challenge. This entails balancing two conflicting objectives: (1) Reducing the number of selected features and (2) increasing classification accuracy. Consequently, defining the objective function for feature selection necessitates the use of a classification algorithm.

3.4 Detection of Spam Using Multi Objective Contrastive Binary HHOA

Selection of features comprises a process in four steps, involving the generation of feature subsets, evaluation of these subsets, checking stopping criteria, and validating the results. Initially, a feature subset is generated within the dataset, wherein candidate features are identified based on the search strategy of multi - objective contrastive binary HHOA[21]. Subsequently, evaluation of these candidate subsets is done and compared with the best previous value of the evaluation feature. If a superior subset is discovered, it replaces the previous best subset. This iterative process of generating and evaluating subsets continues until the termination criterion of multi-objective contrastive binary HHOA is met. This process iterates multiple times until reaching the best global solution. Following each iteration, the fitness function computes the classifier's accuracy for the candidate subset. The process of candidate generation, fitness calculation, and evaluation function persists until the final criteria are satisfied. Typically, stopping criteria are determined based on two things: the error rate and the number of iterations. If the error rate falls below a certain value or if the algorithm surpasses the given number of iterations, the algorithm halts[22-23].

3.5 Feature Selection Using HHOA

Feature selection is an essential preprocessing phase in machine learning that entails identifying a

subset of pertinent features to improve model accuracy and diminish computational complexity. In email categorization, proficient feature selection can enhance spam detection by eliminating redundant and irrelevant information. The Horse Herd Optimization Algorithm (HHOA), derived from the social dynamics of equine herds, has been modified into a binary variant for feature selection. In contrast to conventional algorithms, HHOA sustains a dynamic equilibrium between exploration (seeking new solutions) and exploitation (enhancing existing solutions)[24].

3.5.1 Binary conversion of HHOA

The original HHOA, being continuous, was converted into a binary format with a sigmoid-based transfer function. This function transforms continuous values into probabilities indicating the selection status of a feature. The feature selection procedure can be articulated as follows:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{22}$$

where, $S(x)$ is the sigmoid function value, x is the continuous position of the horse in the search space. and e is Euler's number, a mathematical constant approximately equal to: $e \approx 2.71828$. A threshold is applied to the sigmoid output to determine the selection of a feature:

$$X_i = \begin{cases} 1 \text{ if } S(x) > r \\ 0 \text{ other wise} \end{cases} \tag{23}$$

where r is a random number between 0 and 1, 1 indicates that the feature is selected, while 0 indicates that the feature is not selected. i means iteration, ($i = 1, 2 \ldots n$).

3.5.2 Multi-objective optimization

The proposed HHOA follows a multi-objective optimization approach that simultaneously optimizes two key objectives. Minimization of classification error: Ensures the algorithm maintains high accuracy; Minimization of selected features: Reduces model complexity by selecting the most informative features. The fitness function is defined as:

$$F = \alpha.E + (1 - \alpha).\frac{|S|}{|T|} \tag{24}$$

where, F = fitness function value, E = classification error using a base classifier (e.g., SVM, KNN), $|S|$ = number of selected features, $|T|$ = total number of features, α = balancing parameter between error and feature selection. If α is closer to 1, the algorithm prioritizes minimizing classification error. If α is closer to 0, it focuses on minimizing the number of selected features. This adaptive fitness function ensures both accuracy and feature reduction.

3.5.3 Exploration and exploitation

HHOA conducts exploration through dominant equine behaviour and exploitation via herd-following behaviour. These two actions guarantee the algorithm effectively identifies appropriate feature subsets: Leading horse behavior: Promotes global search to explore new areas in the solution space; Following horse behavior: Promotes local search by refining promising solutions. The position update for horses is given by:

$$X_i(t+1) = X_i(t) + C \cdot (X_l - X_i) + R \cdot (X_b - X_i) \tag{25}$$

where, $X_i(t)$ is current position of horse ii, X_l is position of the leading horse, X_b is best position achieved so far, C and R are random coefficients to maintain diversity.

The proposed HHOA method utilizes a termination criterion to guarantee efficient convergence. The algorithm terminates when any of the following conditions are met: a predetermined maximum number of iterations is reached, and the enhancement in the fitness function becomes trivial. It drops below a designated threshold, or the algorithm converges to an optimal solution where further enhancements are no longer substantial. This adaptive termination technique eliminates superfluous computations, guaranteeing prompt and efficient performance while preserving the precision of the classification outcomes. The binary HHOA facilitates efficient feature selection by diminishing the feature count while maintaining classification accuracy. This not only reduces computing expenses but also improves the model's interpretability. The exceptional efficacy of HHOA, as demonstrated by the experimental findings, confirms its proficiency in enhancing email spam detection.

4 Results Analysis

The studies were performed on a Windows 11 machine with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA RTX 3060 GPU. The Binary Horse Herd Optimization Algorithm (BHHOA) was executed in Python 3.10 utilizing NumPy, Pandas, Scikit-Learn, and Matplotlib modules. A binary variation of HHOA was created and combined with conventional classifiers, including KNN, MLP, SVM, and Naive Bayesian, for comparative

evaluation. The algorithm underwent rigorous validation via comprehensive testing and parameter optimization. Performance evaluation was executed utilizing accuracy, precision, recall, F1 score, False Positive Rate (FPR), and False Negative Rate (FNR) to assess classification accuracy, error mitigation, and robustness. The study examined critical research inquiries regarding BHHOA's efficacy, feature selection abilities, error reduction, and computational efficiency. The Kaggle email classification dataset[25-26], consisting of 57 000 emails with a spam-to-non-spam ratio of 30-70%, was utilized for assessment. Following TF-IDF pre-processing for text representation, the dataset was divided into 80% for training and 20% for testing. Five-fold cross-validation guaranteed dependable and impartial outcomes. Table 1 and graphically in Fig. 3 & Fig. 4 show how BHHOA out performed existing methods.

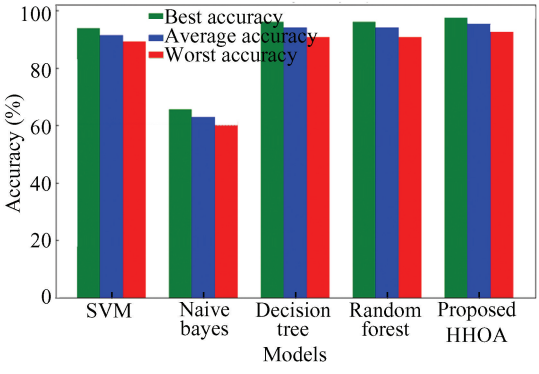

Fig. 3 Bar graph showing accuracy comparison with traditional methods

The Receiver Operating Characteristic (ROC)[26] curve illustrates the balance between the TPR and the FPR. AUC (Area under the Curve) scores approaching 1 signify superior performance as shown in Fig. 5. This demonstrates BHHOA's enhanced categorization capability.

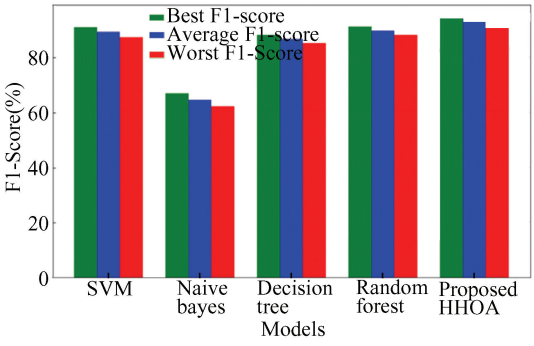

Fig. 4 Bar graph showing F1-score comparison with traditional methods

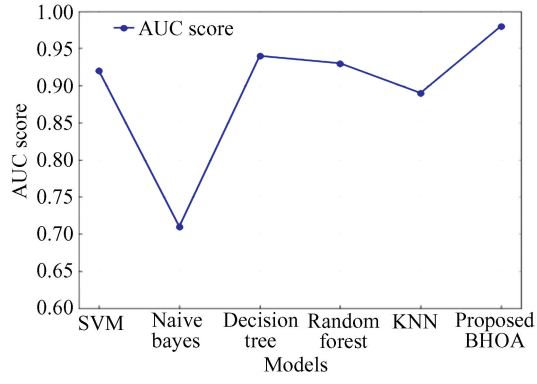

Fig. 5 Line graph showing AUC Score

Table 1 Comparasion of results of proposed methodlogy with tradational methods

Model	Best accuracy (%)	Worst accuracy (%)	Average accuracy (%)	Best F1-score	Worst F1-score	Average F1-score
SVM	93.89	89.12	91.50	90.77	87.10	89.00
Naïve Bayes	65.64	60.02	63.00	66.73	62.01	64.50
Random Forest	96.14	90.78	94.10	90.97	88.00	89.50
Proposed HHOA	97.21	92.60	95.40	94.30	90.50	92.80

4.1 Computational Efficiency

BHHOA shows competitive training and prediction times as shown in Table 2 and Fig. 6, making it a more efficient choice for real-time applications.

4.2 Error Rate Comparison

BHHOA demonstrates the lowest false positive and false negative rates as shown in Table 3 and Fig.7, further justifying its reliability. Fig. 8 is the convergence diagram for the HHOA. The fitness value gradually declines after 50 iterations, signifying successful convergence. The slight variations indicate that the algorithm continues investigating the search space while enhancing the solution.

Table 2 Comparison of training time and prediction time

Model	Training time (s)	Prediction time (s)
SVM	45	10
Naïve Bayes	20	5
Decision Tree	30	7
Random Forest	60	15
KNN	35	12
Proposed BHHOA	38	8

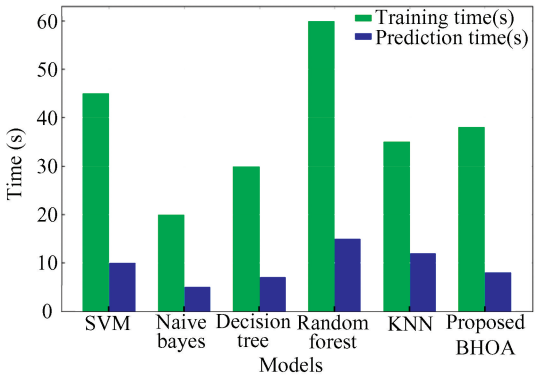

Fig. 6 Bar graph showing computational efficiency

Table 3 Comparison of error rate

Model	FPR	FNR
SVM	0.06	0.11
Naïve Bayes	0.22	0.18
Decision Tree	0.05	0.08
Random Forest	0.04	0.09
KNN	0.10	0.14
Proposed BHHOA	0.03	0.05

4.3 Statically Analysis with Chi-Square Test Results

A chi-square test[27] was performed to assess the statistical significance of the classification performance of the proposed HHOA model. The findings revealed a chi-square statistic of 9070.30 with a p-value of 0.0, which is extremely small, signifying a substantial disparity between the observed and predicted values. With one degree of freedom, the test validated that the model's classification accuracy is not attributable to random chance. The minimal rates of false positives and negatives further substantiate the model's reliability. The results are shown in Table 4.

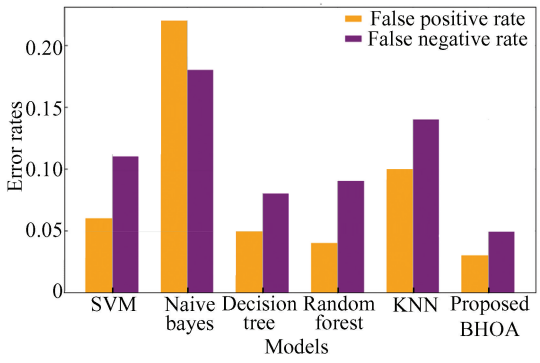

Fig.7 Bar graph showing error rate comparison

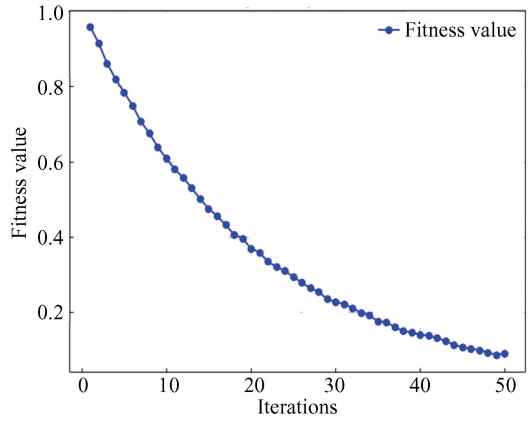

Fig. 8 Convergence graph HHOA

Table 4 Chi-Square Test Results

Chi-square statistic (χ^2)	P-value	Degrees of Freedom	Expected values			
			spam		Non-spam	
			Predicted spam	Non-spam	Predicted spam	Non-spam
9070.30	0.0	1	984.6	2435.4	2297.4	5682.6

This statistical validation reinforces the assertion that the proposed HHOA algorithm proficiently distinguishes spam and non-spam emails, surpassing traditional methods.

5 Conclusions and Future Work

Unsolicited emails, commonly known as spam, pose a significant challenge for both internet users and data centres. They consume substantial storage and resources while also serving as a gateway for intrusion, cyber-attacks, and the user information also accessed unauthorizedly. The objective of this research was to utilize a robust metaheuristic optimization algorithm for identifying emails which are spam in email services. To achieve this goal, the paper utilized the horse herd optimization algorithm, a newly developed nature-inspired metaheuristic optimization approach designed to address exceedingly intricate optimization challenges. This paper outlines the construction of a spam email detection model employing classification with optimization. A comparison between detecting mails which are not legitimate, without optimization and with optimization demonstrates that optimization significantly enhances accuracy.

5.1 Main Challenges of HHOA

The suggested HHOA algorithm encounters multiple obstacles that may affect its performance. It demonstrates parameter sensitivity, necessitating meticulous adjustment to attain optimal outcomes. Moreover, although it is computationally efficient relative to some models, handling extensive datasets may result in prolonged computation time. A further disadvantage is the potential for redundancy, wherein irrelevant or superfluous features may be selected despite the optimization procedure. Moreover, the algorithm may face generalization challenges, as its efficacy can fluctuate when utilized on datasets exhibiting markedly distinct spam characteristics. Mitigating these problems could further augment the algorithm's resilience and utility.

5.2 Limitations of the Study

The suggested HHOA algorithm possesses some drawbacks that may hinder its wider use. The dependence on the Kaggle spam dataset may restrict its applicability to other datasets with varying attributes. The algorithm is also sensitive to initial parameter configurations, which can affect its performance. Evaluating multilingual spam datasets is essential to ascertain their efficacy across various languages. Furthermore, when utilized on exceptionally massive datasets, the technique may encounter computational overhead, affecting efficiency. Mitigating these restrictions can improve the algorithm's resilience and scalability.

References

[1] Shuaib M, Abdulhamid S M, Adebayo O S, et al. Whale optimization algorithm based email spam feature selection method using rotation forest for classification. SN Applied Sciences, 2019, 1: Article number 390. DOI: 10.1007/s42452-019-0394-7.

[2] Abualigah L M, Khader A T, Hanandeh E S. A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. Engineering Applications of Artificial Intelligence, 2018, 73: 111-125. DOI: 10.1016/j.engappai.2018.05.003.

[3] Abualigah L M Q. Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering. Springer: Berlin. 2019.

[4] Raad M, Yeassen N M, Alam G M, et al. Impact of spam advertisement through e-mail: A study to assess the influence of the anti-spam on the e-mail marketing. African Journal of Business Management, 2010, 4(11): 2362-2367.

[5] Karim A, Azam S, Shanmugam B, et al. A comprehensive survey for intelligent spam email detection. IEEE Access, 2019, 7: 168261-168295. DOI: 10.1109/ACCESS.2019.2954791.

[6] Arasteh B, Aghaei B, Farzad B. et al. Detecting SQL injection attacks by binary gray wolf optimizer and machine learning algorithms. Neural Computing and Applications, 2024, 36: 6771-6792. DOI: 10.1007/s00521-024-09429-z.

[7] Majidian Z, Eivazi S T, Arasteh B, et al. An intrusion detection method to detect denial of service attacks using error-correcting output codes and adaptive neuro-fuzzy inference. Computers and Electrical Engineering, 2023, 106: 108600. DOI: 10.1016/j.compeleceng.2023.108600.

[8] Bibi A. Spam mail scanning using machine learning algorithm. Journal of Computer, 2020, 15(2): 73-84. DOI: 10.17706/jcp.15.2.73-84.

[9] Srinivasan S, Ravi V, Alazab M, et al. Spam Emails Detection Based on Distributed Word Embedding with Deep Learning. Machine Intelligence and Big Data Analytics for Cybersecurity Applications. Berlin: Springer, 2021, 161-189.

[10] Sharma P, Bhardwaj U. Machine learning based spam e-mail detection. International Journal of Intelligent Engineering and Systems, 2018, 11(3): 1-10. DOI: 10.22266/ijies2018.0630.01.

[11] Carreras X, Marquez L. Boosting Trees for Anti-Spam Email Filtering. 2001. arXiv: cs/0109015. DOI: 10.48550/arXiv.cs/0109015.

[12] Harisinghaney A, Dixit A, Gupta S, et al. Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. International Conference on Reliability Optimization and Information Technology (ICROIT). Piscataway: IEEE, 2014, 153-155. DOI: 10.1109/ICROIT.2014.6798302.

[13] Soni A N. Spam e-mail detection using advanced deep convolution neural network algorithms. Journal for Innovative Development in Pharmaceutical and Technical Science, 2019, 2(5): 74-80.

[14] MiarNaeimi F, Azizyan G, Rashki M. Horse herd optimization algorithm: A nature-inspired algorithm for high-dimensional optimization problems. Knowledge-Based Systems, 2021, 213: 106711. DOI: 10.1016/j.knosys.2020.106711.

[15] Jayalakshmi N, Sangeeta V, Appala Srinuvasu Muttipati. Taylor horse herd optimized deep fuzzy clustering and Laplace based K-nearest neighbor for web page recommendation. Advances in Engineering Software, 2023, 175: 103351. DOI: 10.1016/j.advengsoft.2022.103351.

[16] Dedeturk B K, Akay B. Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing, 2020, 91: 106229. DOI: 10.1016/j.asoc.2020.106229.

[17] Marinos L, Lourenco M B. ENISA threat landscape report 2018. European Union Agency for Network and Information Security (ENISA). 2018. DOI: 10.2824/622757.

[18] Mafarja M, Aljarah I, Heidari A A, et al. Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems. Knowledge-Based Systems, 2018, 145: 25-45. DOI: 10.1016/j.knosys.2017.12.037.

[19] Shajideen N M, Bindu V. Spam filtering: A comparison between different machine learning classifiers. Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). Piscataway: IEEE, 2018, 1919-1922. DOI: 10.1109/ICECA.2018.8474778.

[20] Zouache D, Arby Y O, Nouioua F, et al. Multi-objective chicken swarm optimization: A novel algorithm for solving multi-objective optimization problems. Computer and Industrial Engineering, 2019, 129: 377-391. DOI: 10.1016/j.cie.2019.01.055.

[21] Xu Y, Chen H, Heidari A A, et al. An efficient chaotic mutative moth-flame-inspired optimizer for global optimization tasks. Expert Systems with Applications, 2019, 129: 135-155. DOI: 10.1016/j.eswa.2019.03.043.

[22] Mohmmadzadeh H. Case study email spam detection of two metaheuristic algorithm for optimal feature selection. Preprints. org. 2020. DOI: 10. 20944/preprints202001.0309.v3.

[23] Pandey A C, Rajpoot D S. Spam review detection using spiral cuckoo search clustering method. Evolutionary Intelligence, 2019, 12: 147-164. DOI: 10.1007/s12065-019-00204-x.

[24] Pashiri R T, Rostami Y, Mahrami M. Spam detection through feature selection using artificial neural network and sine cosine algorithm. Mathematical Sciences, 2020, 14: 193-199. DOI: 10.1007/s40096-020-00327-8.

[25] Rajamohana S P, Umamaheswari K. Hybrid approach of improved binary particle swarm optimization and shuffled frog leaping for feature selection. Computers and Electrical Engineering, 2018, 67: 497-508. DOI: 10.1016/j.compeleceng.2018.02.015.

[26] Razmjooy N, Ramezani M, Ghadimi N. Imperialist competitive algorithm-based optimization of neuro-fuzzy system parameters for automatic red-eye removal. International Journal of Fuzzy Systems, 2017, 19: 1144-1156. DOI: 10.1007/s40815-017-0305-2.

[27] Zhang L, Zhang J, Gao W, et al. A deep learning outline aimed at prompt skin cancer detection utilizing gated recurrent unit networks and improved orca predation algorithm. Biomedical Signal Processing and Control, 2024, 90: 105858. DOI: 10.1016/j.bspc.2023.105858.