Krill Herd Algorithm Based Medical Data Classification Using Hybrid Adaboost KNN Classification

D. Mahammad Rafi
Research Scholar
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology,
Chennai, India
mahammadrafi0780@gmail.com

C.R. Bharathi
Associate Professor
Department of Electronics and Communications,
Vel Tech University, Chennai, Tamilnadu, India.

Abstract: **Medical classification is the process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers. The diagnoses and procedures are usually taken from a variety of sources within the health care record, such as the transcription of the physician's notes, laboratory results, radiologic results, and other sources. It is a kind of complex optimization problem and it also needs to provide diagnosis aid accurately. Many different data mining techniques exist for medical data classification But, the classification accuracy of these models is limited often when the relationship of input/output datasets are complex and/or non-linear. Researchers tried to use diverse methods to get better accuracy of data classification. It is essential to perform medical data classification in order to detect the diseases. Hence to overcome those issues our proposed method is used. Here initially the pre-processing will be applied to extract useful data and to convert suitable sample from raw medical datasets. After preprocessing for optimal selection of features krill herd algorithm (KHA) is used. After selection of optimal features for classification of medical data hybrid adaboost and K-Nearest Neighbor (KNN) is used. This hybrid classification algorithm classifies the medical data as normal and abnormal. Hence our proposed method accurately classifies the medical data using optimal features. The performance of the proposed method is evaluated in terms of accuracy, sensitivity and specificity. The proposed method will be implemented in MATLAB.**

Keywords: Medical classification, krill herd algorithm, adaboost, K-Nearest Neighbor, accuracy, sensitivity and specificity.

1. INTRODUCTION

According to the World Health Organization report, the people in Suboptimal Health Status (SHS), also known as “the third state” (between being healthy and falling sick), account for 75% among the world population. A considerable part of them would pay close attention to their health hoping to get preventive health examination or educate themselves with similar patient’s medical records [1]. Annual geriatric medical examination (GME) is now an integral part of elderly healthcare for many developed countries. However, it is always a difficult task for healthcare professionals to provide an overall report on personal health after a comprehensive medical check-up is performed with hundreds of parameters [2]. Clinical data are one of the main sources of information for the diagnosis and treatment of a large number of illnesses and abnormalities [3]. Medical diagnosis has been being considered as one of the most important and necessary processes in clinical medicine that determines acquired diseases of patients from given symptoms. According to Kononenko, diagnosis commonly relates to the probability or risk of an individual developing a particular state of health over a specific time, based on his or her clinical and non-clinical profile. Due to the increased volume of information available to physicians from modem medical technologies, medical diagnosis contains a lot of incomplete, uncertainty, and inconsistent information, which is essential information about medical diagnosis problems [4, 5].

Many different forms of uncertainty in data have been recognized: one comes from conflicting or incomplete information, as well as from multiple interpretation of some phenomenon; another arises from lack of well-defined distinctions or from imprecise boundaries [6]. Irrespective of all these uncertainties in most of the medical diagnosis problems, there exist some patterns, and then experts make decision according to the similarity between an unknown sample and the basic patterns [7]. Utilizing the medical diagnosis process, characteristics of an individual patient are matched to a computerized clinical knowledge base and patient-specific assessment and recommendations are then presented to the clinical or the patient for a decision [8]. It is obvious

that, on the one hand, a great amount of knowledge on various medical activities, such as screening, diagnosis, classification, treatment, prognosis, etc. is hidden within such clinical data as well as, on the other hand, it is not possible to effectively process such data by physicians using conventional (e.g., statistical) techniques. For this reason, various techniques for knowledge discovery in medical data or medical data mining are developed [9]. Medical data classification (MDC) refers to learning classification models from medical datasets and aims to improve the quality of health care. Medical data classification can be used for diagnosis and prognosis purposes. Medical data exhibit unique features including noise resulting from human as well as systematic errors, missing values and even sparseness. The quality of data has a large implication for the quality of the mining results. It is necessary to perform preprocessing steps in order to remove or at least alleviate some of the problems associated with medical data [10].

Then feature selection is done to enhance medical data classification. For the classifier combination two main approaches are used: classifiers fusion and classifiers selection. In the first method, all classifiers in the ensemble contribute to the decision of the MC system, e.g. through sum or majority voting. In the second approach, a single classifier is selected from the ensemble and its decision is treated as the decision of the MC system [11]. SVM is also a classifier. The SVM attempts to determine a tradeoff between minimizing the training set error and maximizing the margin in order to achieve the best generalization ability and remain resistant to over fitting. Due to its advantageous nature, SVM has been applied to a wide range of classification tasks. In particular, SVM has been shown to perform very well on many medical diagnosis tasks. However, there is still a need for improving the SVM classifier's performance [12]. Hence to overcome the issue several new methods learn to diagnose disease in a fully data driven manner, using multivariate classification or regression to directly map from imaging data to diagnosis. These techniques are not restricted by current knowledge on disease-related radiological patterns and often have higher diagnostic accuracy than more traditional quantitative analysis based on simple volume or density measures [13]. However the transfer of critically ill patients requires good coordination to provide the diagnostic tools and the most appropriate treatment for their conditions [14]. Hence it is very important to classify the disease on time and thereby providing treatments to the patients for risk avoidance. Sometimes same properties with the same symptoms of diseases required different treatments. Therefore accurate decisions and classification of data is required [15].

2. Related Works

Narang, et al. [16] indicated that in order to assess the current general acceptance within the medical community of shaken baby syndrome (SBS), abusive head trauma (AHT), and several alternative explanations for findings commonly seen in abused children. That was a survey of physicians frequently involved in the evaluation of injured children at 10 leading children's hospitals. Physicians were asked to estimate the likelihood that subdural hematoma, severe retinal hemorrhages, and coma or death would result from several proposed mechanisms. Of the 1378 physicians surveyed, 682 (49.5%) responded, and 628 were included in the final sample. A large majority of respondents felt that shaking with or without impact would be likely or highly likely to result in subdural hematoma, severe retinal hemorrhages, and coma or death, and that none of the alternative theories except motor vehicle collision would result in those 3 findings.

Mohapatra, et al. [17] proposed two variations of kernel ridge regression (KRR), namely wavelet kernel ridge regression (WKRR) and radial basis kernel ridge regression (RKRR) for classification of microarray medical datasets. Microarray medical datasets contained irrelevant and redundant genes which caused high number of gene expression i.e. dimensionality and small sample sizes. To overcome the curse of dimensionality of the microarray datasets, modified cat swarm optimization (MCSO), a naturally inspired evolutionary algorithm, was used to select the most relevant features from the datasets. The adequacies of the classifiers were demonstrated by employing four from each binary and multi-class microarray medical datasets.

The classification complications in medical area are solved based on the outcome of medical analysis or report of medical treatment by the medical specialist. Kumar, et al. [18] focused on applying Rough set based data mining techniques for medical data to discover locally frequent diseases. The method applied Optimistic Multi-granulation rough set model (OMGRS) for medical data classification. Multi-granulation rough set provided efficient results than single granulation rough set model and soft rough set based classifier model. The results of applying the OMGRS methodology to medical diagnosis based upon selected information. The performance of the proposed optimistic multi granulation Rough set based classification was compared with other rough set based (RS), Kth Nearest Neighbor (KNN) and Back propagation neural network (BPN) approaches using various classification Measures.

The existing methods of fuzzy soft sets in decision making are mainly based on different kinds of level soft sets, and it is very difficult for decision makers to select a suitable level soft set in most instances. Zhaowen, et al. [19] presented an approach to fuzzy soft sets in decision making to avoid selecting a suitable level soft set and to apply that approach to solve medical diagnosis problems. That approach combined grey relational analysis with the Dempster Shafer theory of evidence. That first utilized grey relational analysis to calculate the grey mean relational degree, by which they calculated the uncertain degree of various parameters. Then, on the basis of the uncertain degree, the suitable basic probability assignment function of each independent alternative with each parameter could be obtained.

Medical datasets consume enormous amount of information about the patients, diseases and the physicians. Diseases diagnosis required many expensive tests to predict the diseases. Cost of disease prediction and diagnosis can be reduced by applying machine learning and data mining methods. Disease prediction and decision making plays a significant role in medical diagnosis.

Inbarani and H. Hannah, [20] conveyed a neighborhood rough set classification approach to deal with medical datasets. Five benchmarked medical datasets had been used in that research work for studying the impact of proposed work in decision making.

Hernández-Chan, et al. [21] demonstrated a method of the application of collective intelligence in medical diagnosis by applying consensus methods. They compared the accuracy obtained with that method against the diagnostics accuracy reached through the knowledge of a single expert. They used the ontological structures of ten diseases. Two knowledge bases were created by placing five diseases into each knowledge base. They conducted two experiments, one with an empty knowledge base and the other with a populated knowledge base. For both experiments, five experts added and/or eliminated signs/symptoms and diagnostic tests for each disease. After that process, the individual knowledge bases were built based on the output of the consensus methods. In order to perform the evaluation, they compared the number of items for each disease in the agreed knowledge bases against the number of items in the GS (Gold Standard).

Data mining is the process of extracting hidden information from a large set of database and it can help researchers gain both novel and deep insights of unprecedented understanding of large biomedical datasets. Data mining can uncover new biomedical and healthcare knowledge for clinical decision making. Parvathi, I and Siddharth Rautaray, [22] introduced data mining in general (e.g. Definition, tasks of data mining, application of data mining)and gave a brief summarization of various data mining algorithms used for classification, clustering, and association. Discussion was made to enable the disease diagnosis and prognosis, and the discovery of hidden biomedical and healthcare patterns from related databases was offered along with a discussion of the use of data mining to discover such relationships as those between health conditions and a disease, relationships among diseases.

3. PROPOSED METHODOLOGY

Medical classification is the process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers. The diagnoses and procedures are usually taken from a variety of sources within the health care record, such as the transcription of the physician's notes, laboratory results, radiologic results, and other sources. It is a kind of complex optimization problem and it also needs to provide diagnosis aid accurately. Many different data mining techniques exist for medical data classification But, the classification accuracy of these models is limited often when the relationship of input/output datasets are complex and/or non-linear. Researchers tried to use diverse methods to get better accuracy of data classification. It is essential to perform medical data classification in order to detect the diseases. In existing methods [23] after preprocessing of medical data, orthogonal local preserving projection (OLPP) is used to reduce the feature dimension. Finally modified group search optimizer algorithm is used with Fuzzy Min-Max neural network for classification of diseases. Due to lack of optimal feature selection algorithms there exist many relevant and irrelevant features of medical data. Hence to overcome those issues our proposed method is used. Here initially the pre-processing will be applied to extract useful data and to convert suitable sample from raw medical datasets. After preprocessing for optimal selection of features krill herd algorithm (KHA) is used. Optimal feature selection, also known as optimal variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Krill Herd algorithm (KHA) is a class of nature-inspired algorithm, which simulates the herding behavior of krill individuals. It has been successfully utilized to tackle many optimization problems in different domains and found to be very efficient. After selection of optimal features for classification of medical data hybrid adaboost and KNN is used. This hybrid classification algorithm classifies the medical data as normal and abnormal. The overall flow diagram of the proposed technique is shown in fig.1.

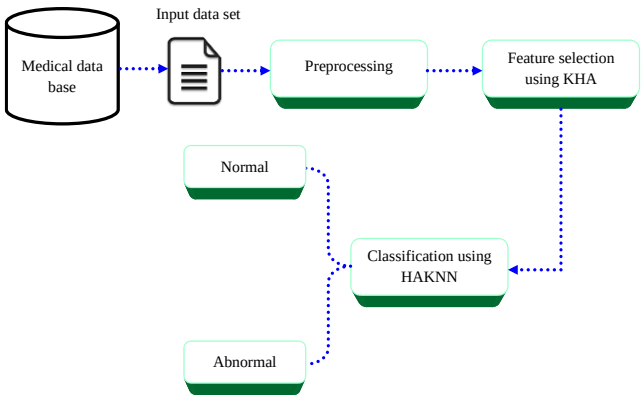

Figure.1 The proposed flow diagram

The proposed methodology has three main stages namely,

- Preprocessing
- Optimal feature selection

❖ Classification

3.1 Preprocessing

Data pre-processing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset. The raw medical dataset is input for the preprocessing step. This raw data is highly susceptible to noise, missing values and inconsistency. The quality of raw data affects the results of the implemented method. In order to improve the quality of the medical data and consequently, of the results raw data is pre-processed so as to improve the efficiency and ease of mining process. In preprocessing, the non-numerical data and missing values are removed from the input dataset and obtained the numerical dataset for proceeding further. The preprocessed output is fed to the further process.

3.2 Feature Selection

Subsequent to the preprocessing the recommended technique is employ to choose the features from the input. Here the features are optimally elected by krill herd optimization algorithm. The comprehensive formula of krill herd optimization algorithm is illustrated in the subsequent segment,

3.2.1 Krill Herd Optimization Algorithm (KHA)

KHA is a novel meta-heuristic swarm insight streamlining technique for tackling enhancement issues, which depends on the simulation of the herding of the krill swarms because of particular organic and ecological forms. A combination of the least distances of the position of food and the highest density of the herd is the objective function of this optimization algorithm. Fig.2 shows the basic representation oh KHA. The description and mathematical expression of these operational processes are provided in the following section,

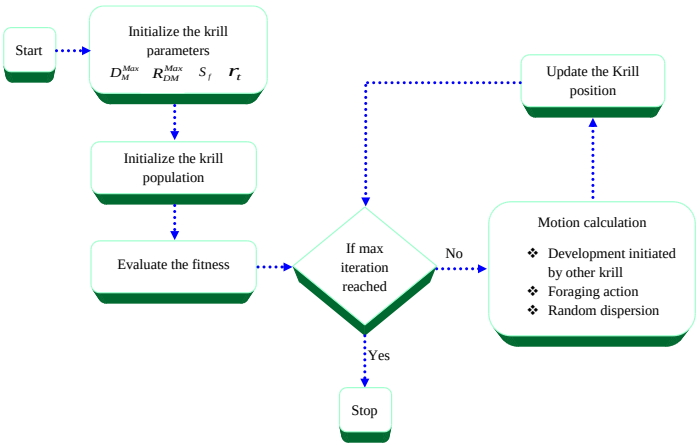

Figure.2 The flowchart for krill herd algorithm

The step by step procedure of KHA algorithm is illustrated in below section,

- Initialization

The main parameters of the KHA are the total evolution number, the population size, $D^{\max}$, S_f, r_t and $R_D^{\max}$. In our suggested technique krill herd represent the feature value.

- Fitness calculation

Evaluate the fitness utility depend on the equation (1) and moreover select the finest result.

$$Fitness = \max accuracy \quad (1)$$

To enhance the objective function value, KHA follows the search directions and repeats the implementation of three movements. By three main processes, the movement of each individual krill is determined.

(a) Development initiated by other krill individuals,

(b) Foraging action,

(c) Random dispersion.

- Development initiated by other krill individuals

In this procedure, while the speed of every individual is impacted by the development of the others, the krill individual attempt to keep up a high thickness. Three impacts namely (x) neighborhood impact (y) target impact and (z) repulsive impact are used to evaluate the direction of motion induced (ξ_m). This motion may be formulated for krill individual m as

$$D_m^{new} = \xi_m D^{\max} + \chi_b D_m^{old} \quad (2)$$

Where,

$$\xi_m = \xi_m^{current} + \xi_m^{target} \quad (3)$$

$$\xi_m^{current} = \sum_{n=1}^{N} F_{mn} P_{mn} \quad (4)$$

$$P_{mn} = \frac{P_m - P_n}{abs(P_n - P_m) + rand} \quad (5)$$

$$F_{mn} = \frac{F_m - F_n}{F^w - F^b} \quad (6)$$

$$\xi_m^{target} = K^{best} F_m^{best} P_m^{bset} \quad (7)$$

$$K^{best} = 2\left(random + \frac{M}{M_{\max}} \right) \quad (8)$$

$$\xi_m = \sum_{n=1}^{N} \left[\frac{F_m - F_n}{F^w - F^b} \times \frac{P_m - P_n}{abs(P_n - P_m) + rand} \right] + 2\left(random + \frac{M}{M_{\max}} \right) F_m^{best} P_m^{best} \quad (9)$$

$D^{\max}$ - Maximum induced motion,

χ_b - Inertia weight of the motion induced in the range [0, 1],

D_m^{old} - Previous induced motion of the mth krill individuals,

$F^w and F^b$ -The worst and the best position among all the krill individuals of the population

$P_m and P_n$ -Current position of the mth and the nth individuals

N - Number of krill individuals other than the particular krill

$M and M_{\max}$ - Number of current iteration and maximum number of iterations

F_m^{best} -The best fitness value of the mth and the nth individuals

P_m^{best} -The best related position of the mth and the nth individuals

Here a parameter named as sensing distance S_d is used for the distance between the individual krills and the neighbors and it is formulated by,

$$S_d = \frac{1}{5N}\sum_{n=1}^{N-1}\left|F_m - F_n\right| \qquad (10)$$

N - Total number of the krill individual

$F_m - F_n$ -Position of the mth and nth krill

- Foraging action

This action based upon two main factors. Initially is the present food area and the second is the data about the past food area. For the mth krill individual, the foraging velocity can be expressed by,

$$F_{Fm}^{new} = S_f \varsigma_m + \chi_x F_{Fm}^{old} \qquad (11)$$

$$\varsigma_m = \varsigma_m^{so\ln} + \xi_m^{bestso\ln} \qquad (12)$$

$$\varsigma_m^{so\ln} = C_m F_m P_m^{so\ln} \qquad (13)$$

$$P_m^{so\ln} = \frac{\sum_{n=1}^{N}\frac{P_m}{F_m}}{\sum_{n=1}^{N}\frac{1}{F_m}} \qquad (14)$$

$$C_m = 2\left(1 - \frac{M}{M_{\max}}\right) \qquad (15)$$

$$\varsigma_m^{bestso\ln} = F_m^{best} P_m^{best} \qquad (16)$$

$$F_{Fm}^{new} = S_f \times \left[2\left(1 - \frac{M}{M_{\max}}\right) \times F_m \times \frac{\sum_{n=1}^{N}\frac{P_m}{F_m}}{\sum_{n=1}^{N}\frac{1}{F_m}} + F_m^{best} \times P_m^{best}\right] + \chi_x \times F_{Fm}^{old} \qquad (17)$$

Where,

χ_x -Inertia weight of the foraging motion,

F_{Fm}^{new} *and* F_{Fm}^{old} -Foraging motions of the new and the old mth krill

- Random dispersion

To enhance the population diversity random diffusion process is mainly considered and it is expressed by,

$$R_{Dm}^{new} = \beta \times R_D^{\max} \qquad (18)$$

Where,

$R_D^{\max}$ - Maximum diffusion speed

β - Random directional vector lies between [-1, 1].

- Updating the position

In this procedure, the individual krill changes its present positions also, moves to better positions in light of induction movement, foraging movement and random dispersion movement. As indicated by the three above examined movements, the upgraded position of the mth krill individuals during the interval of t and Δt might be communicated by

$$P_m(t+\Delta t) = P_m(t) + \Delta t \frac{dP_m}{dt} \quad (19)$$

Where,

An n-dimensional decision space in Lagrangian model is used to express basic KHA technique as shown below,

$$\frac{dP_m}{dt} = D_m^{new} + F_{Fm}^{new} + R_{Dm}^{new} \quad (20)$$

D_m^{new} -the motion induced by other krill individuals

F_{Fm}^{new} -the foraging motion

R_{Dm}^{new} -the physical diffusion of the krill individuals

$$\Delta t = r_t \sum_{n=1}^{N} (UL_n - LL_n) \quad (21)$$

$UL_n \; and \; LL_n$ -Upper and lower limits

r_t -Random number uniformly distributed between 0 to 2.

Based on the above procedure, we select the optimal features and then the selected features are fed to the classification process.

3.3 Classification

Finally, the optimal features are furnished to Hybrid adaboost KNN classifier for the purpose of classification. Here, the selected feature from the earlier process is effectively employed for the segregation of the two classes such as normal or abnormal.

3.3.1 Hybrid adaboost KNN algorithm

Adaboost is a popular boosting technique which helps to combine multiple “weak classifiers” into a single “strong classifier”. Each weak classifier should be trained on a random subset of the total training set. Adaboost assigns a “weight” to each training sample, which determines the probability that each sample should appear in the training set. After training a classifier, Adaboost increases the weight on the misclassified examples so that these examples will make up a larger part of the next classifiers training set, and hopefully the next classifier trained will perform better on them. In this method, the probability of selecting a sample of X to be located in data set of classifier is determined based on the error probability of classifier. If sample X is not classified properly, the probability of being selected for the next classifier is increased and if the sample X is classified properly, it is less likely to be selected for the next classifier. All the learners are simple and weak and must have error less than 0.5. Otherwise, the process is stopped since its continuation makes the learning become difficult for the next classifier. Also, the initial probability of selecting sample is considered to be uniform. In fact, the weight of sample shows the importance of the sample. The final hypothesis is obtained through weighted voting of T number of weak hypotheses. The steps involved in this algorithm are shown below.

Step 1: Initialization of weight W, $W_i = 1/N,$ for $i = 1,....N$, $T = 0, err = 0$

Step 2: In the case, $t \le T \; and \; err^t < 0.5$, then normalize weight W^t, so that,

$$\sum_{t=1}^{N} W_i^t = 1 \tag{22}$$

Step 3: Call KNN, providing with the weight W^t, get hypothesis $h^t : X -> \{-1,1\}$

KNN classifier

The K nearest neighbor algorithm is a method which is used for classification of any objects or other elements based on the closest training data which are available in the feature space. The feature values from the inputs are used for training, the algorithm where these features are stored along with the labels so that while testing, we get accurate classified output. The classification in KNN is based on the labels of its K nearest neighbor by majority. The values nearest to the k value will be chosen for the classification results. Once the classification of nearest neighbor is done, we convert it into vector values with fixed length by utilizing the Euclidean distance function in KNN which is given in below expression,

$$E_D(x,y) = \left(\sum_{k=1}^{N} (x_k - y_k)^2 \right)^{\frac{1}{2}} \tag{23}$$

Where x, y are feature values;

The basis of the KNN classifier is the small neighborhood in the similar features. These processes will give better accuracy in classifying the results.

Step 4: Compute $err^t = \sum_{i=1}^{N} W_i^t e_i^t$,

Where $e_i^t = 1,$ if $h^t(x_i) \neq v_i,$ and 0 otherwise

Step 5: Set $\alpha^t = 0.5 \log[(1 - err^t)/err^t]$

Step 6: Update the weights to be as follows,

$$W_i^{t+1} = W_i^t \exp(2\alpha^t e^t) \tag{24}$$

Step 7: Put $T = t + 1$ and the process repeats until $err = 0$.

After each classifier is trained, the classifier's weight is calculated based on its accuracy. More accurate classifiers are given more weight. Finally we classify the medical data with high accuracy value.

4. Results and Discussion

Our proposed method is implemented using MATLAB 2014 and the experiment is done using i5 processor with 3GB RAM.

4.1 Dataset description

Our proposed method is tested with the help of four datasets such as chronic, mammographic mass, Pima Indians diabetes, Cleveland, Hungarian and Switzerland dataset. The datasets are made available using the UCI machine learning repository.

(i) Chronic dataset

Chronic kidney dataset can be used to predict the kidney disease. It can be collected from the hospitals. The characteristics of the dataset are real time attribute characteristics. This dataset consist of 400 numbers of instances and 25 numbers of attributes. The main task of this data set is used for the classification purpose.

(ii) Mammographic Mass Data Set

The mammographic mass dataset used here has been accumulated at the Institute of Radiology of the University Erlangen-Nuremberg in the region of 2003 as well as 2006. The data set is available using http access of the University of California at Irvine (UCI) machine learning repository. Digital Database for Screening Mammography (DDSM) has been used to evaluate the

proposed structure. The database comprises of around 2,620 cases. For each case, dual images of every breast, inter related patient information, like age, period of the tumor, subtlety rating for varieties from the standard, American College of Radiology (ACR) breast thickness rating are considered. The Mammograms are digitized by various scanners depending upon the wellspring of the data.

(iii) Pima Indians Diabetes Data Set

The wellspring of Pima Indian diabetes enlightening accumulation is the UCI machine learning file. The data source uses 768 cases with two class problems to examine whether the patient would examine positive or negative for diabetes. Each one of the patients in this database is Pima Indian women at least 21 years old and surviving near Phoenix Arizona, USA. This dataset is most normally used for testing of diabetes detecting algorithms. The dataset contains 9 properties.

(iv) Cleveland data

This dataset comprises of 76 qualities; however all passed on tests suggest utilizing a subset of 14 of them. Extraordinarily, ML investigators use only the Cleveland dataset till today. The "goal" field insinuates the closeness of coronary ailment in the patient. It is entire number regarded from 0 (no proximity) to 4. Tests utilizing Cleveland dataset have concentrated on fundamentally attempting to perceive closeness (values 1, 2, 3, 4) from nonappearance (regard 0). The names and government disability amounts of the patients were eliminated from the database, supplanted with sham qualities. Six of the cases have been discarded in light of the way that they had missing qualities. Classes scatterings are 54% coronary ailment truant, 46% coronary disease exist.

(v) Hungarian data

Owing to a boundless rate of missing qualities three of the qualities have been expelled however the arrangement of the data is precisely the near as that of the Cleveland data. Thirty-four instances of the database were expelled in view of missing qualities and 261 cases were accessible. Class scatterings are 62.5% coronary disease not present and 37.5% coronary ailment exist.

(vi) Switzerland data

More amount of missing qualities is in Switzerland data. It encases 123 data cases and 14 qualities. Class appointments are 6.5% coronary sickness not present and 93.5% coronary disease exist.

4.2 Evaluation metrics

In order to assess the efficiency of our proposed method various evaluation metrics are utilized. The metrics consists of group of esteems that contains normal primary evaluating methods. The evaluation metrics used here contains True Positive, True Negative, False Positive and False Negative, Sensitivity, Specificity and Accuracy.

$$Sensitivity = \frac{T(P)}{T(P) + F(N)} \tag{25}$$

$$Specificity = \frac{T(N)}{F(P) + T(N)} \tag{26}$$

$$Accuracy = \frac{T(P) + T(N)}{T(P) + F(N) + F(P) + T(N)} \tag{27}$$

4.3 Performance Analysis

The results of our method assist in analyzing the effectiveness of the prediction method. The results are tabulated below. The results of four datasets are provided in below table,

TABLE.1: PERFORMANCE ANALYSIS OF THE PROPOSED METHOD

Dataset	Accuracy	Sensitivity	Specificity
Chronic data	87.24	0.843	0.976
Mammographic Mass data	95.86	0.985	0.965
Pima Indians Diabetes data	98.587	0.989	0.987
Cleveland data	97.56	0.994	0.916

Hungarian data	98.85	0.985	0.986
Switzerland data	98.55	0.989	0.986

From table.1, it is clear that the accuracy value obtained using the proposed method for chronic data is 87.24%, similarly the sensitivity and specificity value obtained is 0.843 and 0.976 respectively. For mammographic mass data the accuracy value using the proposed method is 98.86%, sensitivity value obtained is 0.995 and specificity value obtained is 0.985. The accuracy value obtained using the Pima Indians Diabetes dataset is 98.587%, sensitivity value obtained is 0.989 and specificity value obtained is 0.987. 97.56% is the accuracy value obtained, 0.994 is the sensitivity value obtained and 0.916 is the specificity value obtained using Cleveland dataset. For Hungarian dataset 98.85% is the accuracy value obtained, 0.965 is the sensitivity value obtained and 0.966 is the specificity value obtained using the proposed method. The accuracy, sensitivity and specificity value obtained using the proposed method for Switzerland dataset is 98.55%, 0.989 and 0.985 respectively.

4.4 Comparative Analysis

The literature review works are contrasted in this area with the proposed method to demonstrate that our proposed method is superior to anything the condition of-fine arts. We can build up that our proposed method achieves great exactness for the medical data classification. Hybrid adaboost algorithm together with KNN algorithm is utilized for medical data classification in our proposed strategy. And furthermore we can set up this forecast precision result by contrasting different classifiers. Here the proposed classifier is contrasted with the fuzzy min max neural system alongside with modified group search algorithm classifier [23], existing KNN and Adaboost classifiers. The Comparison results are introduced in the accompanying table.2.

TABLE.2: SENSITIVITY COMPARISON ANALYSIS OF THE PROPOSED AND EXISTING METHOD

Datasets	Proposed Sensitivity values	Existing Sensitivity values [23]	KNN	Adaboost
Chronic data	0.843	0.651	0.653	0.743
Mammographic Mass data	0.985	0.974	0.50	0.857
Pima Indians Diabetes data	0.989	0.97	0.472	0.695
Cleveland data	0.994	0.851	0.653	0.786
Hungarian data	0.995	0.952	0.653	0.743
Switzerland data	0.989	0.946	0.472	0.727

From the above table it is clear that for chronic data the sensitivity value obtained using the existing method [23], KNN and adaboost classifier is 0.651, 0.653 and 0.743 while the sensitivity value obtained using the proposed method is 0.843. Similarly for mammographic Mass dataset using the proposed method the sensitivity value obtained is 0.985 while 0.974, 0.5 and 0.857 is the sensitivity value obtained using the existing methods [23], KNN and Adaboost classifier. 0.989 is the sensitivity value obtained using the proposed method while 0.970, 0.472 and 0.695 is the sensitivity value obtained using the existing method [23], KNN and adaboost classifier for Pima Indians Diabetes dataset. For Cleveland dataset the sensitivity value obtained using the proposed method is 0.994 while the sensitivity value obtained using the existing method [23] is 0.851, 0.653 is the sensitivity value obtained using the existing method KNN and 0.786 is the sensitivity value obtained using the existing adaboost classifier. For Hungarian dataset the sensitivity value obtained using the proposed method is 0.995 while 0.952, 0.653 and 0.743 is the sensitivity value obtained using the existing [23], KNN and adaboost classifiers. For Switzerland dataset the value of sensitivity using the proposed classifier is 0.989 while the value of sensitivity using the existing [23], KNN and adaboost classifier is 0.946, 0.472 and 0.727 respectively. The values are plotted in the form of graph below,

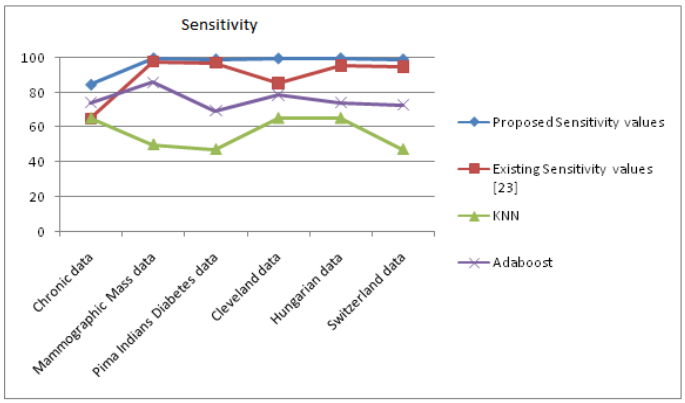

Figure.3 comparison of sensitivity values

Hence our proposed method has better sensitivity value for all the datasets than the existing method.

TABLE.3: SPECIFICITY COMPARISON OF PROPOSED AND EXISTING METHOD

Datasets	Proposed Specificity values	Existing Specificity values [23]	KNN	Adaboost
Chronic data	0.976	0.698	0.743	0.869
Mammographic Mass data	0.965	0.915	0.437	0.656
Pima Indians Diabetes data	0.987	0.933	0.258	0.517
Cleveland data	0.916	0.851	0.869	0.766
Hungarian data	0.986	0.698	0.472	0.762
Switzerland data	0.9865	0.707	0.685	0.9678

From the above table it is clear that for chronic dataset using the proposed method the value of specificity obtained is 0.976 while the specificity value obtained using the existing method [23], KNN and adaboost is 0.698, 0.743 and 0.869. The value of specificity obtained using the proposed method 0.965 and the value of specificity obtained using the existing method [23], KNN and adaboost classifier is 0.915, 0.437 and 0.656 respectively for Mammographic Mass dataset. For Pima Indians Diabetes dataset the specificity value obtained using the proposed method is 0.987 while the value of specificity obtained using the existing classifiers [23], KNN and adaboost classifiers is 0.933, 0.258 and 0.517. The specificity value obtained using the proposed classifier is 0.916 while the specificity value obtained using the existing classifiers [23], KNN and adaboost is 0.851, 0.869 and 0.766 respectively for Cleveland dataset. For Hungarian dataset the specificity value obtained using the proposed method is 0.986 while the specificity value obtained using the existing classifier [23], KNN and adaboost classifier is 0.698, 0.472 and 0.762 respectively. The specificity value obtained using the existing classifier [23], KNN and adaboost classifier is 0.707, 0.685 and 0.9678 while the value of specificity obtained using the proposed method is 0.9865 respectively for Switzerland dataset. The values are plotted in the form of graph below,

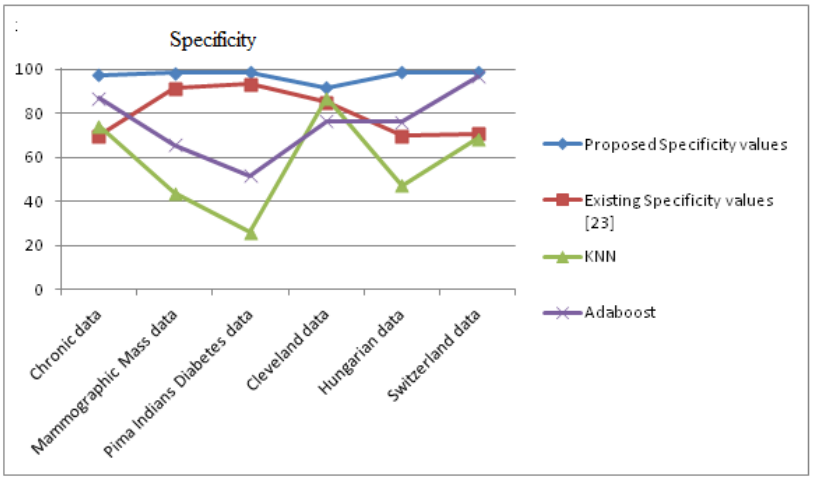

Figure.4 comparison of specificity values

Hence our proposed method has better specificity value for all the datasets than the existing method.

TABLE.4: ACCURACY COMPARISON OF PROPOSED AND EXISTING METHOD

Datasets	Proposed Accuracy values	Existing Accuracy values [23]	KNN	Adaboost
Chronic data	87.24	65.14	76.66	69.53
Mammographic Mass data	95.86	94.21	92.66	89.66
Pima Indians Diabetes data	98.58	95.3	92.58	92.41
Cleveland data	97.56	85.14	89.21	93.52
Hungarian data	98.85	83.33	88.23	94.25
Switzerland data	98.55	84.01	91.66	96.66

For chronic dataset the accuracy value obtained using the proposed method is 87.24% while the accuracy value obtained using the existing classifier [23], KNN and adaboost classifier is 65.14%, 76.66% and 69.53%. For Mammographic Mass dataset the value of accuracy obtained using the proposed classifier is 95.86% while the value of accuracy obtained using the existing classifier [23], KNN and adaboost classifier is 94.21%, 92.66% and 89.66%. The accuracy value obtained using the proposed classifier for Pima Indians Diabetes dataset 98.58% while the accuracy value obtained using the existing classifier [23], KNN and adaboost classifier is 95.30%, 92.58% and 92.41%. For Cleveland dataset the value of accuracy obtained using the proposed classifier is 97.56% while the accuracy value obtained using the existing classifier [23], KNN and adaboost classifier is 85.14%, 89.21% and 93.52%. The accuracy value obtained using the proposed classifier is 98.85% for Hungarian dataset while the accuracy value obtained using the existing classifier [23], KNN and adaboost classifier is 83.33%, 88.23% and 94.25%. For Switzerland dataset the accuracy value obtained using the proposed classifier is 98.55% while the accuracy value obtained using the existing classifier [23], KNN and adaboost is 84.01%, 91.66% and 96.66% respectively. The values are shown in the form of figure below,

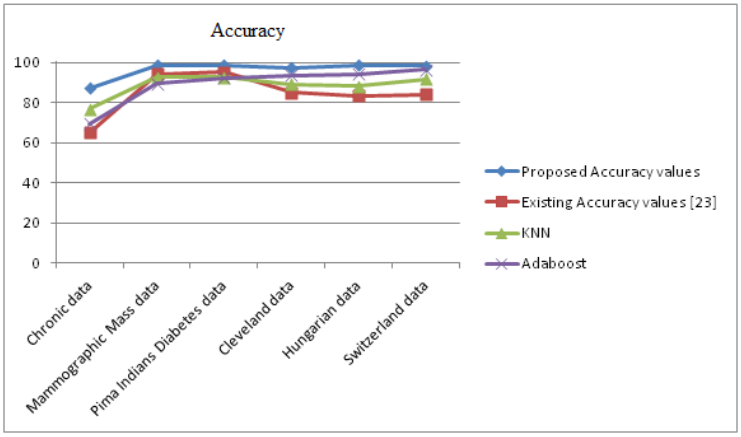

Figure.5 comparison of accuracy values

From the figure it is clear that our proposed method has high accuracy value than the existing method for all the datasets. Hence our proposed method has better outcomes for all the datasets.

CONCLUSION

In this paper we have recommended a technique for medical data classification with the help of optimal feature selection. Initially preprocessing is applied to the input dataset. Then the optimal features are selected from the input dataset by means of krill herd algorithm. Next the selected optimal features are classified using hybrid adaboost KNN classification algorithm. The implementation of the proposed method was done in MATLAB. For experimentation, the dataset given in the UCI machine learning repository such as, chronic, Mammographic Mass data, Pima Indians Diabetes data, Cleveland, Hungarian and Switzerland etc., will be subjected to analyze the performance of the proposed technique in class imbalance problem utilizing accuracy, sensitivity and specificity. The results of our proposed method have shown that, our hybrid classifier achieves better result when compared to other method. Our method achieves the maximum accuracy value for Hungarian dataset and Switzerland dataset. Both dataset achieves 97.85% and 98.25% of accuracy value for medical data classification.

Reference

[1] Lin, Wenmin, Wanchun Dou, Zuojian Zhou and Chang Liu, "A cloud-based framework for Home-diagnosis service over big medical data", Elsevier on Journal of Systems and Software, Vol.102, pp.192-206, 2015.

[2] Chen, Ling, Xue Li, Yi Yang, Hanna Kurniawati, Quan Z. Sheng, Hsiao-Yun Hu and Nicole Huang, "Personal health indexing based on medical examinations: a data mining approach", Elsevier on Decision Support Systems, Vol.81, pp.54-65, 2016.

[3] Garcia-Hernandez, Jose Juan, Wilfrido Gomez-Flores and Javier Rubio-Loyola, "Analysis of the impact of digital watermarking on computer-aided diagnosis in medical imaging", Elsevier on Computers in biology and medicine, Vol.68, pp.37-48, 2016.

[4] [4] Ye and Jun, "Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses", Elsevier on Artificial intelligence in medicine, Vol.63, No.3, pp.171-179, 2015.

[5] Thong and Nguyen Tho, "Intuitionistic fuzzy recommender systems: an effective tool for medical diagnosis", Elsevier on Knowledge-Based Systems, Vol.74, pp.133-150, 2015.

[6] Wójtowicz, Andrzej, Patryk Żywica, Anna Stachowiak and Krzysztof Dyczkowski, "Solving the problem of incomplete data in medical diagnosis via interval modeling", Elsevier on Applied Soft Computing, pp.1-14, 2016.

[7] Ye, Jun and Jing Fu, "Multi-period medical diagnosis method using a single valued neutrosophic similarity measure based on tangent function", Elsevier on Computer methods and programs in biomedicine, Vol.123, pp.142-149, 2016.

[8] Thong and Nguyen Tho, "HIFCF: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis", Elsevier on Expert Systems with Applications, Vol.42, No.7, pp.3682-3701, 2015.

[9] Gorzałczany, Marian B and Filip Rudziński, "Interpretable and accurate medical data classification-a multi-objective genetic-fuzzy optimization approach", Elsevier on Expert Systems with Applications, pp.1-17, 2016.

[10] AlMuhaideb, Sarab and Mohamed El Bachir Menai, "An Individualized Preprocessing for Medical Data Classification", Elsevier on Procedia Computer Science, Vol.82, pp.35-42, 2016.

[11] Kurzyński, Marek, Marcin Majak and Andrzej Żołnierek, "Multiclassifier systems applied to the computer-aided sequential medical diagnosis", Elsevier on Biocybernetics and Biomedical Engineering, Vol.36, No.4, pp.619-625, 2016.

[12] Shen, Liming, Huiling Chen, Zhe Yu, Wenchang Kang, Bingyu Zhang, Huaizhong Li, Bo Yang and Dayou Liu, "Evolving support vector machines using fruit fly optimization for medical data classification", Elsevier on Knowledge-Based Systems, Vol. 96, pp.61-75, 2016.

[13] De Bruijne and Marleen, "Machine learning approaches in medical image analysis: From detection to diagnosis", Elsevier on Medical Image Analysis, Vol.33, pp.94-97, 2016.

[14] Castellano, Nuria N., Jose A. Gazquez, Rosa M. García Salvador, Antonio Gracia-Escudero, Manuel Fernandez-Ros and Francisco Manzano-Agugliaro, "Design of a real-time emergency telemedicine system for remote medical diagnosis", Elsevier on Biosystems Engineering, Vol.138, pp.23-32, 2015.

[15] Garcia-Hernandez, Jose Juan, Wilfrido Gomez-Flores and Javier Rubio-Loyola, "Analysis of the impact of digital watermarking on computer-aided diagnosis in medical imaging", Elsevier on Computers in biology and medicine, Vol.68, pp.37-48, 2016.

[16] Narang, Sandeep K., Cynthia Estrada, Sarah Greenberg and Daniel Lindberg, "Acceptance of shaken baby syndrome and abusive head trauma as medical diagnoses", The Journal of Pediatrics, Vol.177, pp.273-278, 2016.

[17] Mohapatra, P., Sreejit Chakravarty and P. K. Dash, "Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system", Swarm and Evolutionary Computation, Vol.28, pp.144-160, 2016.

[18] Kumar, S. Senthil and H. Hannah Inbarani, "Optimistic multi-granulation rough set based classification for medical diagnosis", Procedia Computer Science, Vol.47, pp.374-382, 2015.

[19] Li, Zhaowen, Guoqiu Wen and Ningxin Xie, "An approach to fuzzy soft sets in decision making based on grey relational analysis and Dempster–Shafer theory of evidence: An application in medical diagnosis", Artificial intelligence in medicine, Vol.64, No.3, pp.161-171, 2015.

[20] Inbarani and H. Hannah, "A novel neighborhood rough set based classification approach for medical diagnosis", Procedia Computer Science, Vol.47, pp.351-359, 2015.

[21] Hernández-Chan, Gandhi S., Edgar Eduardo Ceh-Varela, Jose L. Sanchez-Cervantes, Marisol Villanueva-Escalante, Alejandro Rodríguez-González and Yuliana Pérez-Gallardo, "Collective intelligence in medical diagnosis systems: A case study", Elsevier on Computers in biology and medicine, Vol.74, pp.45-53, 2016.

[22] Parvathi, I and Siddharth Rautaray, "Survey on data mining techniques for the diagnosis of diseases in medical domain", International Journal of Computer Science and Information Technologies, Vol.5, No.1, pp.838-846, 2014.

[23] D. Mahammad Rafi and Chettiar Ramachandra Bharathi, "Optimal Fuzzy Min-Max Neural Network (FMMNN) for Medical Data Classification Using Modified Group Search Optimizer Algorithm", International Journal of Intelligent Engineering and Systems, Vol.9, No.3, pp. 1-10, 2016.