

TOPICAL REVIEW

Bio-Inspired Feature Selection Algorithms With Their Applications: A Systematic Literature Review

TIN H. PHAM  **AND BIJAN RAAHEMI**

Knowledge Discovery and Data Mining Laboratory, Telfer School of Management, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Tin H. Pham (tpham104@uottawa.ca)


This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN/341811-2012.

ABSTRACT Based on the principles of the biological evolution of nature, bio-inspired algorithms are gaining popularity in developing robust techniques for optimization. Unlike gradient descent optimization methods, these metaheuristic algorithms are computationally less expensive, and can also considerably perform well with nonlinear and high-dimensional data. Objectives: To understand the algorithms, application domains, effectiveness, and challenges of bio-inspired feature selection techniques. Method: A systematic literature review is conducted on five major digital databases of science and engineering. Results: The primary search included 695 articles. After removing 263 duplicated articles, 432 studies remained to be screened. Among those, 317 irrelevant papers were removed. We then excluded 77 studies according to the exclusion criteria. Finally, 38 articles were selected for this study. Conclusion: Out of 38 studies, 28 papers discussed Swarm-based algorithms, 2 papers studied Genetic Algorithms, and 8 papers covered algorithms in both categories. Considering the application domains, 21 of the articles focused on problems in the healthcare sector, while the rest mainly investigated issues in cybersecurity, text classification, and image processing. Hybridization with other BIAs was employed by approximately 18.5% of papers, and 13 out of 38 studies used S-shaped transfer functions. The majority of studies used supervised classification methods such as k-NN and SVM for building fitness functions. Accordingly, we conclude that future research should focus on applying bio-inspired feature selection to a diverse area of applications such as finance and social networks. And further exploration into enhancement techniques such as quantum representation, rough set theory, chaotic maps, and Lévy flight is necessary. Additionally, we suggest investigating other transfer functions besides S-shaped, such as V-shaped and X-shaped. Moreover, clustering and deep learning models for constructing fitness functions in bio-inspired feature selection algorithms need to be investigated further.

INDEX TERMS Bio-inspired optimization, feature selection, metaheuristics, systematic literature review, swarm intelligence.

I. INTRODUCTION

Advancement in the Information Technology (IT) has ushered in a new era of data analysis, where a vast amount of data is generated daily with high dimensionality in various fields of applications such as business intelligence, healthcare, social media, transportation, online education, government, marketing, financial. Extracting hidden patterns

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenzhou Tang .

or insights from such data remains a challenging task in machine learning and data mining models, as the existing data analytics techniques are not ideal for extracting the latent data insights and knowledge due to their inability to handle large-scale data with high dimensionality [1]. High-dimensional data are likely to consist of redundant and unrepresentative features that greatly increase memory storage needs, add additional processing time to analytics techniques, and thus negatively affect their performance and increase data processing expenses.

Dimensionality reduction is the key task and powerful technique for data pre-processing [2], which attempts to map space of high dimensionality into lower dimensionality without substantial loss of information. Dimensionality reduction methods include two types: Feature Extraction (FE) and Feature Selection (FS). A feature is a measurable property of the problem under observation. The FE methods tend to map the original features into lower dimensions with a new feature space; the newly added features generally are a combination of original features. On the other hand, FS tends to pick up a small significant subset of features from the original dataset by removing irrelevant, redundant, or noisy features based on a predefined evaluation measure [3]. Both methods aim at reducing the dimensionality in a large dataset, but the FE methods do this by generating new feature combinations, whereas FS chooses a feature subset from all the available features without any modifications. Therefore, the FS methods are often recommended for several domains and applications such as text mining and genetic analysis, where the readability and interpretability of the data are necessary because the meaning of the original features is kept in the reduced subset.

Bio-inspired computing optimization algorithms (BIA) is an emerging approach that is based on the principles and inspiration of the biological evolution of nature to develop new and robust competing techniques. The behavior of some insects or groups of animals in nature such as colonies of ants, flocks of birds, swarms of bees, and schools of fish, has attracted the attention of computer science researchers to solve several problems in science and engineering. Swarm intelligence is a subfield of artificial intelligence which is concerned with the intelligent behavior of biological swarms by the interaction of individuals in such environments to solve real-world problems by simulating such biological behaviors [4]. Bio-inspired algorithms applied for feature selection (BIA-FS) are a promising technique to deal with non-linear high-dimensional data. FS methods mainly include three categories that are the filter, wrapper, and embedded approaches [2]. The accuracies of the wrapper methods are better than the filter methods [5]. However, they may consume more computing resources. Moreover, the embedded methods are actually special cases of the wrapper methods since the feature selections are regarded as part of the training phase in the machine learning models [6]. This will be discussed in subsequent sections. Generally, feature selections can be regarded as optimization problems in which a subset of the original dataset is represented by a solution to the optimization problem, and these problems can be solved by exhaustive and heuristic search approaches [7]. Swarm intelligence algorithms are efficient metaheuristic search methods for wrapper-based feature selection problems [8].

A. SUMMARY OF PREVIOUS SURVEY WORKS

The current literature reveals that only a handful of review papers specifically evaluate BIAs in the context of feature

selection [7], [8], [9], [10], [11], [12], [13], [14]. A limited number of these reviews focus on a particular algorithm [10], [11] or a category of algorithms [7], [8], while others provide a broad analysis of various BIAs without limiting the scope to feature selection [9], [12], [13], [14].

Brezocnik et al. [7] summarized the advantages and application areas of 22 BIA-FS and 64 of their variants that can be divided into insect, bacteria, bird, mammal, fish, frog, and other algorithms. Xue et al. [8] discussed GA, PSO, ACO, GP, DEA, and memetic algorithms with their applications and open issues of BIA-FS in general. They also briefly analyzed other less popular methods, including ABC, variants of LCS, ES, and AIS. Shami et al. [11] detailed the principles, developments, and application areas of various PSO variants, both continuous and discrete search space. Applications in feature selection, wireless communication, image processing, and electrical power system were discussed. Deb et al. [10] analyzed variants of CSO and briefly discussed their application for feature selection. In addition, Almgren et al. [9] and Almazrua et al. [13] provided in-depth analysis of multiple BIAs hybridized with other statistical methods that were employed for feature selection. Del Ser et al. [12] comprehensively reviewed the broad landscape of BIAs and identified open challenges concerning theoretical perspective and high-level future directions of bio-inspired optimization.

To sum up, all the above survey works provide good references for BIA-FS, but these reviews are not comprehensive and in-depth. The survey work [7] merely introduced the principles, variants and applications for different BIOAs, with no analysis of datasets employed and performance comparison, which provides an important basis for the improvement of BIA-FS. The review in [8] only focused on six algorithms, which is narrow. Besides the above shortcoming, these problems also exist with no discussion about datasets, transfer functions, and performance comparison. Other papers either focused on one BIA with multiple application domains [10] or only one category of BIA-FS [9], [13], which did not particularly address common challenges of BIA-FS.

B. SCOPE OF DISCUSSION

The primary focus of this systematic literature review is on identifying the BIA techniques, application domains, data types, and datasets employed for feature selection. Furthermore, we aim to analyze their performance, highlight gaps, and uncover the trend of research in this area. To the best of our knowledge, no systematic literature review in BIA-FS investigates data types and datasets. This work focuses on swarm-based and evolutionary algorithms. Thus, physical-inspired methods such as Simulated Annealing and Harmony Search are excluded.

C. OUR CONTRIBUTIONS

This systematic literature review (SLR) will identify the most recent state-of-the-art bio-inspired algorithms applicable to feature selection and investigate their applications. This work will benefit future research in this area because it aims to

provide a big picture of BIA-FS and their corresponding applications in business, healthcare, engineering, and the like. Below are the main contributions of this SLR.

- We present a comprehensive and critical review of 21 bio-inspired algorithms employed for feature selection and make a preliminary analysis of the basic information for the chosen BIAs, which can provide a big picture for the BIA-FS study.
- We analyze and summarize the application domains, machine learning techniques for constructing fitness functions, data types, and datasets employed by the chosen BIAs to provide a clear insight into their applications in business and engineering.
- We compare and analyze the accuracy and efficiency in terms of feature reduction.
- Most importantly, we highlight the most common techniques employed to improve BIA-FS performance. To the best of our knowledge, this SLR is the first to analyze and categorize improvement techniques.
- We discuss the challenges and future directions of the BIA-FS field, which can provide a referencing framework for future research.

II. BASIC CONCEPTS

A. DIMENSIONALITY REDUCTION

The curse of dimensionality refers to a set of problems that arise when working with high-dimensional data. It describes the explosive nature of increasing data dimensions, leading to an exponential increase in computational efforts to analyze such data. This term was first introduced by Richard Bellman [15] to explain the increase in the volume of Euclidean space associated with adding extra dimensions. As the dimensionality increases, the number of data points required for good performance of any machine learning algorithm increases exponentially. Dimensionality reduction (DR) is a technique for transforming the high-dimensional representation of data into lower-dimensional representations. Specifically, DR transforms the original dataset having high dimensionality and converts it into a new dataset representing low dimensionality while preserving the original meanings of the data as much as possible [16]. The low dimensional representation of the original data helps to overcome the issue of the curse of dimensionality [17]. The low-dimensional data can be easily processed, analyzed, and visualized.

DR techniques transform dataset X with dimensionality D into a new dataset Y with dimensionality d , while retaining the information of the data as much as possible. The ideal case is $d \ll D$. There are two major categories of DR: Linear and Non-linear. Linear techniques assume that the data lie on or near a linear subspace of the high-dimensional space. Nonlinear techniques for dimensionality reduction do not rely on the linearity assumption, resulting in more complex embedding of the data in the high-dimensional space that can be identified [18].

- **Linear techniques:** perform dimensionality reduction by projecting X into a linear subspace of lower

dimensionality. Principal Component Analysis (PCA) [19] is by far the most popular linear technique. PCA constructs a low-dimensional representation of X , which is Y , that contains as much of the variance of X as possible. In mathematical terms, PCA attempts to find a linear mapping M that maximizes $M^T \text{cov}(X)M$, where $\text{cov}(X)$ is the covariance matrix of X [18]. M is found by solving the equation $\text{cov}(X)M = \lambda M$. The low-dimensional data representations Y are computed by mapping them using $M : Y = (X - \bar{X})M$. The main drawback of PCA is that the size of the covariance matrix increases exponentially as the dimension of data increases. As a result, the computation of M might be infeasible for very high-dimensional data.

- **Non-linear techniques:** over the past few decades, a variety of non-linear DR techniques (NLDR) have been developed to work with applications having complex nonlinear structures [20]. To model relations present in the data in a nonlinear manner, kernel methods, kernel methods also known as “Kernel Trick” can be used [21]. The kernel trick avoids explicit mapping to learn a nonlinear function. One of the most discussed methods is Kernel PCA (KPCA). KPCA is the reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function [22]. KPCA transforms the input data X from the original input space to kernel space for each data point using a non-linear transformation. The inner product of new feature vectors is used to form a kernel matrix K . Then, PCA is used on the centralized K to estimate the covariance matrix of the new feature vectors. An example of constructing a Kernel covariance matrix using radial basis kernel function is computed as: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2b^2)$, where b is the bandwidth of the kernel.

B. THE FUNDAMENTALS OF BIO-INSPIRED ALGORITHMS

1) COMMON FRAMEWORK

Despite being inspired by different nature phenomena, all BIAs in this review follows the same high-level framework as in Figure 1.

In step S1, the initial population and other parameters such as the number of iterations, termination threshold, etc. are initialized. Usually, the initial population is generated randomly, aiming to cover as many regions in the solution space as possible. This parameter together with its generating method plays a crucial role in BIA performance, and its values depend heavily on optimization problems. There are no mathematical formulas for finding these numbers rather than iterations of trial and error. Most BIAs use iterative methods, and the maximum iteration times and precision threshold are two common conditions of algorithm termination, which should also be initialized in step S1 [23].

The fitness value is a measure to evaluate how good individual solutions perform. For example, the output of each

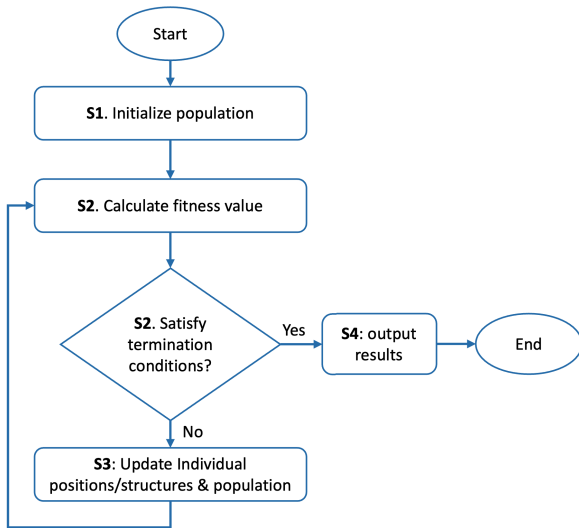


FIGURE 1. Common framework of bio-inspired algorithms.

individual is used as input to a classifier, and the accuracy of the classifier is used as the fitness value for the corresponding solution. In step S2, the fitness values of the population in each iteration are computed, and if the global best solution satisfies the termination conditions, the algorithm stops (in step S4); otherwise, every individual updates their positions or structures, which is step S3. Then the workflow jumps to step S2 to execute the next iteration.

2) CATEGORIES OF FEATURE SELECTION METHODS

Regarding strategies for FS, Bolón-Canedo et al. [24] broadly categorized into three approaches: filters, wrappers and embedded methods.

Filters work based on the characteristics of the features, where relevant features are retained and irrelevant (redundant) features are excluded from the datasets. A typical approach to measuring features' relevancy is mutual information [25]. Many mutual information feature selection methods have been proposed in the last 25 years [26]. Therefore, filter methods are independent of any learning algorithms. As a result, filter methods are more computationally efficient than wrapper methods. However, due to the lack of a specific learning algorithm guiding the feature selection phase, the selected features may not be optimal for the target learning algorithms [27].

Wrapper methods rely on machine learning models to evaluate the quality of selected features. A high-level framework of wrapper methods consists of two phases: (1) search for a subset of features; and (2) evaluate the selected features. It repeats (1) and (2) until some stopping criteria are satisfied. Firstly, different subsets of features are determined; then chosen learning algorithms evaluate the quality of these features based on the learning performance. This process repeats until such as the highest learning performance is achieved or the desired number of selected features is obtained. Thanks to

this unique characteristic, wrapper methods usually produce higher quality subsets (more relevant features), but of course more costly in computational terms than filters. However, one downside of wrapper methods is that exhaustive search can become computationally intensive for large datasets. For example, a search space for d features is 2^d , which is impractical when d is huge. Therefore simplified algorithms such as sequential search or evolutionary algorithms such as Genetic Algorithm (GA) or Particle Swarm Optimization (PSO) which yield local optimum results are employed which can produce good results and are computationally feasible [28].

Embedded methods are a trade-off between filter and wrapper methods which embed the feature selection into the model learning [27]. Thus, they inherit the merits of wrapper and filter methods - (1) they include the interactions with the learning algorithm; and (2) they are far more efficient than the wrapper methods since they do not need to evaluate feature sets iteratively.

3) CHARACTERISTICS OF BIO-INSPIRED ALGORITHMS

The two most discussed BIA categories are Evolutionary Algorithms (EA) and Swarm-based Algorithms (SWA), inspired by natural evolution and animals' collective behavior, respectively. According to Wang et al. [23], Genetic algorithms (GA) and Particle Swarm Optimization (PSO) are the most discussed BIAs, measured by the total number of published papers per year until October 2020.

In 1995, Kennedy and Eberhart introduced the PSO method for continuous optimization problems [29]. Each candidate solution is represented by a particle, and each particle has two main properties: position and velocity. For an optimization problem with n dimensions, the i^{th} particle moves with a certain velocity v_i , and the position of the particle is expressed as x_i . A solution in PSO algorithm can be written as: $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n})$ and $v_i = (v_{i,1}, v_{i,2}, \dots, v_{i,n})$ stands for its velocity. In each iteration, the position and velocity of each particle are updated based on three forces: its own inertia, its best position in previous iterations, and the swarm's best particle. Weights are used to control the influence of these forces, which balance the exploration and exploitation of the swarm. Exploration refers to searching the unexplored area of the solution space, while exploitation refers to the search in the neighborhood of a promising region. The main advantages of PSO include simple implementation and fewer controlling parameters compared to other BIAs. One of the major performance problems of PSO is premature convergence, as pointed out in [30]. This problem occurs due to the lack of population diversity, especially in complex multimodal functions [30]. Although only three parameters in PSO, it is difficult to control them and find their appropriate setting at each iteration.

Genetic algorithms are based on natural selection. The process of genetic algorithms consists of four operations: initialization, selection, crossover, and mutation. Initialization

involves building a population of potential chromosomes (solutions) by random creation or some other methods. The selection process picks high-quality chromosomes as parents, which maintains the exploration of the population. Crossover is the process of generating new chromosomes by combining aspects from previous solutions chosen by the selection algorithm via one of several possible crossover algorithms (single-point, multi-point, uniform, partially mapped crossover, etc.) in the hope of producing a “child” chromosome fitter than either of its “parent” chromosomes. The mutation introduces a small random modification of the children solutions to make new solutions different from their parents and hopefully better. It is used to maintain exploitation as the mutation aims to search in the local area. GA has several strengths and weaknesses. It is able to search for the optimal solution in a very large search space, and it does not require gradient information. Due to its parallel nature, GA offers a large variety of options for acceleration across multiple cores or machines and can take advantage of large amounts of computing power where it is available [31]. However, GA does not guarantee finding the global optima. It can be computationally expensive and take a long time to converge, as it is often necessary to have a good-sized population and a large number of generations to achieve a good result [32].

In general, BIAs are flexible and can be applied to a wide range of optimization problems, including problems that are highly nonlinear, non-convex, or have multiple objectives. Since they are gradient-free, they can find better solutions faster and are computationally cheaper than gradient-based methods. However, BIAs do not guarantee convergence to a global optimum. The key aspect for BIAs to perform well is to balance the exploration and exploitation phases, which typically requires parameter tuning to achieve optimal performance. This can be time-consuming and requires expertise in the algorithm being used.

4) TRANSFER FUNCTION

Selecting optimal subsets of features from a larger pool of features is a discrete optimization problem. Originally, most BIAs was designed to address continuous optimization issues. Therefore, adapting continuous BIAs into discrete domains requires several modifications such as normalizing, rounding, and utilizing binary operators [33]. The transfer function can also be used to convert continuous components into binary values. Due to the binary nature of feature selection, transfer functions are efficient yet easy ways to limit the result such that 0 means the feature is redundant and not chosen, and 1 represents the feature is useful and chosen [34]. Transfer functions based on their shape have been divided into three groups S-shaped, V-shaped, and U-shaped transfer functions [35], [36], and [37].

In the original version of BIAs, individuals can move around the search space utilizing position vectors within the continuous real domain. Consequently, the concept of position updating can be easily implemented for individuals by

adding velocities to positions [37]. However, the meaning of position updating is different in a discrete binary space [38]. There are only two values, 0 and 1, for binary space, hence updating position using the aforementioned approach is not practical. Therefore, a novel approach must be applied to convert velocities for changing positions from 0 to 1 or vice versa. For example, a sigmoid function can be used to transform all real values of velocities to probability values in the interval [0,1] as in Eq. (1).

$$T(v_i) = \frac{1}{1 + e^{-v_i}} \quad (1)$$

where v_i indicates the velocity of individual i^{th} . After converting velocities to probability values, position vectors could be updated with the probability of their velocities as in Eq. (2).

$$x_i = \begin{cases} 0, & \text{if } threshold < T(v_i) \\ 1, & \text{if } threshold \geq T(v_i) \end{cases} \quad (2)$$

Examples of several S-shaped and V-shaped functions with their graphs are shown in Figure 2 [37].

5) HYBRID BIO-INSPIRED ALGORITHMS

Hybridization is a strategy for improving the performance of BIAs. The basic principle of this mechanism is to use the advantages of some methods to overcome the disadvantages of the BIAs of interest. These methods can be other BIAs such as Zahara and Kao [39] hybridize GA and PSO for the global optimization of multimodal functions. This hybrid technique incorporates concepts from GA and PSO to create individuals in a new generation not only by crossover and mutation operations as found in GA but also by mechanisms of PSO. In addition, some other methods are also employed to improve the performance of BIAs, such as the Lévy flight function [40], [41] to improve local search as well as exploration stage. Another method is the chaotic maps [42], [43], which has been widely utilized throughout the initialization period of the optimizers due to the following characteristics: closeness to the initial condition, semi-randomness, and ergodicity. However, in this review, we do not consider the hybridization between a BIA and a non-BIA method a hybrid BIA, but a different category of improvement techniques. This will be discussed in subsequent sections.

III. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

This section introduces the methodology employed to select and review the research related to bio-inspired algorithms for Feature Selection techniques and their application domains. A systematic literature review was undertaken based on the guidelines of Kitchenham and Charters [44]. Accordingly, we followed the following steps to develop our SLR protocol.

A. RESEARCH QUESTIONS

This study attempts to answer the following questions:

- **RQ1:** What bio-inspired optimization algorithms are employed in feature selection? What are their domains of application?

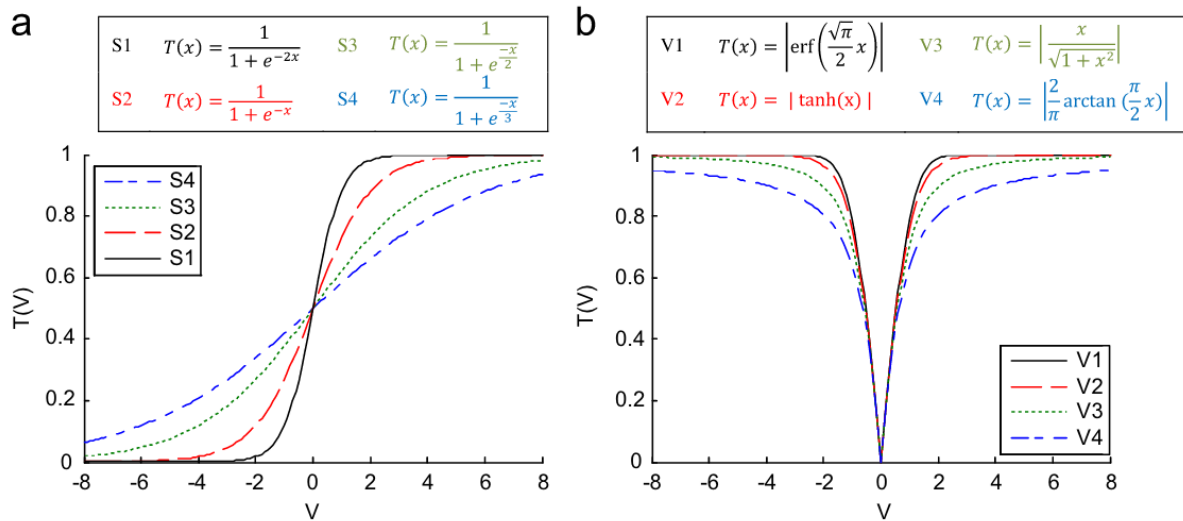


FIGURE 2. (a) s-shaped and (b) v-shaped family of transfer functions.

- **RQ2:** How effective are bio-inspired feature selection algorithms?
- **RQ3:** What are the challenges and research gaps in existing bio-inspired algorithms for feature selection?

The first question helps us to compile and analyze the state-of-the-art bio-inspired algorithms used to reduce the number features and in which application. The second question aims to assess the effectiveness of those methods. The last one guides us to identify the present studies’ challenges, which hints at the future research direction in this domain.

B. SEARCH STRATEGY

Identifying correct and sufficient keywords is a crucial task in developing an SLR. We initially ran search queries using fundamental keywords on Google to identify the major papers in this area. The purpose of these queries was to extract the main keywords and key databases from those papers. The next stage of searching is selecting the main keywords from the above papers and running search queries on major databases. This search aims to find exemplars on this topic within the article title, abstract, and keywords. Once exemplar articles are located, another screening step is conducted to determine the final list of keywords. Lastly, we ran the queries using the above keywords on five different databases from the University of Ottawa library databases.

The query results were exported to .RIS files, and then imported to Covidence for de-duplication. The abstracts were initially screened in Covidence to exclude all papers that did not meet the inclusion criteria. The results were exported back from Covidence to Zotero to extract the papers for a full-text review.

C. SEARCH QUERY

The query consists of 2 components. The first part captures “bio-inspired” related terms, and the second one captures “feature selection” associated terms.

(“bio-inspired” OR “bioinspired” OR “biologically inspired” OR “nature-inspired” OR “nature inspired” OR “naturally inspired”)

AND

(“feature selection” OR “feature reduction” OR “dimensionality reduction”)

D. DATA SOURCE

We searched the above search string in five major digital libraries, which are ACM Digital Library, IEEE Xplore, Scopus, Web of Science, and ScienceDirect. The search string is modified and translated to the proper input query for searching each digital library. We only focused on peer-reviewed journals and conference articles and excluded the book chapters and other types of publications. The search was conducted on February 21st, 2023, with the publication year restricted between 2020 and 2023. Table 1 shows the search results by database.

E. INCLUSION AND EXCLUSION CRITERIA

The retrieved articles from the digital libraries were excluded based on the exclusion criteria. Table 2 presents both inclusion and exclusion criteria.

The paper selection steps are represented in the PRISMA diagram [45] as in Figure 3.

- Duplicate removal: Our primary search included 695 articles. After removing the 263 duplicated articles, 432 studies remained to be screened.
- Title and abstract screening: Articles are screened out by filtering titles, abstracts, and keywords based on the inclusion and exclusion criteria, and 317 irrelevant papers are removed.
- Full-text screening: From the 115 articles that were left, we excluded 77 studies according to the exclusion criteria.

TABLE 1. The number of retrieved articles.

Database	Website	Search in	Search results
Scopus	https://www.scopus.com/	Title, Abstract and Key Words	241
IEEE Xplore	http://ieeexplore.ieee.org/	Title, Abstract and Key Words	118
ScienceDirect	http://www.sciencedirect.com/	Title, Abstract and Key Words	51
Web of Science	https://webofknowledge.com/	Topic	196
ACM Digital Library	http://dl.acm.org/	All Fields	89
Total Number of Retrieved Articles			695

TABLE 2. Inclusion & exclusion criteria.

Inclusion Criteria	Exclusion Criteria
<p>I1: Peer-reviewed conference proceedings and journal publications.</p> <p>I2: Only English articles.</p>	<p>E1: Papers that do not include applications in any domains (business, engineering, healthcare, etc.)</p> <p>E2: Papers that do not include any technical improvements of the bio-inspired method.</p> <p>E3: Papers about bio-inspired algorithms that do not focus on feature selection or dimensionality reduction.</p> <p>E4: Papers that only review other studies.</p>



FIGURE 3. PRISMA diagram.

F. QUALITY ASSESSMENT

As part of the review protocol stage, it is important to assess the quality of the primary and the final selected studies used in this SLR. We have based the quality assessment of our primary selected studies on the following criteria:

- Is the paper published in non-predatory publishers?
- Does the study have a clear and sound research methodology?

G. DATA EXTRACTION

This step entails deriving relevant data and information from the selected papers. Table 3 shows our data extraction form. The extracted data is then analyzed to answer our research questions. In the last column of Table 3, we specified

the research question that the corresponding data helps answer. We used the details about algorithms and datasets to answer RQ1. We analyzed this information to group similar studies together. Extracting the objective, contribution and conclusion of each study will help us recognize the key improvements and trends of the works, analyze the gaps, and determine future research (RQ2). Therefore, we summarized the articles according to their goals and conclusions to find the gaps and recognize the direction of future research.

IV. DATA SYNTHESIS AND RESULTS

In this section, we investigate all final selected articles (38 articles). The data is synthesized to address the three mentioned research questions.

RQ1: What bio-inspired optimization algorithms are employed in feature selection? What are their domains of application?

A. BIO-INSPIRED ALGORITHMS

Generally, two categories of bio-inspired algorithms are studied for feature selection: Swarm-based Algorithms (SWA) and Evolutionary Algorithms (EA).

- SWA methods are inspired by the natural and artificial systems composed of many individuals that coordinate using decentralized control and self-organization [46]. In particular, the discipline focuses on the collective behaviors resulting from the local interactions of the individuals with each other and their environment. For example, colonies of ants and termites, schools of fish, flocks of birds, herds of land animals, etc.
- EA uses mechanisms inspired by biological evolution, which include reproduction, mutation, recombination,

TABLE 3. Data extraction form.

No.	Information	Description	Target Questions
1	Publication Metadata	Authors, title, abstract, publication year and publication type and venue of the studies.	Meta-analysis and supplementary information.
2	Algorithm name	Name of the BIA-FS.	RQ1
3	Algorithm category	Category of the BIA.	RQ1
4	Fitness function	The function used to evaluate the quality of individual candidate solutions.	RQ1
5	Transfer function	The function to convert individual positions from continuous-value search space to binary search space.	RQ1
6	Hybrid mode	Whether the BIA is hybridized with other BIAs.	RQ1
7	Application domains	The area(s) and the dataset that the algorithms applied to.	RQ1
8	Specific applications	More detailed information on applications such as network intrusion detection, chronic liver disease prediction, etc.	RQ1
9	Data structure	Datasets are structured or unstructured.	RQ1
10	Data type	Specific domains of the datasets (microarray data, medical dataset, image, etc.)	RQ1
11	Number of features	The number of features (dimensions) of the original datasets.	RQ2
12	Number of features after FS	The number of features (dimensions) of the output subsets.	RQ2
13	Machine learning type	Types of machine learning models employed in evaluating BIA-FS performance such as classification, clustering, etc.	RQ2
14	Machine learning technique	Specific technique such as k-means, SVM, etc.	RQ2
15	Performance	The improvement in accuracy, running time, and storage usage, compared to other methods or baseline techniques.	RQ2
16	Key improvements	Changes or modifications from the original version of the BIA being investigated. If a novel BIA is being proposed, improvements over other BIAs are highlighted.	RQ1, RQ3
17	Future directions	Future directions, trends, and gaps.	RQ3

and selection. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function determines the quality of the solutions. The population's evolution occurs after the repeated application of the above operators.

Figure 4 shows more details regarding the categories of algorithms and the frequency of each technique applied in the reviewed articles.

Figure 5 illustrates that SWAs were employed over three times as many as EAs, 78% versus 22%, for Feature Selection tasks.

SWA algorithms are based on behaviors and interactions animals have while they are searching for food and mates. Also, they have several characteristics such as adaptation, scalability, speed, autonomy, parallelism and fault tolerance. The key characteristics of this family of methods are self-organization and working division. As per animals' and birds' biological behavior, each in the group is responsible for a specific task individually, and sometimes they work together to achieve a given task.

In Feature Selection, the objective is to find an optimal subset of features that minimizes a cost (fitness) function. This function is also called the optimization function. The optimization objective is usually to minimize the difference in the variance of data points, before and after applying Feature Selection. An exhaustive search for the optimal subset is computationally expensive and virtually unattainable for high-dimensional data. Heuristics refers to experience-based techniques for problem-solving and learning. It gives a satisfactory solution in a reasonable amount of computational time, which may not be optimal. Specific heuristics are problem-dependent and designed only to solve a particular problem. Examples of this method include using a rule of thumb, an educated guess, an intuitive judgment, or even common sense [47].

The term metaheuristic was coined by Glover in 1986 [48] to refer to a set of methodologies conceptually ranked above heuristics in the sense that they guide the design of heuristics. A metaheuristic is a higher-level procedure or heuristic designed to find, generate, or select a lower-level procedure or

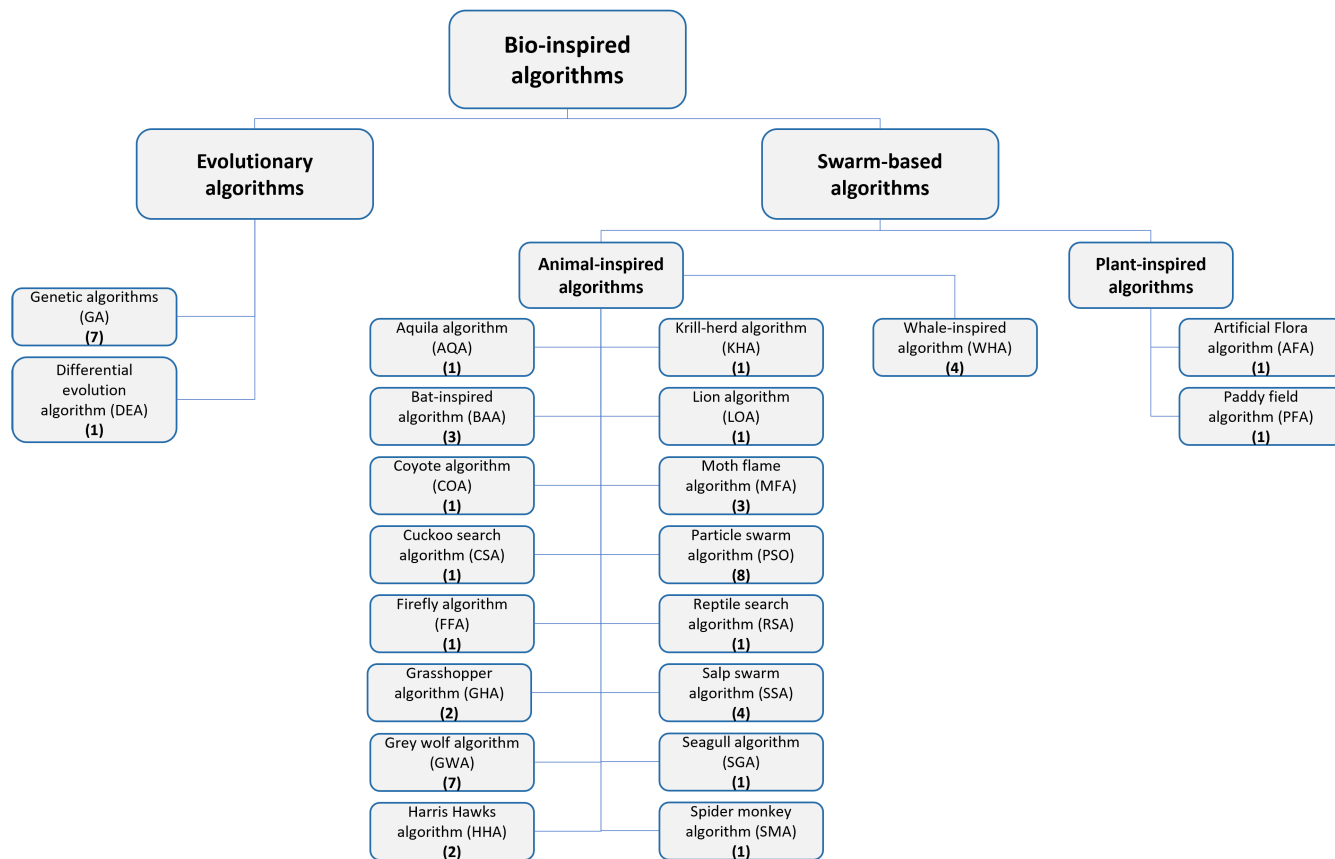


FIGURE 4. A taxonomy of bio-inspired algorithms (with frequency of occurrence).

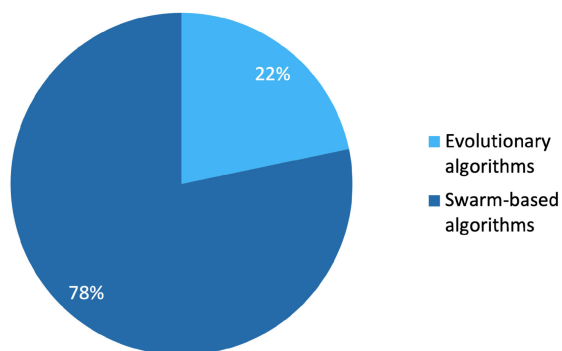


FIGURE 5. Category of bio-inspired feature selection algorithms.

heuristic (partial search algorithm) that may provide a sufficiently good solution to an optimization problem. By searching over a large set of feasible solutions, metaheuristics can often find good solutions with less computational effort than calculus-based methods or simple heuristics can. Metaheuristic optimization algorithms are becoming increasingly well-known in different applications because of their nature:

- Based on simple ideas to be easy for implementation.

- Can find optimal neighborhood solution.
- Can be applied in different areas of applications.

Figure 6 provides the detailed composition of each category. For the EA category, only GA and DEA were used. In contrast, a wide variety of SWAs (19 methods) was investigated to optimize Feature Selection models. This is interesting since the most frequently applied EA method, GA, in Feature Selection was also one of the most used techniques in other fields, which is in line with the findings in this study [4].

The family of GA techniques accounted for up to 18% (seven papers) of bio-inspired techniques for Feature Selection. There is only one research [49] that used the original version of GA, while the rest six papers [50], [51], [52], [53], [54], [55] either improved various aspects of the original GA or hybridized GA with other BIAs to exploit the strengths and overcome weaknesses of the two techniques. For example, Tahir et al. [55] applied various chaotic maps to improve the initialization and mutation phase of GA. They provide a fast convergence rate and are used to avoid the local minima, which are weaknesses of most of metaheuristics search algorithms. Chaotic maps were also used to improve DEA [43], which is in the EA family as well. The main reason why chaotic maps are employed is that initialization largely impacts BIAs performance, and a well-distributed

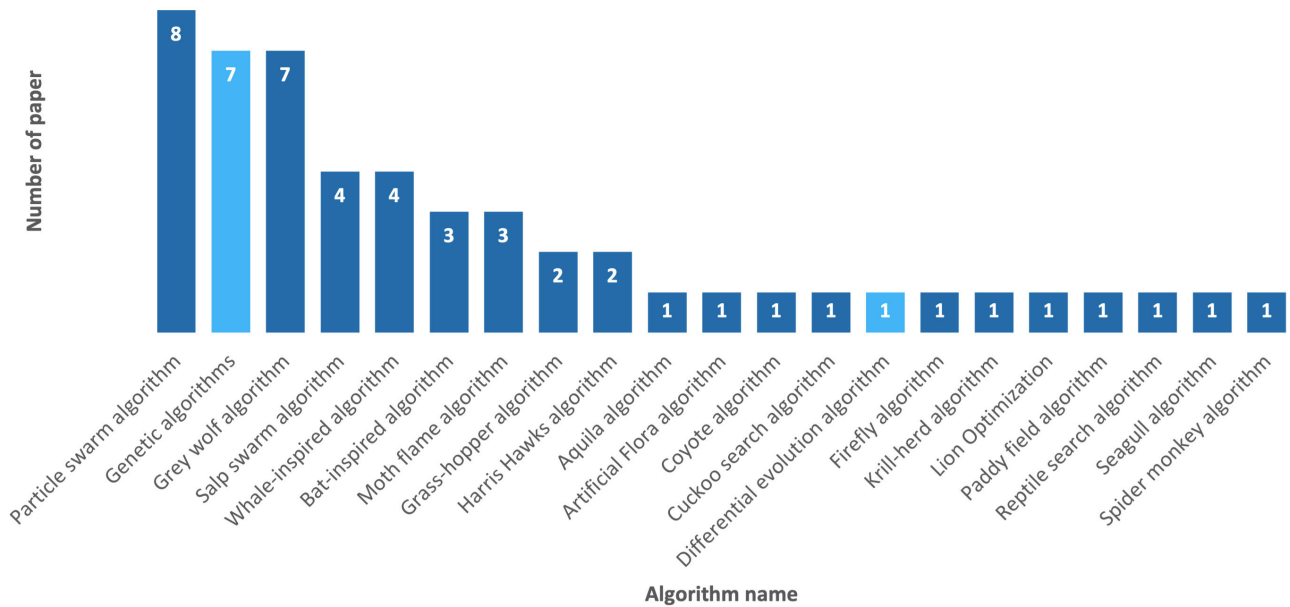


FIGURE 6. Number of research by algorithm.

initial population is always useful [56]. The chaotic map has been widely utilized throughout the initialization period of the optimizers due to the following characteristics: closeness to the initial condition, semi-randomness, and ergodicity [43]. Another notable application of GA is hybridization, where its unique feature, crossover and mutation were employed to improve three different BIAs [50], [51], [52]. Similarly, Zhang et al. [43] integrated the mutation mechanism of DA into SSA, aiming to prevent SSA from dropping into premature convergence. All authors of those research reported a better performance in terms of classification accuracy, the number of features reduced, and processing time compared with the original BIAs.

Regarding SWAs, PSO and GWA were the most common algorithms that appeared in eight and seven studies, respectively. GWA is another powerful population-based bio-inspired optimization technique that solves an optimization problem via performing leadership and hunting behavior of grey wolves mathematically [57]. Each wolf in the population represents a candidate solution for the problem in the search space. Afterward, three processes, including searching for prey, encircling prey, and attacking prey, are carried out based on different kinds of wolves in the search space to hunt prey (discovering the best solution). The encircling and attacking prey processes are repeated until a termination criterion is satisfied. There are four kinds of wolves i.e., alpha (α), beta (β), delta (δ) and omega (ω). The hunting processes are guided by alpha, beta, delta wolves and omega wolves following these three candidates. This is because α , β , and δ represent the fittest wolves in the whole population,

which are updated with the best new solutions after each iteration. The influence of the top 3 candidates on the whole population leads to strong exploitation (local search) and fast convergence speed. On the other hand, the lack of information sharing among individuals results in weak exploration (global search), low diversity and consequently faces premature convergence [57]. To overcome this drawback, researchers utilized different techniques. Alzaqebah et al. [58] implemented an initialization phase based on Information Gain (IG), a feature with a high IG value means it is significant for classifying the instance instead of random initialization. Similarly, Zenboud et al. [59] modified initialization phase using the k-means clustering method. Features are grouped into clusters. Each cluster represents an initial individual. The distance between each feature and its cluster centroid is used to determine whether it is selected. In comparison, Preeti and Deep [60] used random walks in leader wolves to spread the population in the whole space uniformly instead of sticking around a local optimum. This change speeds up the algorithm's convergence as well as exploration capability. Another strategy is observed in [61], Rough Set theory is integrated into GWA to enhance the potential to discover a minimal subset of features. They observed a higher classification accuracy over baseline GWA.

PSO was demonstrated as an effective method for optimization problems since it is powerful while easy to be implemented [62]. However, conventional PSO algorithms may have certain drawbacks, such as a lack of exploration, which leads to the possibility of falling into local optima [40]. Several enhancement approaches for PSO observed in the

reviewed literature include adapting Lévy flights function to improve local search [40]; hybridization with other BIAs to increase population diversity [40], [53]; cooperative learning within swarm intelligence [63]. Lévy flight is a random walk method that follows a heavy-tailed distribution. In this method, the short-distance and occasional long-distance searching appear alternately, such that expands the search scope and enhances the local search performance [40]. Hybridization seems to be on the rise when it comes to BIAs. Ji et al. [40] introduced a mutation mechanism (inspired by GA) to mutate some of the solutions and the corresponding P_{best} values (local best solutions) of them in the population, thereby enhancing the development efficiency while ensuring the population diversity. This modification balances the influence of exploration (global search guided by global best solutions G_{best}) and exploitation (local search guided by local best solutions P_{best}). Martarelli and Nagano [53] hybridized WHA with an improved version of PSO called Adaptive PSO [64], which uses additional adaptive parameters to improve the algorithm's convergence speed and achieve a balance between exploitation and exploration of the search space. Sarhani and Voß [63] followed a unique approach, which is worth noting as the authors claimed that an actual trend in bio-inspired optimization is to re-iterate the existing knowledge in a different form, so they aim to fill this gap. Instead of having one swarm (of n particles) trying to find the optimal d -dimensional vector, the vector is split into clusters of features that we can consider independent of the others. In other words, a solution vector of the selected features is a combination of the different solutions provided by each swarm according to the same principle of the cooperative PSO [65].

Employing multi-objective fitness functions is another approach for solving FS problem. Multi-objective optimization involves optimizing two or more (normally conflicting) objective functions simultaneously, and it frequently arises in many application domains, such as business and engineering [66], [67]. Feature selection can be considered a multi-objective optimization problem since its two main objectives: maximizing the classification accuracy and minimizing the number of selected features, are likely to conflict. In multi-objective feature selection, an archive is a set of non-dominated solutions obtained during the optimization process. Non-dominated solutions are those that are not inferior to any other solutions in all objectives. The archive keeps track of the best solutions found so far and provides a diverse set of solutions to choose from rather than just a single optimal solution. The archive can be updated as new solutions are found, and the selection of solutions from the archive can be based on various criteria, such as diversity, distance, or preference of the decision-maker. Three articles discussed this approach, with PSO and HHA appearing two and one times, respectively. Han et al. [68] introduced two modifications to PSO. Firstly, an adaptive penalty mechanism is incorporated into the archive updating mechanism to maintain the diversity

of the archive. Secondly, a novel adaptive leading particle selection based on feature information combines feature frequencies and the opposite mutation to enhance the diversity of the swarm. Feature information reflects current search information of the archive, thus, incorporating feature frequencies and opposite mutation into the leader particle selection can avoid the duplicated search of known space and change the selection pressure of particles adaptively. Zhou et al. [69] proposed a binary PSO with a two-level particle cooperation strategy. In the first level, randomly generated ordinary particles and strict particles filtered by ReliefF [70] are combined as the initialized particles to maintain rapid convergence. In the second level, under the decomposition multi-objective optimization framework MOEA/D [71], cooperation between particles is conducted to search for Pareto solutions more efficiently during the update process. Dabba et al. [72] presented a study on multi-objective binary HHA for gene selection, integrating several filter-based ranking methods to filter out redundant features before applying feature selection.

Other less popular SWAs, including SSA and WHA, both were studied in four articles; BAA, GHA, and MFA each appeared in three papers. Balasubramanian and N.P. [73] attempted BIA-FS in a collaborative strategy. Implementing three BIAs (GWA, SSA, and LOA) independently, FS based on the correlation between each of the features selected from the BIAs are calculated to find optimal features from the three feature sets. Sahlol et al. [74] applied two-layer dimensionality reduction architecture: (1) VGGNet [75] to extract features from medical images, (2) SSA to select the most relevant features. The SSA selected only 1,000 out of 25,000 features extracted with VGGNet, while improving accuracy simultaneously. A similar design was also observed in [76]; however, both layers employed a different BIA. Integrating quantum mechanisms into BIAs is also a novel approach with quantum WHA [55], [77] and quantum MFA [78]. This is based on quantum computing principles and takes advantage of a probabilistic representation of the Q-bits and enhances the population diversity. Instead of binary representation, Q-bit is used as a probabilistic representation, defined as the smallest unit of information. It should be noted that although this approach is based on the concept of quantum computing, it is not a quantum algorithm but a novel evolutionary algorithm for a classical computer [79].

Regarding the rest of the BIAs, each of them appeared in only one study. Lévy flight and chaotic maps were the prominent techniques to improve the original BIAs. Ewees et al. [41] applied Lévy distribution to improve the walk of the original SGA method and thus improve the exploration stage. In another research [42], the chaotic maps are used in the initialization phase of RSA to improve its solutions diversity.

Figure 7 reveals the techniques used to enhance BIAs' performance over their original versions. Thirty-three articles (87%) modified the original version of the interested BIAs before applying them for FS, and only five studies did not

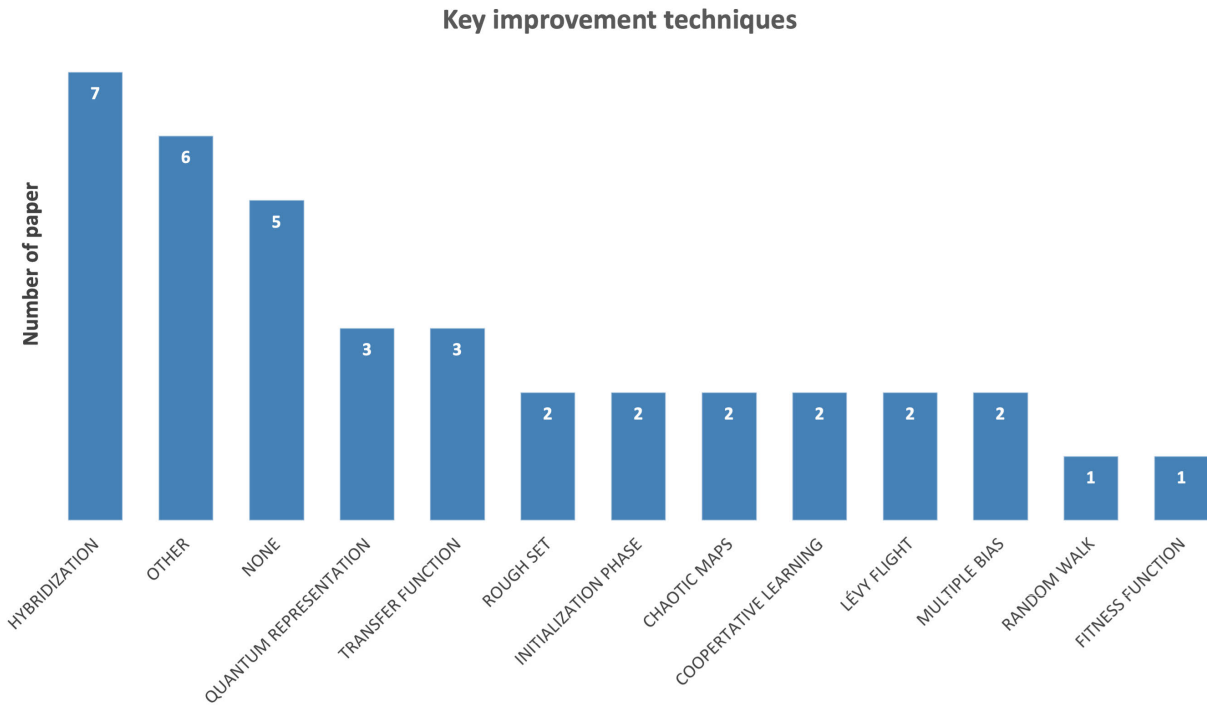


FIGURE 7. Category of improvement techniques for BIAs.

introduce any enhancement, which is the “None” category. Hybridization between BIAs dominates the list with seven articles (18.5%). Other notable strategies include replacing random steps in BIAs with a different distribution such as “Lóvy flight”, “Chaotic maps”, etc. Employing quantum representation instead of binary representation is also an interesting approach, which is a potential future research direction.

We decided to delve into the “Hybridization” category for more insight. Figure 8 shows the proportion of BIAs used for improving other BIAs. We classified research as hybridization when the authors borrowed ideas from BIA(s) to improve other baseline BIAs. For example, Ji et al. [40] adapted the mutation mechanism from GA to increase the population diversity of PSO. Thus, PSO is the base algorithm and GA was employed to enhance PSO. In this review, papers using techniques that are not BIAs to boost the base BIAs’ performance are not considered hybridization. Ewees et al. [52] integrated two algorithms, GA and DEA, into GHA, thus, this research is counted for the two methods. GA tops the list with four articles, helping EA surpass SWA, five over three. Thom de Souza et al. [80] claimed that COA is a population-based algorithm classified as both swarm intelligence and evolutionary heuristics. Thus, the original version of COA is already a hybrid algorithm. As a result, it falls into the “None” category.

B. APPLICATION DOMAINS

A list of the BIAs with their corresponding application domains employed in the reviewed articles is shown in

BIAs employed for Hybridization

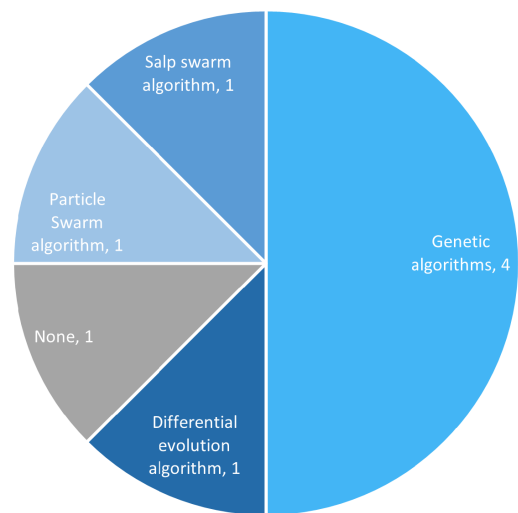


FIGURE 8. The BIAs employed for hybridization.

Table 4. Interestingly, real-life healthcare problems attracted a remarkable amount of research. Figure 9 shows that 55% of bio-inspired feature selection studies (21 out of 38) in this SLR belong to this domain. 29% of papers (eleven) proposed different BIAs, either enhanced variants over the original versions or novel BIAs, and generally assessed their performance without a specific real-life application. Other areas

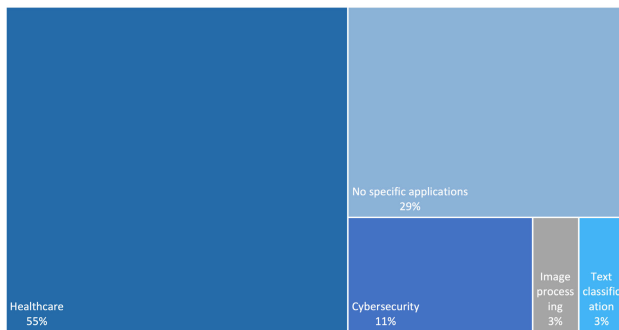


FIGURE 9. BIA-FS application domains.

that showed much less attention from researchers are Cybersecurity with three studies, Text classification and Image processing, each has one paper.

Regarding Healthcare, a majority of articles worked on disease classification such as Cancer [55], [61], [72], [78], [83], [84], [85]; Parkinson disease [54], [82]; Leukemia [74]; COVID-19 [34]; Multiple chronic diseases [60]. GWA and PSO are the most commonly used BIAs in this domain, each with four papers. Chronic disease classification, especially cancer, is a repetitive task that needs the greatest consideration to prevent misclassification. A huge amount of data in a wide array of forms (medical images, voice, historical medication, etc.) is usually required for this task. However, human interpretation of the data heavily relies on training and involvement. As a result, a computer-aided diagnosis (CAD) system is a prevalent and effective technique that assists doctors in interpreting medical images. For example, the CAD framework generally involves two fundamental stages in breast cancer detection: image acquisition and tumor detection. Initially, the breast X-ray image is acquired with digital mammography. Moreover, the detection and characterization of tumors in a mammogram image can be accomplished through various phases: extracting the cancer area, computing certain features representing every extracted cancer region, and classifying these features to distinguish the mammogram images [61]. Medical data is a prolific source of high-dimensional data. This is understandable due to several reasons.

- Medical data is a valuable resource that can be used for many purposes including managing and planning for future health needs and clinical research. However, the heterogeneity and complexity of medical data can be an obstacle in applying data mining techniques. A patient's electronic health records contain imaging data, speech samples, clinical variables, information about activity levels and vital signs from wearables, genomic data, and other data streams. This leads to a high-dimensional and potentially rich representation of the patient's health.
- Due to the non-linear nature of healthcare data, which is attributed by the combination of various sources and data types, traditional feature reduction techniques such

as PCA [93], ICA [94], LDA [95], SVD [96], etc. are not proper approaches. This is because these statistical methods linearly project dimensions from the original space onto a lower dimensional space; therefore they strongly rely on the distribution of data (most of these methods require Gaussian distribution of the input data). This is where bio-inspired feature selection methods have an advantage over projection-based algorithms. Since they use metaheuristic optimization functions, they can cope with the non-linearity of the data and be less computationally expensive.

Another domain that BIA-FS shows high potential application is Cybersecurity. Moizuddin and Jose [88], Alzaqebah et al. [58], and Davahli et al. [50] proposed three variants of GWA for the problem of network intrusion detection. Security of client data, the privacy of clients, intrusion detection and protection against intentional and accidental attacks are major concerns of Internet-based service providers. An Intrusion Detection System (IDS) [97] is an integral component of an organization's security infrastructure, which prevents unauthorized access to data. With the evolution of the Internet of Things (IoT), cloud architectures, Long-Term Evolution (LTE), 5G and smart grid technologies, there is an explosive increase in network traffic characterized by multi-dimensional data and sophisticated attack scenarios [88]. Feature selection for the deployment of IDSs is a non-trivial problem due to the multitude of features and redundancies among them. In addition, computing time is a critical criterion to evaluate IDSs, however some of the BIA-FS algorithms such as binary GWA still have a high computational time [98]. Davahli et al. [50] presented a lightweight, intelligent and more accurate detection model for IoT wireless networks (IoTIDS) with a low computational time by hybridizing GWA and GA. GA provides favorable information sharing between individuals. Therefore, it is an effective global search but a weak local search. In contrast, GWA lacks information sharing among search agents, which causes low diversity and premature convergence. The outcome method overcomes weaknesses of the two, to a certain degree, while retaining their strengths.

Sharaff et al. [90] attempted to classify spam messages using KHA. Being online has become a global trend and Email platform has become the highest prone to spam attacks. Spam is an activity by which hackers use electronic messaging systems to send unsolicited messages in mass content to unknown users. It can also be taken as one of the major attractions of attackers in the form of short message service (SMS) messages. To identify spam messages, KHA vectorizes messages into clusters and employs the distance (similarity) between each individual (message) and its cluster centroid as the fitness function of each candidate solution.

For image processing, Ansari et al. [49] designed an architecture in which GA was responsible for feature selection for detecting text in diverse natural scene images. The goal is to classify text and non-text regions in natural scene images using GA considering diversified set of images having

TABLE 4. Distribution of BIA-FS application domains.

Application Domain	Algorithm	Reference	Frequency
Healthcare	Bat-inspired algorithm (BAA)	[81]	1
	Harris Hawks algorithm (HHA)	[72]	1
	Genetic algorithms (GA)	[54], [55]	2
	Grey wolf algorithm (GWA)	[59]–[61], [73]	4
	Grasshopper algorithm (GHA)	[82]	2
	Whale-inspired algorithm (WHA)	[77], [83], [84]	3
	Aquila algorithm (AQA)	[34]	1
	Moth flame algorithm (MFA)	[78], [83], [85]	3
	Particle swarm algorithm (PSO)	[54], [69], [76], [84]	4
	Cuckoo search algorithm (CSA)	[76]	1
	Lion algorithm (LOA)	[73]	1
	Salp swarm algorithm (SSA)	[73], [74]	2
	Spider monkey algorithm (SMA)	[86]	1
	Paddy field algorithm (PFA)	[86]	1
	Reptile search algorithm (RSA)	[42]	1
	Firefly algorithm (FFA)	[87]	1
Cybersecurity	Grey wolf algorithm (GWA)	[50], [58], [88]	3
	Genetic algorithms (GA)	[50]	1
	Harris Hawks algorithm (HHA)	[89]	1
Text classification	Krill-herd algorithm (KHA)	[90]	1
Image processing	Genetic algorithms (GA)	[49]	1
No specific applications	Bat-inspired algorithm (BAA)	[91]	1
	Coyote algorithm (COA)	[80]	1
	Artificial Flora algorithm (AFA)	[51]	1
	Particle swarm algorithm (PSO)	[40], [53], [63], [68]	4
	Genetic algorithms (GA)	[51]–[53]	3
	Salp swarm algorithm (SSA)	[43]	1
	Differential evolution algorithm (DEA)	[43]	1
	Bat-inspired algorithm (BAA)	[91]	1
	Grasshopper algorithm (GHA)	[52]	1
	Seagull algorithm (SGA)	[41]	1
Whale-inspired algorithm (WHA)	[92]	1	

noise, low contrast/resolution, and random appearance of foreground (font, style, sizes, orientations) and background properties. Multiple techniques were used to extract features such as appearance, contour, texture, etc. before applying GA to them.

C. DATASETS

Data can be classified into two categories: structured and unstructured. Structured data has a predefined schema and can be stored and searched in relational databases with structured query language (SQL). In contrast, there is no

predefined schema for unstructured data at the time of storing the data in the database. Unstructured data is much more challenging to aggregate, process, and analyze. Figure 10 shows that structured data was employed more frequently than unstructured data to evaluate the performance of BIA-FS, with 25 (66%) and 11 (29%) studies, respectively. Two articles (5%) tested both types of data.

We categorized the datasets in terms of the data type into six categories. Figure 11 presents the proportion of the articles over these categories. It should be noted that medical data attracted the attention of more than half of the research (52%),

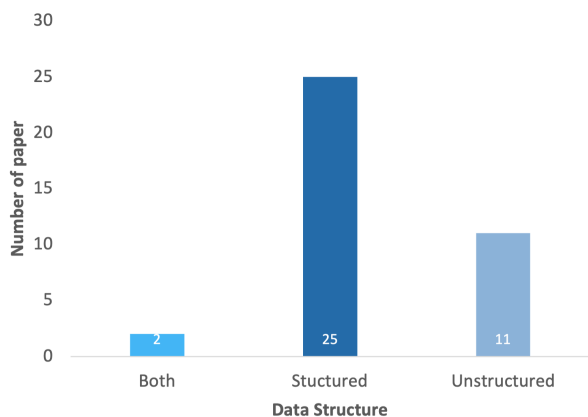


FIGURE 10. Distribution of data structure employed in reviewed papers.

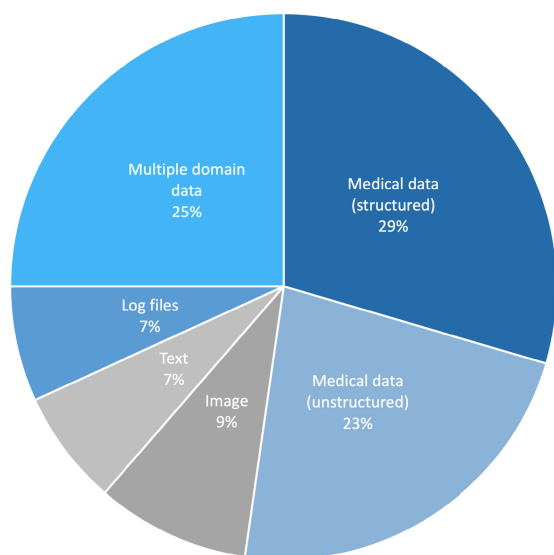


FIGURE 11. Data type studied by the review BIA-FS.

either in structured or unstructured format. A quarter (25%) of the reviewed literature did not focus on a particular data type but employed datasets in various domains. The following is a description of these data types.

1) MEDICAL DATA

- **Structured:** all medical datasets contain information of a particular disease such as heart disease, a type of cancer or eye illness, etc. They vary in the number of instances and features. However, most of them have target variables (both binary and multiple classes). Researchers tend to use publicly available datasets in the UCI Machine Learning repository [99]. In this SLR, microarray data is also classified as structured medical data. A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time [100]. The output of microarray is a table containing

information of thousand genes, thus implying the name microarray data. Microarray datasets are commonly very large, and analytical precision is influenced by many variables. So it is extremely useful to reduce the dataset to those genes that are best distinguished between the two cases or classes (e.g. normal vs. diseased) [101].

- **Unstructured:** An EHR (electronic health record) is a digital version of patient information that usually contains diagnostic, prescription, and medical images. It is a prolific source of unstructured high-dimension data. The major unstructured medical data investigated in the reviewed papers consists of medical images (lung CT images [84], Retinal images [76], etc.), audio [54], and medical signals [55]. Interestingly, prior to conducting FS on these types of data, all authors used different techniques to extract key features in the first place. This is due to the very high-dimensional nature of these data types, may exceed hundreds of thousands of features for medical images.

2) MULTIPLE DOMAIN DATA

This refers to structured datasets from different domains. For example, Bacanin et al. [51] employed 18 datasets from UCI Machine Learning Repository [99]. These datasets are Breast Cancer containing extracted features of breast tumors, Zoo providing features for classifying different animals, and Wine is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars, etc. It is worth noting that out of nine studies that utilized multiple domain datasets, eight purely evaluated BIAs performance with no specific applications. Only one of them aimed to solve the problem of network intrusion detection.

3) IMAGES

These are general images, medical images are excluded. Ansari et al. [49] proposed a framework to recognize text in diversified natural scene images. They used various image datasets containing text and non-text scenes. For example, one dataset is a collection of training character patches and word patches annotated by the bounded box and their text contents. Another dataset provides variations in font, color, layout, size and inclusion of noise, distortion, blur and varying illuminations. Sehgal et al. [82] predicted Parkinson’s disease based on a dataset of handwritten tests of patients and healthy individuals. The participants give the required information by filling a form for the purpose of research and by drawing spirals and meanders.

4) TEXT

One paper [90] attempted to reduce the dimensionality of text data. The dataset, available on UCI [99], contains a collection of SMS spam messages that were manually extracted from multiple websites, with binary labels: Spam and Ham (or not Spam). Spam classification models were built to measure the efficacy of the KHA. The dataset underwent typical Natural

Language Processing (NLP) preprocessing techniques such as tokenization, normalization, stop word removal, part of speech (POS) to make it compatible with KHA.

5) LOG FILES

In this SLR, log files are classified as structured data because the log files were converted into tabular-form datasets in the reviewed papers. There are two articles employing log files for network intrusion detection. A log file is an auto-generated data file that contains information about usage patterns, activities, and operations within an operating system, application, server, or another device and is the primary data source for network observability.

D. MACHINE LEARNING TECHNIQUES

As we mentioned, the filter feature selection methods are based on the statistical properties of the individual features. As these are independent of the classifier, they are scalable but do not guarantee to perform well with a given learning algorithm. Wrapper methods select a subset of features based on the evaluation criteria of a learning algorithm (e.g., classification). In BIAs' realm, the evaluation criteria are named fitness functions. There is a wide array of machine learning (ML) methods for building fitness functions. In the following, we explain each category and the ML methods in these categories. Figure 12 demonstrates a hierarchy of the algorithms in the reviewed literature and the frequency of each technique type.

(* *Linear Discriminant Analysis*, ** *Quadratic Discriminant Analysis*, *** *Gaussian Process Classifier*)

1) CLASSIFICATION

Classification methods are the most commonly applied techniques for evaluating BIA-FS. Different classification methods are applied. K-nearest neighbors (k-NN) algorithm is a handy, straightforward supervised machine learning method that attempts to address both regression and classification problems. k-NN method decides the class label based on a predetermined number of nearest samples. In particular, it determines the label for the test instances according to the k adjacent neighbors' labels of that instance. Thanks to its simplicity, k-NN is the dominant classification method that appeared in 14 studies. Eleven publications solely employed k-NN classification accuracy to construct fitness functions. Three papers evaluated multiple classifiers when building their fitness functions. Dabba et al. [72] proposed two objective functions, each one is a combination of classification accuracy of SVM or k-NN with the number of subset's features, to evaluate each individual and optimize them simultaneously. Agrawal et al. [92] had an in-depth investigation into the effect of four different classifiers (SVM, k-NN, Decision Tree (DT) C4.5, and Linear Discriminant Analysis classifier (LDA)) on the performance of WHA. Classification accuracy, fitness value, AUC, and the number of features were used to compare the four methods. C4.5 outperformed other methods marginally. Thus, it seemed that the choice

of classifier for fitness function is unlikely to impact WHA performance. Pasha et al. [54] independently assessed eleven classifiers namely Logistic Regression (LR), linear Support Vector Machine (LSVM), radial basis function Support Vector Machine (rSVM), Gaussian Naïve Bayes (GNB), Gaussian Process Classifier (GPC), k-Nearest Neighbor (kNN), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Ada Boost (AB) and Quadratic Discriminant Analysis (QDA). They reported three best GA-inspired classifiers: MLP, GPC and LR; and one best PSO-inspired classifier: MLP; that can be recommended for classifying the Parkinson's disease data.

Support vector machines (SVM) is a supervised classification approach seeking a maximum margin hyperplane that categorizes input samples into two classes. In particular, it classifies new data points based on a labeled training dataset for each category. SVM is the second most popular classification technique in the reviewed articles (ten articles). The kernel of SVM refers to a set of mathematical functions that take the input data and map it to high dimensional space. Therefore, SVM is able to perform both linear and non-linear (using kernel function) classification tasks. However, a majority of research [50], [63], [76], [78], [85], [86] did not report the kernel they used. Three papers [49], [72], [92] employed linear kernel, and one article [54] evaluated both linear and radial basis function kernel. Ansari et al. [49] showed that the proposed GA feature selection for diversified natural scene text classification works well compared to benchmark feature selection/optimization and existing methods in terms of binary classification.

Artificial Neural Networks (ANN) attracted attention from six articles. In [88], the fitness function is based on the classification error rate of an Auto Encoders [102]. An AE follows an unsupervised learning approach to learn the representation of unlabeled data by encoding for creating new data models or for dimensionality reduction. In contrast, Alzaqebah et al. [58] and Pasha et al. [54] employed multilayer perceptron (MLP) architecture. MLP-based fitness function was reported to outperform other techniques [54]. It is interesting to note that two publications about network intrusion detection had their fitness functions constructed from ANN classifiers.

Regarding other classification methods, four articles applied decision tree (DT) algorithms, Bayesian classifiers and Rough Set classifiers. Each appeared in two papers. LDA, GPC and QDA were used in one research. Two variants of DT were used: Classification and Regression Trees [81], C4.5 [92]. Except for [92], where C4.5 performed slightly better than other methods, no other papers reported similar results. Bayesian classification uses the Bayes theorem to predict the occurrence of any event. Bayesian classifiers are the statistical classifiers with the Bayesian probability understandings. The theory expresses how a level of belief is expressed as a probability. Thom de Souza et al. [80] coupled COA with Naïve Bayes classifier, demonstrating a good balance between exploration and exploitation during its

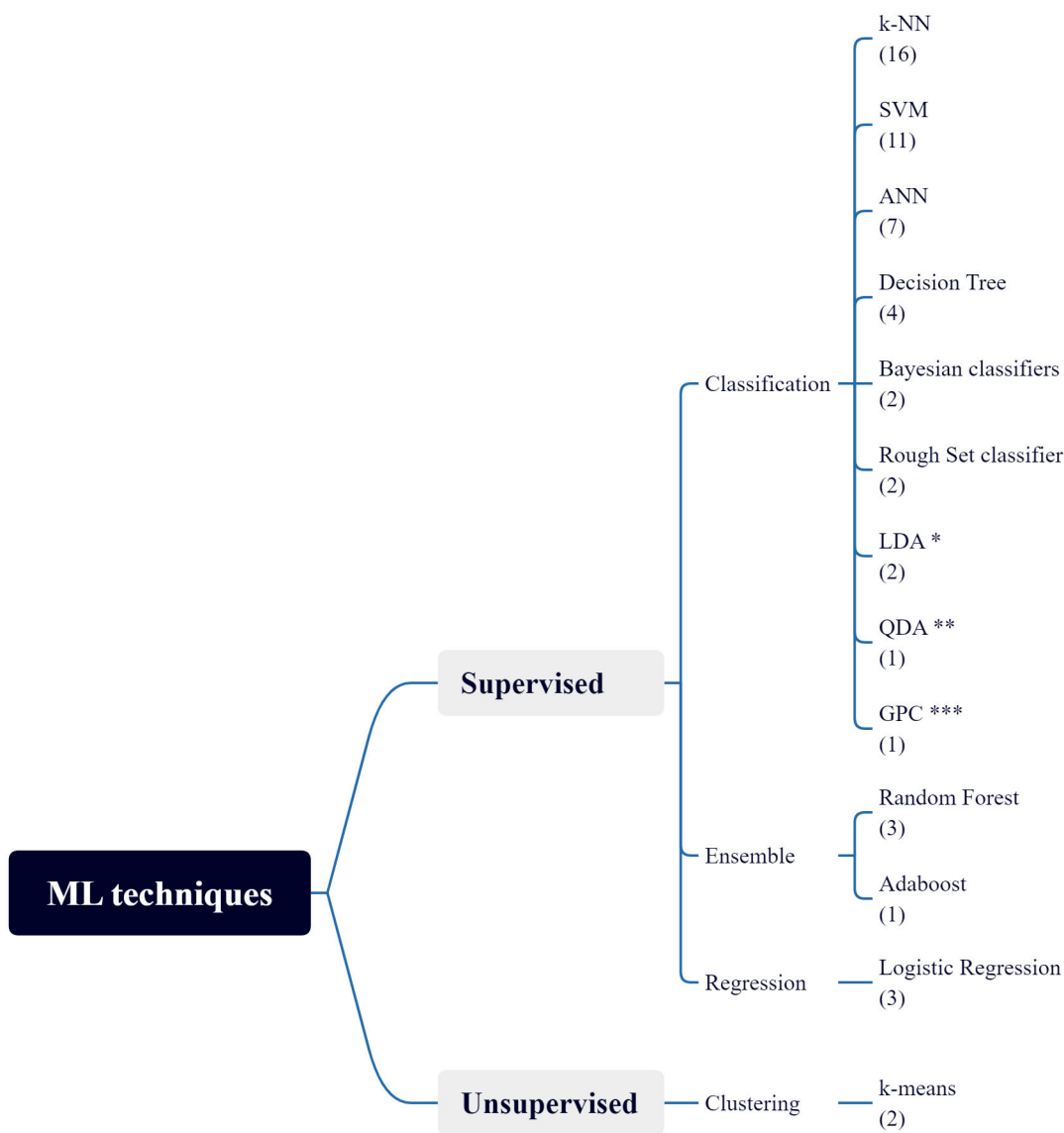


FIGURE 12. ML techniques for constructing fitness functions (with frequency of occurrence).

search for the best solution, avoiding random searches while escaping from local optima. Rough set theory was introduced by Pawlak [103]. The prime objective of this theory is to handle uncertainties and impreciseness while analyzing an information system without any additional information, such as the membership function of a fuzzy set. The BAT which is based on the rough set classifier outperformed DT classifier, as reported in [81].

2) ENSEMBLE

Ensemble methods are algorithms that aggregate multiple intelligent models into one model. Its purpose is to accumulate individual learners’ strengths to create stronger and more

robust learners. Every ensemble method follows different goals; for instance, bagging tries to decrease variance, boosting manages to decrease bias, and stacking wants to improve predictions [104]. Random Forest (RF) was employed three times [54], [76], [82] and Adaboost [54] appeared once.

3) REGRESSION

The studies in this SLR utilized Logistic Regression (LR) method in two papers [54], [76]. Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome, which outputs the probabilities of two values such as true or false, yes or no, etc. Multinomial

logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems. Both articles compared multiple classifiers and LR was not reported to outperform other methods.

4) CLUSTERING

Clustering is the technique of organizing identical instances into the same groups. This unsupervised machine learning technique searches for similarities in the instances and groups them into clusters. K-means was the only method used in this category. Clustering methods were implemented considerably less than classification approaches with two papers [53], [90]. Sharaff et al. [90] found that KHA as feature selection significantly increased the accuracy of spam message classification by 10% to 20%.

RQ2: How effective are bio-inspired feature selection algorithms?

E. PERFORMANCE ANALYSIS

Table 5 presents a summary of the improvement techniques, types of statistical tests, and common parameters that are required for all studies on BIAs within the scope of our Systematic Literature Review (SLR). These parameters include population size and the number of iterations. However, it has been observed that the BIA-FS methods presented in [40], [51], [54], [58], [61], [72], [74], [76], [82], [83], [84], and [86] lack the necessary statistical analysis to demonstrate the significance and superiority of these variants, which is a crucial component of empirical research. In table 5, “None” means the corresponding column is not reported.

While most articles reported the iterations at which the BIAs have reached a stable solution, it is not sufficient to compare their convergence ability based solely on this information. The reason for this is that various BIAs utilize diverse search strategies and employ different control parameters. This variability arises from the different underlying principles and mechanisms of each algorithm, which can lead to different performance characteristics and outcomes. Hence, the exact approach to conducting convergence analysis may differ depending on the specific BIAs used.

It is worth noting that several papers employed alternative approaches to performance analysis due to different reasons. Davahli et al. [50], and Moizuddin et al. [88] opted to fix the iteration number and compare running time instead. This approach is primarily dictated by the nature of the problem under investigation, where the BIA-FS must converge more rapidly within a limited number of iterations. In multi-objective BIA-FS papers [68], [69], the comparison of solutions often involves the use of the Hypervolume (HV) metric. HV is a widely adopted metric that can estimate both the convergence and diversity of the Pareto optimal set. Its popularity stems from its ability to capture the trade-offs between multiple objectives without the need for a priori specification of their relative importance. As such, HV has become a standard tool for assessing the quality of the approximation sets generated by multi-objective algorithms.

Regarding wrapper-based FS, an ML method, for example, a classifier or a clustering method, is required for building fitness function. Once the optimal feature subset is retrieved, another ML method is employed to assess its performance. It should be noted that the two ML methods play different roles and may not be the same algorithm. However, most articles in this SLR applied the same ML techniques for constructing fitness function and assessing the performance of the corresponding BIA-FS. Several typical measures, such as classification accuracy, F1 score, specificity, sensitivity, and the number of features reduced, were employed to evaluate BIA-FS performance. However, the most commonly used criteria were classification accuracy and the number of features reduced. Therefore, we only analyzed these two measures. All research used more than one dataset for evaluation; thus, our analysis is based on the results of the largest dataset (highest number of features). Two papers [84], [90] did not report a reduction ratio, and three papers [53], [55], [87] did not use a classifier, but unsupervised methods, for assessing the performance of their BIA-FS. That explains the missing data of the corresponding BIAs in Figure 13. The “Reduction ratio” is calculated by Eq. (3), as shown at the bottom of page 20.

Interestingly, the number of features before reduction dramatically varies from dataset to dataset. The lowest is a ten-feature dataset of medical data (structured) [81], while the largest has more than 25,000 features which are of medical image [74]. Eleven articles employed high-dimensional datasets (more than one thousand features). Several other high-dimensional datasets are microarray data of 15,000 features [72], text messages of 6,000 features [90], medical data (structured) of 10,000 features [60], etc.

Regarding the reduction ratio, 90% of the BIA-FS were able to eliminate more than 50% of the features. The lowest reduction rate is 30% [81], which was the performance of BAA on a ten-feature dataset. However, ten features of this dataset were picked manually before feeding into BAA. As a result, those are highly relevant features that explain the poor reduction performance. The top reduction ratio reported is 99.83% [85] and 99.91% [78] with MFA on a very high-dimensional dataset. It is the “Breast” dataset [105], a binary dataset containing information of 24,481 genes (or features). In these two papers, Dabba et al. attempted two variants of MFA and found that the quantum MFA was more effective than the other one, with the number of features after reduction being 22 and 40, respectively. SSA achieved a reduction rate of 95.65% [74] on a binary image dataset (25,008 features were extracted from white blood cell leukemia images using VGGNet [75], and 1,087 features were selected by SSA). GA [49] also reached a 99.61% reduction ratio, from 4,051 features to 16 features, on an image dataset for a binary classification task. Another top performer was WHA [92] with a 98.1% reduction rate on a 6,430-feature multi-class text dataset. Overall, the highest reduction ratio is observed on high-dimensional datasets of image and microarray data. This is likely due to several reasons:

TABLE 5. A summary of parameters and statistical tests.

Ref.	BIA	Improvement technique	Statistical test	Population	Max iteration
Moizuddin et al. [88]	GWA	Transfer function	t-test	50	100
Acharjya et al. [81]	BAA	Rough set theory	Friedman test, Wilcoxon signed-rank test	None	900
Alzaqebah et al. [58]	GWA	Initialization phase	None	10	100
Dabba et al. [72]	HHA	Transfer function	None	50	30
Tahir et al. [55]	GA	Chaotic maps	Wilcoxon rank-sum test	30	50
Sathiyabhama et al. [61]	GWA	Rough set theory	None	22	None
Thom de Souza et al. [80]	COA	Hybridization	Wilcoxon signed-rank test	30	100
Preeti et al. [60]	GWA	Random walk	post-hoc test	10	100
Ansari et al. [49]	GA	None	Friedman test, None	200	100
Ji et al. [40]	PSO	Hybridization	None	20	200
Nadimi-Shahraki et al. [34]	AQA	Lévy flight	Friedman test	300	300
Vijh et al. [84]	WHA	Hybridization	None	25	35
Sayed et al. [83]	WHA, MFA	None	None	50	50
Pasha et al. [54]	GA, PSO	None	None	50	100
Davahli et al. [50]	GWA	Hybridization	None	8	20
Sarhani et al. [63]	PSO	Cooperative learning	t-test	None	100
Singh et al. [76]	CSA, BAA, PSO	Multiple BIAs	None	50	10,000
Balasubramanian et al. [73]	LOA, GWA, SSA	Multiple BIAs	ANOVA	50	100
Sahlol et al. [74]	SSA	Other	None	10	100
Isaac et al. [86]	SMA, PFA	Cooperative learning	None	50	100
Dabba et al. [85]	MFA	Other	Wilcoxon signed-rank test	Other	None
Bacanin et al. [51]	AQA	Hybridization	None	40	250
Zhang et al. [43]	SSA	Hybridization	Wilcoxon signed-rank test	10	50
Dabba et al. [78]	MFA	Quantum representation	None	50	30
Ewees et al. [52]	GHA	Hybridization	Wilcoxon rank-sum test	30	200
Elgamal et al. [42]	RSA	Chaotic maps	Wilcoxon rank-sum test	10	100

- They contain an abundant amount of redundant, irrelevant features. Consequently, it is easier to identify those features than in low-dimensional datasets.
- In general, machine learning tasks for processing image do not require all features (pixels). In fact, only several features such as contour, shape, texture, etc. are needed.
- Regarding microarray data, predicting a particular disease relies on some genes among thousands of genes.

Classification accuracy also witnessed a big gap between the best and the worst performers, nearly 100% and 70%. A different pattern is seen in the accuracy rate where several best performances were achieved more on low-dimensional datasets than high-dimensional ones. GWA [88] and AFA [51] reported classification accuracy up to 99.99% and 99.38%, respectively. The former employed a 41-feature multi-class network traffic log files and the latter tested on a 325-feature dataset. Vijn et al. [84] achieved 97.18% accuracy rate with WHA on a 60-feature medical image dataset. Sehgal et al. [82] also reported 100% accuracy when applying GHA with a 26-feature voice dataset to predict Parkinson's disease. Another article [50] got a high accuracy rate with GWA on network traffic log files at 99.1%.

Interestingly, Alzaqebah et al. [58] also employed GWA on network traffic data. However the accuracy ratio was only 80.1%. Hybridization with GA may contribute to the success of the GWA variant in [50]. Although RSA only reached 70% accuracy rate, it was superior to other algorithms in this paper [42] (GA, PSO, GHA). Furthermore, the results also indicated that RSA could improve computational accuracy and accelerate the convergence rate.

RQ3: What are the challenges and research gaps in the existing bio-inspired algorithms for feature selection?

The answers to this research question are elaborated on in the following section.

V. GAP ANALYSIS AND FUTURE WORK

A. APPLICATION DOMAINS

This study categorized these techniques into two major groups and five application domains, which helped highlight dominant techniques and domains. Interestingly, 55% of bio-inspired feature selection studies (21 out of 38) in this review tried to solve problems in the healthcare domain, which is a prolific source of high-dimensional data. And 78% of articles investigated SWA techniques. This leads to several potential future works to study the applications of BIA-FS in other domains such as finance, which is also a good source of high-dimensional data. In particular, financial fraud attempts have increased drastically, which makes fraud detection more important than ever [104]. Another observation of the

limitations of these methods is the ability to meet the requirements of real-time applications such as network intrusion detection. This resulted in a mere number of studies about bio-inspired feature selection methods in this field, which also presents an opportunity for future research. Potential future research into big data and distributed data platform are also needed as no papers in this SLR investigated this direction.

B. IMPROVEMENT TECHNIQUES

Several techniques were employed to boost the performance of BIA-FS. Hybridization with other BIAs attracted significant attention from researchers, with 18.5% of those articles proposing at least one modification. Within the "Hybridization" category, half of the papers chose GA to hybrid with their BIAs of interest. Thus, hybridization with other popular BIAs such as GWA, PSO, etc. may be a potential research direction. In addition, quantum representation appeared in three articles, while rough set theory, chaotic maps, and Lóvy flight were used in two papers for each method. Significant improvements in reduction ratio and classification were observed in these techniques. Therefore, they appear to be interesting areas for future research.

C. TRANSFER FUNCTIONS

Most bio-inspired optimization algorithms are developed to solve continuous problems, while feature selection is a binary optimization problem (each candidate solution is a d-dimensional vector consisting of 0 or 1 values, where 0s indicates excluded features and 1s represents selected ones). Therefore, adopting a binary representation is an important step. Transfer or discretization function, which is a family of techniques, is employed for this task. Different discretization methods affect feature selection performance differently in subsequent steps.

S-shaped transfer functions (family of sigmoid techniques) were used in 13 out of 38 papers, either exclusively or partially, making it the dominant technique. Typically, researchers tended not to focus on transfer functions and opt for sigmoid functions by default. V-shaped functions are also worth noting with three appearances. Only one article, Nadimi-Shahraki et al. [34], specifically evaluated the effect of various transfer functions by testing 4 S-shaped and 4 V-shaped functions with AQA. Alzaqebah et al. [89] employed a novel transfer function called X-shaped [106], which combines S-shaped and V-shaped functions. The X-shaped transfer function is used to improve the searching capability of the HHA. The additional advantage of the X-shaped function is balancing the exploration and exploitation phases. Dabba et al. [72] used a combination of a discretization technique based on clustering (K-means) and the sigmoid function as

$$\frac{\text{no.of features of original dataset} - \text{no.of features of the optimal subset}}{\text{no.of features of original dataset}} \quad (3)$$

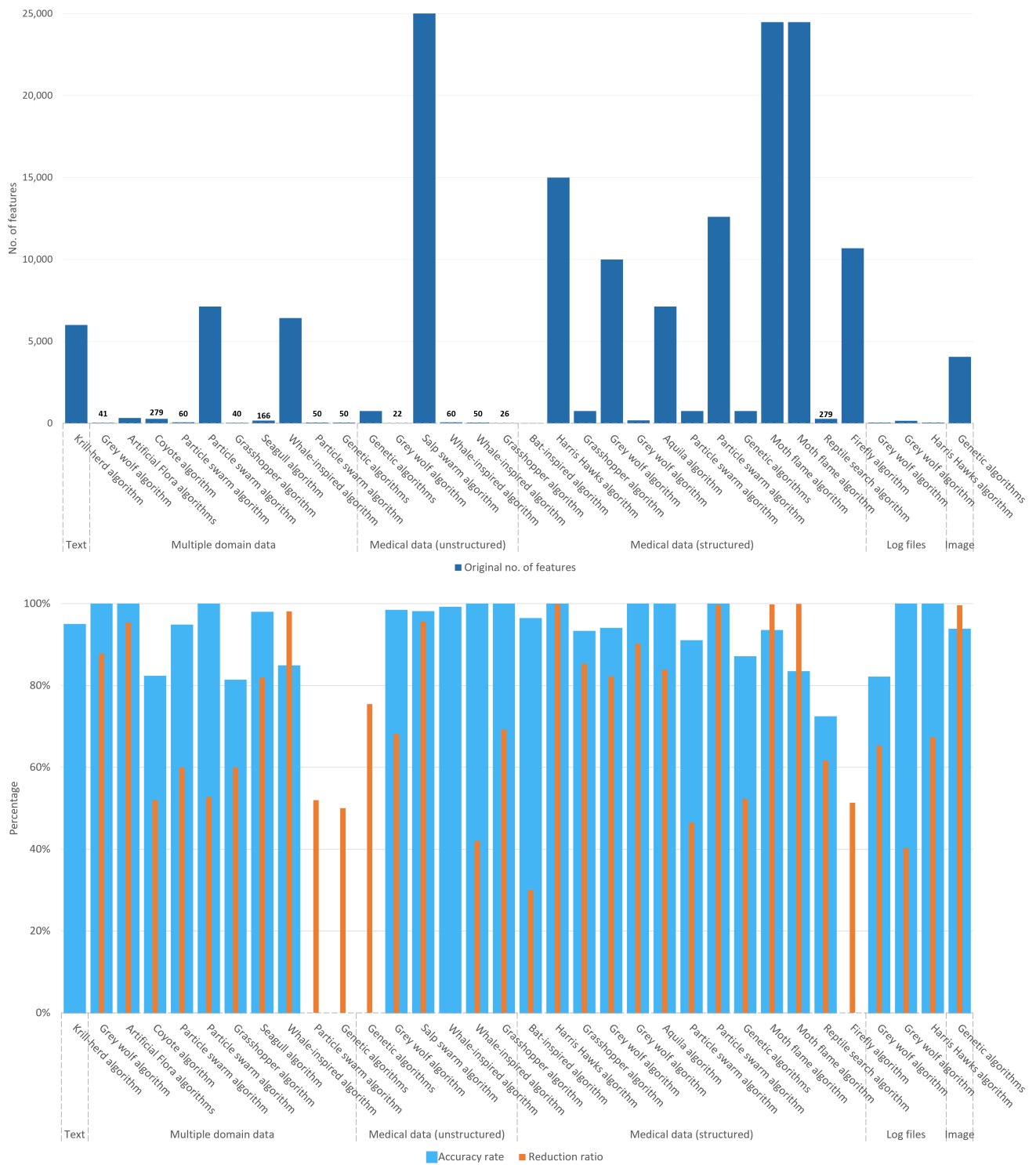


FIGURE 13. Feature size, classification accuracy, and reduction ratio by BIA-FS and data type.

a transfer function for HHA. Interestingly, Dabba et al. [85] employed another transfer technique with MFA, discretization based on the ranking of minimum redundancy-maximum relevance measure. However, those two articles did not

evaluate how different transfer functions influence feature selection performance. With this in mind, future studies about other transfer functions besides S-shaped are needed.

TABLE 6. The list of all queries explored in five databases.

Ref.	BIA	Improvement technique	Statistical test	Population	Max iteration
Ewees et al. [41]	SGA	Lévy flight	Friedman test, post-hoc test	30	100
Sehgal et al. [82]	GHA	Fitness function	None	None	50
Agrawal et al. [92]	WHA	Quantum representation	Wilcoxon signed-rank test	50	100
Martarelli et al. [53]	GA, PSO	Other	Wilcoxon signed-rank test	30	15
Alzaqebah et al. [89]	HHA	Transfer function	F-test	10	100
Kaur et al. [77]	WHA	Quantum representation	Wilcoxon signed-rank test	20	100
Zenbout et al. [59]	GWA	Initialization phase	Chi-squared test	11	300
Han et al. [68]	PSO	Other	Wilcoxon rank-sum test	30	100
Zhou et al. [69]	PSO	Other	Wilcoxon signed-rank test	300	70

D. NEIGHBORHOOD TOPOLOGY

The individuals in the bio-inspired algorithms must exchange their information to obtain the global optimum. For instance, genetic algorithms adopt the crossover operator to exchange information, while particle swarm algorithms such as bat or cat-swarm, each particle is attracted to the global optimum at each iteration to update its position. Different algorithms employ different neighborhood topologies.

Dabba et al. [72], each Hawk utilizes the prey (global optimum) to adjust its speed, direction, and position but does not interact with other Hawks (local optimum). In contrast, Acharjya et al. [81], the best bat in the population (global optimum) and the best bat in the group (local optimum) will influence each bat's position. In contrast, GA [50] can only exchange information between two genes in each iteration, excluding the global optimum's impact. Kennedy [107] defines four neighborhood types for particle swarm optimization methods: circles, stars, wheels and random edges. Its findings show that the topological structures of the particle swarm have a significant effect on its ability to find optima: the optimal pattern of connectivity among individuals depends on the problem being solved.

Due to the differences in the topology, each BIA-FS has its pros and cons. Therefore, hybridizing two or even more algorithms is a crucial research direction. There have been numerous attempts to hybridize algorithms. Davahli et al. [50] hybridize GA and GWA to solve network intrusion detection problems. Agrawal et al. [92] combine WHA with quantum theory to feature selection in general. However, the studies in this SLR did not examine the effect of neighborhood topology on the results. Since the topological structures are likely to influence how a method is easier to fall into local optimum, while others have slower convergence speed, a deep analysis of this area will help to gain better insight into more efficient hybridization mechanisms.

E. MACHINE LEARNING TECHNIQUES

Regarding ML techniques for fitness function, researchers focused on supervised classification methods such as k-NN, SVM (71% of papers). On the other hand, neural networks and clustering techniques were employed only seven and two times, respectively. In recent decades, neural networks have been very successful in a wide array of applications. The more data is fed into neural networks, the better results they produce. Therefore, a future research direction is to investigate which BIA-FSs are optimized for deep learning models. This will benefit both supervised and unsupervised tasks as deep learning techniques are not limited to only classification. Additionally, clustering is beneficial for investigating latent space or uninterpretable patterns. Future studies focusing on unsupervised practices can help uncover new insights.

VI. THREATS TO VALIDITY AND LIMITATION

According to Perry et al. [108], three types of threats to validity that any SLRs may encounter: construct validity, external validity, and external validity.

- **Construct Validity:** this refers to the quality of the methodology in terms of being helpful in answering the research questions. Even though the essential keywords were included in the query that was ran across the most popular databases, it could be possible that some relevant papers were not included in the SLR. In systematic literature review papers, identifying relevant papers plays a crucial role. Note that the research questions were answered based on the papers selected from the set found in the searching stage. The grey literature was not reviewed and it may have included relevant research. Such threat to validity was mitigated by searching for more papers through snowballing. Moreover, language bias is considered a threat because, in this systematic

TABLE 7. Glossary.

General term	Description
BIA	Bio-inspired Optimization Algorithm
BIA-FS	Bio-inspired algorithms applied for feature selection
DR	Dimensionality Reduction
EA	Evolutionary Algorithms
FE	Feature Extraction
FS	Feature Selection
PCA	Principal Component Analysis
SVM	Support Vector Machine
SWA	Swarm-based Algorithms

Algorithm	Description
ABC	Artificial Bee colony
ACO	Ant Colony algorithm
AFA	Artificial Flora algorithm
AIS	Artificial Immune systems
AQA	Aquila algorithm
BAA	Bat-inspired algorithm
COA	Coyote algorithm
CSA	Cuckoo search algorithm
CSO	Chicken swarm optimization
DEA	Differential Evolution algorithm
ES	Evolution strategy
FFA	Firefly algorithm
GA	Genetic algorithms
GP	Genetic programming
GHA	Grasshopper algorithm
GWA	Grey Wolf algorithm
HHA	Harris Hawks algorithm
KHA	Krill-herd algorithm
LCS	Learning classifier systems
LOA	Lion algorithm
MFA	Moth Flame algorithm
PFA	Paddy Field algorithm
PSO	Particle swarm algorithm
RSA	Reptile search algorithm
SGA	Seagull algorithm
SMA	Spider Monkey algorithm
SSA	Salp swarm algorithm
WHA	Whale-inspired algorithm

literature review, only English language papers were selected.

- **External Validity:** it considers whether applying the conclusion of this study and the results to other cases or situations is possible. The focus of this SLR is limited to the application of bio-inspired algorithms to feature selection and their application in business, engineering, healthcare and so on. The results and challenges discussed here were based specifically on the papers related to these topics.
- **Internal Validity:** examines any bias in performing the research. The major threat to the internal validity here is that the literature review was done by one person. Some aspects might have been overlooked or misrepresented. To partially mitigate this threat, the supervisor was involved during the whole review process; a librarian, peers, and another professor was consulted to review the protocol and the systematic literature review to minimize this threat.

VII. CONCLUSION

The field of bio-inspired techniques is an emerging area that can solve complex real-world optimization problems.

Research interests in bio-inspired methods for feature selection are rising due to their superiority over projection-based algorithms in dealing with non-linear high-dimensional data. There are, however, research gaps that have not been investigated thoroughly. This systematic literature review analyzed and synthesized 38 peer-reviewed papers out of 695 articles retrieved from well-known search databases.

The contributions of this systematic review include the identification of the bio-inspired algorithms that are applied in feature selection and their domains of application. There are two major groups and six application domains. It is worth noting that around 55% of bio-inspired feature selection studies (21 out of 38) were applied in the healthcare domain, which is a prolific source of high-dimensional data. Also, 78% of articles investigated SWA techniques. Those methods significantly increase classification accuracy and reduction ratio in various classification and clustering tasks. In 87% of the reviewed articles, the BIAs were customized to improve their performance in feature selection. Hybridization with other BIAs was a popular approach, employed by approximately 18.5% of the papers. Of those, half utilized GA in combination with their BIAs of interest. Regarding transfer function, 13 out of 38 papers utilized S-shaped functions, which belong to the family of sigmoid techniques. This made S-shaped transfer functions the most commonly used technique, either exclusively or partially, in the selected papers. Moreover, researchers in this area focused on supervised classification methods such as k-NN and SVM (71% of papers) for building fitness functions, a crucial component of BIA-FS. On the other hand, neural networks and clustering techniques were employed only in six and two papers, respectively.

Based on our analysis, researchers should expand their studies in other application domains besides healthcare, such as finance and social networks, where high-dimensional data exists in various forms, from text and image to audio. BIA-FS for big data on distributed platforms likely merits further investigation. In addition, cybersecurity applications, where real-time feature selection is a critical requirement, could also be of further attention. Notably, BIA-FS that incorporated enhancement techniques such as quantum representation, rough set theory, chaotic maps, and Lóvy flight demonstrated notable improvements in reduction ratio and classification performance. Thus, these techniques may offer promising areas for future research. Moreover, further exploration of other transfer functions, such as V-shaped and X-shaped, is necessary. Finally, another future direction is to conduct in-depth studies on the impact of clustering and deep learning models for constructing fitness functions in BIA-FS. Exploring these unsupervised methods may provide a better understanding of the mechanisms of bio-inspired feature selection algorithms.

APPENDIX

- Table 6 presents the search string for each digital library.
- Table 7 presents the abbreviations for key concepts.

REFERENCES

- [1] B. K. Patra and S. Nandi, "Effective data summarization for hierarchical clustering in large datasets," *Knowl. Inf. Syst.*, vol. 42, no. 1, pp. 1–20, Jan. 2015.
- [2] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, and L. Gu, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Appl. Soft Comput.*, vol. 74, pp. 634–642, Jan. 2019.
- [3] H. Liu and H. Motoda, Eds., *Computational Methods of Feature Selection*. New York, NY, USA: Chapman & Hall, Oct. 2007.
- [4] A. Darwish, "Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications," *Future Comput. Informat. J.*, vol. 3, no. 2, pp. 231–246, Dec. 2018.
- [5] M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, "A wrapper-filter feature selection technique based on ant colony optimization," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 7839–7857, Jun. 2020.
- [6] J. Li, J. Tang, and H. Liu, "Reconstruction-based unsupervised feature selection: An embedded approach," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 2159–2165.
- [7] L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: A review," *Appl. Sci.*, vol. 8, no. 9, p. 1521, Sep. 2018.
- [8] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.
- [9] N. Almgren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*, vol. 7, pp. 78533–78548, 2019.
- [10] S. Deb, X.-Z. Gao, K. Tammi, K. Kalita, and P. Mahanta, "Recent studies on chicken swarm optimization algorithm: A review (2014–2018)," *Artif. Intell. Rev.*, vol. 53, no. 3, pp. 1737–1765, Mar. 2020.
- [11] T. M. Shami, A. A. El-Saleh, M. Alswaiti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, "Particle swarm optimization: A comprehensive survey," *IEEE Access*, vol. 10, pp. 10031–10061, 2022.
- [12] J. Del Ser, E. Osaba, D. Molina, X.-S. Yang, S. Salcedo-Sanz, D. Camacho, S. Das, P. N. Suganthan, C. A. C. Coello, and F. Herrera, "Bio-inspired computation: Where we stand and what's next," *Swarm Evol. Comput.*, vol. 48, pp. 220–250, Aug. 2019.
- [13] H. Almazrua and H. Alshamlan, "A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data," *IEEE Access*, vol. 10, pp. 71427–71449, 2022.
- [14] F. Fausto, A. Reyna-Orta, E. Cuevas, Á. G. Andrade, and M. Perez-Cisneros, "From ants to whales: Metaheuristics for all tastes," *Artif. Intell. Rev.*, vol. 53, no. 1, pp. 753–810, Jan. 2020.
- [15] R. Bellman, *Dynamic Programming*, 1st ed. Princeton, NJ, USA: Princeton Univ. Press, 1957.
- [16] A. Juvonen, T. Sipola, and T. Hämäläinen, "Online anomaly detection using dimensionality reduction techniques for HTTP log analysis," *Comput. Netw.*, vol. 91, pp. 46–56, Nov. 2015.
- [17] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," in *Computational Intelligence and Bioinspired Systems (Lecture Notes in Computer Science)*, J. Cabestany, A. Prieto, and F. Sandoval, Eds. Berlin, Germany: Springer, 2005, pp. 758–770.
- [18] L. van der Maaten, E. Postma, and H. Herik, "Dimensionality reduction: A comparative review," *J. Mach. Learn. Res.*, vol. 10, pp. 66–71, Jan. 2007.
- [19] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, Sep. 1933.
- [20] X.-L. Zhang, "Nonlinear dimensionality reduction of data by deep distributed random samplings," in *Proc. 6th Asian Conf. Mach. Learn.*, Feb. 2015, pp. 221–233.
- [21] Q. K. Weinberger, F. Sha, and K. L. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*. New York, NY, USA: Association for Computing Machinery, Jul. 2004, p. 106.
- [22] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.
- [23] Z. Wang, C. Qin, B. Wan, and W. W. Song, "A comparative study of common nature-inspired algorithms for continuous function optimization," *Entropy*, vol. 23, no. 7, p. 874, Jul. 2021.
- [24] V. Bolón-Canedo, N. Sánchez-Marroño, and A. Alonso-Betanzos, "Feature selection for high-dimensional data," in *Artificial Intelligence: Foundations, Theory, and Algorithms*. Cham, Switzerland: Springer, 2015.
- [25] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, no. 2, p. 40, 2012.
- [26] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.
- [27] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2017.
- [28] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
- [29] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw. (ICNN)*, vol. 4, Nov. 1995, pp. 1942–1948.
- [30] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 240–255, Jun. 2004.
- [31] A. S. Akopov, L. A. Beklaryan, M. Thakur, and B. D. Verma, "Parallel multi-agent real-coded genetic algorithm for large-scale black-box single-objective optimisation," *Knowl.-Based Syst.*, vol. 174, pp. 103–122, Jun. 2019.
- [32] P. V. Paul, N. Moganaragan, S. S. Kumar, R. Raju, T. Vengattaraman, and P. Dhavachelvan, "Performance analyses over population seeding mutation of the permutation-coded genetic algorithm: An empirical study based on traveling salesman problems," *Appl. Soft Comput.*, vol. 32, pp. 383–402, Jul. 2015.
- [33] S. Taghian, M. H. Nadimi-Shahraki, and H. Zamani, "Comparative analysis of transfer function-based binary metaheuristic algorithms for feature selection," in *Proc. Int. Conf. Artif. Intell. Data Process. (IDAP)*, Sep. 2018, pp. 1–6.
- [34] M. H. Nadimi-Shahraki, S. Taghian, S. Mirjalili, and L. Abualigah, "Binary Aquila optimizer for selecting effective features from medical data: A COVID-19 case study," *Mathematics*, vol. 10, no. 11, p. 1929, Jun. 2022.
- [35] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. Comput. Simulation*, vol. 5, Oct. 1997, pp. 4104–4108.
- [36] S. Mirjalili, H. Zhang, S. Mirjalili, S. Chalup, and N. Noman, "A novel U-shaped transfer function for binary particle swarm optimisation," in *Soft Computing for Problem Solving 2019 (Advances in Intelligent Systems and Computing)*, A. K. Nagar, K. Deep, J. C. Bansal, and K. N. Das, Eds. Singapore: Springer, 2020, pp. 241–259.
- [37] S. Mirjalili and A. Lewis, "S-shaped versus V-shaped transfer functions for binary particle swarm optimization," *Swarm Evol. Comput.*, vol. 9, pp. 1–14, Apr. 2013.
- [38] M. F. Tasgetiren, P. N. Suganthan, and Q.-Q. Pan, "A discrete particle swarm optimization algorithm for the generalized traveling salesman problem," in *Proc. 9th Annu. Conf. Genetic Evol. Comput. (GECCO)*. New York, NY, USA: Association for Computing Machinery, Jul. 2007, pp. 158–167.
- [39] Y.-T. Kao and E. Zahara, "A hybrid genetic algorithm and particle swarm optimization for multimodal functions," *Appl. Soft Comput.*, vol. 8, no. 2, pp. 849–857, Mar. 2008.
- [40] B. Ji, X. Lu, G. Sun, W. Zhang, J. Li, and Y. Xiao, "Bio-inspired feature selection: An improved binary particle swarm optimization approach," *IEEE Access*, vol. 8, pp. 85989–86002, 2020.
- [41] A. A. Ewees, R. R. Mostafa, R. M. Ghoniem, and M. A. Gaheen, "Improved seagull optimization algorithm using Lévy flight and mutation operator for feature selection," *Neural Comput. Appl.*, vol. 34, no. 10, pp. 7437–7472, May 2022.
- [42] Z. Elgamal, A. Q. M. Sabri, M. Tubishat, D. Tbaishat, S. N. Makhadmeh, and O. A. Alomari, "Improved reptile search optimization algorithm using chaotic map and simulated annealing for feature selection in medical field," *IEEE Access*, vol. 10, pp. 51428–51446, 2022.
- [43] H. Zhang, T. Liu, X. Ye, A. A. Heidari, G. Liang, H. Chen, and Z. Pan, "Differential evolution-assisted salp swarm algorithm with chaotic structure for real-world problems," *Eng. Comput.*, pp. 1–35, Jan. 2022, doi: 10.1007/s00366-021-01545-x.
- [44] B. A. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ., Keele, U.K., Durham Univ., Durham, U.K., Joint Rep. EBSE 2007-001, Jan. 2007.

- [45] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Mar. 2021, Art. no. n71.
- [46] G. Beni and J. Wang, "Swarm intelligence in cellular robotic systems," in *Robots and Biological Systems: Towards a New Bionics?* (NATO ASI Series), P. Dario, G. Sandini, and P. Aebischer, Eds., Berlin, Germany: Springer, 1993, pp. 703–712.
- [47] K.-L. Du and M. N. S. Swamy, *Search and Optimization by Metaheuristics*. Cham, Switzerland: Springer, 2016.
- [48] F. Glover, "Future paths for integer programming and links to artificial intelligence," *Comput. Oper. Res.*, vol. 13, no. 5, pp. 533–549, Jan. 1986.
- [49] G. J. Ansari, J. H. Shah, M. C. Q. Farias, M. Sharif, N. Qadeer, and H. U. Khan, "An optimized feature selection technique in diversified natural scene text for classification using genetic algorithm," *IEEE Access*, vol. 9, pp. 54923–54937, 2021.
- [50] A. Davahli, M. Shamsi, and G. Abaei, "Hybridizing genetic algorithm and grey wolf optimizer to advance an intelligent and lightweight intrusion detection system for IoT wireless networks," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 5581–5609, Nov. 2020.
- [51] N. Bacanin, T. Bezdan, F. Al-Turjman, and T. A. Rashid, "Artificial flora optimization algorithm with genetically guided operators for feature selection and neural network training," *Int. J. Fuzzy Syst.*, vol. 24, no. 5, pp. 2538–2559, Jul. 2022.
- [52] A. A. Ewees, M. A. Gaheen, Z. M. Yaseen, and R. M. Ghoniem, "Grasshopper optimization algorithm with crossover operators for feature selection and solving engineering problems," *IEEE Access*, vol. 10, pp. 23304–23320, 2022.
- [53] N. J. Martarelli and M. S. Nagano, "Unsupervised feature selection based on bio-inspired approaches," *Swarm Evol. Comput.*, vol. 52, Feb. 2020, Art. no. 100618.
- [54] A. Pasha and P. H. Latha, "Bio-inspired dimensionality reduction for Parkinson's disease (PD) classification," *Health Inf. Sci. Syst.*, vol. 8, no. 1, Dec. 2020.
- [55] M. Tahir, A. Tubaishat, F. Al-Obeidat, B. Shah, Z. Halim, and M. Waqas, "A novel binary chaotic genetic algorithm for feature selection and its utility in affective computing and healthcare," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11453–11474, Jul. 2022.
- [56] Q. He and L. Wang, "An effective co-evolutionary particle swarm optimization for constrained engineering design problems," *Eng. Appl. Artif. Intell.*, vol. 20, pp. 89–99, Feb. 2007.
- [57] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.
- [58] A. Alzaqebah, I. Aljarah, O. Al-Kadi, and R. Damaševičius, "A modified grey wolf optimization algorithm for an intrusion detection system," *Mathematics*, vol. 10, no. 6, p. 999, Mar. 2022.
- [59] I. Zenboud, A. Bouramoul, S. Meshoul, and M. Amrane, "Efficient bioinspired feature selection and machine learning based framework using omics data and biological knowledge data bases in cancer clinical endpoint prediction," *IEEE Access*, vol. 11, pp. 2674–2699, 2023.
- [60] Preeti and K. Deep, "A random walk grey wolf optimizer based on dispersion factor for feature selection on chronic disease prediction," *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117864.
- [61] B. Sathiyabhama, S. U. Kumar, J. Jayanthi, T. Sathiya, A. K. Ilavarasi, V. Yuvarajan, and K. Gopikrishna, "A novel feature selection framework based on grey wolf optimizer for mammogram image analysis," *Neural Comput. Appl.*, vol. 33, no. 21, pp. 14583–14602, Nov. 2021.
- [62] S. Yadav, A. Ekbal, and S. Saha, "Feature selection for entity extraction from multiple biomedical corpora: A PSO-based approach," *Soft Comput.*, vol. 22, no. 20, pp. 6881–6904, Oct. 2018.
- [63] M. Sarhani and S. Voß, "Chunking and cooperation in particle swarm optimization for feature selection," *Ann. Math. Artif. Intell.*, vol. 90, nos. 7–9, pp. 893–913, Sep. 2022.
- [64] Z.-H. Zhan, J. Zhang, Y. Li, and H. S.-H. Chung, "Adaptive particle swarm optimization," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 6, pp. 1362–1381, Dec. 2009.
- [65] F. van den Bergh and A. P. Engelbrecht, "A cooperative approach to particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 225–239, Jun. 2004.
- [66] H. Li, Y.-S. Ong, M. Gong, and Z. Wang, "Evolutionary multitasking sparse reconstruction: Framework and case study," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 733–747, Oct. 2019.
- [67] W.-N. Chen, D.-Z. Tan, Q. Yang, T. Gu, and J. Zhang, "Ant colony optimization for the control of pollutant spreading on social networks," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 4053–4065, Sep. 2020.
- [68] F. Han, W.-T. Chen, Q.-H. Ling, and H. Han, "Multi-objective particle swarm optimization with adaptive strategies for feature selection," *Swarm Evol. Comput.*, vol. 62, Apr. 2021, Art. no. 100847.
- [69] Y. Zhou, J. Kang, and H. Guo, "Many-objective optimization of feature selection based on two-level particle cooperation," *Inf. Sci.*, vol. 532, pp. 91–109, Sep. 2020.
- [70] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Machine Learning: ECML-94* (Lecture Notes in Computer Science), F. Bergadano and L. De Raedt, Eds. Berlin, Germany: Springer, 1994, pp. 171–182.
- [71] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, Dec. 2007.
- [72] A. Dabba, A. Tari, and S. Meftali, "A new multi-objective binary Harris Hawks optimization for gene selection in microarray data," *J. Ambient Intell. Hum. Comput.*, vol. 14, pp. 1–20, Aug. 2021.
- [73] K. Balasubramanian and N. P. Ananthamoorthy, "Correlation-based feature selection using bio-inspired algorithms and optimized KELM classifier for glaucoma diagnosis," *Appl. Soft Comput.*, vol. 128, Oct. 2022, Art. no. 109432.
- [74] A. T. Sahlol, P. Kollmannsberger, and A. A. Ewees, "Efficient classification of white blood cell leukemia with improved swarm optimization of deep features," *Sci. Rep.*, vol. 10, no. 1, p. 2536, Feb. 2020.
- [75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Apr. 2014, *arXiv:1409.1556*.
- [76] L. K. Singh, M. Khanna, S. Thawkar, and R. Singh, "Collaboration of features optimization techniques for the effective diagnosis of glaucoma in retinal fundus images," *Adv. Eng. Softw.*, vol. 173, Nov. 2022, Art. no. 103283.
- [77] B. Kaur, S. Rathi, and R. K. Agrawal, "Enhanced depression detection from speech using quantum whale optimization algorithm for feature selection," *Comput. Biol. Med.*, vol. 150, Nov. 2022, Art. no. 106122.
- [78] A. Dabba, A. Tari, and S. Meftali, "Hybridization of moth flame optimization algorithm and quantum computing for gene selection in microarray data," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 2731–2750, Feb. 2021.
- [79] K.-H. Han and J.-H. Kim, "Genetic quantum algorithm and its application to combinatorial optimization problem," in *Proc. Congr. Evol. Comput.*, vol. 2, Jul. 2000, pp. 1354–1360.
- [80] R. C. T. de Souza, C. A. de Macedo, L. dos Santos Coelho, J. Piezran, and V. C. Mariani, "Binary coyote optimization algorithm for feature selection," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107470.
- [81] D. P. Acharjya and P. K. Ahmed, "A hybridized rough set and bat-inspired algorithm for knowledge inferencing in the diagnosis of chronic liver disease," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 13489–13512, Apr. 2022.
- [82] S. Sehgal, M. Agarwal, D. Gupta, S. Sundaram, and A. Bashambu, "Optimized grass hopper algorithm for diagnosis of Parkinson's disease," *Social Netw. Appl. Sci.*, vol. 2, no. 6, pp. 1–18, Jun. 2020.
- [83] G. I. Sayed, A. Darwish, and A. E. Hassanien, "Binary whale optimization algorithm and binary moth flame optimization with clustering algorithms for clinical breast cancer diagnoses," *J. Classification*, vol. 37, no. 1, pp. 66–96, Apr. 2020.
- [84] S. Vijh, P. Gaurav, and H. M. Pandey, "Hybrid bio-inspired algorithm and convolutional neural network for automatic lung tumor detection," *Neural Comput. Appl.*, pp. 1–14, 2020, doi: 10.1007/s00521-020-05362-z.
- [85] A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114012.
- [86] A. Isaac, H. K. Nehemiah, S. D. Dunston, V. R. E. Christo, and A. Kannan, "Feature selection using competitive coevolution of bio-inspired algorithms for the diagnosis of pulmonary emphysema," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103340.
- [87] Y. Dou and H. Jiang, "Informative gene identification for single-cell RNA-seq data with mutual information based firefly algorithm," in *Proc. 8th Int. Conf. Biomed. Bioinf. Eng.* New York, NY, USA: Association for Computing Machinery, 2021, pp. 57–64.
- [88] M. Moizuddin and M. V. Jose, "A bio-inspired hybrid deep learning model for network intrusion detection," *Knowl.-Based Syst.*, vol. 238, Feb. 2022, Art. no. 107894.

- [89] A. Alzaqebah, I. Aljarah, and O. Al-Kadi, "A hierarchical intrusion detection system based on extreme learning machine and nature-inspired optimization," *Comput. Secur.*, vol. 124, Jan. 2023, Art. no. 102957.
- [90] A. Sharaff, C. Kamal, S. Porwal, S. Bhatia, K. Kaur, and M. M. Hassan, "Spam message detection using danger theory and Krill herd optimization," *Comput. Netw.*, vol. 199, Nov. 2021, Art. no. 108453.
- [91] A. K. Naik, V. Kuppili, and D. R. Edla, "Efficient feature selection using one-pass generalized classifier neural network and binary bat algorithm with a novel fitness function," *Soft Comput.*, vol. 24, no. 6, pp. 4575–4587, Mar. 2020.
- [92] R. K. Agrawal, B. Kaur, and S. Sharma, "Quantum based whale optimization algorithm for wrapper feature selection," *Appl. Soft Comput.*, vol. 89, Apr. 2020, Art. no. 106092.
- [93] F. R. S. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philosoph. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.
- [94] J. Hérault and C. Jutten, "Space or time adaptive signal processing by neural network models," *AIP Conf.*, vol. 151, no. 1, pp. 206–211, Aug. 1986.
- [95] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Commun.*, vol. 30, no. 2, pp. 169–190, Jan. 2017.
- [96] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Rev.*, vol. 35, no. 4, pp. 551–566, Dec. 1993.
- [97] K. A. Scarfone and P. M. Mell, "Guide to intrusion detection and prevention systems (IDPS)," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NIST SP 800-94, 2007.
- [98] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, Jan. 2016.
- [99] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/citation_policy.html
- [100] M. Schena, D. Shalon, R. Davis, and P. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, Oct. 1995.
- [101] S. Selvaraj and J. Natarajan, "Microarray data analysis and mining tools," *Bioinformatics*, vol. 6, no. 3, pp. 95–99, Apr. 2011.
- [102] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot, "Higher order contractive auto-encoder," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Berlin, Germany: Springer, 2011, pp. 645–660.
- [103] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, Oct. 1982.
- [104] M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review," *IEEE Access*, vol. 10, pp. 72504–72525, 2022.
- [105] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognit.*, vol. 40, no. 11, pp. 3236–3248, Nov. 2007.
- [106] K. K. Ghosh, P. K. Singh, J. Hong, Z. W. Geem, and R. Sarkar, "Binary social mimic optimization algorithm with X-shaped transfer function for feature selection," *IEEE Access*, vol. 8, pp. 97890–97906, 2020.
- [107] J. Kennedy, "Small worlds and mega-minds: Effects of neighborhood topology on particle swarm performance," in *Proc. Congr. Evol. Comput.*, vol. 3, Jul. 1999, pp. 1931–1938.
- [108] D. E. Perry, A. A. Porter, and L. G. Votta, "Empirical studies of software engineering: A roadmap," in *Proc. Conf. Future Softw. Eng. (ICSE)*. New York, NY, USA: Association for Computing Machinery, May 2000, pp. 345–355.



TIN H. PHAM received the B.Sc. degree in information systems from the University of Economics, Ho Chi Minh City, Vietnam, in 2010, and the M.Sc. degree from Heilbronn University, Germany, in 2019. He is currently pursuing the Ph.D. degree in digital transformation and innovation with the University of Ottawa, Canada. He has done researches in the areas of machine learning, data mining, natural language processing, and fintech. He is also a member of the Knowledge Discovery and Data Mining (KDD) Laboratory, Telfer School of Management, University of Ottawa, where he is participating in various academic and industrial projects. His current research interests include design and development of machine learning and data mining techniques and algorithms for dimensionality reduction in high-dimensional data, with applications in business and engineering.



BIJAN RAAHEMI received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, in 1997. He is currently a Professor in information systems and analytics and the Director of the Knowledge Discovery and Data Mining (KDD) Laboratory, University of Ottawa, Canada. His current research interests include artificial intelligence, machine learning, data mining, big data analytics and their emerging applications in engineering, business, and healthcare. Then, he was appointed as a Senior Researcher with Telecommunications Industry focusing on computer networks architectures, data and multimedia communications, and services. His work has appeared in more than 85 peer-reviewed journals and conference proceedings. He also holds eight patents in data communications. He is a Co-Editor of the *Handbook of Research on Data Science for Effective Healthcare Practice and Administration*. He is a registered member of the Professional Engineers of Ontario (PEO) and a member of the Association for Computing Machinery (ACM).

...