A Hybrid Classification Model using Genetic Algorithm and Support Vector Machine combined with Consistency-based subset evaluation for Feature Selection

[1]M. S. Padmavathi, [2]C.P Sumathi

[1,2] Research Department of Computer Science, S.D.N.B Vaishnav College for Women, Chennai, India. [1] padmanivas_2002@yahoo.co.in, [2] drcpsumathi@gmail.com

Abstract : Medical data mining is an area of application where classification accuracy is important. Specifically in the area of disease diagnosis, researchers have concentrated on hybrid classifiers to efficiently improve the accuracy of their model. The proposed hybrid model involves Consistency-based subset evaluation method in conjunction with re-ranking algorithm to find the best feature subset. The dataset with reduced features is subjected to three new hybrid classifiers: (1) Fuzzy-rough instance selection method with Support Vector Machine (SVM) as classifier. (2) Genetic Algorithm (GA) to remove the wrongly classified instances and Fuzzy-rough nearest neighbor for classification. (3) GA for selecting the instances and classification using SVM. The experimental results prove that all three suggested hybrid models provide better accuracy compared to the Fuzzy-rough instance selection and Fuzzy-rough nearest neighbor classifier as described in the literature. Among the proposed hybrid models, GA with SVM combination yields better result with a classification accuracy of 99.3% – 99.8%. Also the percentage of instance removal is considerably reduced using GA compared with Fuzzy-rough instance selection method.

Keywords — Classification, Consistency-based subset evaluation, Fuzzy-rough instance selection, Fuzzy-rough nearest neighbor, Genetic Algorithm, Support Vector Machine.

I. INTRODUCTION

Health care industry is accumulated with enormous amount of data which is too large and complex for processing and analysis. Hence, it is necessary to develop an effective computer-aided disease diagnosis and classification system for medical informatics [1]. In this research paper, experiments have been carried out on datasets of breast cancer, diabetes and heart disease to evaluate the proposed hybrid models. By performing suitable medical tests, earlier diagnosis of the disease followed by appropriate treatment is considered important for reduction in death rate. Therefore it is necessary to identify the presence of the disease at the initial stage and necessary precautions to be taken [2].

Missing data can have a significant effect on the conclusions drawn from the data. Even though there are several ways to handle missing data, deciding the best analysis strategy to yield the least biased estimates is important. Different ways of missing data treatment includes: (1) deletion methods (2) single imputation (3) multiple imputation, etc [3]. To improve the quality of data in the proposed work, data cleaning is done using deletion method and single imputation method. A comparative study is done between these two methods of data cleaning and its effect on classification accuracy. Feature selection is used to select a subset of variables which describes the input data while reducing effects from noise variables and still provide better prediction results [4]. To remove an irrelevant feature, it is necessary to have a feature selection criterion which can measure the relevance of each feature with the output class [5]. In this article, a selection criterion called Consistency measure is used which does not concentrate on maximizing the class separability, instead tries to retain the discriminating power of the data defined by original features [6]. Re-ranking algorithm is utilized in Consistency-based feature selection to minimize the number of wrapper evaluations and for effective search of feature subsets [1].

The current focus of the researchers is to combine several classifier systems to perform information fusion of classification decisions at various levels overcoming the limitations of single classifiers [7]. Fuzzy-rough instance selection method is used for selecting the instances based on the Fuzzy-rough positive region [8]. Genetic algorithms are adaptive, heuristic and robust which indicates that they can be applied to problems of any domain with slight modification of the representation, fitness evaluation and the parameters of the genetic operators [9]. Genetic algorithms are computationally powerful to remove the noisy instances as outliers. Fuzzy-rough nearest neighbor

 DOI : 10.35291/2454-9150.2019.0202

classifier enhances the traditional K-nearest neighbor classifier by utilizing Fuzzy-rough uncertainty. Owing the advantages of the conventional K-nearest neighbor method Fuzzy-rough nearest neighbor can be used as a viable tool for classification [10]. For classification tasks, SVM constructs hyper planes in a multidimensional space that separates cases of different class labels. Several recent studies have reported that the classification accuracy of SVM yields better results than the other data classification algorithms including statistical classifiers, decision tree algorithms, neural network classifiers and instance based learning methods [11].

The proposed framework involves (1) cleaning data by deleting null values and replacing null values by median (2) selecting the optimal subset of features through Consistency-based subset evaluation and (3) three different combinations of hybrid classifiers (Fuzzy-rough instance selection + SVM, GA + Fuzzy-rough nearest neighbor and GA + SVM) for classifying the medical datasets. The rest of the paper is organized as follows: Section 2 summarizes the existing machine learning techniques for disease diagnosis. Section 3 explains the preliminaries while section 4 details the datasets used. The proposed hybrid approaches and evaluation metrics are described in section 5 and 6. Experimental analysis and comparison is made in section 7 followed by concluding remarks and future work.

II. RELATED WORK

A hybrid intelligent classification model was presented, which utilized re-ranking search algorithm in conjunction with Consistency-based feature selection for obtaining subset of features and Fuzzy-rough instance selection to select appropriate instances. Finally, the Fuzzy-rough nearest neighbor algorithm is applied on Wisconsin Breast Cancer (WBC) dataset for building classification model. The experimental results proved that the proposed classification model obtained an accuracy of 99.71% using the 10-fold cross validation scheme [1]. From the study, it is observed that the method is experimented only for one dataset (WBC). The present study aims to analyze the performance of the model suggested in the existing work [1] to other medical datasets like Pima diabetes, WDBC and Heart (Stalog). In addition, experiments have been conducted to analyze the performance of the three proposed hybrid models on the above medical datasets.

The concept of applying Fuzzy classifiers in the proposed method is obtained using the following research papers: A new hybrid classifier was proposed using fuzzy-rough instance selection and SVM for credit scoring. Fuzzy-rough instance selection is applied instead of clustering algorithms to eliminate isolated and inconsistent instances and SVM for classification [12]. A medical classification model was introduced combining Wavelet Transform (WT) and interval type-2 Fuzzy logic system to deal with high dimensional dataset. WT was employed to extract significant features and Interval type-2 Fuzzy logic system consists of Fuzzy c-means clustering and GA based parameter tuning for classification. The proposed classification model achieved a classification accuracy of 97.88% for breast cancer diagnosis [13]. A genetic search fuzzy rough (GSFR) feature selection algorithm was proposed by applying evolutionary sequential genetic search technique and fuzzy rough set to select features. The dataset with minimal features are applied to the different classifiers of Fuzzy-rough nearest neighbor (FRNN) classifier, which provided better classification accuracy and less computation time [14].

To analyze the effect of hybrid algorithm using GA for outlier removal the following studies are considered: A hybrid algorithm was proposed to detect outliers. The experimental reports proved that, GA was better for detecting outliers and providing optimized data. A comparison with the other optimization techniques such as Ant Colony Optimization and Particle Swam Optimization is also done to support the results obtained [15]. A novel approach of was introduced, to compare GA with Inter Quartile Range and K-Means clustering for removing the misclassified instances. It has been tested with University of California Irvine (UCI) datasets and proved that GA has a less data reduction percentage compared with statistical and clustering methods [16]. A new approach of introducing GA to detect outliers is implemented which proved that GA resulted in better calculation of the number of outliers for a particular period of time [9].

The use of SVM classifier is popular among medical datasets and it is supported by the following researches: A classification model was proposed using GA to obtain the optimal feature set and optimizes the parameter values of SVM. The proposed algorithms achieved better classification accuracy using SVM for all the tested datasets [17]. Clustering based classification was very common in which K-Means algorithm is used to detect outliers and SVM to achieve a classification accuracy of 97.38% [18].

III. METHODOLOGIES

A. Consistency-based subset evaluation

A probabilistic approach to feature selection called Consistency-based subset evaluation is introduced to evaluate the subset of attributes by the level of consistency in the class values [19]. A pattern is a part of an instance without class label describing the values of feature subset. For a feature subset S with $n_{f1}, n_{f2}, n_{f3} \ldots\ldots n_{f|s|}$ number of values for features $f1, f2, f3 \ldots\ldots f|s|$ respectively, there are at most $n_{f1} * n_{f2} * n_{f3} \ldots * n_{f|s|}$ patterns.

 DOI : 10.35291/2454-9150.2019.0202

Consistency measure is calculated in terms of inconsistency rate which is calculated as follows [6]:

(1) The pattern is inconsistent, for the occurrence of at least two instances with the same values except for their class labels. For example, the features $f1, f2$ *and* class variable takes the value of (0 1, *1*) and (0 1, *0*) where except the class attribute, the two features take the same values.

(2) The number of patterns inconsistent for a feature subset is calculated as: for a feature subset S a pattern p appears in n_p instances out of which c_1 instances has category label 1, c_2 has label 2, and c_3 with label 3 where $c_1 + c_2 + c_3 = n_p$. If c_2 is the highest among the three, then inconsistency count is $n - c_2$.

(3) The rate of inconsistency for a feature subset $S(I_R(S))$ is defined as the sum of all the inconsistency counts for all the patterns of the feature subset that appears in the data divided by P (total number of instances).

For a candidate feature subset S, its inconsistency rate is calculated as $I_R(S)$. The subset S is said to be consistent, if $I_R(S) \leq \delta$ where δ is a user given inconsistency rate threshold. Consistency measure can work with discrete valued features hence, continuous feature should be first discretized and then to be used [20].

Re-ranking algorithm: Re-ranking is a meta-search algorithm [21] that creates a univariate ranking for all the attributes in decreasing order based on attributes evaluation metric such as information gain. The ranking is split into blocks of size B = 20 and an attribute selection search is run for the first block. Given the selected attributes and the rest of the attributes are modified based on conditional information gain of each attribute. Again attribute selection search is run again on the first block and so on. Search stops when a new block does not alter the selected subset. The process of attribute selection is done by using greedy stepwise search algorithm. Three different approximation methods such as the conditional mutual information maximization, mutual information based feature selection and the max-relevance & min-redundancy are used to approximate the re-ranking of remaining attributes. Among the three, conditional mutual information maximization is used here.

B. Fuzzy-rough instance selection

Fuzzy-rough instance selection is done based on the conflicts with other instances present in the fuzzy-rough positive region [22]. The instances that negatively affect the fuzzy positive region are removed to reduce the training time of classifiers [23]. Fuzzy-rough set theory is based on the hybridization of Fuzzy set and rough set theory.

For a set of training samples $S \subset X$, a decision system $(X, A \cup \{d\})$ can be modeled such that a is a quantitative attribute in $A \cup \{d\}$ and $l(a)$ as range. The approximate equality between two objects x and y in S with respect to a is defined as:

$$R_a^{\alpha}(x,y) - \max\left(0, 1 - \alpha\frac{|a(x) - a(y)|}{l(a)}\right) \quad (1)$$

Where, α determines the granularity of R_a^{α}. For any subset B of A, the lower approximation $R_B^{\alpha} \downarrow^S A$ of a Fuzzy set A in S and the Fuzzy B-positive region $POS_B^{\alpha,S}$ for y in S can be defined as:

$$(R_B^{\alpha} \downarrow^S A)(y) - \inf_{x \in S} I(R_B^{\alpha}(x,y), A(x)) \quad (2)$$

$$POS_B^{\alpha,S}(y) - (R_B^{\alpha} \downarrow^S R_d^{\alpha} y)(y) \quad (3)$$

Fuzzy-rough positive region is used for both feature and instance selection. Fuzzy-rough instance selection uses Fuzzy-rough lower approximation techniques and select instances with a high membership degree in the Fuzzy rough positive region [1]. In Fuzzy-rough instance selection algorithm, the degree of membership of each object x to the positive region is evaluated. If the membership is less than the given threshold ($\tau = 1$) then the object can be removed [23]. Thus resulting instances contains no inconsistencies.

C. Genetic Algorithm

Genetic Algorithm is a popular stochastic search method used based on Darwin's theory of natural selection and survival of the fittest [24]. The group of individuals involved in the algorithm is called population and each individual is characterized as chromosomes. Chromosomes constitute sequence of genes comprised of bits, characters or sequences indicating the presence of the element in the set. Reproduction is performed through the following operators:

- Selection – Selection of the individuals with the best fitness values that can reproduce.
- Crossover – A single point cross over is done to create offspring by combining the partial characteristics (genes) from each parent.
- Mutation – Choosing a 7% of mutation rate is done to make random changes in the genes of individuals.

The fitness function plays a key role in evaluating the chromosome to find the best fit for the environment that continues to exist in the next generation. Fitness function involves hybrid classifier with boosting technique for classification and can be determined using the formula:

$$Fitness = \frac{\text{Total No. of Correctly classified instances}}{\text{Total No. of Training samples}} \quad (4)$$

In the application of GA, a gene represents an instance. The number of genes in a chromosome is equal to the number of initial instances. Each gene gets a value of either 0 or 1 where 0 means the instance is omitted and 1 means the instance is retained. The chromosome that is retained through a series of evolving generations represents the

 DOI : 10.35291/2454-9150.2019.0202

optimal set of instances to be selected. Selection of individuals is done through roulette wheel selection. Based on the fitness function (high value), parents are selected. The above process is iterated many times for a number of generations until optimal solution is reached. Threshold for fitness function is to minimize the number of incorrectly classified instances at which the system converges. After the last iteration, the optimum fitness value is obtained and the number of incorrectly classified instances is removed as outliers thereby reducing the outlier percentage.

D. Fuzzy-rough nearest neighbor classifiers

Fuzzy K-Nearest Neighbor (FNN) as an extension of the K Nearest Neighbor algorithm (KNN). It classifies an object to different classes by considering the relative importance (closeness) of each neighbor with respect to the test instance [25]. However, FNN has problems in handling data with insufficient knowledge. To solve this problem, Fuzzy-rough ownership function was introduced to handle both Fuzzy uncertainty (caused by overlapping classes) and rough uncertainty (caused by insufficient information of attributes). The Fuzzy-rough ownership function τ_c of class C can be defined as, for an object y,

$$\tau_c\,(y) = \frac{\sum_{x\epsilon X} R(x,y)C(x)}{|X|} \quad (5)$$

Where, $\tau_c(y)$ is interpreted as the confidence with which y can be classified to class C and $\mathrm{R}(x,y)$ is calculated as:

$$R(x,y) = \exp\left(-\sum_{a\epsilon C} K_a\,(a(y) - a(x))^{2/(m-1)}\right) \quad (6)$$

Here, m is used to control the weighting of similarity and K_a denotes the bandwidth of membership.

$$k_a = \frac{|U|}{2\sum_{x\epsilon U}\|a(y) - a(x)\|^{2/(m-1)}} \quad (7)$$

At the initial stage, the parameter K_a is calculated for each attribute and all the memberships of decision classes for test object y are set to 0. Equation (5) is used to calculate the weighted distance of y from all objects in the universe and is also used to update the class memberships of y. Finally, the algorithm outputs the class with the highest membership after considering all the training objects.

E. Support Vector Machine

SVM is a supervised learning algorithm proposed by Boser, Guyon, and Vapnik [26]. It can perform classification and regression tasks for the datasets with multiple continuous and categorical variables. The subset of data instances called "Support Vectors" is used to define a hyperplane such that it separates the classes and maximize the margin between two classes. SVM is classified as Linear and Non-Linear: (1) Linear SVM is used for linearly separable datasets. The discriminant function of the hyper plane can be written as:

$$g(x) = \omega^{T} x + b \quad (8)$$

$$L_p(\omega, b, \alpha) = \frac{1}{2}\,||\omega||^2 - \sum_{i=1}^{n} \alpha_i\,\{y_i\,(W^T x_i + b) - 1\} \quad (9)$$

Where, α_i denotes Lagrange multipliers. The optimization equation to minimize L_p for determining optimal ω *and* b is as follows:

$$Maximize\left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\,\alpha_j y_i\,y_j\,x_i^T x_j\right] \quad (10)$$

(2) Non-Linear SVM uses kernel function to solve non-linear classification problem. The kernel function maps data points on a higher dimensional space constructing a hyper plane to separate the classes.

$$g(x) = W^{T}\emptyset(X) + b \quad (11)$$

Where, $\emptyset(X)$ represents the mapping of input vectors to the kernel space X. The optimization equation can be written as:

$$Maximize\left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\,\alpha_j y_i\,y_j\,K(x_i\,x_j)\right] \quad (12)$$

Where, $K(x_i\,x_j)$ representing the kernel function is equal to $\{\emptyset\ (x_i),\ \emptyset(x_j)\ \}$. The kernel function [27] can be polynomial, Radial Basis Function (RBF), or any symmetric function that satisfies the Mercel conditions [28]. Sequential Minimal Optimization (SMO) algorithm with poly kernel is used for training the support vector classifier.

IV. DATA SOURCE

Four datasets with and without null values from the UCI machine learning repository is selected for this study and is described below:

Wisconsin Breast Cancer (WBC) Dataset: The WBC dataset was collected from the patients of University of Wisconsin-Madison for the cytological diagnosis of fine needle aspiration. The dataset contains 699 records out of which 16 instances contain missing values. There is a class variable to identify the tumor as benign / malignant (takes the value of 2 or 4), where benign occupies 65% of the whole dataset and the rest 35 % is malignant. It contains id number and 8 ordinal attributes (takes value from 1-10) to describe the cell nuclei. They are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland chromatin, Normal Nucleoli and Mitoses [1]. The attribute id number for the patients is not considered for the experiment.

 DOI : 10.35291/2454-9150.2019.0202

Pima Diabetes Dataset: Pima Diabetes dataset was introduced by Blake in 1998 [29] for diagnosing the presence of diabetes in pregnant women. Out of 768 samples present in the dataset, 268 cases indicates the presence (class '1') and 500 cases indicate the absence (class '0') of the disease. Nearly 48 % of the instances contain missing values which is a serious problem to be handled. The output variable is "Diagnosis" and the remaining are numeric attributes like Number of times pregnant, Plasma glucose concentration based on 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), Efficacy of 2-hour post glucose insulin levels (mu U/ml), Body Mass Index, Diabetes pedigree function, Age [2]..

Wisconsin Diagnostic Breast Cancer (WDBC) Dataset: WDBC dataset was created by William Street and Wolberg, [30] from the University of Wisconsin. The purpose of the dataset is to automate the diagnosis of breast cancer from a digitized image of a fine needle aspirate of a breast mass. There are 569 instances with 357 benign and 212 malignant samples. It consists of 30 real valued attributes describing the characteristics of the cell nuclei such as (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension) and a class value to predict benign / malignant [31].

Heart (Statlog) Dataset:The Statlog heart dataset was developed by Clevand Clinc Foundation containing 270 samples with no null values. The dataset contains 13 attributes such as age, sex (male, female), chest pain type (angina, asympt, notang, abnang), resting blood pressure, serum cholesterol measured in mg/dl, fasting blood sugar > 120 mg/dl (0, 1), resting electrocardiographic results (norm, abn, hyper), maximum heart rate achieved, exercise induced angina (0, 1), old peak – ST depression caused by exercise relative to rest, slope of the peak (up, flat, down), number of major vessels (0-3) and thal (normal, fixed defect, reversible defect). The input is a combination of real valued, ordinal and binary attributes. The output task is to predict the presence or absence of disease [2].

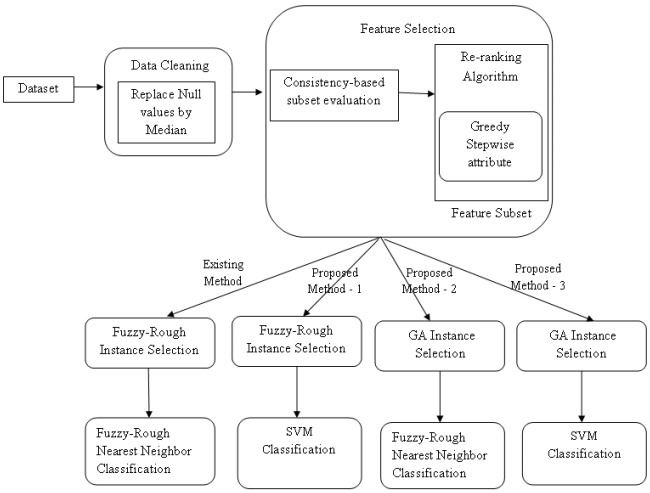

Figure 1: Block diagram for the proposed model

V. PROPOSED MODEL

Data preprocessing is necessary for any dataset to remove noisy, inconsistent and uncertain data. Appropriate methods used for data cleaning and data reduction may considerably increase the accuracy of the model. In this aspect, different combinations of data preparation methods are attempted based on the literature [1]. The four stages of data analysis involved in the experiment are:

- Data cleaning is done in two ways: In the first approach, the instances containing missing values are eliminated [1] for further processing. In second approach, the missing values are replaced by median [16]. The performances of both the methods are compared in order to prove the effect of imputation on the proposed model.
- The optimal feature set is chosen by utilizing the re-ranking search combined with Consistency-based subset evaluation method [1].
- Instances are selected using Fuzzy-rough instance selection [1] and GA (proposed).
- The resultant dataset is classified using Fuzzy-rough ownership function [1] and SVM (proposed).

Figure 1 depicts the block diagram of the research work in which novelty is introduced at instance selection and classification. The three different combinations (proposed) based on instance selection and classifications are:

Proposed Method 1: Fuzzy-rough instance selection for selecting the instances followed by SVM for classifying the instances.

Proposed Method 2: GA for removing the misclassified instances and the resultant is classified using Fuzzy-rough ownership function.

Proposed Method 3: GA for instance selection and SVM for classification.

The performance of the proposed model(s) is compared with the work of Aytug Onan[1] based on classification accuracy and other proposed evaluation metrics.

 DOI : 10.35291/2454-9150.2019.0202

VI. EVALUATION METRICS

Dividing the data into training and test set, generally 70:30 or 60:40 would be suitable for large datasets and N-fold cross-validation is well of use with small datasets, where the data used to train the classifier can be maximized [32]. Since the size of the dataset used in the experiment is small, an improvement of cross-validation namely stratified 10-fold cross-validation is used in which the class distribution in each fold is approximately similar to the initial dataset [33]. The entire dataset is divided into K (K=10) folds in which each fold is used once as a test set and has a training set (K-1) times. For each K = 1, 2….10, the classifier performance is evaluated. Finally, the average classification accuracy obtained from all 10 folds is calculated. The metric used to evaluate the experiment is given below:

Classification accuracy: One important metric to evaluate the model is classification accuracy. It measures the ratio of correct predictions over the whole range of instances evaluated.

$$Classification\ Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (13)$$

Where
True positive (TP) - range of positive samples correctly predicted.
False negative (FN) - range of positive samples wrongly predicted.
False positive (FP) - range of negative samples correctly predicted as positive.
True negative (TN) - range of negative samples wrongly predicted

True Positive Rate / Sensitivity / Recall: Sensitivity is used to measure the fraction of positive patterns that are correctly classified. It is the ability of the test to correctly identify the patients with the disease among the total number of diseased person in the dataset.

$$Sensitivity = \frac{TP}{TP+FN} \quad (14)$$

True Negative Rate / Specificity: Specificity is used to measure the fraction of negative patterns that are correctly classified. It correctly identifies the patients without disease among the non-diseased persons in the dataset.

$$Specificity = \frac{TN}{TN+FP} \quad (15)$$

F-Measure: F-Measure can be calculated as the harmonic mean between precision and recall. The value of F-Measure lies between 0 and 1 and the performance of classification algorithm increases for higher values of F-Measure.

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (16)$$

Area under curve (AUC): The AUC is a commonly used evaluation metric for binary classification problems [1]. It is defined as the probability that a classifier provides a higher chance for ranking a randomly chosen positive instance than a randomly chosen negative one. The area under Receiver Operating Characteristic (ROC) curve quantifies the overall ability of the test to discriminate between those individuals with the disease and those without the disease. The range of values lies between 0 and 1. AUC equals 1 when all the test data is assigned to true class labels. If its value is between 0.5 and 1, there is a 50% chance that a classifier can distinguish the classes. The value can't be less than 0.5 and if it is equal to 0.5, then the test made is of no use. ROC is one widely used performance metric for imbalanced datasets.

Kappa Statistics: Kappa is a statistical measure of agreement between the predicted class and actual class values. It is considered as an important measure for imbalanced datasets. Kappa can be calculated as:

$$Kappa = \frac{Pr(a)-Pr(e)}{1-\Pr(e)} \quad (17)$$

Where, $Pr(a)$ is the percentage of agreement between the classifier and the underlying truth and $Pr(e)$ is the chance of agreement calculated. The value of Kappa lies between -1 and 1, the possible interpretation of Kappa is as given below:

- 1 : perfect agreement
- 0.80 – 1.00 : very good agreement
- 0.60 – 0.80 : good agreement
- 0.40 – 0.60 : moderate agreement
- 0.20- 0.40 : fair agreement
- 0 - 0.20 : poor agreement

In rare situations, the value of Kappa can be negative (< 0) indicating that there is no effective agreement between two rates.

VII. EXPERIMENTAL ANALYSIS

The experiments were conducted on an open source tool WEKA (Waikato Environment for Knowledge Analysis) version 3.7.2 developed and maintained by University of Waikato, New Zealand. Table 1 describes the characteristic of dataset. The number of features selected based on re-ranking search combined with Consistency-based feature selection is given in table 2. From diabetes dataset, it is noted that the effect of imputation causes a change in selecting the features.

Table 1: Characteristics of Dataset

Dataset	No. of Instances	Input variables	Output variable	Missing Values
WBC	699	9	1	16
Pima	768	8	1	376

 DOI : 10.35291/2454-9150.2019.0202

Diabetes				
WDBC	569	30	1	NIL
Heart(Stalog)	270	13	1	NIL

Table 2: No. of features selected using Consistency-based subset evaluation

Dataset	Total no. of features	No. of features selected
WBC (1) (Deleting instances with null values)	9	7
WBC (2) (Replacing null values with median)	9	7
Pima Diabetes (1) (Deleting instances with null values)	8	7
Pima Diabetes (2) (Replacing null values with median)	8	8
WDBC	30	8
Heart (Stalog)	13	10

For each trial, GA has a tendency to select different instances as outliers. To achieve consistent result for removing the wrongly classified instances, the experiment was repeated 50 times and the corresponding classification accuracy obtained by SVM is recorded. Among the 50 test runs performed one trail / run which is closer to the average classification accuracy value is selected for further analysis. Table 3 reports the number of instances obtained after instance selection algorithm using Fuzzy-rough and GA. The results obtained clearly states that, in most of the cases the number of instances reduced using GA is very less compared to Fuzzy-rough instance selection. Figure 2 shows the comparison of instance reduction percentage obtained by Fuzzy-rough instance selection and GA. It is clearly visible that GA has a much less data reduction percentage compared to Fuzzy-rough instance selection for all the tested datasets irrespective of imputations.

Table 3: No. of instances selected using Fuzzy-rough instance selection and GA

Dataset	Total no. of instances	No. of instances selected	
		Fuzzy-rough	GA
WBC (1)	683	351	651
WBC (2)	699	358	662
Pima Diabetes (1)	392	298	287
Pima Diabetes (2)	768	258	512
WDBC	569	269	524
Heart (Stalog)	270	192	214

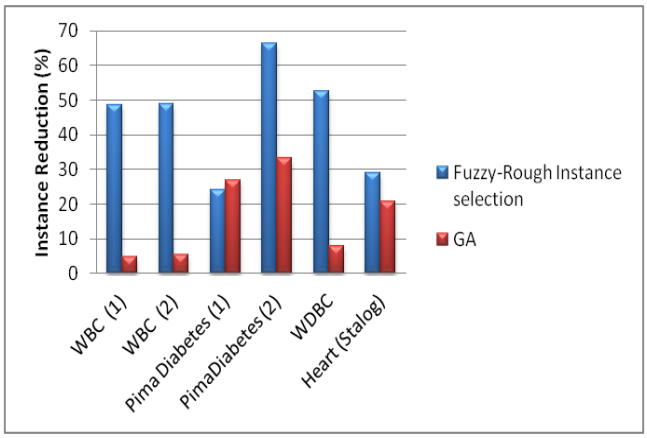

Figure 2: Comparison of instance reduction (%) using Fuzzy-rough Instance selection and GA

Table 4 reports the classification accuracy obtained by the proposed models. It can be observed from the results that in majority of the cases, all the three proposed methods obtain better accuracy compared to the work of existing study [1]. It is also evident that among the proposed methods, the combination of GA and SVM provides the highest accuracy for all the tested datasets. From table 5, it is proved that the evaluation metrics of the proposed method 3 (GA+SVM) shows better results than achieved in the study of [1].

Table 4: Comparison of Proposed Methods with the Existing Method in terms of Classification Accuracy (%)

Dataset	Existing Study	Proposed Method 1	Proposed Method 2	Proposed Method 3
WBC (1)	99.71	99.14	99.38	99.84
WBC (2)	99.16	98.32	99.84	99.84
Pima Diabetes (1)	80.20	86.24	98.25	99.30
Pima Diabetes (2)	81.78	85.27	98.24	99.80
WDBC	98.51	99.25	99.61	99.80
Heart (Stalog)	85.41	90.62	99.06	99.53

Table 5: The evaluation metrics for proposed method 3 and Existing Study [1] for WBC dataset

Evaluation Metrics	Existing Study [1]	Proposed Method 3 WBC (1)	Proposed Method 3 WBC (2)
Accuracy (%)	99.7151	99.8464	99.8489
Sensitivity	1.0000	1.0000	0.9954
Specificity	0.9947	0.9976	1.0000
F-Measure	0.9970	0.9980	0.9980
AUC	1.0000	0.9990	0.9980
Kappa	0.9943	0.9966	0.9966

Further, the efficiency and stability of the models is checked by considering three more datasets (i) Pima Diabetes (with null values) (ii) WDBC and Heart (without null values). Table 6 shows the evaluation metrics for the above datasets and it is observed that the performance of the proposed models is consistent across tested datasets.

 DOI : 10.35291/2454-9150.2019.0202

Table 6: The evaluation metrics obtained by proposed method 3 for Pima Diabetes, WDBC and Heart dataset

Evaluation Metrics	Proposed Method 3 (GA + SVM)			
	Pima Diabetes (1)	Pima Diabetes (2)	WDBC	Heart *(Stalog)*
Accuracy (%)	99.3031	99.8047	99.8092	99.5327
Sensitivity	0.9911	1.0000	0.9944	0.9878
Specificity	1.0000	0.9917	1.0000	1.0000
F-Measure	0.9930	0.9980	0.9980	0.9950
AUC	0.9960	0.9960	0.9970	0.9940
Kappa	0.9792	0.9946	0.9958	0.9901

The process of selecting the features brings efficiency at each stage of evaluation process. Consistency measure is monotonic and used to remove redundant / irrelevant features. It is capable of handling noise and efficiently rejects irrelevant data as a percentage of inconsistencies. Selected features improve the performance in terms of instance reduction and classification accuracy. The proposed model (GA + SVM) with and without feature selection is compared in figure 3 and 4. It is evident that for all the tested datasets, the instance reduction percentage is low and the classification accuracy is high when the model is subjected to feature selection.

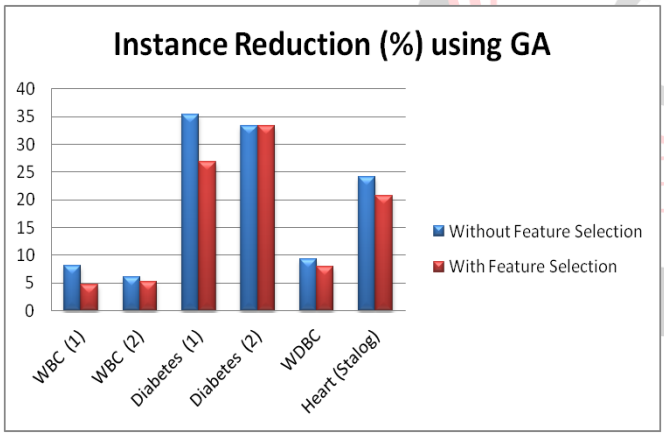

Figure 3: Instance selection using GA with and without feature selection

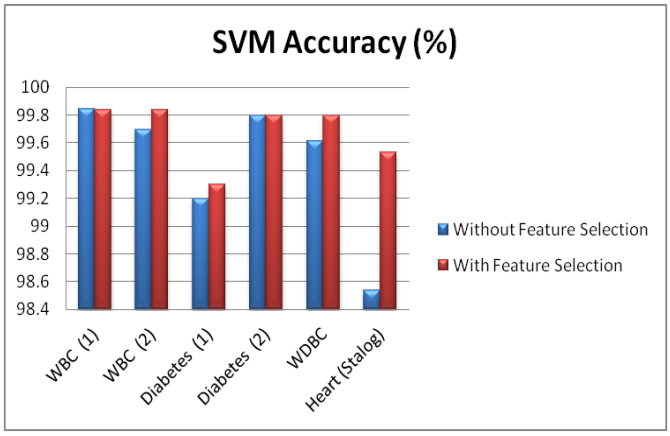

Figure 4: Classification Accuracy using SVM with and without feature selection

Research Findings:

- There is a considerable increase in the classification accuracy percentage, when the null values are replaced by median rather than deleting the instances with null values.
- GA for instance selection has less data reduction percentage compared with Fuzzy-rough instance selection.
- For all the experimented datasets, irrespective of the instance selection algorithms used, SVM provides better classification accuracy than Fuzzy-rough nearest neighbor classifier.
- Among the proposed methods, GA for instance selection followed by SVM for classification yields highest classification accuracy across all the tested datasets.
- In comparison with the models proposed in literature, the proposed method yields a less data reduction percentage and improved classification accuracy.

VIII. CONCLUSION

This article presents a classification model based on the SVM classifier, Consistency-based feature selection and GA in support of instance selection for medical diagnosis. The two important tasks to build robust model are feature selection and instance selection. Sometimes in the medical domain, the dataset may consist of erroneous and noisy instances. Therefore, appropriate selection of instances and determination of an optimal subset with relevant features is necessary to construct classification model.

The proposed framework can be summarized as given below. Initially, missing values are cleaned by removing the instances and replacing the missing values with their median. Next, selecting the optimal feature set is done using Consistency-based subset selection followed by GA to remove the misclassified instances. Finally, SVM is used for classifying the datasets. The model provides better classification accuracy of well over 99 % for the tested datasets. On an average, the instance reduction percentage of GA and Fuzzy-rough instance selection is 16.45 % and 44.89 %. It is proved that GA instance reduction rate is 28% less compared with Fuzzy-rough instance selection.

The experimental results on WBC indicates that the proposed classification model provide promising classification results in terms of evaluation metrics like classification accuracy, sensitivity, specificity, etc. WBC dataset contains only 2.28% of missing values and hence it is better to drop the null values rather than imputation. But in the case of Pima diabetes, the percentage of missing value is 48.9%; hence it is always better to impute the

 DOI : 10.35291/2454-9150.2019.0202

values. From the experimental results, it is very clear that the proposed classification model can be used as a tool for an automated diagnosis of diseases such as breast cancer, diabetes and heart. Compared to the state-of-the-art methods reported in the literature, this model provides better results with respect to evaluation metrics.

There may be limitations in this research study which are outlined below. Health care being a rich domain in terms of datasets with different features, development of a classification model that would be used as a viable tool to classify various medical datasets could be a challenging research problem. However, steps have been taken to test the performance of the model on four datasets with different diseases. The scope of this work is limited to the automated diagnosis of breast cancer, diabetes and heart. It is better to carry out more experimental work to obtain a generic classification model for medical domain with better diagnostic ability and stability. It can be concluded that the datasets used in this research study can properly work with Consistency-based feature selection, GA and SVM. In addition, the behavior of feature and instance selection and the other machine learning techniques should also be taken into consideration. Imbalanced classification problem occurs for WBC and Pima diabetes which is not eliminated. Further, the model has to be pruned to convert imbalanced datasets to balance dataset.

The future work will focus on the following aspects: Initially, at the stage of preprocessing WBC and Pima diabetes dataset can be balanced using the sampling techniques like SMOTE, SMOTE+ENN, etc. Next, the feature selection algorithms like RFE, regularization methods can be attempted to select significant attributes that can enhance the model accuracy. Finally, the other classification and regression models suggested in the literature like neural network, decision trees and logistic regression can be explored.

REFERENCES

[1] Aytug Onan, "A Fuzzy-rough nearest neighbor classifier combined with Consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer", *Expert Systems with Applications* 42, 6844-6752, 2015.

[2] Nihat Yilmaz, Onur Inan and Mustafa Serter Uzer, "A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases", *Springer:Transaction Processing Systems:J Med Syst*, 38:48, 2014.

[3] Mellisa Humphries, "Missing Data & how to deal: an overview of missing data", Presentatión in internet, http://www.utexas.edu/ cola/centers/prc/_files /cs/ Missing-Data.pdf, 2012.

[4] I. Guyon and A. Elisseeff, " An introduction to variable and feature selection", *J Mach Learn*, 3, 1157–82, 2003.

[5] Girish Chandrashekar and Ferat Sahin, "A Survey on feature selection methods", *Computers and Electrical Engineering*, 40, 16–2, 2014.

[6] Manoranjan Dash and Huan Liu, "Consistency-based search in Feature Selection", *Artificial Intelligence : Elsevier*, 151(1-2), 155-176, 2003.

[7] Michal Wozniaka, Manuel Grañab and Emilio Corchadoc, "A survey of multiple classifier systems as hybrid systems. Information Fusion", *Special Issue on Information Fusion in Hybrid Intelligent Fusion Systems*, 16, 3–17, 2014.

[8] R. Jensen, "Fuzzy-rough data mining". In S. O. Kuznetsov et al. (Eds.), Lecture notes in artificial intelligence, Berlin Heidelberg: *Springer-Verlag*, 31–35, 2011.

[9] P. Vishnu Raja and V. Murali Bhaskaran, "An Effective Genetic Algorithm for Outlier Detection", *International Journal of Computer Applications*, 38(6), 2012.

[10] M. Sarkar, "Fuzzy-rough nearest neighbors algorithm". *Fuzzy Sets and Systems*, 158, 2123 – 2152, 2007.

[11] T. Van Gestel, J.A Suykens, B. Baesens, et al. "Machine Learning", 54: 5. https://doi.org/10.1023/B:MACH.0000008082.80494.e0, 2004.

[12] Z. Liu and S. Pan, "Fuzzy-Rough Instance Selection Combined with Effective Classifiers in Credit Scoring", *Neural Process Letters: Springer*, 47 - 193. https://doi.org/10.1007/s11063-017-9641-3, 2018.

[13] T. Nguyen, A. Khosravi, D. Creighton and S. Nahavandi, " Medical data classification using interval type-2 Fuzzy logic system and wavelets", *Applied Soft Computing*, 30, 812–822, 2015.

[14] L. Meenachi and S. Ramakrishnan, "Evolutionary sequential genetic search technique-based cancer classification using fuzzy rough nearest neighbour classifier", *Healthcare Technology Letters : IEEE*, 5(4),130-135, 2018.

[15] Amandeep Kaur and Kamaljit Kaur, "Implementing Outlier Detection using Greedy Based Information Theoretic Algorithms and its

 DOI : 10.35291/2454-9150.2019.0202

Comparison with PSO and ACO Optimization Techniques", *International Journal of Current Engineering and Technology*, E-ISSN 277 –4106, INPRESSCO, Article ID 418060, 2015.

[16] T. Santhanam, and M.S Padmavathi, "An Efficient Model by Applying Genetic Algorithms for Outlier Detection in Classifying Medical Datasets", *Australian journal of Basic and Applied Sciences, 2015.*

[17] G. Ravi Kumar, G.A Ramachandra, and K. Nagammai, "An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets", *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(2), 2014.

[18] B. Zheng, S.W Yoon, S. W and S.S Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms". *Expert Systems with Applications*, 41, 1476–1482, 2014.

[19] H. Liu and R. Setiono, " A probabilistic approach to feature selection: A filter solution", InL.Saitta(Ed.), *Proceedings of the thirteenth international conference on machine learning*, San Francisco: Morgan Kaufmann, (pp. 319–327), 1996.

[20] H. Liu, F. Hussain, C.L Tan, et al, "Data Mining and Knowledge Discovery", 6: 393. https://doi.org/10.1023/A:1016304305535, 2002.

[21] P. Bermejo, L. Ossa, J.A Gamez and , J. M Puerta, "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking", *Knowledge-Based Systems*, 25, 35–44, 2012.

[22] R. Jensen and Chris Cornelis, "Fuzzy-rough nearest neighbour classification and prediction", *Theoretical Computer Science*, 412.42, 5871-5884, 2011.

[23] R. Jensen and Chris Cornelis, "Fuzzy-rough Instance Selection". *WCCI, IEEE World Congress on Computational Intelligence CCIB*, Barcelona, Spain, 18-23, 2010.

[24] D.E Goldberg, "Genetic Algorithm in Search, Optimization, and Machine Learning", *Addison-Wesley*, Boston, 1989.

[25] J.M Keller, M.R Gray and J.A Givens, "A Fuzzy K-nearest neighbor algorithm", *IEEE Trans. Systems Man Cybernet*, 15(4), 580-585, 1985.

[26] B.E Boser, I.M Guyon and V.N Vapnik, " A training algorithm for optimal margin classifiers", *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144–152, 1992.

[27] R. Courant and D. Hilbert, "Methods of Mathematical Physics", *Wiley*, New York, USA, 1953.

[28] John Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training support Vector Machines", Technical Report. MSR-TR-98-14, *Microsoft Research*, 1998.

[29] C.L Blake and M.C.J, *UCI repository of machine learning, 1998*.

[30] N. William Street, W.H Wolberg and O.L Mangasarian, "Nuclear feature extraction for breast tumor diagnosis", *International Symposium on Electronic Imaging: Science and Technology*. *San Jose, CA*, 1905, 861-870, 1993.

[31] Ahmet Mert, Niyazi Kiliç, Erdem Bilgili and Aydin Akan, " Breast Cancer Detection with Reduced Feature Set", *Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine*, Article ID 265138. http://dx.doi.org/10.1155/2015/265138, 2015.

[32] D. J Hand, H. Mannila and P. Smyth, Principles of data mining. MIT press, 2001.

[33] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees". *Wadsworth*, Belmont, CA, 1984.

 DOI : 10.35291/2454-9150.2019.0202