Feature Selection Method Based on Honeybee-SMOTE for Medical Data Classification

Shobha Aswal[1*], Neelu Jyothi Ahuja[2], Ritika Mehra[3]
E-mail: shobha.swl@gmail.com, neelu@ddn.upes.ac.in, riti.arora@gmail.com
*Corresponding author
[1] Department of Computer Science and Engineering, Uttarakhand Technical University, India and Graphic Era Hill University, India
[2] Department of Systemics, School of Computer Science, University of Petroleum and Energy Studies, India
[3] Department of Computer Science and Engeneering, Dev Bhoomi Uttarakhand University, India

Keywords: Feature selection, classification, honeybee algorithm, SMOTE, C4.5 algorithm

Received: April 1, 2022

Bio-Medical data analysis has an important role in clinical practices. Usually, bio-medical data have complex issues like skewedness, redundant and irrelevant attributes etc. Several redundant and unrelated features frequently degrade the accuracy of the classifier while using with imbalanced datasets. The selection of features becomes critical in this situation. The key goal of feature selection is to establish a feature subspace that maintains classifier accuracy even as reducing the excessive computational learning cost and casting off noise. Appropriate feature selection approaches are highly dependent on their ability to match the issue context and uncover fundamental patterns within the data. This study's main goal is to construct a disease detection model that uses a hybrid feature-selection strategy based on Honeybee-SMOTE and classification using the c4.5 algorithm. The empirical results establish the suggested hybrid methodology's superiority over competing methods regarding the accuracy parameter, precision-parameter, recall-parameter, f1-score parameter and G-Mean parameter. The statistical analysis of the collected findings demonstrates that the suggested hybrid method outperforms and is competitive with existing state-of-the-art algorithms.

Povzetek: Metoda izbire atributov za strojno učenje je prilagojena zdravstvenim domenam.

1 Introduction

Data mining applications have proven their importance in the clinical decision-making process. Clinical data mining comprises five steps, the first step is the processing of data, second step is data transformation, then apply various datamining techniques in the third step, fourth step is pattern evaluation and final step is knowledge representation [1] [2]. Among these steps, data-preprocessing is used for cleaning the raw data, and the pre-processed data is used as input to the machine learning model. The difficulty in medical data mining is the shape and dimensionality of the dataset. Therefore, significant research is ongoing to handle it through efficient feature extraction, feature selection, and resampling techniques.

Feature selection is the essential step of data pre-processing and is used to find the relevant features to reduce redundancy. Feature selection techniques are classified into three parts: filter, wrapper, and embedded techniques [3]. The filter method is used to bring out the essential features. The wrapper method determines a specific learning algorithm to find the feature subset. It is a sequential feature selection approach and has a high computational cost. Finally, the embedded method combines the filter and wrapper approach to find the optimal solution.

Data mining has applications in various domains, and it plays a vital role in medical data analysis. For example, feature selection is essential in medical data classification due to the data's imbalanced, skewed, and asymmetric nature. As a result, the raw medical data contains irrelevant and redundant features, critical for classification models. To deal with this problem, feature selection and extraction procedures are used so that only relevant data can be used to enhance the classification model's performance. Feature extraction derives new feature space from original features with reduced dimensions without losing data, and feature selection picks only the relevant features from original dataset.

Feature selection is a kind of NP-Hard Problem. Metaheuristic based algorithms can handle multi-dimensional data and provide solutions near to the global optima. Several metaheuristic algorithms like Genetic Algorithms (G.A.), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Grey Wolf Optimizer (GWO), Whale-Optimization Algorithm (WOA), Bat Algorithm (B.A.), Crow Search Algorithm (CSA) have been used for feature selection.

Medical data classification has been done at the data and algorithmic levels in the past research. Data level is at the pre-processing level where resampling techniques

such as under-sampling and over-sampling have been used. Various classification models are designed to offer the best solution to the given problem at the algorithmic level. Resampling uses under-sampling and over-sampling of class instances to increase or decrease class size. SMOTE (Synthetic Minority Oversampling Technique) is a widely used resampling technique proposed by Chawla et al. in 2002 [4]. Due to its success rate, many improvised versions like borderline-SMOTE [5], safe-level-SMOTE [6], RSMOTE [7] have been proposed in recent years.

The research has shown that hybrid algorithms effectively improve medical applications' accuracy. Hence, a hybrid classification model is proposed; the proposed model hybrids the SMOTE with Honeybee inspired feature selection for classifying medical datasets.

SMOTE works in feature space by increasing the sample size of the minor class by creating new synthetic instances so that minor classes can't be ignored and the correct decision can be taken. The new samples are generated by identifying the percentage of oversampling and the number of K-nearest neighbors. To produce new instances for continuous features, compute the spacing between a feature vector of minority class and one of its k-nearest neighbors. Then, compute the product of the result of the previous step and a number in-between 0 and 1 at random. Then add the obtained result to the original feature's feature value. It is shown in equation 1. [8]

$$x_n = x_0 + \delta . (x_{0i} - x_0) \quad (1)$$

Where, x_n is the new feature vector produced, x_0 is the feature vector of each instance, x_{0i} is the selected nearest neighbour of x_0 and $\boldsymbol{\delta}$ is the random number chosen between 0 and 1.

The Honeybee Mating Optimization Algorithm is inspired by the swarm intelligence techniques developed by Abbas [9][10]. It describes the marriage procedure of honey bees. A swarm of honey bees consists of three bees' types: queens, drones, and workers. The milky-white colour substance called royal jelly is used to feed the queens. Queen bees are more giant due to their royal jelly diet. The queen bee's life is five or more years, while drones and worker bees live only for 6 months. A queen meets several hundred drones in her life span. The job of the drone is to provide sperm to the Queen. Workers are skilled with brood care. For mating process, the queens take flight away from the hive, dance with the drones, and then mate. The Queen is given some energy at the start of her flight, and after mating, she returns to the hive when her energy level reaches a certain level or her spermathecal supply is exhausted.

In the process of mating, the queen mates with many drones and store the genotype of drones in her spermathecal to fertilize the eggs. Each brood is created by using the genotype of various drones and the Queen. The reproduction process forms new colonies in which the workers have differed genetically. The work of worker bees is to collect the pollen and nectar; after some time, the stored nectar becomes honey. Pollen is used for instant purposes to feed the broods, while nectar is stored for long-term purposes. The flowchart of the Honey-bee Optimization Algorithm is shown in figure 1.

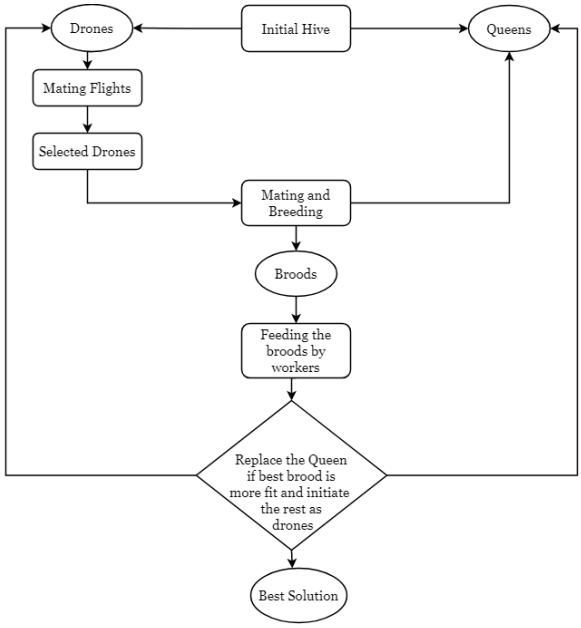


Figure 1: Flowchart of HBO algorithm [11].

The probability that a drone mate with a queen is shown in equation 2.

$$prob(Q, D) = e^{-\Delta f / s(t)} \quad (2)$$

Where prob (Q, D) is the probability of successful mating of Queen and drone; Δf is the absolute difference between the fitness of drone D and fitness of queen Q. $s(t)$ is the speed of Queen at time t. After each mating in the space, the queen speed is calculated using the following equation 3.

$$s(t+1) = \alpha * s(t) \quad (3)$$

And the fall in energy is calculated by the following equation 4.

$$E(t+1) = E(t) - \gamma \quad (4)$$

Where α is a factor between 0 to 1; and γ is the energy drop after each transition.

The rest of the paper is organized as follows: Section 2 discusses the related work, Section 3 presents the proposed methodology and algorithm, Section 4 describes the experimental setup, Data set description and results obtained and finally Section 5 concludes the work with some future scope.

2 Related work

Swarm intelligence-based metaheuristic algorithms have been extensively considered to resolve optimization problems. Susana M. Vieira et al. [12] suggested a modified binary particle swarm optimization (MBPSO)

method for feature selection with an SVM classifier to predict the mortality in septic patients. The SVM used the wrapper approach of feature selection to predict the outcome. Kung-Jeng Wang et al. [8] proposed a hybrid classifier by combining Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO). It improves breast cancer patients' 5-year survival rates. Finally, Douglas Rodrigues et al.[13] presented a combination of BAT Algorithm and Optimum Path Forest Classifier, a wrapper based feature selection method.

Ibrahim Aljarah et al. [14] proposed a hybrid method based on Grasshopper Optimization Algorithm (GOA) and SVM optimizer for feature selection. Hossam M. Zawbaa et al.[15] proposed hybrid GWO-ALO by applying GWO's (grey wolf optimization) searchability and ALO's (Antlion optimization) exploitation ability to select the relevant features subset. [16] M. Mafarja et al. proposed two binary variants of the WOA approach for feature selection. The first variant used Tournament and Roulette Wheel selection mechanism for exploration, and the second variant used crossover and mutation operators for the exploitation phase.

Another feature selection approach, known as the Chaotic crow search algorithm (CCSA), is proposed by Gehad Ismail Sayed et al.[17] and applied on 20 benchmark datasets, including medical datasets from the UCI machine learning repository. [18] A.P. Engelbrecht et al. proposed a set-based particle swarm optimization algorithm with KNN classifier and compared it with other PSO wrapper algorithms and found it compelling. Sankalp Arora et al. [19] proposed a binary Butterfly Optimization Algorithm (BOA) for better accuracy prediction to deal with the feature selection approach. Lalit Kumar et al. (Kumar and Bharti 2019) proposed a modified-binary particle swarm optimization (modified-BPSO) algorithm to choose the relevant features for classification. Their work has been compared with binary particle swarm optimization (BPSO) algorithm and Moth Flame Optimization (MFO) algorithm. Ah. E.Hegazy et al. [21] proposed an improvised salp swarm algorithm for the selection of features using a KNN classifier on 23 benchmark datasets from the UCI library. In their research, Jaime Lynn Speiser (Speiser, 2021) reveals that the Binary Mixed Model (BiMM) forest with eliminated background gives maximum accuracy compared with other feature selection techniques. However, it predicts the same accuracy as other techniques predict in a few cases. Adamu et al.[22] proposed an improvised crow search algorithm to deal with feature selection. Mohammad Tubishat et al. [23] introduced an enhanced version of the Salp swarm algorithm (SSA)known as Dynamic Salp swarm algorithm (DSSA) with a KNN classifier to handle the feature selection. In [24] Monika Arya et. al proposed a deep learning-based ensemble classification method for early detection of diabetes. In [25] the authors proposed a Rough-Mereology framework for the classification of Hepatitis-C-Virus and Coronary Heart Disease. Their work analyzed medical datasets by using rough-based granular computing. U. Ramasamy et.al [26] proposed a decision tree based classifier for the prediction of Rheumatoid-Arthritis disease. Ali A. Abaker et. al [27] have analyzed the prediction algorithms and found logistic regression techniques performed better among random forest and KNN classifier.

3 Proposed methodology and algorithm

The proposed hybrid framework for classifying medical data set contains the following steps: -

Step 1: Data pre-processing- In this step, data cleaning, dimension reduction, data transformation and data integration is done. First, the missing value imputation is done by applying the mean strategy. Then, after data imputation, the data distribution is converted to Gaussian distribution by applying the Yeo-Johnson power transformation technique. Finally, data standardization has is done using standard scaling techniques.

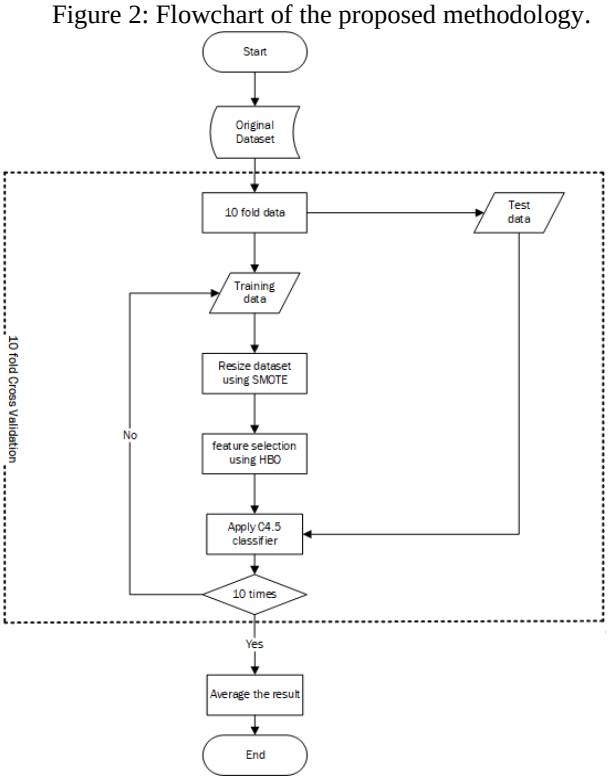
Step 2: Applying 10-fold cross-validation - In this step, the cross-validation applied to evaluate the performance of the proposed model is 10-fold cross-validation. Among 10 independent folds, randomly chosen 9 folds are used as training data, and the rest are used as test data.

Step 3: Feature selection- SMOTE algorithm is applied on each fold of training data, and the sample size increased by 900%. After that, the Honey bee optimization algorithm is applied for the feature selection.

Step 4: Classification-The classification model has been constructed using the c4.5 algorithm.

The implementation of the proposed algorithm is shown in figure 2.

Figure 2: Flowchart of the proposed methodology.



Input: original imbalanced training dataset I
Output: return selected features and their performance values

Step1:Train <- I.training_odd_sample
Test <- I.training_even_samples
Opt_train1 ← 10 fold cross-validation (Train)
Oversamp_train <- SMOTE (Opt_train1)
Step2: Initialize population
While (Maximum number of iteration encountered)
Step3: Choose Queen Q randomly
Step4: Select maximum number of mating flights M
do while i<=M
Initialize Queen's Spermetheca (Sp), energy (E) and speed (S)
Select α
Do while (E>0 and Sp ! = full)
Select a drone (D)
If D passes the probabilistic condition
Add D's sperm in Sp
End if
S (t+1) = α * S(t)
E (t+1) = α * E(t)
end do
Step5: do j=1, Size of Spermetheca (Sp)
Select sperm from Spermetheca
Apply a crossover operator between Queen's genotype and selected sperm generate a brood.
Select a worker randomly
Improve brood fitness by using the selected worker
If (brood_fitness > queen_fitness)
Q=brood
Else If the brood_fitness is better than one of the drones_fitness
D=brood
Endif
Endif
For (pop_size)
{Update Q, D
Calculate the fitness of all broods}
End do

Step 6: Initialize empty tree
If(S=0) {Return single node of failure value}
End if
If(S=T)
Return single node of attained target value,
End if
If (R is empty) then
Return single node of majority attribute in T
End if
For all R_i
Select the attribute D of highest gain value,
D← $\{d_1, d_2, d_3, \ldots\ldots, d_n\}$
The subset of S ← $\{S_1, S_2, S_3, \ldots\ldots, S_n\}$ made with records of D
Return the tree made of all attached subtrees
End for

Step 7: Return the best features and their performance

The model's performance is calculated based on various parameters such as precision, recall, f1-score and g-mean values.

4 Experimental setup, dataset description and results

The proposed algorithm is implemented on a device with the following technical specifications: Intel-i5 7200U, 8GB RAM, 1 T.B. Hard disk, and NVIDIA GTX 760MX Graphics and the study was simulated on MATLAB version 7. The initial assumed parameters are summarized in Table 1.

Table 1: Initial assumed parameters

Parameters	Values
Search Population $Y_i(i = 1,2, \ldots., n)$	30
Upper Bound	100
Lower Bound	0
No. of Iterations	100
$Queen_{initial\_val}$	$f\_Score_{init\_fit}$
$Queen_{initial\_pos}$	[1, T_v]

Table 2 provides a description of the medical data sets that were acquired from the UCI machine learning library. Various data pre-processing techniques like data cleaning, data integration, data transformation and data reduction have been applied before validation.

Table 2: Data set description

Dataset	No. of Instances	No. of Attributes	% of minor class	% of major class
Indian Liver Patient	579	11	20	80
Lung Cancer	27	57	20	80
Pima Indians Diabetes	768	9	20	80
Thyroid Disease	1275	15	20	80
Hepatitis	706	9	20	80

Performance Evaluation parameters

The confusion matrix is used to analyze the performance of the classifier. It counts the number of predicted data over actual data. The parameters of the confusion matrix are described in table 3.

Table 3: Confusion matrix

		Predicted	
		Negative	*Positive*
Actual	*Negative*	T.N.	F.P.
	Positive	F.N.	T.P.

True Positive (T.P.): a score of how many cases were correctly categorised while the real class was affirmative.
False Positive (F.P.): the number of instances where the genuine class is positive but the classification was incorrect.
True Negative (T.N.): the number of cases classified correctly when the real class is negative.
False Negative (F.P.): the number of cases classified incorrectly when the real class is negative.
Accuracy: the accuracy of the classifier is the total number of instances classified by the classifier correctly over a total number of instances and is calculated in equation 5.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (5)$$

Recall: the ratio of positive class classified successfully to the total number of classes predicted correctly. Recall is calculated in equation 7.

$$\text{Recall} = \frac{\text{TP}}{TP+FN} \quad (6)$$

Precision: It's the proportion of accurately anticipated positive classes to the total number of positive classes.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

F_1Score: The F_1Score, commonly known as the F-measure, is calculated by averaging precision and recall. The formula to calculate F1Score is as follows:

$$F1Score = \frac{2*Precision*Recall}{(Precision+Recall)} \quad (8)$$

G-mean:

The G-mean metric is the widely accepted parameter to calculate the overall accuracy of the classifier. It is calculated by using the formula below:

$$Gmean = \sqrt{(Recall * \frac{TN}{TN+FP})} \quad (9)$$

Results and discussions

The implementation results of the hybrid algorithm over five medical data stes are summarized in table Table 4.

Table 4: Implementation results of Hybrid (Proposed) Algorithm

Data set	Accuracy	Precision	Recall	F1-Score	G-Mean
Indian Liver Patient	0.98	0.77	0.67	0.74	0.75
Lung Cancer	0.96	0.67	0.65	0.69	0.72
Pima Indians Diabetes	0.97	0.69	0.63	0.68	0.77
Thyroid Disease	0.98	0.78	0.68	0.76	0.69
Hepatitis	0.97	0.73	0.65	0.78	0.68

The graph in figure 3 shows the performance of the Hybrid (proposed) Algorithm.

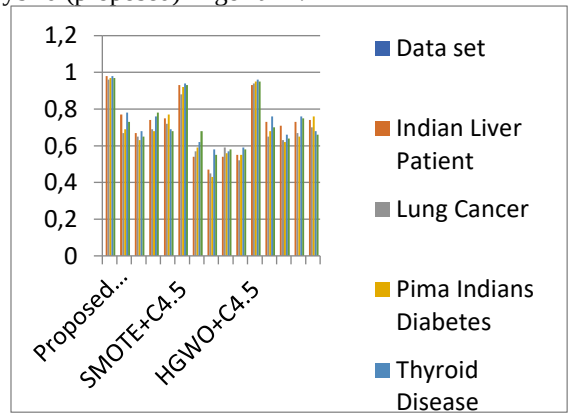


Figure 3: Performance of Hybrid (proposed) Algorithm.

Table 5: Performance comparison

	Proposed (SMOTE+HBO+C4.5)					SMOTE+C4.5					HGWO+C4.5				
Data set	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**G-Mean**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**G-Mean**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**G-Mean**
Indian Liver Patient	0.98	0.77	0.67	0.74	0.75	0.93	0.54	0.47	0.54	0.55	0.93	0.73	0.71	0.73	0.74
Lung Cancer	0.96	0.67	0.65	0.69	0.72	0.88	0.57	0.45	0.59	0.52	0.94	0.65	0.63	0.67	0.70
Pima Indians Diabetes	0.97	0.69	0.63	0.68	0.77	0.92	0.59	0.43	0.56	0.55	0.95	0.68	0.62	0.65	0.76
Thyroid Disease	0.98	0.78	0.68	0.76	0.69	0.94	0.62	0.58	0.57	0.59	0.96	0.76	0.66	0.76	0.68
Hepatitis	0.97	0.73	0.65	0.78	0.68	0.93	0.68	0.55	0.58	0.58	0.95	0.70	0.64	0.75	0.66

The suggested hybrid algorithm is compared with the SMOTE+C4.5 and HGWO+C4.5 algorithms in terms of performance. The results are summarized in Table 5.

The suggested hybrid algorithm's performance is compared to two recently published related algorithms for feature selection in medical data sets. The two works that are compared with proposed hybid approach are proposed by S. Sundaramurthy et al. 2020 [28] and Rostami, M.et al. 2020 [29] . The graph in figure 4 shows the comparison in the accuracy of proposed hybrid algorithm with SMOTE+C4.5 and HGWO+C4.5 algorithm. The comparison is done on the basis of the efficiency of the compared and the proposed technique to select the relevant features to form a feature subset. Feature ranking of the available features is done based on the gini importance. The Indian Liver Patient Dataset is chosen for comparison. The description of the data set is given in table 2. The data set contains 11 attributes or features. The equation for Gini importance is given in Eq. 10.

$$Gini\ (P) = \sum_{i=1}^{n} pi\ (1 - pi) \qquad (10)$$

Table 6 shows the feature rank of each feature.

Table 6: Feature ranking according to Gini Importance of feature.

Feature rank	Feature number	Gini importance
1	feature5	(0.06192)
2	feature 8	(0.055762)
3	feature 3	(0.054146)
4	feature 2	(0.051976)
5	feature 4	(0.050384)
6	feature 7	(0.048218)
7	feature 1	(0.048100)
8	feature 4	(0.047880)
9	feature 6	(0.0478027)
10	feature 9	(0.047667)
11	feature 10	(0.038753)

The graph in figure 4 shows the feature rank and gini importance of each feature present in the data set.

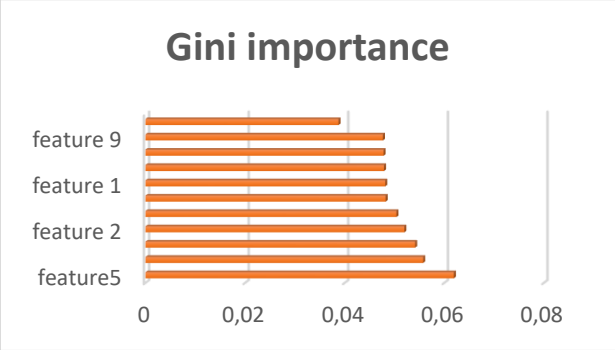


Figure 4: Feature rank and Gini importance

Further, feature selection using the proposed hybrid approach and two recent techniques was done to form a feature subset. The top five features of the feature subset, formed using the three approaches are summarized in Table 7.

Table 7: Top five features of feature subset

Feature selection approach	Top five features of feature subset
Hybrid (Proposed)	feature 5, feature 8, feature 3, feature 2, feature 4
PSO-based multi objective feature selection	feature 8, feature 7, feature 4, feature 9, feature 10
Ensemble based feature selection	feature 5, feature 2, feature 4, feature 7, feature 10

The above table shows that proposed hybrid approach forms the feature subset of most relevant features. While the compared recent approaches are not able to identify the most relevant features to form the feature subset. Thus, it can be suggested that the proposed hybrid approach is more efficient in selecting relevant features for classification and hence improving the overall performance of the classifier.

5 Conclusion and future scope

In this study, a hybrid feature selection approach SMOTE+HBO+C4.5 is proposed. Experimental findings showed that in all five evaluation criteria, the proposed approach significantly outperformed. Furthermore, the suggested SMOTE+HBO+C4.5 prediction model was examined using two state of art algorithms SMOTE+C4.5 and HGWO+C4.5. The suggested approach performed better than the other approaches for medical data classification. In terms of enhancing the classification performance of medical datasets, the framework can be a useful tool for doctors and researchers alike. In future, the work can be extended with advanced optimization functions along with hybrid classifiers.

6 References

[1] U. Fayyad, P. Stolorz, Data mining and KDD: Promise and challenges, Futur. Gener. Comput. Syst. 13 (1997) 99–115. https://doi.org/10.1016/s0167-739x(97)00015-0.

[2] J.H. Holmes, Knowledge Discovery in Biomedical Data: Theory and Methods, Error, Elsevier Inc., 2013. https://doi.org/10.1016/B978-0-12-401678-1.00007-5.

[3] B. Remeseiro, V. Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 25–29. https://doi.org/10.1016/j.compbiomed.2019.103375.

[4] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, snopes.com: Two-Striped Telamonia Spider, J. Artif. Intell. Res. 16 (2002) 321–357. https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.

[5] H. Han, W.Y. Wang, B.H. Mao, Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 3644 LNCS (2005) 878–887. https://doi.org/10.1007/11538059_91.

[6] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 5476 LNAI (2009) 475–482. https://doi.org/10.1007/978-3-642-01307-2_43.

[7] B. Chen, S. Xia, Z. Chen, B. Wang, G. Wang, RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise, Inf. Sci. (Ny). 553 (2021) 397–428. https://doi.org/10.1016/j.ins.2020.10.013.

[8] K.J. Wang, B. Makond, K.H. Chen, K.M. Wang, A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients, Appl. Soft Comput. J. 20 (2014) 15–24. https://doi.org/10.1016/j.asoc.2013.09.014.

[9] H.A.H. Abbass, A monogenous MBO approach to satisfiability, Proceeding Int. Conf. Comput. Intell. Model. Control Autom. CIMCA. (2001). https://www.researchgate.net/publication/2481231_A_Monogenous_MBO_Approach_to_Satisfiability.

[10] H.A. Abbass, MBO: Marriage in honey bees optimization a haplometrosis polygynous swarming approach, Proc. IEEE Conf. Evol. Comput. ICEC. 1 (2001) 207–214. https://doi.org/10.1109/cec.2001.934391.

[11] O.B. Haddad, A. Afshar, M.A. Mariño, Multireservoir optimisation in discrete and

continuous domains, Proc. Inst. Civ. Eng. Water Manag. 164 (2011) 57–72. https://doi.org/10.1680/wama.900077.

[12] S.M. Vieira, L.F. Mendonça, G.J. Farinha, J.M.C. Sousa, Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients, Appl. Soft Comput. J. 13 (2013) 3494–3504. https://doi.org/10.1016/j.asoc.2013.03.021.

[13] D. Rodrigues, L.A.M. Pereira, R.Y.M. Nakamura, K.A.P. Costa, X.S. Yang, A.N. Souza, J.P. Papa, A wrapper approach for feature selection based on Bat Algorithm and Optimum-Path Forest, Expert Syst. Appl. 41 (2014) 2250–2258. https://doi.org/10.1016/j.eswa.2013.09.023.

[14] I. Aljarah, A.M. Al-Zoubi, H. Faris, M.A. Hassonah, S. Mirjalili, H. Saadeh, Simultaneous Feature Selection and Support Vector Machine Optimization Using the Grasshopper Optimization Algorithm, Cognit. Comput. 10 (2018) 478–495. https://doi.org/10.1007/s12559-017-9542-9.

[15] H.M. Zawbaa, E. Emary, C. Grosan, V. Snasel, Large-dimensionality small-instance set feature selection: A hybrid bio-inspired heuristic approach, Swarm Evol. Comput. 42 (2018) 29–42. https://doi.org/10.1016/j.swevo.2018.02.021.

[16] M. Mafarja, S. Mirjalili, Whale optimization approaches for wrapper feature selection, Appl. Soft Comput. 62 (2018) 441–453. https://doi.org/10.1016/j.asoc.2017.11.006.

[17] G.I. Sayed, A.E. Hassanien, A.T. Azar, Feature selection via a novel chaotic crow search algorithm, Neural Comput. Appl. 31 (2019) 171–188. https://doi.org/10.1007/s00521-017-2988-6.

[18] A.P. Engelbrecht, J. Grobler, J. Langeveld, Set based particle swarm optimization for the feature selection problem, Eng. Appl. Artif. Intell. 85 (2019) 324–336. https://doi.org/10.1016/j.engappai.2019.06.008.

[19] S. Arora, P. Anand, Binary butterfly optimization approaches for feature selection, Expert Syst. Appl. 116 (2019) 147–160. https://doi.org/10.1016/j.eswa.2018.08.051.

[20] L. Kumar, K.K. Bharti, An improved BPSO algorithm for feature selection, Springer Singapore, 2019. https://doi.org/10.1007/978-981-13-2685-1_48.

[21] A.E. Hegazy, M.A. Makhlouf, G.S. El-Tawel, Improved salp swarm algorithm for feature selection, J. King Saud Univ. - Comput. Inf. Sci. 32 (2020) 335–344. https://doi.org/10.1016/j.jksuci.2018.06.003.

[22] A. Adamu, M. Abdullahi, S.B. Junaidu, I.H. Hassan, An hybrid particle swarm optimization with crow search algorithm for feature selection, Mach. Learn. with Appl. 6 (2021) 100108. https://doi.org/10.1016/j.mlwa.2021.100108.

[23] M. Tubishat, S. Ja'afar, M. Alswaitti, S. Mirjalili, N. Idris, M.A. Ismail, M.S. Omar, Dynamic Salp swarm algorithm for feature selection, Expert Syst. Appl. 164 (2021) 113873. https://doi.org/10.1016/j.eswa.2020.113873.

[24] M. Arya, H. Sastry G, A. Motwani, S. Kumar, A. Zaguia, A Novel Extra Tree Ensemble Optimized DL Framework (ETEODL) for Early Detection of Diabetes, Front. Public Heal. 9 (2022) 1–13. https://doi.org/10.3389/fpubh.2021.797877.

[25] M.M. Eissa, M. Elmogy, M. Hashem, Rough-mereology framework for making medical treatment decisions based on granular computing, Inform. 40 (2016) 343–352.

[26] U. Ramasamy, S. Sundar, An Illustration of Rheumatoid Arthritis Disease Using Decision Tree Algorithm, Inform. 46 (2022) 109–119. https://doi.org/10.31449/inf.v46i1.3269.

[27] A.A. Abaker, F.A. Saeed, A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications, Inform. 45 (2021) 117–125. https://doi.org/10.31449/inf.v45i1.3111.

[28] S. Sundaramurthy, P. Jayavel, A hybrid Grey Wolf Optimization and Particle Swarm Optimization with C4.5 approach for prediction of Rheumatoid Arthritis, Appl. Soft Comput. J. 94 (2020) 106500. https://doi.org/10.1016/j.asoc.2020.106500.

[29] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, Integration of multi-objective PSO based feature selection and node centrality for medical datasets, Genomics. 112 (2020) 4370–4384. https://doi.org/10.1016/j.ygeno.2020.07.027.