Examining the Various Neural Network Algorithms Considering the Superiority of Mouth Brooding Fish in Data Classification

Lang Liu*, Yong Zhu
Department of Information Engineering, Gongqing College of Nanchang University, Gongqingcheng 332020, Jiangxi, China

Abstract—**Data classification, a crucial practice in information management, involves categorizing data based on its sensitivity to determine appropriate access levels and protection measures. This paper explores the utilization of novel algorithms, including mouth-brooding fish (MBF), alongside machine learning techniques, for the analysis of medical health data. The SVM exhibits suboptimal performance in the task of data categorization. Therefore, Adaboost may be considered a viable substitute for MBF due to its superior performance in terms of F-score, accuracy, specificity, and sensitivity. The accuracy of MBF, which stands at about 95%, surpasses that of Adaboost by a significant margin of 77%. The F-score, accuracy, and specificity values obtained for MBF are exceptional when compared to the other chosen models, with values of 97.17%, 93.6%, and 96.5%, respectively. The proposed algorithm exhibits promising advancements in health data categorization, offering a potential breakthrough in data classification methodologies. Leveraging this innovative approach could facilitate more accurate and efficient management of sensitive medical data, thereby enhancing healthcare systems' capabilities for data protection and analysis. The main novelty of this study lies in the introduction and evaluation of the MBF algorithm for data classification within the medical domain. Unlike traditional algorithms, MBF draws inspiration from the collective behavior of mouth-brooding fish, offering a unique optimization strategy that enhances both exploration and exploitation of the solution space. This novel approach presents a promising avenue for advancing healthcare analytics and decision-making processes.**

Keywords—Medical data analysis; clinical decision support; dataset classification; Mouth Brooding Fish; Support Vector Machine (SVM)

I. Introduction

Different signaling pathways for various biological activities are formed inside the cell by the interconnection and interaction of various signals. Mutations in the gene that controls these processes result in cellular malfunction and may potentially cause cancer [1]. The term "driver pathway" or "driver gene set" often refers to the group of altered genes highly influential in cell signaling pathways. In addition to deepening our knowledge of the rules of molecular action and the processes behind cancer development, the discovery of driver pathways may potentially point to novel molecular targets for cancer therapy. It is commonly recognized that several genetic variants can affect the same pathways [2]. To better capture the diverse patterns of malignancies, it is, essential to go from the gene to the pathway level. At the route level, several investigations have discovered patterns of mutations [3]. One method used to forecast the state of civil infrastructure is the health monitoring of structures [4]. The weather and functional condition fluctuations threaten the accuracy of damage detection work during continuous monitoring in the bridge structural health monitoring system [5].

Digital medical technology has matured due to information technology advancements, medical data is expanding at a never-before-seen rate, and biomedical research has transformed into a typical data-intensive discipline, giving rise to the phenomena known as "big data." The significant data age has transformed biomedical research, human thought processes, and way of life. Data is becoming a new strategic resource and a significant driver of innovation. Relevant medical industry departments can be guided to strengthen the collection and management of big data related to medical health through the integration analysis and application requirements description of big data in the medical service field. This will lay the groundwork for future data development and application [6, 7].

Thousands or even hundreds of thousands of MAs have been developed during the decades-long history of modern optimization for use in various sectors; natural phenomena inspire most of these MAs. Since its inception in the 1960s, genetic algorithms (GAs) have undergone three stages of development: the concept-proposal stage, the OP-growth stage, and the mature stage of evolving towards depth [8]. The traditional medical health big data classification algorithms face challenges, including high sample size and delayed processing, as the amount of medical and health care data continues to expand steadily. The Mouth Brooding Fish (MBF) algorithm is adjusted to more accurately categorize the imbalanced data set. The MBF algorithm replicates the mutualistic Organisms that use biotinteraction strategies to live and spread across the environment. In this study, the MBF algorithm is studied. Overfitting will not occur since the MBF eliminates noise from the training data set based on the ensemble learning concept. According to the simulation findings, this approach outperforms Gaussian Kernel, Random Forest (RF), Adaboost, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) in spotting dishonest behaviors [2]. This is a crucial point of reference for developing the medical credit scheme. The primary objective of our study was to construct an appropriate model for the provided professorial scenario. Indeed, given the potential for a model or structure to exhibit superiority in any given application or case study, the primary objective was to ascertain the most suitable fit for the given dataset. In addition, we attempted to use the most renowned and extensively utilized machine learning models as comparator models. The superiority of the current work over its counterpart in the previous years is highlighted as follows:

- Introduction of novel algorithms, particularly MBF, for the analysis of medical health data, demonstrating superior performance compared to traditional methods like SVM.

- Comparative evaluation of MBF and Adaboost algorithms, revealing MBF's exceptional accuracy of approximately 95%, surpassing Adaboost by a substantial margin of 77%.

- Detailed analysis of performance metrics including F-score, accuracy, specificity, and sensitivity, showcasing MBF's outstanding performance with F-score, accuracy, and specificity values of 97.17%, 93.6%, and 96.5% respectively, thereby highlighting its superiority over other selected models.

- Significance of the proposed algorithm in advancing health data categorization, offering promising advancements in data classification methodologies, and facilitating more accurate and efficient management of sensitive medical data, thereby enhancing the capabilities of healthcare systems in data protection and analysis.

The rest of the paper is organized as follows: The second section reviews the related works to highlight the significant limitations and drawbacks tackled in the current work. The methodology and dataset adopted for reaching the conclusions are explained in the third section. The results are discussed in the fourth section, and the conclusions are drawn in the fifth section.

II. Literature Review

Researchers have been experimenting with various data mining approaches in the medical and health domains to increase the accuracy of medical diagnoses. Additional reliable and accurate methods would yield additional supporting information for identifying potential patients through precise sickness forecasting. Data mining techniques play a significant part in clinical decision-making by creating various models that give doctors precise, dependable, and timely forecasts [9]. Reducing the number of datasets in the healthcare industry while considering data categorization methods based on meta-heuristic algorithms has drawn much interest in recent years. A few examples are the enhanced KNN method presented by Xing and Bei [10] and their comparison with the conventional KNN algorithm. Weights are allocated to each class, and the classification is carried out in the standard KNN classifier's query instance neighborhood. The method considers the distribution of classes surrounding the query to guarantee that the allocated weight does not negatively impact the outliers. Boyapati et al. [11] concluded that the Support Vector Machine approach was better than the Decision Tree algorithm, providing a preferred dataset distribution or categorization. By accounting for the multimodal distribution of the numerical variables, Khanmohammadi and Chou's novel Gaussian Mixture Model-based Discretization Algorithm (GMBD) maintained the most common patterns from the original dataset [12]. Six publicly accessible medical datasets confirmed the GMBD algorithm's efficacy. The experimental findings showed that the GMBD algorithm performed better than regarding the number of rules produced and the classification precision in the associative classification algorithm; there are five more static discretization techniques. Chang et al. presented a model that combines a cross-validation technique, a classification algorithm, and recursive feature removal. The authors ranked each feature's relevance using the recursive feature elimination approach in the first stage, and then they utilized cross-validation to identify the best feature subset. In order to reliably forecast patient outcomes using their ideal features subset, four classification algorithms—SVM, C4.5 decision tree (RF), extreme gradient boosting (XGBoost), and others—were examined in the second stage. Of the quartet of classifiers, using the optimum features subset, XGBoost demonstrated the best prediction performance with accuracy, F1, and area under receiver operating characteristic curve (AUC) values of 94.36%, 0.875, and 0.927, respectively. Table I also summarizes similar research according to the methods and objectives employed.

TABLE I. A Brief Review of the Related Works Based on the Used Techniques and Purposes

No.	References/Year	Method	Aim	Features
1	[13]/2019	Random Forest classifier	Medical data classification	Highly accurate predictors were provided for ten different diseases, along with a sufficiently generic technique that should work well for other diseases with comparable datasets. Highly accurate predictors were provided for ten different diseases, along with a sufficiently generic technique that should work well for other diseases with comparable datasets.
2	[14]/2021	Decision tree classifiers	Medical data classification	In terms of authenticity and correctness, the suggested approach seemed appropriate.
3	[15]/2020	Modified nearest neighbor (ENN) based on RF and misclassification-oriented synthetic minority over-sampling approach (M-SMOTE)	addressing the blindness of the over-sampling method for synthetic minorities while creating samples	Comprehensive tests on 10 UCI datasets show that RFMSE helps address unbalanced data categorization. The suggested technique is more effective in improving F-value and MCC than standard methods.
4	[16]/2020	Grey Wolf Optimization (GWO) method with Hybrid Kernel SVM	Classification of data for chronic renal illness	According to the latest results, the intended classification scheme outperformed, achieving improved 97.26% accuracy for the renal chronic dataset compared to the 94.77% achieved by the existing SVM approach and the 93.78% achieved by the fuzzy min–max GSO neural network (FMMGNN) classifier.

5	[17]/2019	A unique code division multiplexing (CDM) and block classification-based reversible data hiding (RDH) method.	Block categorization for healthcare system image processing	The suggeAccording to experimental data, the approach can produce a superior overall performance on medical photos than other cutting-edge RDH systems, accordi
	[18](Yadav and Jadhav 2019)[18]2019	Deep convolutional neural networks for the categorization of medical images	Classifying pneumonia by analyzing a dataset of chest X-rays	When applied to a short dataset, transfer learning outperforms support vector machines with oriented fast and rotated binary (ORB) robust independent elementary features and capsule networks regarding classification accuracy.
7	[19]/2021	An approach for adaptive harmony search	Selecting genes and categorizing high-dimensional medical data	According to the simulation results, the suggested hybridization has great promise for high-dimensional database feature subset prediction and sample classification.
8	[20]/2023	An algorithm for the modified Hunger Games search (mHGS)	Selection of features and worldwide optimization	The experimental findings imply that the suggested mHGS can improve convergence time and produce useful search results without adding to the computing burden. Additionally, it has enhanced SVM classification performance.

III. Methodology

A. Selected Algorithms

Support Vector Machine (SVM), AdaBoost, Multilayer Perceptron (MLP), Gaussian Kernel, and Random Forest (RF) have been selected for data classification here.

1) Support Vector Machine (SVM): Since the margin in SVM is calculated using the points closest to the hyperplane (support vectors), it is unnecessary to worry about additional observations; in logistic regression, on the other hand, the classifier is defined over all of the points. As a result, SVM naturally speeds faster. SVMs are a group of supervised learning techniques used in regression analysis, outlier identification, and classification. Among support vector machines' benefits are efficient in places with several dimensions. It is still useful when there are more dimensions than samples. The spots that are nearest to the hyperplane are these. These data will be used to define a separation line. The distance between the hyperplane and the observations (support vectors) that are closest to it is known as the margin. A big margin is considered good in SVM [21].

One sparse approach is SVM. Like nonparametric techniques, SVM necessitates the availability of all training data, meaning that it must be kept in memory during the training phase when the SVM model's parameters are discovered. Nevertheless, SVM relies solely on a subset of these training examples—referred to as support vectors—for subsequent prediction once the model parameters have been determined. The support vectors specify the hyperplanes' boundaries. Following Support vectors are identified following phase with an objective function regularized by an error term and a constraint, supporting relaxation is used. Rather than the dimensionality of the input space, the number of support vectors determines the complexity of the SVM classification job. Data-dependent and variable, the number of support vectors that are eventually kept from the original dataset depends on the data complexity, represented by the data dimensionality and class separability. Although, in reality, this is rarely the case, the maximum constraint for the number of support vectors is half the size of the training dataset [22].

2) Adaboost: The AdaBoost algorithm, also called Adaptive Boosting, is a machine-learning ensemble method that uses boosting techniques. Because the weights are reassigned to each instance—higher weights are given to instances that are mistakenly classified—it is known as adaptive boosting. AdaBoost builds the model sequence using a different method than XGBoost, an improved version of Gradient Boosting with various enhancements and improvements. The particular challenge and the application's needs will determine which solution is best [23].

*3) Multilayer Perceptron (MLP):*An MLP neural network is used to model the system. The inputs of an artificial neural network (ANN) are represented by $u(t - z_i)$, where $i=1,2,\dots,n$, and the delay is indicated by z_i [24]. This research reveals the relevant parameters of the first neuron in the hidden layer as $w_{11}^1, w_{12}^1, \dots, w_{1n}^1$. The h-th neuron's related parameters and the hidden layer output are represented by $w_{h1}^1, w_{h2}^1, \dots, w_{hn}^1$, and $w_{21}, w_{22}, \dots, w_{2h}$. Fig. 1 displays the suggested output of the Artificial Neural Network (ANN) [25].

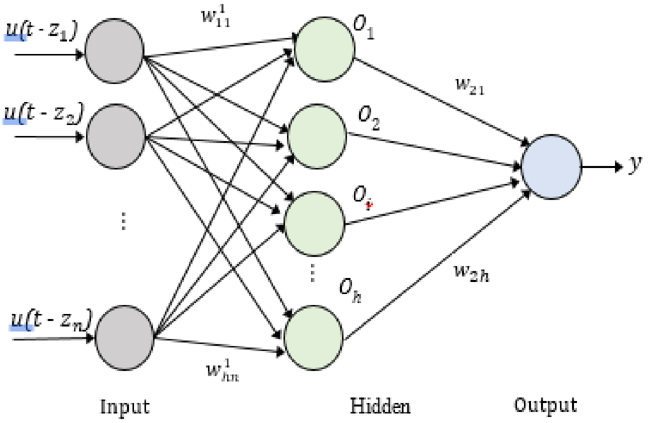


Fig. 1. System modeling utilizing an MLP neural network.

$$n_{ti} = w_i^1 U$$
$$O_i = g(n_{ti}), \quad i = 1,\dots,h \tag{1}$$

Accordingly,

$$w_i^1 = [w_{i1}^1, w_{i2}^1, \dots, w_{in}^1]$$
$$g(n_{ti}) = \frac{1 - exp(-n_{ti})}{1 + exp(-n_{ti})} \tag{2}$$

As stated in Equation 3, the output of ANN is specified.

$$y = w_2 o \tag{3}$$

According to the components of Equation 4, the main parameters are defined as follows:

$$\begin{gathered} o = [o_1, o_2, \dots, o_h]^T \\ w_2 = [w_{21}, w_{22}, \dots, w_{2h}] \end{gathered} \tag{4}$$

According to Equation 5, the major parameters of ANN are adjusted:

$$\begin{gathered} o = [o_1, o_2, \dots, o_h]^T \\ w_2 = [w_{21}, w_{22}, \dots, w_{2h}] \end{gathered} \tag{5}$$

Using Equation 6, the parameters of ANN are adjusted:

$$E = \frac{1}{2} e_{est}^2 = \frac{1}{2}(y_d - y)^2 \tag{6}$$

The approximated /real outputs indicate $/y_d$. According to which the updating law is [26]:

$$w_2(t+1) = w_2(t) + \eta e_{est} o \tag{7}$$

The first layer with the weights adaptive principle is represented by Equation 8:

$$w_i^1(t+1) = w_i^1(t) + \eta e_{est} \dot{g}(n_{ti}) w_{2i} U \tag{8}$$

Assuming that η remains constant, we can represent the vector of weights in the ith neuron as w_i^1 and the vector of weights for the ith neuron output as w_{2i}. The differential of $g(n_{ti})$ is represented by $\dot{g}(n_{ti})$ (concerning the input n_{ti}). Equation 9 is also used to determine the Jacobian of the system.

$$\frac{\partial \Delta f}{\partial u_c} = ([w_{11}^1, w_{21}^1, \dots, w_{h1}^1] diag[\dot{g}(n_{t1}), \dots, \dot{g}(n_{th})] w_2) \tag{9}$$

4) GK: The Gaussian kernel (GK) is defined as follows in one-dimensional, two-dimensional, and neuronal dimensions:

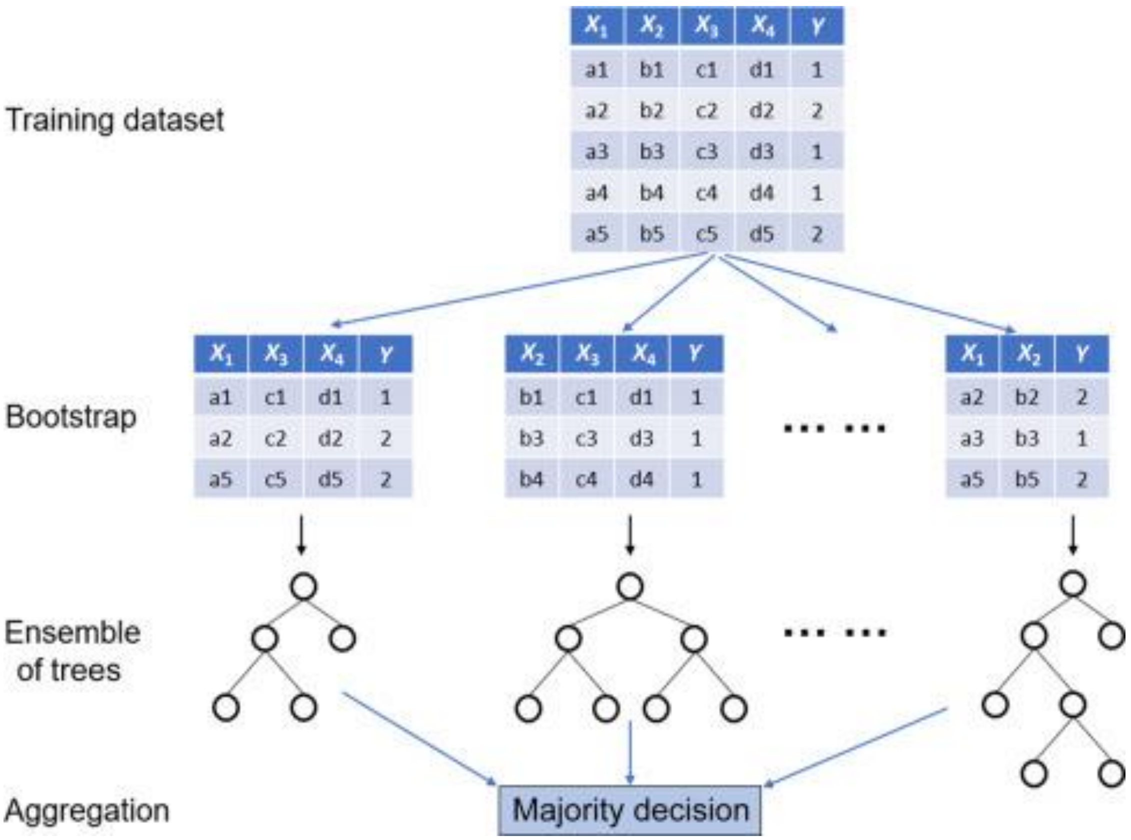
$$\begin{gathered} G_{1\ D}(x;\ \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \\ G_{2\ D}(x, y', \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \\ G_{ND}(\vec{x};\ \sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^N} e^{-\frac{|\tilde{x}|^2}{2\sigma^2}} \end{gathered} \tag{10}$$

The σ value determines the width of the Gaussian kernel. In statistics, the Gaussian probability density function is referred to as the standard deviation, while its square, σ^2, is the variance. When we discuss the Gaussian as an aperture function in observations, we will use "s" to refer to the inner scale or simply the scale. This paper's scale is limited to positive values, where σ > 0. During the observation process, s can never be reduced to zero. This implies observing through a tiny aperture, which is practically impossible. The inclusion of the factor of 2 in the exponent is merely a matter of convention. It allows us to have a more simplified formula for the diffusion equation, which we will discuss in more detail later. The convention is to include a semicolon between the spatial and scale parameters to distinguish between them clearly.

5) RF: The Random Forest (RF) classifier is a method that concurrently trains multiple decision trees using bootstrapping and then aggregates the results through a process known as bagging (Fig. 2) [27]. Bootstrapping involves training distinct decision trees simultaneously on various subsets of the training dataset, utilizing different subsets of the available features. This ensures that each decision tree within the random forest is unique, thereby reducing the overall variance of the RF classifier. The RF classifier amalgamates the decisions of individual trees to arrive at the final decision, enabling it to exhibit robust generalization. Compared to other classification methods, the RF classifier typically attains higher accuracy without succumbing to overfitting issues.

Like the Decision Tree (DT) classifier, the RF classifier does not require feature scaling. However, the RF classifier demonstrates greater resilience in selecting training samples and noise in the training dataset than the DT classifier. Despite being more challenging to interpret, the RF classifier offers ease of hyper parameter tuning compared to the DT classifier.

6) Mouth Brooding Fish (MBF): According to Fig. 3, the MBF algorithm simulates organisms' strategies to ensure their survival and proliferate within an ecosystem through symbiotic interactions [29]. It consists of five control parameters that the user determines. The key factors that influence the cichlid population are the number of cichlids in the group, the location where the mother cichlid originates from (source point or SP), the extent of dispersion, the likelihood of dispersion, and the damping effect on the mother's source point. It is advisable to analyze the problem and review the outcomes of parameter tuning to select the optimal values for the control parameters. In order to compare the MBF algorithm with CMAES, JADE, SaDE, and GL-25, we need to assume that the controlling parameters are constant. The MBF algorithm is population-based, so the number of individuals in the population is one of the parameters that can be controlled. The population size indicates the number of fish that will undergo the problem-solving process in the Mouth Brooding Fish algorithm [30]. The primary foundation of the Mouth Brooding Fish algorithm lies in the behaviors of cichlids as they navigate around their mother, as well as the impact of natural elements or threats on these behaviors. The MBF algorithm consists of several main parts to find the best possible results for the given problems.



X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

Fig. 2. A dataset with two classes (Y = 1) and four features (X1, X2, X3, and X4) is employed to build a Random Forest (RF) classifier. The RF classifier is an ensemble method that simultaneously uses bootstrapping and aggregation to train multiple decision trees. Each tree is trained on unique subsets of training samples and features [28].

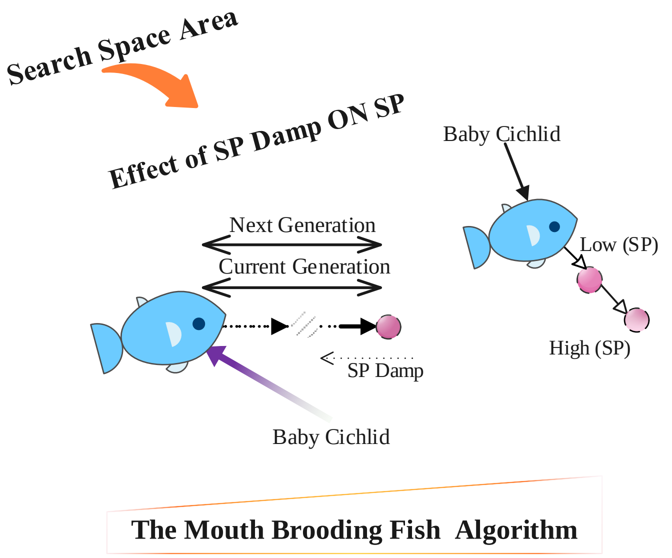


Fig. 3. Mouth Brooding Fish Algorithm [31].

In nature, marriage is a crucial mechanism that aids colonies or populations in achieving optimal outcomes by promoting convergence. However, it only sometimes yields favorable outcomes when it occurs. Mouth-brooding fish allow their best cichlids to mate. Thus, the MBF algorithm selects one pair of parents from each cichlid using a probability distribution or Roulette Wheel selection (where higher point values have a higher likelihood). Cichlids that hatch in a new position replace their parents in the population without moving [32]. Before assessing the fitness of the newly hatched fish using a fitness function, we need to ensure that the new positions for the offspring are within the boundaries of the search space. The mathematical equations of this algorithms are defined below:

1) Objective function: $f(x)$ represent the objective function to be minimized or maximized, where x denotes the vector of decision variables.

2) Mouth-Brooding fish model: The position of each fish (solution) in the search space can be represented as $x_i=[x_{i1},x_{i2},...,x_{id}]$, where i denotes the index of the fish and d is the dimensionality of the problem.

3) Fish movement: The movement of fish i at iteration t is governed by $x_{it} = x_{it-1} + \Delta_{it}$ whew Δ_{it} represents the change in position of fish i at iteration t.

4) Local search mechanism: The local search mechanism could involve exploring the neighborhood of each fish i to find better solutions. This can be represented as adjusting the position of fish i based on its local surroundings: $\Delta x_{it} =$

$\alpha\nabla f(x_i) + \beta\Delta x_{it-1} + \epsilon_t$ where α and β are parameters controlling the influence of the gradient and previous movement, respectively, and ϵt is a random perturbation.

5) Updating rules: The updating rules determine how the positions of fish are updated iteratively. One common approach is to use a simple update rule such as: $x_{it} = x_{it} + \Delta x_{it}$

B. Dataset

The reason for creating this dataset is the necessity for practical and varied healthcare data that can be used for educational and research purposes. Accessing healthcare data for learning and experimentation can be challenging due to its sensitivity and the privacy regulations surrounding it. In order to fill this gap, the Faker library in Python is used to create a dataset that closely resembles the structure and attributes typically seen in healthcare records [33]. We have created this healthcare dataset as a valuable resource for those interested in data science, machine learning, and data analysis. The purpose of this tool is to imitate authentic healthcare data, allowing users to practice, enhance, and demonstrate their abilities in manipulating and analyzing data within the healthcare sector. We can find additional details about the data set in reference [33].

Moreover, the dataset available at the provided Kaggle link offers comprehensive insights into healthcare demographics and outcomes, encompassing various attributes crucial for medical analysis and decision-making. It includes data from diverse sources, capturing demographic information such as age, gender, and ethnicity, alongside clinical details including medical conditions, diagnosis codes, and medication usage. Moreover, the dataset incorporates vital signs measurements, laboratory test results, and insurance details, providing a holistic view of patients' health status and treatment journeys. Additionally, the dataset likely contains information on healthcare utilization, including hospital admissions, procedures performed, and associated costs, facilitating in-depth analysis of healthcare resource allocation and patient care pathways. With its rich and diverse array of variables, this dataset presents a valuable resource for exploring patterns, trends, and associations within the healthcare domain, enabling researchers and practitioners to derive actionable insights for improving patient outcomes and healthcare delivery.

C. Evaluation Criteria

The primary factors for comparing the results are F-score, accuracy, specificity, sensitivity, and precision [34]. Precision refers to a slight variation between two or more measurements, whereas accuracy represents the disparity between a result and its actual value. The end outcomes should align well, as indicated by precision. The F1 score is the weighted average of precision and recall, including false positives and negatives. Specificity is the test's ability to identify unstick people correctly. Mathematically, a test with high specificity that produces a positive result can confirm a disease because it rarely produces positive results in healthy people. A test's sensitivity determines whether it detects a disease. High-sensitivity tests have few false negatives, reducing disease cases missed. The specificity of a test refers to its capability to correctly identify someone who does not have a disease as being negative. To put it differently, Specificity refers to the percentage of individuals who do not have Disease X and receive a damaging result on their blood test. A particular test ensures that all healthy individuals are accurately recognized as healthy, meaning there are no incorrect positive results.

Accuracy is one of the most often utilized measures for classifying data. A confusion matrix determines a model's accuracy by employing the following equation [35].

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

Moreover, precision (P), sensitivity (Sn), also known as true positive rate (TPR), specificity (Sp), and F-score values considered for the calculations based on the values of the confusion matrix are as follows [35]:

$$P = \frac{TP}{FP + TP} \tag{12}$$

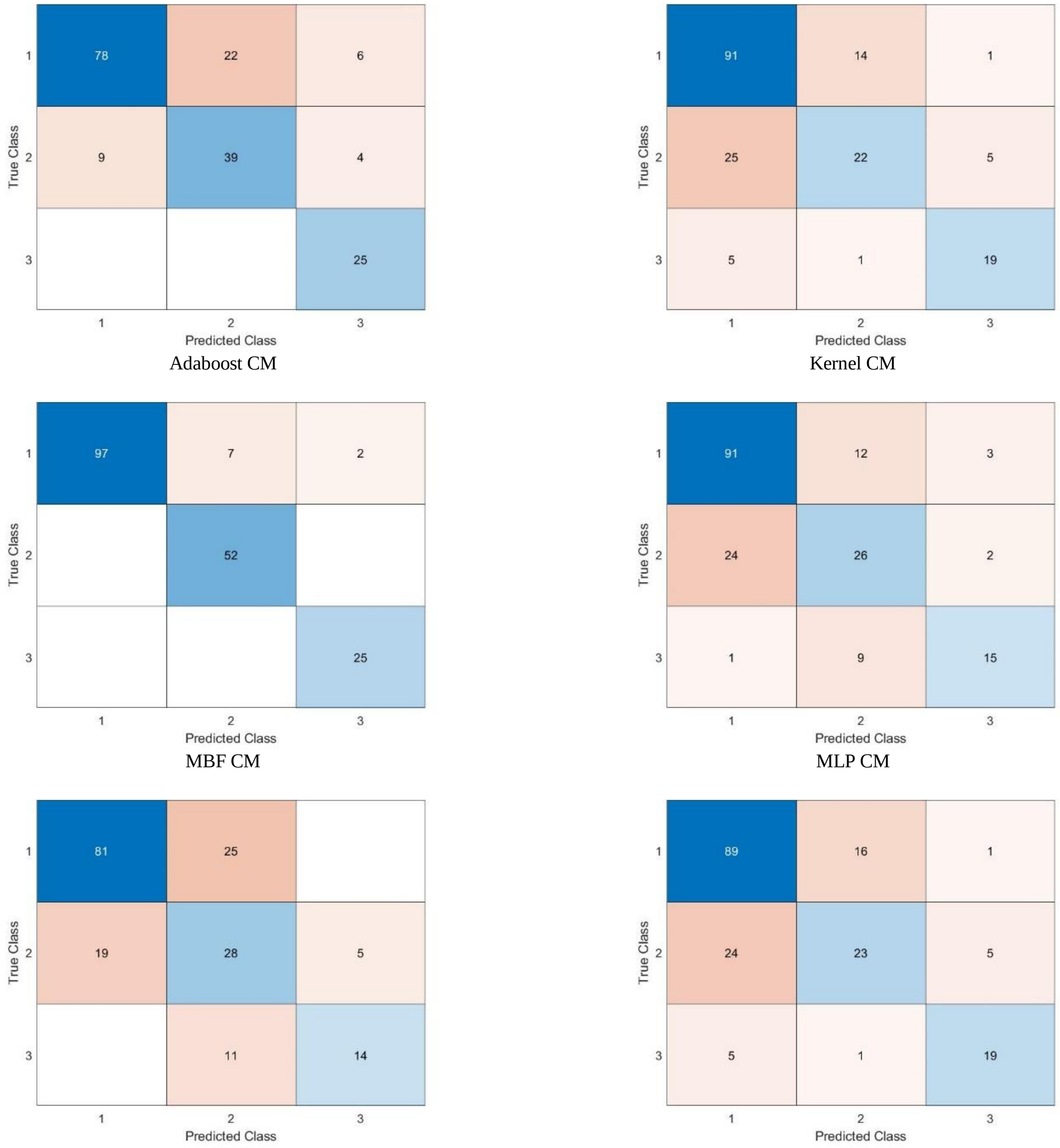
$$Sn = \frac{TP}{FN + TP} \tag{13}$$

$$Sp = \frac{TN}{FP + TN} \tag{14}$$

$$F - score = 2 \times \frac{P \times Sn}{P + Sn} \tag{15}$$

IV. Results and Discussion

The main results obtained in the work are discussed in this section. Also, the superiority of the proposed algorithm in data classification is validated by considering the related works. As shown in Fig. 4, a classification model's performance can be assessed by a confusion matrix in statistics and machine learning. It provides an overview of the categorization findings by displaying the numbers of true positive, true negative, false positive, and false negative estimations. As seen from Fig. 4, the proposed algorithm, MBF, performs better than the rest. Confusion matrices are a widely used metric in classification problem-solving. Both binary and multiclass classification issues can benefit from its use. Confusion matrices show the counts of the actual and expected values. True Negative, or "TN," is the output that indicates how many negative cases were correctly categorized. Similarly, "TP" stands for True Positive and represents the proportion of correctly identified positive cases. False Positive value, or the number of actual negative instances categorized as positive, is represented by the phrase "FP." In contrast, the False Negative value, or the number of real positive examples classified as negative, is represented by the term "FN."



Adaboost CM

Kernel CM

MBF CM

MLP CM

RF CM

SVM CM

Fig. 4. Confusion matrix for the selected algorithms.

As shown in Fig. 5, MBF has better sensitivity, which means that the percentage of real positive cases that the model accurately detected or categorized as positive is remarkable. In terms of TPR, the weakest performance is attributed to SVM. Also, the accuracy of MBF is acceptable according to the values given in Fig. 6.

Fig. 7 to 11 demonstrate the values of F-score, accuracy, specificity, and sensitivity obtained for the various selected models. MBF is superior in terms of the criteria values obtained in the work. The SVM does not have acceptable performance in data classification. Accordingly, Adaboost can be an excellent alternative to MBF as it has the highest values of F-score, accuracy, specificity, and sensitivity after that. The results reported in Table II match those in Fig. 7 to 11. MBF, with a value of about 95%, is by far more accurate than Adaboost by 77%. Compared to the other selected models, the F-score, accuracy, and specificity values obtained for MBF are remarkable, with values of 97.17%, 93.6%, and 96.5%, respectively.

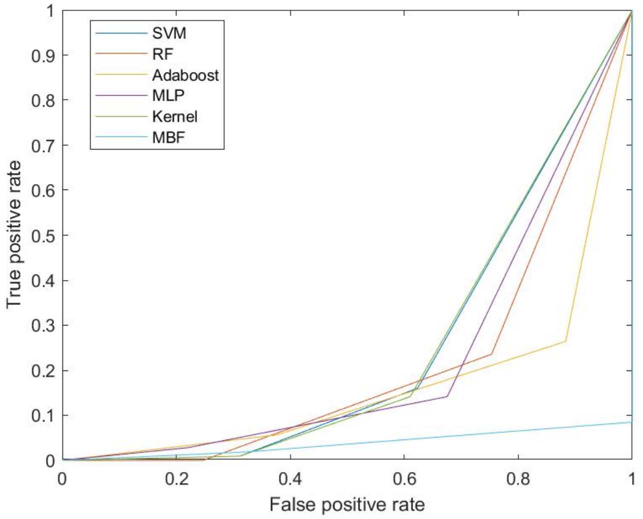


Fig. 5. The true positive rate for the selected models.

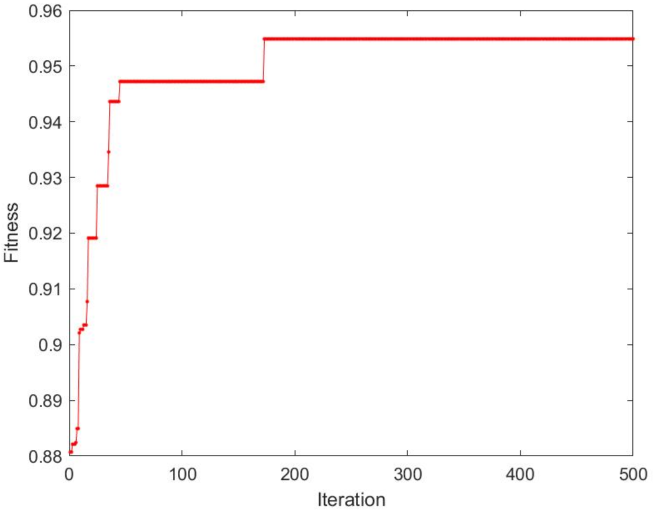


Fig. 6. The accuracy of the proposed method based on iteration and fitness.

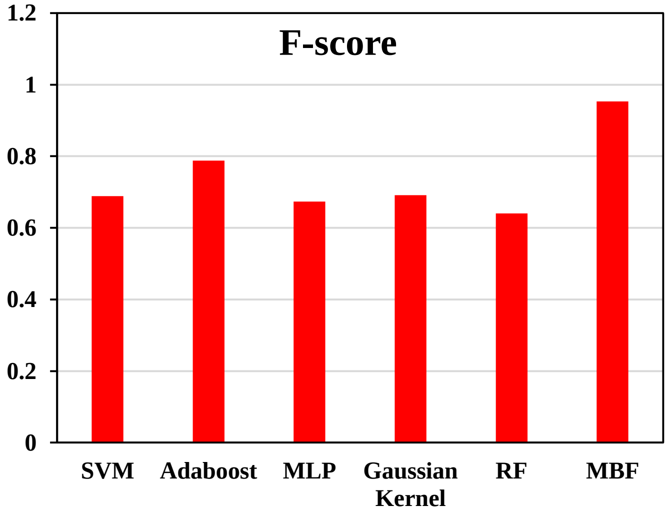


Fig. 7. F-score values of the selected models.

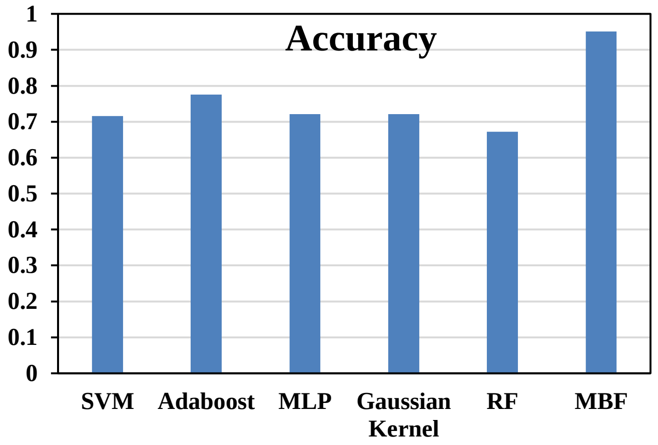


Fig. 8. Accuracy values of the selected models.

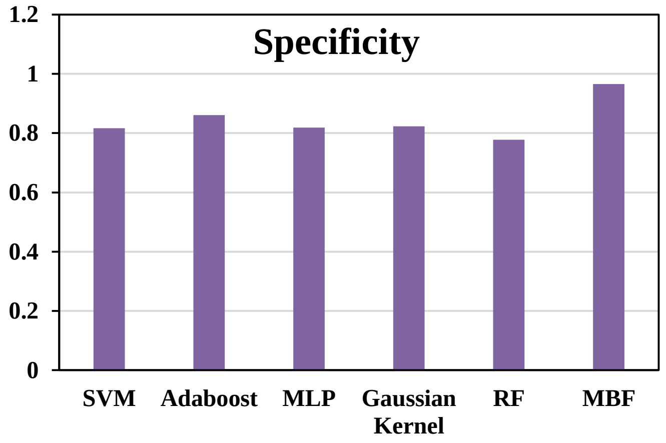


Fig. 9. Specificity values of the selected models.

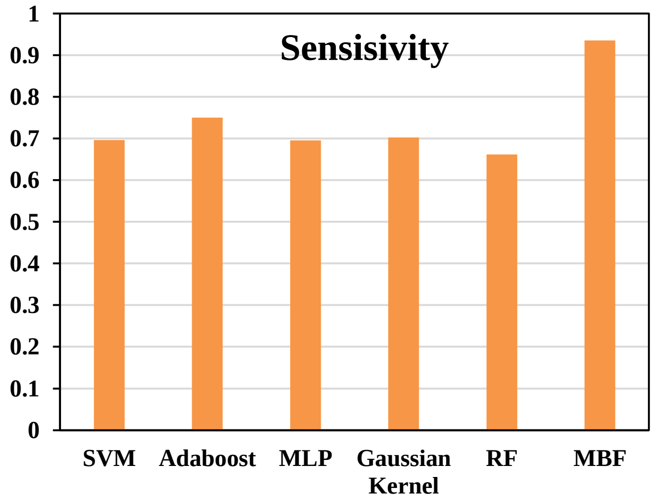


Fig. 10. Sensitivity values of the selected models.

Based on Fig. 7 to 11, the performance metrics, including F-score, accuracy, specificity, sensitivity, and precision, obtained for the various selected models in the study. Each figure provides a visual representation of the values achieved by the models across these metrics. Notably, Fig. 7 depicts the F-score values, which represent the harmonic mean of precision and recall, showcasing the balance between these two metrics. Fig. 8 presents the accuracy values, indicating the proportion of correctly classified instances among the total instances. Specificity values, representing the true negative rate, are displayed in Fig. 9, indicating the ability of the model to correctly identify negative instances.

Fig. 10 showcases sensitivity values, also known as the true positive rate, indicating the model's ability to correctly identify positive instances. Finally, Fig. 11 illustrates the precision values, which represent the proportion of true positive predictions among all positive predictions made by the model. Together, these figures provide a comprehensive overview of the performance of each model across multiple evaluation metrics, facilitating comparisons and insights into their effectiveness in data classification tasks.

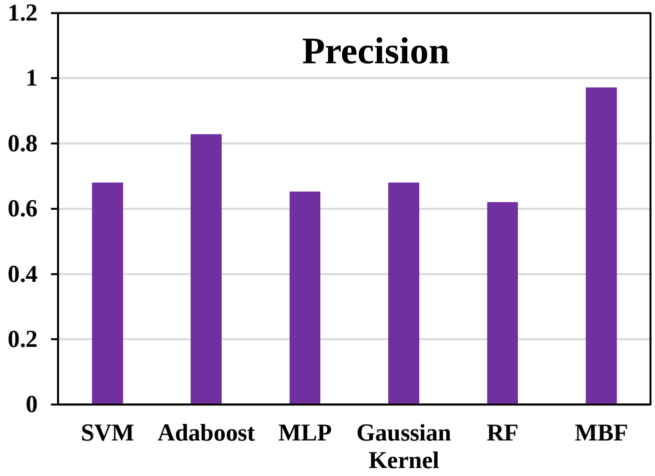


Fig. 11. Precision values of the selected models.

TABLE II. OBTAINED STATISTICAL RESULTS

	SVM	Adaboost	MLP	Gaussian Kernel	RF	MBF
Accuracy	0.715847	0.775956	0.721311	0.721311	0.672131	0.950820
F_score	0.688438	0.787384	0.673673	0.691201	0.640517	0.953391
Precision	0.680643	0.828616	0.65283	0.680522	0.620870	0.971698
Sensitivity	0.696412	0.750061	0.695891	0.702220	0.661447	0.935761
Specificity	0.816446	0.861194	0.818811	0.822478	0.777839	0.965116

V. CONCLUSION

In summary, the current work examines the performance of MBF, SVM, Adaboost, MLP, GK, and RF for data classification in the medical field. The outcomes of the work were examined based on F-score, accuracy, specificity, and sensitivity. The results indicated that the selected algorithms' performance in data classification was acceptable, as the SVM was the weakest and MBF was the strongest. The outputs of the confusion matrix demonstrated that MBF, with an accuracy of 95%, outperforms the rest, and after that, Adaboost, with 77%, can be a good alternative. The F-score, accuracy, and specificity values obtained for MBF are comparable to those of the other models that were chosen, with respective values of 97.17%, 93.6%, and 96.5%. The gap between the MBF and the rest was remarkable in terms of precision as MBF has the precision of 97.17% while SVM, MLP, GK, and RF have the precision of 68%, 65.28%, 68.05%, and 62% respectively. Accordingly, SVM, MLP, GK, and RF performance are identical. However, Adaboost and MBF show desirable capability inaccurate data classification, which can be improved in future work. Future investigations are necessary to validate the kinds of conclusions that can be drawn from this study.

REFERENCES

[1] C. R. Farrar and K. Worden, "An introduction to structural health monitoring," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 365, no. 1851, pp. 303-315, 2007.

[2] F. Chen, Y. Wang, X. Zhang, and J. Fang, "Five hub genes contributing to the oncogenesis and trastuzumab-resistance in gastric cancer," Gene, vol. 851, p. 146942, 2023.

[3] N. H. Binti Rosli and P. Keikhosrokiani, "Chapter 18 - Big medical data mining system (BigMed) for the detection and classification of COVID-19 misinformation," in Big Data Analytics for Healthcare, P. Keikhosrokiani Ed.: Academic Press, 2022, pp. 233-244.

[4] P. Selvaprasanth, J. Rajeshkumar, R. Malathy, D. Karunkuzhali, and M. Nandhini, "Nature Inspired Algorithm for Placing Sensors in Structural Health Monitoring System - Mouth Brooding Fish Approach," Simulation and Analysis of Mathematical Methods in Real - Time Engineering Applications, pp. 99-130, 2021.

[5] H. Huang et al., "Contrastive learning-based computational histopathology predict differential expression of cancer driver genes," Briefings in Bioinformatics, vol. 23, no. 5, p. bbac294, 2022.

[6] G. Alimjan, T. Sun, Y. Liang, H. Jumahun, and Y. Guan, "A new technique for remote sensing image classification based on combinatorial algorithm of SVM and KNN," International Journal of Pattern Recognition and Artificial Intelligence, vol. 32, no. 07, p. 1859012, 2018.

[7] D. Chen, R. Ma, and H. Du, "A fast incomplete data classification method based on representative points and K-nearest neighbors," in 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), 2022: IEEE, pp. 423-428.

[8] G. Hu, Y. Guo, G. Wei, and L. Abualigah, "Genghis Khan shark optimizer: A novel nature-inspired algorithm for engineering optimization," Advanced Engineering Informatics, vol. 58, p. 102210, 2023/10/01/ 2023, doi: https://doi.org/10.1016/j.aei.2023.102210.

[9] S. Yang, J.-Z. Guo, and J.-W. Jin, "An improved Id3 algorithm for medical data classification," Computers & Electrical Engineering, vol. 65, pp. 474-487, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.compeleceng.2017.08.005.

[10] W. Xing and Y. Bei, "Medical health big data classification based on KNN classification algorithm," IEEE Access, vol. 8, pp. 28808-28819, 2019.

[11] S. Boyapati, S. R. Swarna, V. Dutt, and N. Vyas, "Big Data Approach for Medical Data Classification: A Review Study," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 3-5 Dec. 2020 2020, pp. 762-766, doi: 10.1109/ICISS49785.2020.9315870.

[12] S. Khanmohammadi and C.-A. Chou, "A Gaussian mixture model based discretization algorithm for associative classification of medical data," Expert Systems with Applications, vol. 58, pp. 119-129, 2016.

[13] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking," Informatics in Medicine Unlocked, vol. 15, p. 100180, 2019.

[14] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20-28, 2021.

[15] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," Journal of Biomedical Informatics, vol. 107, p. 103465, 2020.

[16] L. J. Rubini and E. Perumal, "Hybrid kernel support vector machine classifier and grey wolf optimization algorithm based intelligent classification algorithm for chronic kidney disease," Journal of Medical Imaging and Health Informatics, vol. 10, no. 10, pp. 2297-2307, 2020.

[17] B. Ma, B. li, X.-Y. Wang, C.-P. Wang, J. Li, and Y.-Q. Shi, "A code division multiplexing and block classification-based real-time reversible data-hiding algorithm for medical images," Journal of Real-Time Image Processing, vol. 16, no. 4, pp. 857-869, 2019/08/01 2019, doi: 10.1007/s11554-019-00884-9.

[18] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," Journal of Big data, vol. 6, no. 1, pp. 1-18, 2019.

[19] R. Dash, "An adaptive harmony search approach for gene selection and classification of high dimensional medical data," Journal of King Saud University-Computer and Information Sciences, vol. 33, no. 2, pp. 195-207, 2021.

[20] E. H. Houssein, M. E. Hosney, W. M. Mohamed, A. A. Ali, and E. M. Younis, "Fuzzy-based hunger games search algorithm for global optimization and feature selection using medical data," Neural Computing and Applications, vol. 35, no. 7, pp. 5251-5275, 2023.

[21] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," Mechanical systems and signal processing, vol. 21, no. 6, pp. 2560-2574, 2007.

[22] G. Gui, H. Pan, Z. Lin, Y. Li, and Z. Yuan, "Data-driven support vector machine with optimization techniques for structural health monitoring and damage detection," KSCE Journal of Civil Engineering, vol. 21, pp. 523-534, 2017.

[23] D. Kim and M. Philen, "Damage classification using Adaboost machine learning for structural health monitoring," in Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2011, 2011, vol. 7981: SPIE, pp. 659-673.

[24] M. O. G. Nayeem, M. N. Wan, and M. K. Hasan, "Prediction of disease level using multilayer perceptron of artificial neural network for patient monitoring," International Journal of Soft Computing and Engineering (IJSCE), vol. 5, no. 4, pp. 17-23, 2015.

[25] M. W. Moreira, J. J. Rodrigues, N. Kumar, J. Al-Muhtadi, and V. Korotaev, "Nature-inspired algorithm for training multilayer perceptron networks in e-health environments for high-risk pregnancy care," Journal of medical systems, vol. 42, pp. 1-10, 2018.

[26] H. Yu, Y. Liu, G. Zhou, and M. Peng, "Multilayer Perceptron Algorithm-Assisted Flexible Piezoresistive PDMS/Chitosan/cMWCNT Sponge Pressure Sensor for Sedentary Healthcare Monitoring," ACS sensors, 2023.

[27] Y. M. Abd Algani, M. Ritonga, B. K. Bala, M. S. Al Ansari, M. Badr, and A. I. Taloba, "Machine learning in health condition check-up: An approach using Breiman's random forest algorithm," Measurement: Sensors, vol. 23, p. 100406, 2022.

[28] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," in SoutheastCon 2016, 2016: IEEE, pp. 1-6.

[29] D. S. Shayegan, A. Lork, and S. Hashemi, "Mouth brooding fish algorithm for cost optimization of reinforced concrete one-way ribbed slabs," Int. J. Optim. Civil Eng, vol. 9, no. 3, pp. 411-422, 2019.

[30] E. Jahani and M. Chizari, "Tackling global optimization problems with a novel algorithm–Mouth Brooding Fish algorithm," Applied Soft Computing, vol. 62, pp. 987-1002, 2018.

[31] K. Ota, M. Aibara, M. Morita, S. Awata, M. Hori, and M. Kohda, "Alternative reproductive tactics in the shell-brooding Lake Tanganyika cichlid Neolamprologus brevis," International Journal of Evolutionary Biology, vol. 2012, 2012.

[32] M. Babazadeh, O. Rezayfar, and E. Jahani, "Interval reliability sensitivity analysis using Monte Carlo simulation and mouth brooding fish algorithm (MBF)," Applied Soft Computing, vol. 142, p. 110316, 2023.

[33] "https://www.kaggle.com/datasets/prasad22/healthcare-dataset/."

[34] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in Australasian joint conference on artificial intelligence, 2006: Springer, pp. 1015-1021.

[35] A. Kulkarni, D. Chong, and F. A. Batarseh, "5 - Foundations of data imbalance and solutions for a data democracy," in Data Democracy, F. A. Batarseh and R. Yang Eds.: Academic Press, 2020, pp. 83-106.