Received: January 21, 2023. Revised: April 11, 2023. 470

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Enhanced Fuzzy Logic Pre-Processing Technique using Hybridized Bat and Particle Swarm Optimization Algorithm for Feature Selection

C. Saranya Jothi[1]* **Carmel Mary Belinda[2]**

[1]*Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India*
* Corresponding author's Email: saranyajothi22@gmail.com

Abstract: In the modern era, the number of diseases is increasing and it is very crucial to diagnose them. One among them is the hepatitis virus in which the proper identification of symptoms is tedious. Machine learning algorithms are essential in recognizing the disease in the earlier stages. In this paper, a novel pre-processing technique and feature selection (FS) method are proposed to achieve improved accuracy with less time complexity. The proposed methodology is divided into three phases: In the first phase, the pre-processing technique such as Multiple Imputation with weight based fuzzy logic (MI-WFL) is applied to decrease the classification error. In the second phase, the correlation-based hybridized bat and particle swarm optimization (CHBPSO) FS algorithm is used to minimize the time complexity and maximize the accuracy rate. In the third phase, the poly kernel support vector machine (PK-SVM) classifier is implemented to avoid overfitting issues. According to the results of the experiments, the CHBPSO algorithm outperforms in the accuracy rate of 96.15%, and classification error for the following metrics mean absolute error (MAE) is 0.17, root mean squared error (RMSE) is 0.23, relative absolute error (RAE) is 49.66, and root relative squared error (RRSE) is 54.62 are lower than the existing algorithms. Furthermore, the proposed solution has attained less processing time of 1.6 sec due to the reduced optimal features.

Keywords: Bat algorithm, Feature selection, Fuzzy logic, Hepatitis, Machine learning.

1. Introduction

Hepatitis is the most common infection in the liver that occurs because of the virus. This virus frequently damages the immune cells in the liver that produce a few signs in the immune system that create hepatitis advances such as fever, malaise, and nausea [1]. There are five different types of hepatitis with slightly different characteristics. The three common methods which identify the virus in the earlier stages are pre-processing [2], feature selection (FS) [3], and classification method [4]. The pre-processing method [2] is used to clean the data. Generally, the original dataset contains big data (i.e., noisy data, irrelevant data, and very large data sizes), too small data (i.e., missing attributes, missing attributes value, and a small amount of data), and fractured data (i.e., incompatible data, multiple data sources and data from multiple levels of granularity). Traditionally, pre-processing methods are used to resolve the data issues by manually filling the data using techniques such as mean [5], median [1], and mode imputation [5]. Further, the researchers used some advanced pre-processing techniques namely min-max scalar [6], standard scalar [6], normalization [2], data discretization [7], and maximum absolute scaling [8].

After cleaning the data, it is fed as the input for FS. As many researchers continue to exert their resolve to develop FS approaches based on various optimization methodologies, still FS approaches remain a mystery to the research field. Traditional FS methods are backward feature elimination [9], recursive feature elimination [3], correlation coefficient (CC)[10], chi-square test [11], fisher's score [12], and random forest algorithm [8] which does not concentrate on the optimal feature extraction. To address the drawbacks of earlier heuristic-based strategies for FS, several meta-heuristic algorithms have been used. Some of

the metaheuristic algorithms with FS include bee colony [13], ant colony [14], harmony search [4], particle swarm optimization (PSO) [13], bat algorithm (BA) [15], and crow search algorithm [16]. The reduced data set with prominent features are taken for further classification. Classification is a supervised machine learning technique in which the model attempts to predict the label of given input data. The data is categorized into two types namely training and testing. The training data is used to train the given model by using the input data and the test model is used to evaluate the data. There are two types of classification: i) binary classification and ii) multi-class classification. In binary classification, two outcomes are predicted. In the multi-classification, more than two outcomes are predicted. Logistic regression [11], decision trees [1], simple bayes [2], and support vector machines (SVM) [1] are some examples of binary classification techniques.

In this manuscript, a novel hybrid multiple imputation with weight based fuzzy logic (MI-WFL) pre-processing technique is implemented by addressing the problem misclassification error, correlation-based hybridized bat and particle swarm optimization (CHBPSO) FS algorithm to improve the processing time and to accelerate the classification accuracy Poly kernel -support vector machine (PK-SVM) is adopted. The following are the significant contributions of this work:

- Selection of MI-WFL pre-processing technique for predicting the missing values along with the original instance that maintains the minimum misclassification error.
- Weight based correlation method is applied to improve the searching speed than searching in random space.
- Identification of features using the CHBPSO algorithm improves the search for optimal features efficiently.
- To achieve better accuracy results for disease prediction, we applied the PK-SVM algorithm to avoid the overfitting problem.

The remaining text is as follows: section 2 explains the previous work; it provides a technical overview of the pre-processing, FS, and classification methods. The proposed framework is displayed in section 3, and in section 4 we present the experimental findings and performance assessment. The paper's conclusion and future work are also covered in section 5.

2. Literature survey

In the existing work, pre-processing, FS methods, and classification models are discussed below. In the past two decades, all the researchers concentrated on the three combined techniques to improve the performance they are (i) pre-processing with classification techniques, (ii) FS with classification methods, and (iii) pre-processing, FS, and classification methods. The pre-processing and FS techniques are combined to improve the accuracy rate. Some of the pre-processing with classification techniques are discussed in detail. Some of the pre-processing with classification techniques are discussed in detail. The common existing pre-processing techniques are filter and wrapper method (FWM) [2], standard and min-max scalar (SMS) [6], genetic algorithm partial least squares (GA-PLS) [7] and mean imputation method (MIM) [8]. In [2] authors described the data, when there is a lot of irrelevant, redundant information or noisy and inaccurate data, during the training phase it becomes more challenging. It is commonly known that data preparation and filtering stages in machine learning issues consume a significant amount of processing time. Therefore, the data processing technique is important to implement. In this paper filter approach and wrapper method are applied to select the important instance and also detect the outliers. SMS [6] pre-processing method is to fill the null value in the dataset. Also, various algorithms are selected using the methods such as relief and the least absolute shrinkage selection operator to make this method attain good accuracy of 84% but it maximizes the computational complexity. GA-PLS method [7] is implemented for data pre-processing. This pre-processing technique provides the lowest error using root relative squared error (RRSE) and also multi-dataset was applied by the GA-PLS algorithm. The complexity time is increased due to overfitted solutions and concentrating on cross-validation procedures to avoid the risk. MIM [8] is utilized to reject the outliers. Then for classification, the random forest algorithm is implemented. This paper attains low performance in accuracy 79.8% by considering the mean filling of data in the dataset.

Next, a lot of FS with classification methods are considered for reducing the data size. In [9] authors develop a sequential FS algorithm to select the most dominant features to detect the death rate that reduces the processing time. Further, the random forest algorithm is applied which has a less accuracy level of 86.67% due to the lack of pre-processing technique. Bio-inspired algorithm [12] using the improved salp swarm algorithm (ISSP) based on the swarming

mechanism to reduce the features. This paper mainly focuses on accuracy, reliability, and speed to modify the current best solution, a novel control parameter, inertia weight, is added. This paper doesn't support the large size of datasets. Also, [15] a hybrid wolf-bat algorithm is proposed for the FS problem. To minimize the classification error, the weight-based multi-layer perceptron neural network is applied. The various datasets are implemented in the proposed methodology. Therefore, the convergence speed is high and also the multi-classification problem is not supported for this work. An improved chaotic crow search algorithm [16] is developed which has a superior search approach and the PSO which produces the best global solution in the search field. To avoid the overfitting problem K-Nearest Neighbour (KNN) classifier is used which gives 89.67 % accuracy. In this work, execution time is more because of the noisy and redundant data.

The pre-processing, FS, and classification methods are combined and used in identifying the disease in the early stages. Haider Banka, and Suresh Dara [17] used a quartile-based rapid heuristic technique for pre-processing, where less relevant in categorizing as a group using a raw domain attribute. To choose the critical feature subsets in the PSO framework the hamming distance is added as a proximity indicator to update the velocity. In this paper, they didn't focus on error prediction like RRSE, root mean squared error (RMSE), and mean absolute error (MAE). In [18], authors proposed a hybridized bio-inspired algorithm like the lion optimization algorithm and glow-worm swarm optimization algorithm for selecting the most important features. Before the FS, pre-processing techniques are used such as hot deck imputation applied for filling the missing values, and for data transformation, min-max normalization is implemented. The advantage of these methods is increased precision rate, specificity, and sensitivity. This paper does not concentrate on the classification method which attains high error. To avoid the anomaly data for giving the input of classification by introducing two steps i) pre-processing and ii) FS. In the pre-processing, data normalization and data discretization methods are applied to save the training time which has the same scale (i.e., it ranges between 0 and 1). Then, the FS method merges the correlation FS with PSO to select the best feature set, and for evaluating the performance SVM classifier is implemented [3]. In this paper, improper FS may reduce performance. Computer-aided diagnosis system [4] is proposed to predict if the disease is present or absent. The input data is pre-processed using mean imputation and smote techniques. The mean imputation method which handles the missing values and smote technique solves the imbalance problem. To select the optimal feature subset three bio-inspired algorithms are used. The selected feature subset is given as an input to the SVM classifier which produces 90.48% accuracy but it fails to produce better classification results. In [19], authors introduced the pre-processing technique such as feature scaling and FS to improve the accuracy rate. In the feature scaling methods are applied to fill the missing values using the techniques like min-max scaling, log1p transformation, standardization, and maximum absolute scaling. In FS, univariate FS is implemented. The ensemble learning algorithm is applied for classification which gets 91.82 % accuracy and this method suffers in computational time because of choosing the pre-processing technique, it takes more time to fill in the missing values. In this paper, the parameters such as MAE and RRSE have high error rates due to the constant filling of missing values.

In the previous work, numerous pre-processing and FS algorithms were developed but still, they have some issues in dimensional space and time to predict the hepatitis virus. To overcome the issues, we introduce a hybrid MI-WFL algorithm with CHBPSO to increase the accuracy rate in less time. So, we combined MI-WFL pre-processing method, CHBPSO FS techniques, and PK-SVM classification method in our proposed methodology to overcome the overfitting problem and also predict hepatitis diseases in early stages with good precision and less error rate.

3. Proposed framework

To forecast diseases in their early stage, the proposed methodology MI-WFL with CHBPSO algorithm along with PK-SVM classifier are presented in-depth as shown in Fig. 1. Initially, the hepatitis dataset is used as the input. The proposed work is divided into three sections. In section 3.1, the data pre-processing technique MI-WFL has been developed to clean up the dataset by filling up the missing values. Section 3.2, introduced a new hybrid CHBPSO algorithm to decrease the high dimensional space to find the best optimal solution. In section 3.3, a PK-SVM classifier is used to forecast whether the class will lie on survival or not.

3.1 Data pre-processing techniques

Pre-processing techniques involve exploring several methods to handle the missing data which include data cleansing, data integration, data reduction, and data transformation [3]. In this paper, a

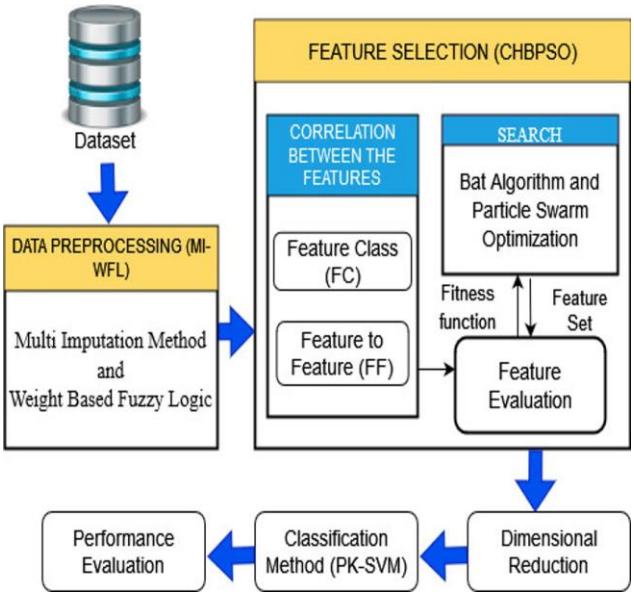


Figure. 1 Proposed framework

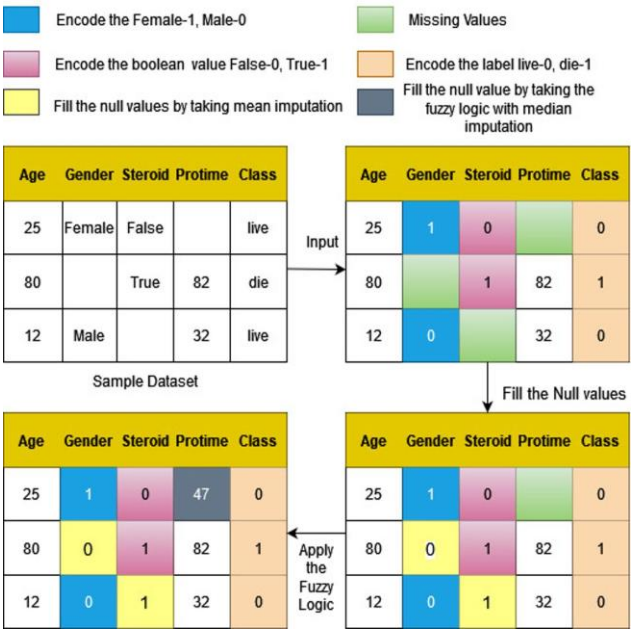


Figure. 2 Example of data pre-processing

novel pre-processing (i.e., data cleansing) technique hybrid MI-WFL is discussed. The principal steps of our approach are as follows:

Step 1: In the Initial stage, collect the hepatitis dataset from the UCI repository. In our dataset, the first row represents the feature's name. It is divided into two types a categorical column and a numeric column which is illustrated in Fig. 2. Consider, the categorical column is stored in a category variable and the numeric column is stored in a number variable. **Ex:** (i) category= (Gender, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver-Big, Liver Firm, Spleen, Spiders, Ascites, Varices, Histology) and (ii) number= (Age, Bilirubin, Alk_phosphate, Sgot, Albumin, Protime).

Step 2: In the first type, the null values (NV) in the category column are filled using the multiple imputation method is briefly illustrated in the following steps:

(i) Identify the number of features NV in the dataset that are highlighted in green colors are shown in Fig. 2. For **example,** Steroid=1 NV, Fatigue=1 NV, Malaise=1 NV, Anorexia=1 NV, Sgot=4 NV, Spleen NV, Ascites NV, Varices=5 NV, Bilirubin=6 NV, Liver-Big=10 NV, Liver-Firm=11 NV, Albumin= 16 NV, Alk_phosphate =29 NV, and Protime =67 NV.

(ii) After identifying the NV, encode all the category columns into an integer value. Consider the following examples:

- The gender feature has two attributes namely male and female. The male attribute is encoded with the binary value 1 and the female attribute is encoded with the binary value 0.
- Convert all the Boolean attributes (i.e., true, false) into integer values. Ex: true-1 and False-0. Similarly change the class label name (i.e., live, die) into integer values. Ex: live-0 and die-1.

(iii) In order to fill the NV, the mean imputation is implemented in the category column. Finally, the entire missing value of the categorial column is filled.

Step 3: In the second type, the missing NV in the numeric column is filled using weight-based fuzzy logic which improves the classification results. The weight of features is classified into Low (L), Medium (M), High (H), and Very high (VH) is projected in Table 1. Here, the age feature attribute is considered for filling the missing NV.

The following example is taken as a sample for predicting filling NV using fuzzy logic. The NV of bilirubin is filled by checking the age feature classification. If the age is 51 (i.e., high classification), then bilirubin NV will be considered to range from 4 to 5.9 using the median imputation method. The 4.95 value is filled in the place of the missing row in the attribute bilirubin.

Similarly, for all the missing NV combinations such as (age, alk-phosphate), (age, sgot), (age, albumin), and (age, protime) the integer values are filled.

Step 4: This process is repeated until all the missing values in the dataset are filled to obtain the integer values.

Table 1. Prediction of filling NV using fuzzy logic

Classification	Age	Bilirubin	Alk phosphate	Sgot	Albumin	Protime
L	7-24	0.3-1.9	26-96	14-98	2.1-3.5	0-39
M	25-45	2-3.9	96.1-147	98.1-182	3.6-4.8	39.1-58
H	46-65	4-5.9	147.1-194	182.1-278	4.9-6.4	58.1-78
VH	>65	6-8	194.1-295	278.1-648	>6.4	78.1-100

Table 2. Glossary of terms

Symbol	Description
x_i	Feature value, where (i=1,2,3, …, 20)
x_i'	Mean of the features
$r_{x_i y_j}$	The correlation coefficient of feature to class
y_j	Class value, where (j=1,2)
y_j'	Mean of the class
z_i	Neighbour feature in the dataset
z_i'	Mean of the neighbor feature
$r_{x_i z_i}$	The correlation coefficient of feature to feature
x_{pbest}	Highest features fitness function
f_i	To find the frequency
f_{min}	Minimum frequency, it takes a random value between 1 to 10
$f_{\max}$	Maximum frequency, it takes between 1 to 10
β	Frequency of pulse rate $\beta\epsilon$ (0,1)
v_i^t	velocity
v_i^{t-1}	Initial velocity is taken as correlation weight
B	Bat population
k	Number of iterations where (k=1,2,3,….B)
T	Maximum number of iterations
x_i^{t+1}	Update the new position
TP	True Positive
P	Positive
FP	False Positive
N	Total number of actual unfavorable incidents
a_i	Predicted value
b_i	Observed value
c_i	Actual value
n	Total number of instances

3.2 Feature selection

Feature extraction and FS are two types of pre-processing techniques. In our work, we focus on the FS by introducing a novel CHBPSO. To improve the accuracy, precision, and processing time improved weight-based correlation FS method is introduced. FS plays an important role in predicting the condition (i.e., class) of the person, so we focus on the reduced FS. The CHBPSO algorithm for the FS method is divided into the following sub-sections they are correlation between the features and framework of hybridized bat and particle swarm optimization (HBPSO). Section 3.2.1 describes the correlation between the features using weight analysis. In section 3.2.2, the HBPSO algorithm has been developed. Table 2 summarizes the glossary of terms.

3.2.1. Correlation between the features

In the correlation technique, weight between the features is used for improving the search (i.e., HBPSO) of the optimal solution. In general, the degree of the CC $r_{x_i y_j}$ which ranges from -1 to +1. It finds the relationship between the feature class (FC) and feature to feature (FF) using Eq. (1).

$$r_{x_i y_j} = \frac{\sum(x_{i-} x_i')(y_{j-} y_j')}{\sqrt{\sum(x_{i-} x_i')^2 \sum(y_{j-} y_j')^2}} \quad (1)$$

Where x_i represents the feature value and x_i' denotes the mean of the features, y_j acts for the class value, y_j' stands for the mean of the class. Here *i =1,2, 3,…20* and *j=1,2*.

1. Consider the sample for FC. Let us assume CC $r_{x_i y_j}$ in Eq. (1), Where $x_1 = age, y_1 = class$. Apply the values in Eq. (1) $r_{(age,class)} = \frac{(30-41.2)\times(2-1.79)}{\sqrt{(30-41.2)^2}\times\sqrt{(2-1.79)^2}} + .. \frac{(39-41.2)\times(2-1.79)}{\sqrt{(39-41.2)^2}\times\sqrt{(2-1.79)^2}} = -0.21$. Similarly, repeat this process for different $r_{x_i y_j}$. Where, x_i = (Sex, Setorid, Antivirals, Fatigue, Malaise, Anorexia, Liverbig, Protime, Varices, Histology, Liverfirm, Spleen-palpable, Spiders, Ascites, Bilirubin, Alkphosphate, Sgot, Albumin) and y_j = class. After applying different $r_{x_i y_j}$ we get,
$r_{(sex,class)}=-0.17, r_{(steriod,class)}=-0.13,$
$r_{(\text{fatigue},class)}=0.30, r_{(\text{malaise,class})}=0.33,$
$r_{(antivirals,class)}=-0.13, r_{(\text{anorexia},class)}=0.13,$
$r_{(\text{liverbig},class)}=0.07, r_{(\text{protime},class)}=0.40,$
$r_{(\text{varices},class)}=0.36, r_{(Histology,class)}=0.33,$
$r_{(\text{liverfirm},class)}=0.06, r_{(\text{spleenpalpable},class)}=0.23$
$r_{(\text{spiders},class)}=0.39, r_{(\text{ascites},class)}=0.47$
$r_{(\text{bilirubin},class)}=-0.46, r_{(\text{alkphosphate},class)}=-0.16$

$r_{(\text{sgot},class)=-0.077}$, $r_{(\text{albumin},class)=0.50}$

2. Assume the FF sample which is to find the distance between an FF apply Eq. (2).

$$r_{x_i z_i} = \frac{\sum(x_{i-}x_i')(z_{i-}z_i')}{\sqrt{\sum(x_{i-}x_i')^2 \sum(z_{i-}z_i')^2}} \quad (2)$$

Here, z_i act as the neighbor feature in the dataset. Where, z_i' stands for the mean of the neighbor feature. Let us assume CC $r_{x_i z_i}$ in Eq. (2), Where $x_1 = age, z_1 = age$. Apply the values in Eq. (2)

$r_{(age,age)} = \frac{(30-41.2)\times(30-41.2)}{\sqrt{(30-41.2)^2}\times\sqrt{(30-41.2)^2}} + \cdots = 1.$

Similarly, repeat this process for different $r_{x_i z_i}$

3.2.2. Framework of hybridized bat and particle swarm optimization (HBPSO)

This work has suggested a novel hybrid optimization method that combines the BA [15] and PSO [13] thereby concentrating the reduced features of data. (i) In BA [15] detect optimal solutions based on the food-searching ways of the bat. The bat is inspired by other categories due to the echolocation behavior. To improve the searchability and accuracy BA algorithm is adopted. (ii) PSO algorithm excels more than other algorithms as it converges at a low speed to reach the optimal solution. At the end of each search, the optimal weighted features are identified to make better computational efficiency. As a result, the hybrid algorithm increases the rate of an optimal solution.

Step 1: The CC feature weight is taken as the initial population size is projected in Fig. 4.

Step 2: Apply the fitness function using Eq. (3) for the feature.

$$fitness = \frac{k\times r_{x_i y_j}}{\sqrt{k+k\times(k-1)\times r_{x_i z_i}}} \quad (3)$$

Where k denotes the feature subset contains the constant value that changes for each iteration, $r_{x_i y_j}$ is an average term of FC and $r_{x_i z_i}$ is an average term of FF.

Step 3: After the first iteration, the highest features fitness function displayed in Eq. (4) value is taken as the x_{pbest} is illustrated in Fig. 4.

$$x_{pbest} = \max(fitness) \quad (4)$$

Step 4: In the second iteration, apply BA to find the frequency f_i that is determined in Eq. (5).

$$f_i = f_{min} + [(f_{max} - f_{min}) \times \beta] \quad (5)$$

Where, f_{min} , $f_{\max}$ takes a random value between 1 to 10, and β defines the frequency of pulse rate which ranges from 0 to 1.

Step 5: After finding the frequency value, update the velocity v_i^t for each feature by applying in Eq. (6)

$$v_i^t = v_i^{t-1} + [(r_{x_i y_j} - x_{pbest}] \times f_i \quad (6)$$

Here, the initial velocity v_i^{t-1} is taken as correlation weight, the parameter $r_{x_i y_j}$ is the correlation of the FC value and *t* represents the number of iterations (i.e., maximum iteration =20)

Step 6: Apply the PSO algorithm to update the new position x_i^{t+1} using Eq. (7).

$$x_i^{t+1} = v_i^t + r_{x_i y_j} \quad (7)$$

After the second iteration, evaluate the fitness function by applying Eq. (3) is displayed in Fig. 5. In each iteration, x_{pbest} values are updated.

Step 7: Compare the first iteration and second iteration x_{pbest} values and fix the maximum x_{pbest} value as the fitness function for the third iteration is shown in Fig. 5. Repeat steps 2 to 4 until reaching the best x_{pbest} the optimal solution up to 20 generations.

3.2.2.1. Example of HBPSO

1) Consider the CC feature weight for each parameter namely age=0.2196, bilibrum=0.4505, sgot=0.0756, alkphoshate=0.141, etc. is taken as the initial position of the HBPSO algorithm using step 1.
2) Calculate the fitness function for each parameter which is defined in step 2. Here, for the first iteration (k=1), the second iteration (k=2), and this process follows up to 20 iterations (k=20). Apply the values from the sub-section 3.2.1 $r_{x_i y_j}$ = -0.2916 and $r_{x_i z_i}$ =1, then apply in the fitness value using Eq. (3) (i.e., $fitness(x_1 = age) = \frac{1*-0.2196}{\sqrt{1+1(1-1)*1}}$ =-0.2196). Similarly, for different features the fitness value namely $fitness(x_2 = bilibrum)$ = -0.4638, $fitness(x_3 = sgot)$ = -0.0777, $fitness(x_4 = alk\_phosphate)$= -0.1612 etc. are generated.
3) From the first iteration, the $fitness(x_2 = bilibrum)$ = -0.4638 has the maximum fitness value. Therefore, it is chosen as the x_{pbest} value.
4) *Apply **BA*:** In the second iteration using Eq. (5), the following frequency f_i values are

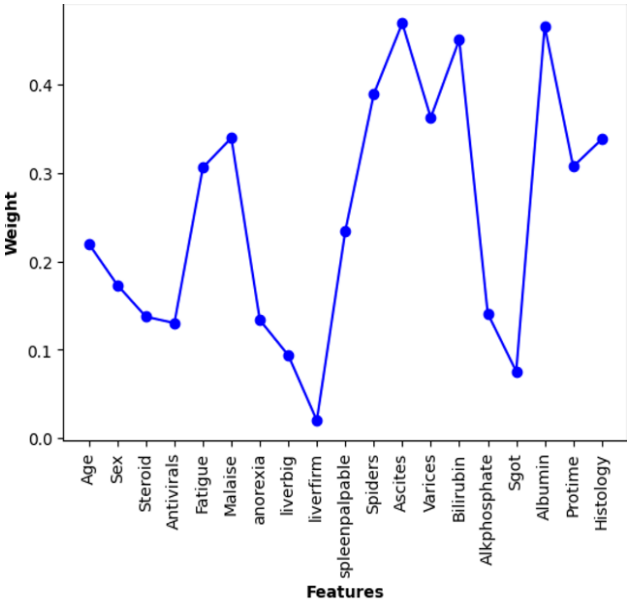
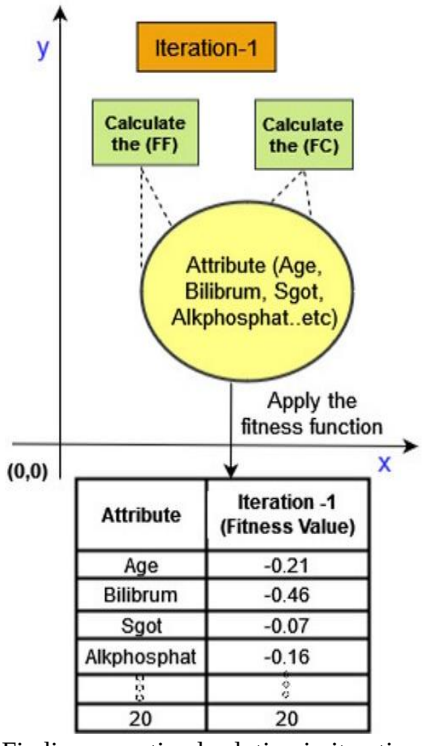


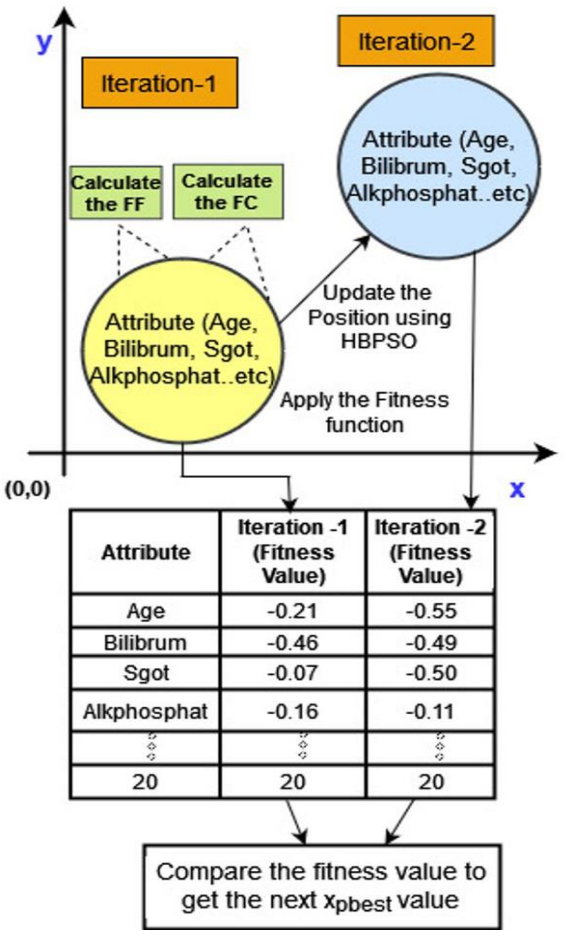
Figure. 3 Initial position with CC feature weight



Attribute	Iteration -1 (Fitness Value)
Age	-0.21
Bilibrum	-0.46
Sgot	-0.07
Alkphosphat	-0.16
⋮	⋮
20	20

Figure. 4 Finding an optimal solution in iteration-1 using the HBPSO optimization algorithm

developed $f_i = f_{min} + [(f_{max} - f_{min}) \times \beta\ ]$ = f_1 = (0.5+ [(8-0.5) × 0.5])= 4, f_2= (1+[(4-1) × 0.5]) = 2.5, f_3 = (1+[(2.5-1) × 1]) = 2.5 , f_4 = (1+[(2.5-1) × 0.5]) = 1.75, etc. Here $f_1 = age, f_2$ = bilibrum, f_3 = sgot, f_4 = alk_phosphate



Attribute	Iteration -1 (Fitness Value)	Iteration -2 (Fitness Value)
Age	-0.21	-0.55
Bilibrum	-0.46	-0.49
Sgot	-0.07	-0.50
Alkphosphat	-0.16	-0.11
⋮	⋮	⋮
20	20	20

Compare the fitness value to get the next x_{pbest} value

Figure. 5 Update the current position for iteration-2 using the HBPSO optimization algorithm

5) Update the velocity v_i^t for each feature by applying in Eq. (6) we get $v_i^t = v_i^{t-1} + [(x_{ij} - x_{pbest}] \times f_i$ = $v_1^1 = (0.21 + (-0.219 + 0.4638) \times 4)$ =1.1892 , v_2^1 = (0.4505+ (-0.4638+0.4638) × 2.5) = 0.4505, v_3^1= (0.756 + (-0.0777+0.4638) × 2.5) = 1.721, v_4^1= (0.141+ (-0.1612+0.4638) × 1.75) = 0.67055. Here, v_1^1 = age , v_2^1 = bilibrum, v_3^1 = sgot, v_4^1 = alk_phosphate.
6) *Apply* ***PSO*****:** To find the new position, the frequency, and the velocity values are taken from the BA algorithm which is represented in the given Eq. (7) $x_i^{t+1} = v_i^t + r_{x_i y_j} = x_1^1$ = (1.1892-0.2196)= 0.9696, x_2^1 = (0.4505-0.4638) = -0.0133, x_3^1 = (1.721-0.0777) = 1.6433, x_4^1 = (0.6705-0.1612)=0.50935, etc. Calculate the fitness value for the updated position and assign the maximum value as the x_{pbest} = 1.004 from the first (x_{pbest} = -0.4638) and the second iteration (x_{pbest} = 1.004).

Algorithm 1: The proposed Hybrid CHBPSO algorithm for FS

Input: The entire feature set $x=(f_1, f_2, f_3, \ldots\ldots, f_n)$
Output: The best optimal subset (x_{pbest})

1. Initialize the bat population *(B)*, velocity (v_i), frequencies (f_i), and the maximum number of iterations *(T)*, and initialize the parameter *k* denotes the iteration (k=*1, 2, 3, …. B*).
2. Initialize the position (x_i) (i =1,2, 3, 4, …B) is taken as correlation weight displayed in Fig. 3.
3. while *(k<T)* do
4. if *(k = 1)* then
5. Calculate the fitness function by Eq. (3)
6. x_{pbest} = *max (fitness)*
7. else
/* update frequency and velocity using bat algorithm*/
8. Update f_i and v_i by using Eq. (5) and Eq. (6)
/* Apply the PSO algorithm to update the new position*/
9. Update the new position by Eq. (7)
10. Evaluate the fitness for the new position in Eq. (3)
11. Assign the value *max = maximum fitness value*
if *(max* > x_{pbest}*)* then
x_{pbest} = max
end if
12. end if
13. update *k =k+1*
14. Repeat step 3
15. End while

7) Repeat steps 4-6 for 20 generations to reach the best optimal solution. The most weighted features after maximum iterations are antivirals, spiders, ascites, varices, and class.

3.3 Classification method (PK-SVM)

SVM is a supervisor model for machine learning [1], it will assist us in locating the best hyper-plane that categorizes fresh samples. It helps us choose the best hyperplane and also called SVM is a maximum margin classifier. The data points that are closest to the hyper-plane are considered to support factors. The maximization of the margin solves the overfitting problem. To solve this, we have three important lines, the middle line of the street H_0, which is a hyper-plane, and two gutters of the street, H_1, and H_2. If the expression for any unknown sample is equal to or greater than zero. We predict it as a positive sample. Otherwise, it can be predicted the unknown is negative.

4. Experimental results with discussion

In this paper, we implement novel technologies such as the data pre-processing method (i.e., MI-WFL) and FS method (i.e., CHBPSO) used in the early prediction of diseases that reduce the death rate. The input hepatitis dataset is taken from the UCI machine learning repository [20] for evaluation. The proposed method is compared with the existing methods namely relief [6], random forest (RF) [8], PSO-KNN [16], PSO-SVM [3], ensemble method [4], artificial bee colony -genetic algorithm (ABC-GA) [13], PSO -firefly (F) [21]. In this experiment, the proposed and the existing methods are compared with 15 metrics. The various metrics used for measuring the performance of the proposed method are explained in Eqs. (8-18) in detail as follows: true positive rate (TPR) [19], false positive rate (FPR) [19], precision value (PV) [4], recall [4], F-measure [4], matthews correlation coefficient (MCC) [6], receiver operating characteristic (ROC) area [19], precision-recall curve (PRC) area [6], processing time [6], accuracy [3], kappa statistic (KS) [6], MAE [17], RMSE [6], relative absolute error (RAE) [17], RRSE [17] is illustrated in Fig. 6.

The TPR [19] is used to predict the percentage of positive values that are successfully identified. Whereas, the FPR [19] is used to predict the percentage of negative values that are mislabelled as good as shown below in Eq. (8) and (9),

$$TPR = \frac{TP}{P} \tag{8}$$

$$FPR = \frac{FP}{N} \tag{9}$$

Where, true positive (TP) represents the actual positive result of the person, positive (P) contains the positive value (i.e., live), false positive (FP) denotes the negative values that are mislabelled and N acts as the total number of actual unfavorable incidents. One of the frequently used performance parameters is precision. It is used to measure the statistical variability of the corrected values is calculated by using the formula Eq. (10),

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

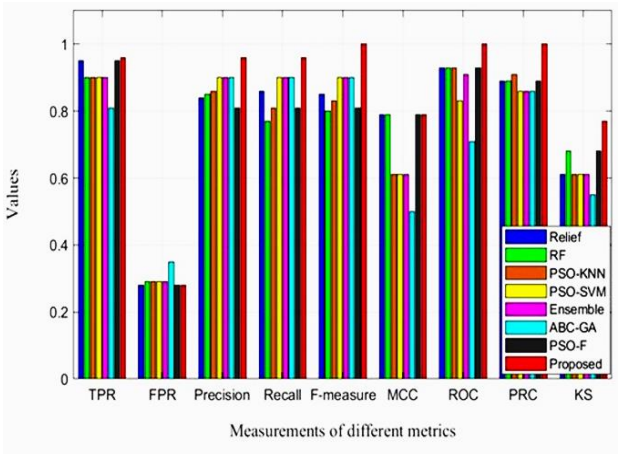


Figure. 6 Performance comparison with an existing algorithm using different metrics

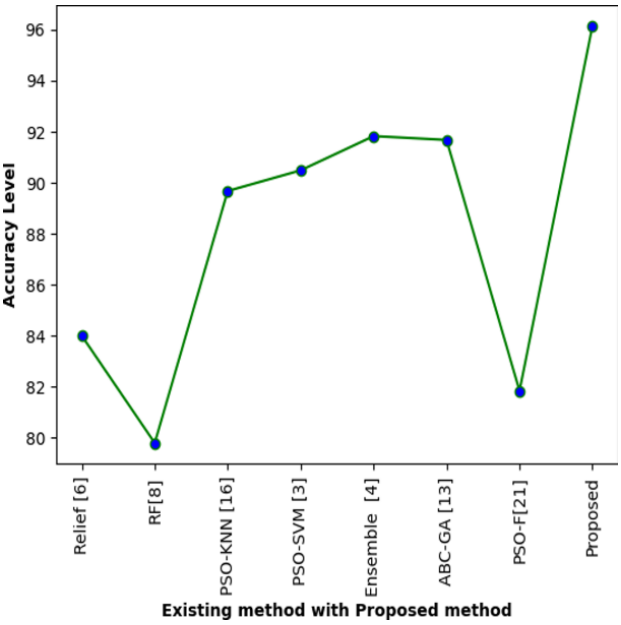


Figure. 7 Comparison with existing and proposed methods in terms of accuracy metric

The recall parameter determines a model's ability to find all relevant positive cases in a dataset calculated below the formula Eq. (11).

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

Where false negative (FN) represents the predicted wrongly in the dataset. The parameter F-measure is estimated using the combination of precision and recall presented in the given Eq. (12)

$$F_{measure} = \frac{2}{\text{Precision}^{-1}+\text{Recall}^{-1}} \tag{12}$$

Where Precision^{-1} denotes the inverse of precision value and Recall^{-1} represents the inverse of recall value. The MCC forecasts the classes for model evaluation and differentiates the projected and actual classes difference as shown in Eq. (13).

$$MCC = \frac{(\text{TP}\times\text{TN})-(\text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FP})\times(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}} \tag{13}$$

Where true negative represents the (TN). The area under the roc curve (AUC) lies under curve=1, AUC is calculated using the formula given below Eq. (14).

$$AUC = \int TPR \times d(FPR) \tag{14}$$

Where d is the differentiation of FPR and the integral of TPR. Whereas the KS is calculated using observed agreement and the chance of agreement, the formula is displayed below in Eq. (15).

$$KS = \frac{2\times(\text{TP}\times\text{TN}-\text{FN}\times\text{FP})}{(\text{TP}+\text{FP})\times(\text{FP}+\text{TN})+(\text{TP}+\text{FN})\times(\text{FN}+\text{TN})} \tag{15}$$

The accuracy is predicted using correctly predicted instances divided by using a total number of test instances given in Eq. (16) as displayed in Fig. 7.

$$Accuracy = \frac{TP}{(TP)+(TN)} \times 100 \tag{16}$$

The proposed method is compared with existing methods such as Relief [6], RF [8], PSO-KNN [16], PSO-SVM [3], Ensemble Method [4], ABC-GA [13], PSO-F [21]. In Fig. 6 the different metrics namely TPR [19], FPR [19], MCC [6], ROC [19], PRC [6], and KS [6] and accuracy [3] are used to evaluate the proposed method. In the Relief algorithm [8], the FPR is high, which makes other parameters' performance low due to not selecting prominent features. In RF [8] algorithm, the accuracy rate, precision, and recall are low because of the lack of FS. The PSO-KNN [16], PSO-SVM [3], Ensemble [4], and PSO-F [21] algorithms achieve the average performance in the metrics TPR, FPR, Precision, Recall, F-measure, Roc, PRC, KS, and accuracy by focusing only on the FS, not in the pre-processing technique. In ABC-GA [13] algorithm attains low-performance metrics in all metrics when compared to other methods due to the poor optimal features selection and also improper outliers' classification. The proposed method MI-WFL with CHBPSO shows outperforming results due to the prominent MI-WFL pre-processing technique where filling the missing values are carefully chosen. Moreover, the accuracy level is increased due to solving the overfitting problem using PK- SVM classifier. MAE error rate is defined by observed data with expected values. The

Table. 3 Comparison between existing algorithm with proposed using error metrics

Algorithm	MAE [17]	RMSE [6]	RAE [17]	RRSE [17]
Relief [6]	0.26	0.29	90.54	84.29
RF [8]	0.16	0.25	54.9	72.55
PSO-KNN [16]	0.17	0.26	60.18	74.34
PSO-SVM [3]	0.15	0.24	53.56	68.82
Ensemble [4]	0.16	0.24	53.79	70.05
ABC-GA [13]	0.14	0.25	49.86	71.09
PSO-F [21]	0.18	0.29	63.58	82.63
Proposed	**0.17**	**0.23**	**49.66**	**54.62**

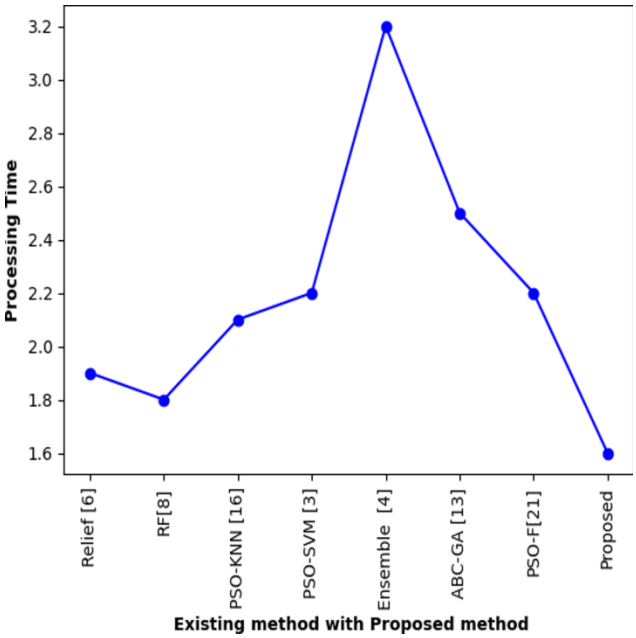


Figure. 8 Processing time analysis with existing algorithm and proposed method in sec

formula is specified below in Eq. (17), where a_i represents the predicted value, b_i represents the observed value, and n indicates the total number of instances (i.e., record).

$$MAE = \frac{\sum_{i=1}^{n}|a_i - b_i|}{n} \quad (17)$$

The RMSE is used to calculate the classification prediction, where c_i denotes the actual value shown in Eq. (18).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - c_i)^2} \quad (18)$$

Table. 3 reveals the error rate of the metrics such as MAE [17], RMSE [6], RAE [17], and RRSE [17] on the prediction results. The Relief [6], and PSO-F [21] algorithms generate a high error rate in the parameters like MAE [17], RMSE [6], RAE [17], and RRSE [17] produce high error rates of 0.26, 0.29, 90.54%, 84.29 and 0.18, 0.29, 63.58, and 82.63 that indicates the improper outlier selection. The PSO-KNN [16], ABC-GA [13], Ensemble [4], and PSO-SVM algorithms yield an average error rate in the MAE [17], RMSE [6], RAE [17], RRSE [17] metric than other approaches due to the insufficient selection of the classifier. The proposed method MI-WFL with CHBPSO gives a lower error rate in MAE [17], RMSE [6], RAE [17], and RRSE [17] error calculating metrics which have 0.17, 0.23, 49.66%, 54.62% compared to an existing algorithm which computes the pre-processing fuzzy logic to reduce the error rate. Moreover, it is reliable to match the dataset using the SVM classifier.

4.1 Processing time

In this section, the time complexity of state of art techniques is compared with the proposed work. Fig. 8 shows the relationship processing time using different algorithms namely relief [6], RF [8], PSO-KNN [16], PSO-SVM [3], Ensemble Method [4], ABC-GA [13], PSO-F [21]. In the ensemble learning algorithm [4] structure is uncertain in training the weak learner's model which consumes more time 3.2 sec but improves the accuracy to a high rate. The difficulties in the optimal FS make the ABC-GA algorithm reduce the processing time by 2.5 sec. It takes a long time to discover the best feature as the complexity of crossover and mutation processes. The hybridized techniques such as PSO-KNN [16], PSO-SVM [3], and PSO-F [21] take approximately 2.1 sec, 2.2 sec, and 2.1 sec due to the complexity of the dataset without pre-processing methods.

In Relief [6] and RF [8] algorithms, the processing time is low 1.9 sec and 1.8 sec, because of improper optimal FS. The proposed MI-WFL with CHBPSO algorithm is hybridized and it is superior to other algorithms in accuracy and reduces the processing time by nearly 1.6 sec by utilizing the minimum features to detect the disease and there it has better control to solve the overfitting problem using PK-SVM classifier.

5. Conclusion

To detect hepatitis disease in the early stage, the MI-WFL and CHBPSO algorithm have been proposed. Initially, MI-WFL pre-processing approach is implemented to fill in the missing values. To lower the error rate, the MI technique is used in the category column and the WFL method is used in the numerical column. Secondly, hybridized CHBPSO algorithm is developed to find the optimal reduced features are

antivirals, spiders, ascites, varices, and class which decreases the processing time and improves the performance of precision level. Finally, the PK-SVM classifier is adopted to minimize the overfitting problems. Experimental results show that the proposed method has a higher level of accuracy, minimum classification error, and less time complexity when compared to the other state-of-the art algorithm. In the future, real-time hospital data sets are concentrated to improve performance.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, methodology, software, formal analysis, resources, data curation, and writing-original draft preparation, writing-review, editing: C. Saranya Jothi, and Supervision: Carmel Mary Belinda

Reference

[1] Sartakhti, J. Salimi, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using novel hybrid method based on support vector machine and simulated annealing (SVM-SA)", *Computer Methods and Programs in Biomedicine*, Vol. 108, No. 2, pp. 570-579, 2012.

[2] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Learning", *International Journal of Computer Science*, Vol. 1, No. 2, pp. 111-117, 2006.

[3] T. Ahmad and M. N. Aziz, "Data preprocessing and feature selection for machine learning intrusion detection systems", *ICIC Express Lett*, Vol. 13, No. 2, pp. 93-101, 2019.

[4] S. Murugesan, R. S. Bhuvaneswaran, H. K. Nehemiah, S. K. Sankari, and Y. N. Jane, "Feature Selection and Classification of Clinical Datasets Using Bioinspired Algorithms and Super Learner", *Computational and Mathematical Methods in Medicine*, pp. 1-18, 2021.

[5] P. Bose, E. Kranakis, P. Morin, and Y. Tang, "Approximate range mode and range median queries", In: *Proc. of STACS 2005: 22nd Annual Symposium on Theoretical Aspects of Computer Science*, Stuttgart, Germany, Springer, Vol. 22, pp. 377-388, 2005.

[6] A.U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", *Mobile Information Systems*, pp. 1-21, 2018.

[7] P. Stefansson, K. H. Liland, T. Thiis, and I. Burud, "Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocessing technique for datasets with many observations", *Journal of Chemometrics*, Vol. 34, No. 3, p. e3195, 2020.

[8] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods", *Computational Intelligence and Neuroscience*, 2022.

[9] R. Aggrawal and S. Pal, "Sequential Feature Selection and Machine Learning Algorithm-Based Patient's Death Events Prediction and Diagnosis in Heart Disease", *SN Computer Science*, Vol. 1, No. 6, p. 344, 2020.

[10] H. Djellali, S. Guessoum, N. G. Zine, and S. Layachi, "Fast Correlation based Filter combined with Genetic Algorithm and Particle Swarm on Feature Selection", In: *Proc. of 2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B)*, IEEE, pp. 1-6, 2017.

[11] A. Famili, W. M. Shen, R. Weber, and E. Simoudis, "Data Preprocessing and Intelligent Data Analysis", *Intelligent Data Analysis*, Vol. 1, No. 1, pp. 3-23, 1997.

[12] A. E. Hegazy, M. A. Makhlouf, and G. S. E. Tawel, "Improved salp swarm algorithm for feature selection", *Computer and Information Sciences*, Vol. 32, No. 3, pp. 335-344, 2020.

[13] H. Djellali, A. Djebbar, N. G. Zine, and N. Azizi, "Hybrid Artificial Bees Colony and Particle Swarm on Feature Selection", In: *Proc. of Computational Intelligence and Its Applications: 6th IFIP TC 5 International Conference*, CIIA 2018, Oran, Algeria, Springer, pp. 93–105, 2018.

[14] P. Shunmugapriya and S. Kanmani, "A Hybrid Algorithm using Ant and Bee Colony Optimization for Feature Selection and Classification (AC-ABC Hybrid)", *Swarm and Evolutionary Computation*, Vol. 36, pp. 27-36, 2017.

[15] U. Agrawal, J. Arora, R. Singh, D. Gupta, A. Khanna, and A. khamparia, "Hybrid Wolf-Bat Algorithm for Optimization of Connection Weights in Multi-layer Perceptron", *ACM Transactions Multimedia*

Computing, Communications, and Applications, Vol. 16, No. 1s, pp. 1-20, 2020.

[16] A. Adamu, M. Abdullahi, S. B. Junaidu, and I. H. Hassan, "An hybrid particle swarm optimization with crow search algorithm for feature selection", *Machine Learning with Applications*, Vol. 6, p. 100108, 2021.

[17] H. Banka, and S. Dara, "A Hamming distance-based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification, and validation", *Pattern Recognition Letters*, Vol. 52, pp. 94-100, 2015.

[18] V. E. Christo, H. K. Nehemiah, B. Minu, and A. Kannan, "Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network", *Computational and Mathematical Methods in Medicine*, pp. 1-17, 2019.

[19] M. A. Quadir, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease", *Biomedicines*, Vol. 11, No. 2, p. 581, 2023.

[20] https://archive.ics.uci.edu/ml/datasets/hepatitis

[21] F. S. Gharehchopogh and S. K. Mousavi, "A Decision Support System for Diagnosis of Diabetes and Hepatitis, based on the Combination of Particle Swarm Optimization and Firefly Algorithm", *Journal of Health and Biomedical Informatics*, Vol. 6, No. 1, pp. 32-45, 2019.