Journal of Scientific & Industrial Research
Vol. 84, June 2025, pp. 672-680
DOI: 10.56042/jsir.v84i6.13842

Feature Selection for Biomedical Data Classification: Statistical vs. Swarm Intelligence Methods

Ulfeta Marovac[1], Aldina Avdić[1]*, Irfan Fetahović[1], Lejlija Memić[1], Nataša Đorđević[2], Zana Dolićanin[3] & Goran Babić[4]

[1]Department of Technical and Technological Sciences, [2]Department of Natural and Mathematical Sciences,
[3]Department of Biomedical Sciences, State University of Novi Pazar, Vuka Karadžića 9, 36300 Novi Pazar, Serbia

[4]Faculty of Medical Sciences, University of Kragujevac, Svetozara Markovića 69, 34000 Kragujevac, Serbia

Received 19 September 2024; revised 08 April 2025; accepted 01 May 2025

Applying machine learning methods to large datasets with numerous features presents challenges in terms of training time and model complexity. Feature selection is crucial for reducing data dimensions, improving classification accuracy, and optimizing model interpretability. This study aims to enhance the classification of integrated biomedical data to identify thrombophilia diagnosis. The dataset consists of 71 features from 35 women (22 healthy, 13 with thrombophilia), and three classification algorithms (K Nearest Neighbors, Random Forest, Support Vector Machine) are used to evaluate model performance. Identifying key features related to thrombophilia diagnosisis performed using both filter methods and wrapper methods based on swarm intelligence algorithms. Those methods are analyzed and compared as potential approaches for the feature selection process. The wrapper method outperformed the filter methods for clinical and biological data, achieving a classification accuracy of 0.97 compared to 0.91, while selecting only 4 key features compared to 10. For demographic data, both methods produced the same classification accuracy (0.83), but the wrapper method reduced the number of features. These findings demonstrate that wrapper methods based on swarm intelligence algorithms improve model performance and facilitate more efficient data management, which holds significant practical applications for thrombophilia diagnostics. Additionally, the study highlights the advantage of applying the Bat Algorithm in the feature selection process for thrombophilia prediction, contributing to both the novelty and utility of the approach.

Keywords: Biomedical data classification, Feature selection, Machine learning, Swarm intelligence

Introduction

Before Machine Learning (ML) techniques are applied to large datasets with numerous features, it is crucial for the most significant ones to be prioritized to reduce the dataset's size and produce more accurate results. Only important features should be chosen, while ensuring no essential features are excluded.[1] Numerous advantages can be offered by a wise choice of characteristics, such as:

- The cost of data collection is lowered.
- The cost of training classifier models is reduced.
- The model's size is minimized.
- Classification models are made easier to understand.

Certain attributes can negatively impact the classification model; therefore, classification performance can be improved through proper feature selection. Feature selection techniques can be characterized using filter, wrapper, and embedding approaches.[2] However, several alternative and hybrid strategies exist that do not fit into these categories. In filter methods, a rating process is used to evaluate the importance of each feature, and features with poor scores are removed. Filter approaches are quick, scalable, computationally simple, and independent of the classifier. The two main categories of commonly used filter techniques are univariate and multivariate. While individual characteristics are examined separately in univariate methods, multivariate methods consider a subset of features. Unlike filter methods, wrapper methods use classification models to estimate specific subsets of features. Embedded methods are used to select the best subsets of features while simultaneously building suitable classification models.

High-dimensional biomedical datasets are composed of hundreds of features that can be exploited for disease identification, but many irrelevant or weakly correlated features affect diagnostic accuracy. Since the collection of biomedical data is exceedingly labor-intensive, every bit of data is considered valuable. A substantial

*Author for Correspondence
E-mail: apljaskovic@np.ac.rs

impact is made on the speed of decision-making and the likelihood of an accurate diagnosis when the number of procedures needed for diagnosis is reduced.[2,3] By selecting features strongly correlated with the target variable, the size of the dataset that requires processing is decreased, making data collection easier. Many characteristics harm classification prediction; therefore, precision is improved by excluding them.

This research is focused on the determination of diagnostic biomarkers in pregnancy. Because pregnant women are one of the most sensitive groups, it is crucial for issues to be recognized as soon as they arise. Thrombophilia in pregnancy is one such issue that may occur. A more accurate method for determining whether a pregnant woman has thrombophilia is sought in this study. Biomedical (clinical and biological) and demographic information was gathered for this purpose.

The aim of this research is to propose a procedure for selecting the most significant features in a dataset and generate an efficient model for predicting thrombophilia in pregnant women. Both univariate filter methods and wrapper methods based on swarm intelligence algorithms are investigated to find the most suitable feature selection procedure. The classification of pregnant women is done using available biomedical and demographic data, with three classification methods: K Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM).

The structure of the paper is as follows: an overview of related studies on feature selection, with a particular focus on biomedical data, is first presented, followed by a description of the dataset and the applied methodology. The results obtained are then presented and discussed, followed by conclusions, study limitations, and an overview of future research.

Literature Review

Feature Selection (FS) has proven to be a critical preprocessing step for handling high-dimensional datasets, reducing the complexity of models and improving their performance. The relevance of FS in biomedical data, particularly for enhancing classification accuracy and interpretability, has been well-established. Previous studies[3–9] provided crucial insights that shaped the direction of this research.

This research was inspired by earlier applications of FS methods, such as those used to handle datasets with a high number of variables.[3] These studies underscored the challenge of weakly relevant or irrelevant features in high-dimensional biomedical data, which can inflate both time and spatial complexity while degrading predictive accuracy. The approach in the current study builds upon this understanding by refining feature selection strategies specifically for clinical datasets.

Research into feature importance of colorectal cancer phenotypes[5] highlighted the role of FS in improving classification models for clinical data.

Furthermore, studies on the role of statistical FS in pregnancy-related diagnoses[9], where fewer selected features led to improved model performance, provided a basis for comparative analysis between filter- and wrapper-based methods. Inspired by their success in achieving high accuracy with reduced features, similar benefits in the current high-dimensional clinical datasets are explored.

During the last decade, increased interest in the application of Swarm Intelligence (SI) algorithms in the feature selection process has been observed among ML researchers and practitioners. Many algorithms are proposed for various application contexts: artificial bee colony[10], bacterial foraging optimization algorithm[11], bat algorithm[12], monarch butterfly optimization[13], crow search algorithms[14], cuckoo search[15], firefly algorithm[16], grey wolf optimizer[17]. Swarm intelligence algorithms are also extensively used for feature selection problems in the biomedical field, for medical diagnosis and predictions of various diseases.[18–20]

A systematic study and review[21] has described the application of metaheuristics for feature selection in disease diagnosis. The study analyzed 173 papers from various databases such as Scopus, Web of Science, and PubMed. Ten metaheuristics algorithms were investigated, including the cuckoo search algorithm, ant lion optimization, bat algorithm, grey wolf algorithm, whale optimization, and dragonfly optimization, for their ability to select the optimal set of features and predict diseases such as heart disease, Alzheimer's disease, diabetes, chronic disease features, liver disease, etc. It was concluded that metaheuristics could be successfully used in medical diagnosis due to their computational efficiency and accuracy. Most studies related to feature selection are based on considering only one swarm intelligence algorithm. In this research, three different algorithms are introduced and compared in terms of their application in the feature selection process. In addition, proposed swarm optimization algorithms are

compared to statistical, filter-based methods, to find the most suitable method for feature selection in biomedical data classification, i.e. for predicting thrombophilia in pregnant women. To the best of authors' knowledge, there has been no single comprehensive research on applying or enhancing a ML approach for thrombophilia prediction in pregnant women or using a swarm intelligence-based approach in the feature selection process that precedes building an efficient ML model.

Materials and Methods

Subjects were selected from pregnant women who were accommodated at the Gynecology-Obstetrics Clinic of the University Clinical Center Kragujevac for treatment or labor. Two groups of pregnant women were included in the study:

1) Healthy pregnant women (control group);
2) Pregnant women with thrombophilia.

Thrombophilia in pregnant women was diagnosed based on confirmed mutations by molecular diagnostic methods: FV Leiden (G1691A), FII (G20210A), MTHFR (C677T) and PAI-1. The Ethics Committee of the University Clinical Center Kragujevac approved the study protocol, and all patients gave informed consent.

The dataset is composed of 35 women for whom demographic and biomedical data were collected, categorized as either healthy pregnant women or pregnant women diagnosed with thrombophilia (22 healthy and 13 with thrombophilia). Biomedical and demographic information from the patients was gathered. Clinical and biological data were obtained through anamnesis, blood and urine analysis, and ultrasound examination. Demographic data was collected using a questionnaire covering demographic and lifestyle aspects. All biomedical and demographic attributes collected in the study are categorized and summarized in Table 1.

This study aims to create an efficient classification model to categorize pregnant women into one of the two groups (healthy women and women with thrombophilia) identified in the initial dataset. Feature selection process is employed to identify relevant features and remove less significant ones, thus enhancing the model's characteristics. Removing redundant data and reducing the number of attributes decreases model complexity, speeds up training, and minimizes the likelihood of making decisions based on noise.

In feature selection methods (FS methods) used for classification, two distinct methodological approaches are employed: filters and wrappers. These methods differ in terms of computing complexity, metric evaluation, and whether feature selection is integrated with or independent from the learning algorithm, as well as their ability to identify feature reduction and interaction.[22] Statistical methods have been applied for the filter approach, while swarm intelligence algorithms have been utilized for the wrapper approach.

Filter Approach

The filter methods are used to select features by estimating feature relevance using statistical and

Table 1 — Categorization of input attributes for feature selection and classification

Category	Attributes
Biomedical numeric data	Maternal age, Week of gestation, Maternal weight, Maternal height, Maternal weight gain during pregnancy, Fetal biparietal diameter (BPD), Fetal head circumference (HC), Fetal abdomen circumference (AC), Fetal femur length (FL), Estimated fetal weight (EFW), Myometrium thickness, Fetal cerebellum length, Placenta thickness, Maternal blood type, Maternal Rh factor, Maternal hematocrit (Hct), Maternal systolic blood pressure, Maternal diastolic blood pressure, Maternal erythrocyte count, Maternal leukocyte count, Maternal hemoglobin concentration, Maternal platelet count, Maternal iron (Fe) concentration, Maternal creatinine concentration, Maternal urea concentration, Maternal prothrombin time, Maternal international normalized ratio (NR), Maternal activated partial thromboplastin time (aPTT), Fetal heart rate, Maternal heart rate, Umbilical cord diameter, Fetal middle cerebral artery resistance index (MCARI), Uterine artery resistance index (AURI), Umbilical artery resistance index (UARI), Umbilical cord wall thickness, Fetal peak systole, Fetal abdominal artery resistance index (AARI)
Numeric demographic data	Number of people in the household, Material status satisfaction, Number of children, Pregnancy order, Spontaneous miscarriages, Length of the partner relationship, Number of sisters and brothers, Number of cigarettes smoked before pregnancy, Number of cigarettes smoked during pregnancy, Covering up, Adaptation, Toleration, Affective style
Nominal demographic data	Marital status, Education, Type of job, Place of living, Type of dwelling, Apartment ownership, Proximity to relatives, Children from marriage, planned pregnancy, Desired pregnancy, Mode of pregnancy, Pregnancy outside of marriage, Condom use, Marital status of parents, Upbringing in the family, Childhood experiences, Smoking before pregnancy, Smoking during pregnancy, Healthy lifestyles, Need for psychological support

probabilistic data properties, independently of ML algorithms. One of the techniques employed is the SelectKBest class from the Python scikit-learn library. This class integrates various statistical tests to identify a specified number of top features. Specifically, the top 10 features were selected. This method is used for the univariate selection of attributes based on their significance to the target variable.

In this approach, different statistical tests were applied based on the type of data:

- Chi-squared (chi2) test: This test was employed to assess the importance of categorical features (such as demographic data). It measures the dependency between variables in contingency tables, helping to pinpoint the most influential demographic attributes.
- Analysis of Variance (ANOVA): ANOVA was utilized for numerical data (biomedical analysis). This statistical test evaluates whether significant differences exist in the means of numerical variables across groups, thereby identifying crucial biomedical attributes.

By integrating these statistical methodologies, attributes that contribute most significantly to the current analyses were robustly identified and selected, ensuring a focused and data-driven approach to feature selection.

Wrapper Approach

In the current study, swarm intelligence algorithms are employed for the implementation of a wrapper method for feature selection. Swarm intelligence algorithms belong to a class of optimization algorithms called metaheuristics. The idea for their implementation is obtained by observing and imitating the behavior of social animals, such as birds, ants, and whales. Due to their efficiency, these algorithms are used to solve NP-hard problems in various fields, including electrical and civil engineering, communication, finance, transportation, logistics, and healthcare.

Selecting the optimal subset of features is recognized as an NP-hard optimization problem. The no free lunch theorem (NFL) proves that there is no universal optimization algorithm suitable for solving all types of problems. Consequently, no single swarm intelligence algorithm is able to find the optimal set of features for all contexts and applications. In the current study, three well-known algorithms that have been extensively used in this context[23] are proposed. These are Ant Lion Optimizer (ALO), Bat Algorithm (BA), and Gray Wolf Optimizer (GWO). Each algorithm is combined with two classifiers commonly used in these scenarios: Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). The classification metric used is the F1 score, and each algorithm is run 30 times.

Ant Lion Optimizer

This algorithm is inspired by the unique hunting behavior of antlions. Ants move in nature looking for food and antlions hunt them using specifically designed traps. Ants' movement in search space is modeled by normalized random walk[24]:

$$X_i^t = \frac{(X_i^t - a_i) * (d_i - c_i^t)}{(d_i^t - a_i)} + c_i \quad \ldots (1)$$

where, a_i and b_i represent the minimum and maximum of random walk in the *i*-th variable, respectively, while c_i^t and d_i^t are the minimum and maximum of the *i*-th variable at the *t*-th iteration, respectively.

Ants and antlions are evaluated using corresponding fitness functions. Antlions build traps proportional to their fitness, and those with higher fitness have a higher probability of catching ants. Ant will be caught by the antlion if it is fitter than him. Afterwards, the antlion's position is changed to allow him to catch another ant in the next iteration. A roulette wheel operator is employed to select antlions. Additionally, every ant's random walk in each iteration is an average of two values: random walk around selected antlion and elite antlion.

Bat Algorithm

BA was developed by imitating the behavior of bats, specifically their echolocation ability. Bats use echolocation to detect prey or avoid obstacles. They emit strong and short pulses of a certain frequency and listen for echoes coming from objects in the environment. BA uses the idealized characteristics of echolocation, which can be expressed by the following three rules[25]: 1) All bats use echolocation for distance detection, and they "know" the difference between prey and obstacles; 2) Bats fly randomly with velocity speed v_i, at position x_i, emitting sound of fixed frequency f_{min}, variable wavelength λ, and strength A_0. They can update the wavelength (or frequency) of the emitted sound signals and rate of pulse emission r; 3) signal strength (loudness) varies between large value A_0 and minimum A_{min}.

Rules for updating frequency, velocity and positions are given by the following equations:

$$f_i = f_{min} + (f_{max} - f_{min}) * \beta \quad \ldots (2)$$

$$v_i^t = v_i^{t-1} + (x_i^t - x^*) * f_i \quad \ldots (3)$$

$$x_i^t = x_i^{t-1} + v_i^t \quad \ldots (4)$$

Grey Wolf Optimizer

GWO[26] is a swarm intelligence optimization algorithm that mimics hunting techniques and the social hierarchy of grey wolves. Grey wolf hunting is conducted in three phases[27]: 1) Chasing, and approaching the prey; 2) Encircling and harassing the prey until it stops moving; and 3) Attacking the prey.

The encircling behavior of the gray wolves is given by the following equations[28]:

$$\vec{D} = \left|\vec{C} * \overrightarrow{X_p}(t) - \vec{X}(t)\right| \quad \ldots (5)$$

$$\vec{X}(t+1) = \overrightarrow{X_p}(t) - \vec{A} * \vec{D} \quad \ldots (6)$$

where, $\vec{A}$ and $\vec{C}$ are coefficient vectors, while $\vec{X}$ and $\overrightarrow{X_p}$ are positions of grey wolf and prey, respectively.

At the start, the GWO algorithm creates a random initial population of grey wolves (potential solutions). During the execution of the algorithm, alpha, beta, and delta wolves determine the position of the prey and each grey wolf updates its distance from the prey.

Classification Methods

After the feature selection process, three different ML models are trained using a dataset containing only the selected features and their classification accuracy is calculated. These models are KNN, RF, and SVM. The same procedure is carried out independently for both medical data and demographic data.

Comparisons are made between the classification results obtained using all available attributes and the classification results obtained on the selected sets of attributes chosen with the feature selection methods proposed above. The implementation of the proposed ML methods is conducted using Python and the corresponding ML packages (pandas, numpy).

To compare classification results, the accuracy measure is used. Accuracy is defined as a performance metric representing the percentage of correctly classified instances out of the total number of instances, calculated as the ratio of the sum of true positives and true negatives to the total number of instances. The calculation formula is provided below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots (7)$$

where, TP = True positive; FP = False positive; TN = True negative; FN = False negative.

Results and Discussion

The proposed methods for the selection of attributes were applied to the described dataset. First, a statistical approach was applied to the demographic and biomedical data separately, and then a swarm intelligence approach was applied to the same data. After that, classification using three different methods (KNN, RF, SVM) was conducted on the whole dataset and then on the reduced dataset, considering only extracted features.

Statistical Approach - Results

Demographic data is composed of numerical and nominal values. Numerical values were discretized into ordinal categories, allowing the entire set of demographic data to be viewed as categorical for analysis. Therefore, a CH2 test was applied to analyze them, and the 10 most significant attributes correlated with one of the two groups of pregnant women were obtained. The extracted demographic features for the dataset are displayed in Fig. 1.

Ten features were selected from the available 34 features, namely (feature - score): spontaneous abortion - 19.62, number of cigarettes before pregnancy - 12.31, ordinal number of pregnancy - 7.01, number of cigarettes in pregnancy - 6.33, number of brothers/sisters - 5.02, planned pregnancy - 3.55, smoking before pregnancy - 3.31, number of years of relationship - 2.84, residential building - 2.20, place of living - 1.96.

Biomedical attributes are presented as numerical values, and an ANOVA test was used for them. These results are shown in Fig. 2.

The most significant 10 features extracted from the 37 available in the dataset, along with their assigned scores, are: Iron Fe - 66.08, activated partial thromboplastin time (aPTT) - 36.73, weight gain - 6.13, placenta thickness - 4.86, gestational week - 4.19, myometrium thickness - 3.78, urea - 3.26, hemoglobin - 3.26, peak systole (PS) - 2.84, age - 2.56.

A classification was carried out using all demographic and biomedical variables, and the results were compared with those obtained when the classification was conducted based only on the 10 selected features. The proposed strategy was evaluated using five-fold cross-validation.

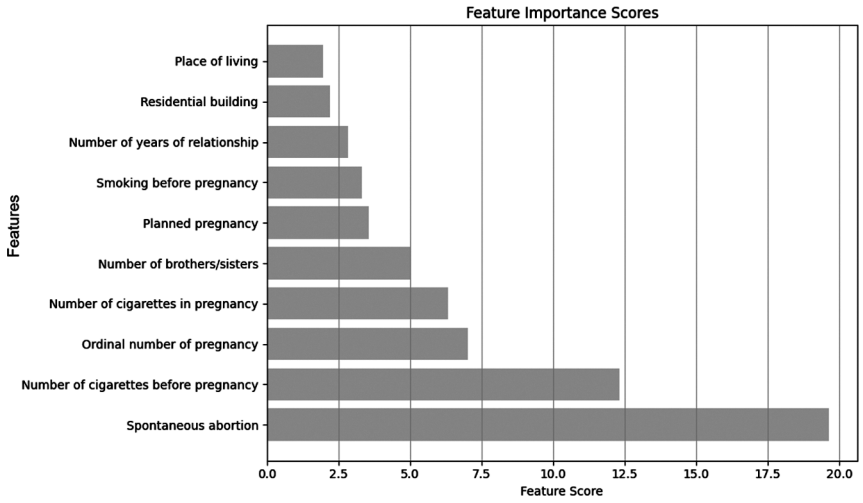

Fig. 1 — Feature score of demographic data

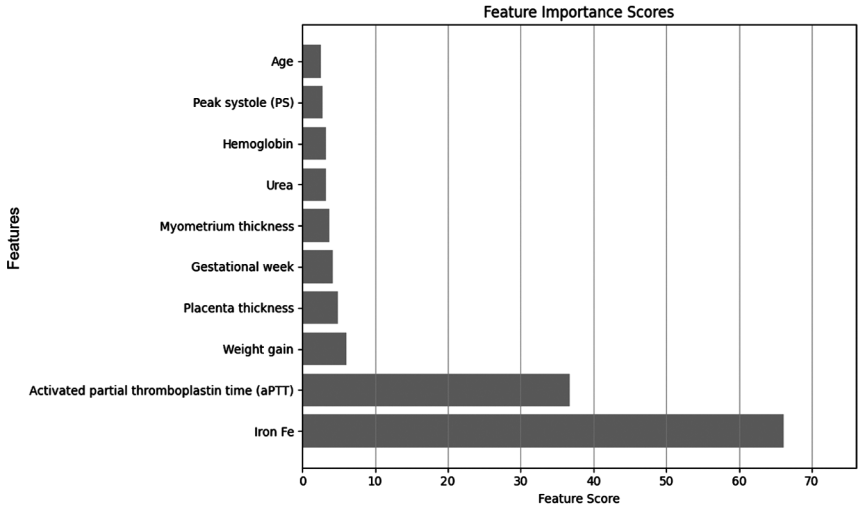

Fig. 2 — Feature's score of biomedical data

Table 2 — Classification accuracy using demographic data with all attributes and selected features

FS method	Features	Classification model based on demographic features			Average results
		KNN	RF	SVM	
chi2	10	**0.8**	**0.83**	**0.8**	0.81
None	34	0.66	0.83	0.63	0.71

Demographic Data

The results of applying statistical methods to two classes of healthy pregnant women and pregnant women with thrombophilia have been considered and described in Table 2. The table presents the model accuracy obtained by applying classification to all attributes compared to classification based only on the selected attributes. The RF method applied to the selected 10 attributes that yielded the best results. In all cases, the classification results were the same or better on the dataset with reduced attributes in comparison to the original dataset. It can also be observed that the average accuracy (0.81) of the three models with a reduced number of attributes is higher than the average accuracy (0.71) of the model based on all attributes.

Biomedical Data

The classification using biomedical attributes is shown in Table 3, where the best results were obtained by the RF method. In the case of the KNN and SVM

Table 3 — Classification accuracy using biomedical data with all attributes and selected features

FS method	Features	Classification model based on clinical and biological features			Average results
		KNN	RF	SVM	
ANOVA	**10**	**0.83**	0.91	**0.63**	0.79
None	**37**	0.46	**0.94**	0.63	0.68

Table 4 — Number of selected features and performance evaluation - demographic data

Algorithm	Features	Classification model based on demographic features			Average results
		KNN	RF	SVM	
ALO-KNN	**51**	0.714	**0.743**	**0.743**	0.733
ALO-SVM	**53**	0.743	0.686	**0.800**	0.743
BA-KNN	**4**	0.600	0.657	**0.686**	0.648
BA-SVM	**2**	0.800	**0.829**	0.800	**0.810**
GWO-KNN	**1**	**0.629**	0.571	**0.629**	0.610
GWO-SVM	**2**	0.771	**0.743**	**0.743**	0.752

methods, better classification results were obtained on the reduced dataset. Less precise results were obtained on the reduced dataset compared to the whole dataset, by the RF method. Additionally, on this data, the average accuracy (0.79) of the three models with a reduced number of attributes is higher than the average accuracy (0.68) of the model based on all attributes. These results justify the application of the feature selection method before the classification model is created.

Swarm Intelligence Approach - Results

Demographic Data

The best results from 30 runs of the SI algorithms, including the number of features selected by each algorithm and performance evaluation results, are presented in Table 4. The BA algorithm selects 4 or 2 features, the GWO algorithm selects 1 or 2 features, while the ALO algorithm selects 51 or 53 features, depending on the classifier used. It can be noticed that the reported number of features in the case of the ALO algorithm is higher than the number of features in the original demographic subset. This is due to the application of the one-hot encoding procedure to categorical features, which increases the total number of features (columns) in the dataset.

The results of the data analysis presented in Table 4 demonstrate that the BA algorithm outperforms other SI algorithms in terms of average classification accuracy when used with an SVM classifier in the feature selection process. To be more precise, the average accuracy for the ALO algorithm is 0.733 and 0.743 across three different ML models; for the GWO algorithm, it is 0.610 and 0.752; and for the BA algorithm, it is 0.648 and 0.810. The BA-SVM algorithm selects only two features: miscarriage and the number of brothers and sisters.

Moreover, the BA algorithm with SVM exhibits equal performance as statistical methods in terms of classification accuracy (0.81), but it selects only 2 features, while the statistical method uses the 10 most significant features.

Biomedical Data

The same approach for data presentation was followed as in the previous section. The BA algorithm and GWO are again more successful in selecting smaller numbers of features, thereby reducing the dataset size. The BA algorithm selects 3 or 4 features, the GWO selects 3 or 1 feature, while the ALO algorithm selects 32 and 33 features, depending on the classifier used. Average classification accuracy calculated from the data provided in Table 5 offers strong evidence of the superiority of the BA algorithm compared to other swarm intelligence algorithms. Specifically, for the ALO algorithm, average accuracy across three different ML models is 0.790 and 0.753; for the GWO algorithm, 0.771 and 0.657 is obtained; while for the BA algorithm, it is 0.876 and 0.924.

It is noted that the best results are obtained when the BA algorithm is used with an SVM classifier. In this case, four features are selected: fetal femur length, blood group, creatinine, and aPTT (activated partial thromboplastin time).

Next, a comparison of these results with those obtained using statistical methods in the previous section is made, concluding that the BA algorithm with SVM outperforms statistical methods in terms of dataset size reduction and prediction quality. BA algorithm selects four features while giving an average prediction

Table 5 — Number of selected features and performance evaluation - biomedical data

Algorithm	Features	Classification model based on clinical and biological features			Average results
		KNN	RF	SVM	
ALO-KNN	**32**	0.743	**0.857**	0.771	0.790
ALO-SVM	**33**	0.686	**0.829**	0.743	0.753
BA-KNN	**3**	0.857	**0.914**	0.857	0.876
BA-SVM	**4**	0.886	**0.971**	0.914	**0.924**
GWO-KNN	**3**	0.771	0.743	**0.800**	0.771
GWO-SVM	**1**	0.657	**0.686**	0.629	0.657

accuracy of 0.924, while the statistical method uses the ten most significant features and yields average classification accuracy of 0.79.

Compared to related studies in the domain of clinical data, the results of this study for feature selection (FS) are comparable, with SI methods demonstrating their superiority over statistical methods. The additional value of these findings lies in their contribution to the detection of thrombophilia, an area with a notable scarcity of studies. This is particularly important given the challenge of accessing large amounts of individual data. The methods employed in the current study not only provide better classification results but also require significantly fewer data.

Moreover, the inclusion of demographic data alongside clinical data in thrombophilia prediction is another novelty of study, distinguishing it from previous research. This combined approach has the potential to improve prediction accuracy with fewer variables, offering a novel direction in thrombophilia research.

Conclusions

This study demonstrates the effectiveness of SI algorithms (particularly BA) in selecting significant features from demographic and biomedical data of pregnant women, outperforming statistical approaches with fewer selected features. The findings suggest that feature selection improves classification performance, especially in predicting thrombophilia, by using a reduced set of variables. An important innovation in this work is the combination of clinical and demographic data, which enhances prediction accuracy despite the limited availability of patient data.

However, the limitation of this study is the relatively small dataset, which affects the generalizability of the results. Future research will focus on training models with larger datasets and exploring enhanced versions of metaheuristic algorithms to further improve classification accuracy. The proposed method holds potential for real-world applications in thrombophilia detection, reducing the need for extensive data collection while maintaining high predictive accuracy.

References

1 Lyu Y, Feng Y & Sakurai K, A survey on feature selection techniques based on filtering methods for cyber attack detection, *Inform*, **14(3)** (2023) 191.

2 Khaire U M & Dhanalakshmi R, Stability of feature selection algorithm: A review, *J King Saud Univ - Comput Inf Sci*, **34(4)** (2022) 1060–1073.

3 Chen R C, Dewi C, Huang S W & Caraka R E, Selecting critical features for data classification based on machine learning methods, *J Big Data*, **7(1)** (2020) 52.

4 Koduru A, Valiveti H B & Budati A K, Feature extraction algorithms to improve the speech emotion recognition rate, *Int J Speech Technol*, **23(1)** (2020) 45–55.

5 Sulistyawan J S, Julian K, Elwirehardja G N, Nugroho K S & Pardamean B, A two-step feature selection approach for identifying SNPs associated with colorectal cancer, *Commun Math Biol Neurosci*, **2024** (2024).

6 Subasi A & Mian Qaisar S, Signal acquisition preprocessing and feature extraction techniques for biomedical signals, In *Adv Non-Invasive Biomed Signal Sens Process Mach Learn* edited by S M Qaisar, H Nisar and A Subasi (Cham: Springer International Publishing) 2023, 25–52.

7 Al-Rajab M, Lu J & Xu Q, A framework model using multifilter feature selection to enhance colon cancer classification, *Plos One*, **16(4)** (2021) e0249094.

8 Dash R, Dash R & Rautray R, An evolutionary framework based microarray gene selection and classification approach using binary shuffled frog leaping algorithm, *J King Saud Univ - Comput Inf Sci*, **34(3)** (2022) 880–891.

9 Marovac U, Memić L, Avdić A, Djordjević N, Dolićanin Z & Babic G, Selecting critical features for biomedical data classification, In *2nd Int Conf Chemo Bioinform* (University of Kragujevac) 28-29 September 2023.

10 Li H, Pun CM, Xu F, Pan L, Zong R, Gao H & Lu H, A hybrid feature selection algorithm based on a discrete artificial bee colony for Parkinson's diagnosis, *ACM Trans Internet Technol*, **21(3)** (2021) 1–22.

11 Niu B, Yi W, Tan L, Geng S & Wang H, A multi-objective feature selection method based on bacterial foraging optimization, *Nat Comput*, **20** (2021) 63–76.

12 Hambali M A, Oladele T O, Adewole K S, Sangaiah A K & Gao W, Feature selection and computational optimization in high-dimensional microarray cancer datasets via InfoGain-modified bat algorithm, *Multimed Tools Appl*, **81(25)** (2022) 36505–36549.

13 Alweshah M, Khalaileh S A, Gupta B B, Almomani A, Hammouri A I & Al-Betar M A, The monarch butterfly optimization algorithm for solving feature selection problems, *Neural Comput Appl*, **34** (2022) 11267–11281.

14 Anter A M & Ali M, Feature selection strategy based on hybrid crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm for medical diagnosis problems, *Soft Comput*, **24(3)** (2020) 1565–1584.

15 Pandey A C, Rajpoot D S & Saraswat M, Feature selection method based on hybrid data transformation and binary binomial cuckoo search, *J Ambient Intel Humaniz Comput*, **11(2)** (2020) 719–738.

16 Zhao J, Lv S, Xiao R, Ma H & Pan J S, Hierarchical learning multi-objective firefly algorithm for high-dimensional feature selection, *Appl Soft Comput*, **165** (2024) 112042.

17 Hu P, Pan J S & Chu S C, Improved binary grey wolf optimizer and its application for feature selection, *Knowl-Based Syst*, **195** (2020) 105746.

18 Anter A M & Ali M, Feature selection strategy based on hybrid crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm for medical diagnosis problems, *Soft Comput*, **24(3)** (2020) 1565–1584.

19 Canayaz M, MH-COVIDNet: Diagnosis of COVID-19 using deep neural networks and meta-heuristic-based feature selection on X-ray images, *Biomed Signal Process Control*, **64** (2021) 102257.

20 Meenachi L & Ramakrishnan S, Metaheuristic search based feature selection methods for classification of cancer, *Pattern Recognit*, **119** (2021) 108079.

21 Kaur S, Kumar Y, Koul A & Kumar Kamboj S, A systematic review on metaheuristic optimization techniques for feature selections in disease diagnosis: Open issues and challenges, *Arch Comput Methods Eng*, **30(3)** (2023) 1863–1895.

22 Pudjihartono N, Fadason T, Kempa-Liehr A W & O'Sullivan J M, A review of feature selection methods for machine learning-based disease risk prediction, *Front Bioinform*, **2** (2022) 927312.

23 Nssibi M, Manita G & Korbaa O, Advances in nature-inspired metaheuristic optimization for feature selection problem: A comprehensive survey, *Comput Sci Rev*, **49** (2023) 100559.

24 Mirjalili S, The ant lion optimizer, *Adv Eng Softw*, **83** (2015) 80–98.

25 Yang X S & Hossein G A, Bat algorithm: a novel approach for global engineering optimization, *Eng Comput*, **29(5)** (2012) 464–483.

26 Karaoglan A D, Optimizing plastic extrusion process via grey wolf optimizer algorithm and regression analysis, *J Sci Ind Res*, **80(01)** (2021) 34–41.

27 Muro C, Escobedo R, Spector L & Coppinger R P, Wolf-pack (canis lupus) hunting strategies emerge from simple rules in computational simulations, *Behav Process*, **88(3)** (2011) 192–197.

28 Mirjalili S, Mirjalili S M & Lewis A, Grey wolf optimizer, *Adv Eng Softw*, **69** (2014) 46–61.