JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 25 Issue 04, 2022 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

MISSING VALUE ESTIMATION METHODS FOR CLASSIFICATION OF ARRHYTHMIA USING DEEP LEARNING

Dipalika Das[1*],Maya Nayak[2], Subhendu Kumar Pani[3]

[1]Dipalika Das, Research Scholar, Department of Computer Science and Engineering, Biju Patnaik University of Technology, Rourkela, Odisha, India; dipalika.das@gmail.com,

[2]Dr. Maya Nayak, Dean School of Computer Studies, Ajay Binay Institute of Technology(ABIT), Cuttack, Biju Patnaik University of Technology(BPUT) Rourkela, Odisha, India; mayanayak3299@yahoo.com,

[3]Dr. Subhendu Kumar Pani, Professor, Krupajal Engineering College(KEC), Bhubaneswar, Biju Patnaik University of Technology(BPUT) Rourkela, Odisha, India; pani.subhendu@gmail.com

Abstract: - Biomedical signals like ECG signals are significant to the classification of heart diseases using deep learning techniques. In reality, the ECG datasets mainly consist of matrix data with missing value because of errors or faults. As many classical classification methods, need a full data matrix for input. Therefore, the apt way to impute the missing data is to alleviate the effectiveness of classification of datasets with few missing values. In this paper, the approach of random forest is used for imbalance dataset and compared with other methods e.g. zero method, mean method and PCA based method. The proposed classification algorithm used is Deep Neural Network. The simulation inference is based on the UCI database reflects that random forest method can manage better accuracy while handling missing values in cardiac arrhythmia dataset. Adaptive Neuro-fuzzy inference system classification model works efficiently with proposed method of imputation with efficiency.

Keywords: - Missing Value Estimation, Arrhythmia Classification, Random Forest, Adaptive Neuro fuzzy inference system (ANFIS)

Introduction

In recent years, some of the most prevalent problems that have jeopardised human health all over the world are cardiac diseases. Among all the cardiac ailments, the one related to the disorder of the rhythm of the heart is known as cardiac arrhythmia. Cardiac arrhythmia is of various types, some of which can cause irreparable long-term damage to the heart, even sudden death [1]. Thus, early detection and proper classification of these fatal arrhythmias early is extremely crucial. This will help choose proper antiarrhythmic drugs and give appropriate medical treatment.

Electrocardiogram or ECG as it is commonly known is mostly used to record the heart signals in medical institutes and hospitals. As it is non-invasive it is the preferred tool for diagnosis and detection of arrhythmias. Heart exhibits bioelectrical activity that can be displayed in the form of a graph and can be recorded which is called the electrocardiograph [2]. Each cycle of ECG cardiac signal consists of P wave, QRS Complex, and T wave components, which are presented as P, Q, R, S, and T. Amplitude and Duration of the heart signal in ECG cardiac cycle of patient are the main parameters to measure or rather for examination of the heart

Submitted: 29/11/2022 **Accepted**: 17/12/2022

disease of the same individual. The examination of heart disease of patients includes inference drawn pertaining to some of the characteristics of ECG, i.e., the peak (P, Q, R, S, T, and U), intervals (PR, RR, QRS, ST, and QT), and segments (PR and ST) [3]. Figure 1 shows a very basic, sample ECG signal that shows the eminent features which pertain to different waves in the heart.

Owing to this structure of the ECG waveform, the condition pertaining to the irregular changes in heart rate can be detected by medical physicians otherwise known as arrhythmia. The visual check of numerous patients having arrhythmia is however tiring, time wasting and extensive hard work. The traditional methods of manual analysis have proved in past are not ideal due to a subjective approach causing errors in diagnosis results in term of accuracy. Therefore, to avoid these flawed diagnoses it is necessary to devise an approach of automated computing which can do detection and classification of arrhythmia effectively, accurately and efficiently. In recent years, a different approach has been devised to help the medical practitioner for better results in treatment of arrhythmia.

Another aspect for devising such models is necessary to pre-process of data available in the dataset for exact recognition of disease. Some of these datasets contain missing value in data which can create problems in classification or recognition. These missing value problems exist in the arrhythmia dataset.

To resolve the missing value problem in arrhythmia, classification by means of solving optimization problems and do research on the performance of matrix completion or principal component analysis for the dataset with different proportions of missing values compared with other state-of-the-art methods for value imputation in arrhythmia classification.

Recent advances in the field of machine learning (ML) and its implementation for biomedicine and bioinformatics has received considerable attention. These studies are focussed on creating an advanced and accurate automatic algorithm for the identification of data and its classification. In order to make the identification and classification more accurate, some machine learning classifiers like Support Vector Machines (SVM), Multilayer Perceptron (MLP), Artificial Neural Network (ANN), K-Nearest Neighbour (KNN), and their variants and combinations have been used [4–8]. For its simplicity and highly adaptive nature,102j KNN is the most commonly used method for arrhythmia classification [9]. In our former work, we also propose a ANFIS having layer of 5 with member function GbellMF and GaussMF.

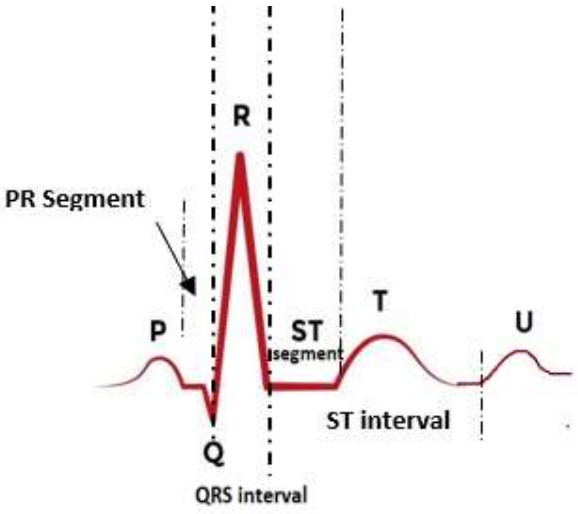

Figure 1 represents Classic ECG signal

As for arrhythmia classification, this paper attempts to figure out the missing value problem under different conditions with uneven number of sample training data in arrhythmia datasets. Some of the major contributions of our work are summarized as follows:

(1) With missing variant value portions (e.g.,10% -70%) for datasets, how the different missing value imputation methods work on the accuracy of selected classifiers has been discussed

(2) The introduction of classifier ANFIS is done with respect to the sample number, which improves the accuracy for the arrhythmia classification

(3) When the missing data volume is larger, UCI database result post the experiment indicates that the Random forest method is effective for the imputation of the missing values in cardiac arrhythmia classification. In terms of accuracy of classification our proposed algorithm is more effective for uneven datasets as compared to other methods.

The two most common ways to deal with the problem of missing values are a) the missing value can be removed or ignored and b) It is filled in or imputed with new values. When the amount of data missing is small then the former solution is applicable. Wherever a bigger amount of data is missing in databases, it is more practical to consider the process of fill in or imputation.

A number of different imputation methods have been reported in literature review. Some rely on simple algorithms such as mean imputation and few on complex methods including regression-based imputation and expectation–maximization (EM) algorithm. [10]. Over the last few years new imputation methods which use machine learning (ML) algorithms have been proposed. Another breakthrough development was in the form of the multiple imputations (MIs) first described by Rubin in the 1970s.

Based on the degree of randomness of deletion Little, R. et al. coined three types of data loss: -a) missing completely at random (MCAR), b) missing at random (MAR), and c) not missing at random (NMAR) [11]. The choice of imputation algorithm and the final effect of imputation are chosen properly for the missing type of a dataset. The MCAR uses the mean-imputation

method as a common missing value imputation method. The imputation of NMAR data depends on the prior knowledge of the dataset itself. The effect of different missing types on data classification is discussed in [12]. Due to the difficulties associated with NMAR data imputation, the current missing value imputation algorithm used is the MAR data, where related feature values are used to estimate the value of the missing data. However, all these methods have their own set of drawbacks. For example, without the knowledge of the distribution of data in the dataset, the linear regression algorithm based on statistical probability and the maximum expectation algorithm will not work. But our understanding of most datasets is relatively low. Bayesian network and k- neighbourhood algorithm is based on data Mining. The Bayesian network should have some knowledge of the domain and of the data. Therefore, it is essential to clear the dependence between various features. Training Bayesian networks directly with the dataset is complex, and KNN algorithm with a high rate of missing cases, the imputation effect will be greatly reduced. In the literature, these processing techniques contain signal processing [13], pattern recognition [14, 15], and machine learning [16–18] methods.

A large number of arrhythmia databases processed by the above techniques have been used as the benchmark by researchers to compare the performance of their research methods with others. In general the arrhythmia databases can be classified into two types: signal (e.g., MIT-BIH) and numeric (e.g., UCI). In this paper, we research on the numeric arrhythmia databases, which have been pre-processed to multidimensional feature vectors by some signal processing and pattern recognition techniques like digital filters and peak analysis [19].

However, some attribute values of the ECG data would inevitably be missing after the pre-processing. Unfortunately, many algorithms used for classification like K-Nearest Neighbor (KNN) are not robust enough to input matrix with missing values, which would make it less effective. Therefore, pre-processing before analysis of the data is an essential task to cope with. For missing values, some studies apply the “case deletion” method directly, i.e., simply removing those instances with missing values. They use only the observed instances to establish the classification models, which may lead to loss of some crucial information especially for small sample datasets [20]. In order to tackle these shortcomings several missing value imputation methods have been proposed in the past decade to fields like DNA microarrays [21–25] and traffic data problems [26, 27]. For example, Troyanskaya et al. present a prevalent imputation method based on KNN, i.e., KNN impute for DNA microarrays [22]. Tan et al. propose the PPCA method, a matrix completion method dealing with missing traffic flow problems [27].

The imputation methods have two divisions which are interpolation based methods and inductive learning-based methods [27]. The former method fills in the vacancies (missing values) according to the mean or median of the rest of the values that belong to the same column, or just simply fill them with zeros. Different from the interpolation one, the inductive learning-based methods refer to assign probabilistic values based on the distribution of the known values.

For the abovementioned arrhythmia databases that consist of numeric, there is not enough literature published on missing value imputation. The “case deletion” [28] and “row average”

[29] methods are most commonly implemented on the numeric arrhythmia datasets regardless of whether the missing values are significant. Since these simple approaches do not consider the information of the missing values and ignore the covariance in the data. This may add a higher degree of uncertainty and shall lead to bias. A modified PCA method was proposed to address the missing value problem in the arrhythmia datasets in the earlier works [30]. In this paper, we want to contemplate further on a more evolved strategy i.e., Robust Principal Component Analysis (RPCA). The Matrix Completion (MC) problem can be viewed as a special case of the RPCA problem [31].

The main algorithm of the two problems is the same. The major difference between the two problems is that while the former aims to recover the matrix that is corrupted, the latter tries to recover the matrix with missing values. In this paper we shall refer to the Matrix Completion (MC) method with RPCA method. In MC method, we can solve the optimization problem under some conditions, as in Equation below, to recover the incomplete matrix [32]:

Minimize rank(Z),

Subject to $Z_{ij} = Pij$, $(i,j) \in \Omega$

where **P** is the observed matrix and the set Ω is the indices of **P**. And, this approach is capable of recovering matrices of rank about 10 with nearly a billion unknowns from just about 0.4% of their sampled entries [33].

This paper deals with the imputation methods of the missing values in ECG and the machine learning methods suitable for arrhythmia classification. There is a brief discussion on the studied problems in the field of arrhythmia classification and review the previous work done in this field. The remainder of this paper is organized as follows. Section 2 describes the arrhythmia dataset. Section 3 relates to different missing value imputation methods, and our algorithm, i.e., Random Forest and Adaptive Neuro fuzzy inference system (ANFIS). Section 4 gives the experimental results on the UCI datasets. Finally, the conclusion and scope of future work are drawn in Section 5.

2. Description of Data Set

The standard multivariate ECG dataset taken here is chosen from the Irvine (UCI) cardiac arrhythmias database of the University of California [33]. This database contains 452 instances of samples with 279 features, of which the first to 4 features refer to the general information of a patient like age and sex, whereas the remaining 275 features are the numeric features selected from the ECG signal waveform by some signal or pattern process techniques.

These 452 participants can be divided into 16 classes according to the ECG data. The first class is “Normal”, and the other 15 classes are “Abnormal”, corresponding to 15 different kinds of arrhythmia. A brief description of the 16 classes is given in Table 1. For more details of the data set, please refer to the download website [34].

The UCI cardiac arrhythmia database contains two significant characteristics. First, there are several missing feature values (about 33%). Secondly the distribution of class labels is not in balance.

As shown in Table 1, the “Normal” class has 245 instances of samples whereas one of the abnormal classes “Supraventricular Premature Contraction” has only 2 cases. For the first,

second, and third-degree atrioventricular block cardiac arrhythmia class, there is no instance of sample due to the insufficiency of the data sample. Note that, certain classes with no samples in the training dataset is not in our consideration. We focus on the imbalance of training data rather than solving the missing sample belonging to a certain class.

Table 1: UCI cardiac arrhythmia database with description of class

Class	Name of the class	Instances number
1	Normal	245
2	Coronary artery disease	44
3	Old anterior infarction	15
4	Old interior infarction	15
5	Sinus Tachycardia	13
6	Sinus Bradycardia	25
7	Ventricular Premature Contraction	3
8	supraventricular Premature Contraction	2
9	Left bundle branch blockage	9
10	Right bundle branch block	50
11	First degree atrioventricular block	0
12	second degree atrioventricular block	0
13	third degree atrioventricular block	0
14	Left ventricular hypertrophy	4
15	Atrial flutter	5
16	others	22

3. Research Methodology

3.1. Combination of Feature Selection and Missing Value Imputation

The process combining feature selection and missing value imputation is illustrated in Figure 2. The incomplete P dimension dataset A is composed of training and test sets, denoted by A_tr and A_te, respectively. For feature selection, A_tr contains a number of complete (i.e., A_complete) and incomplete (i.e., A_incomplete) data samples. The feature selection step is performed on the A_complete subset, leading to a new subset that contains O dimensions (where O < P), denoted as A_complete'.

It should be noted that the feature selection process only considers the data in A_complete, since each of these data contains no missing feature values, which allow feature selection algorithms to successfully select a subset of representative features. However, the issue of whether A_complete represents the population is beyond the scope of this paper. Next, the A_incomplete subset is also reduced to the same N dimensional subset, denoted as A_incomplete'. For missing value imputation, the A_complete' subset is used to construct a learning model. For the example of imputing the missing value of the i-th feature (i = 1, 2, : : : , O) in A_incomplete', the learning model is trained by the data samples of A_incomplete',

where the i-th feature of A_incomplete' is used as the output feature and the rest of the features are the input features.

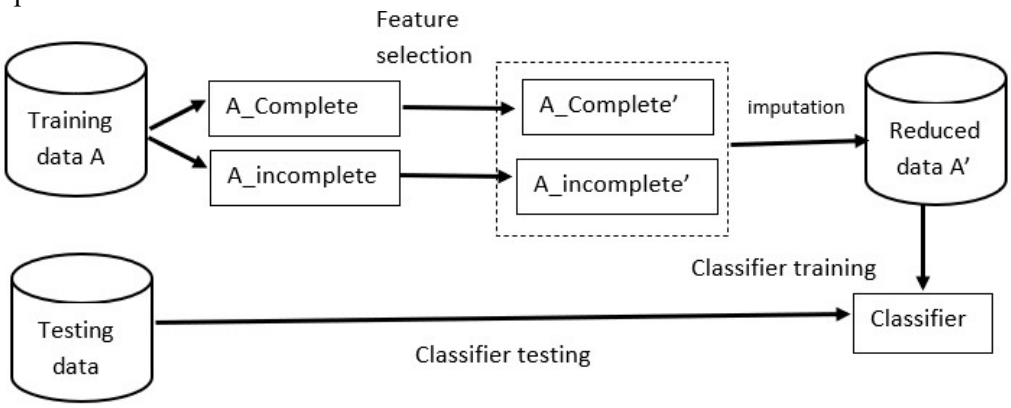

Figure 2 represents as missing value imputation and feature selection combined. A_complete and A_incomplete is related to complete and incomplete training data samples of dataset A. A_complete' and A_incomplete' is related to reduced feature set of the main A_complete and A_incomplete respectively. A' means the reduced features set of the main A without any missing value

The model then produces estimations to replace the missing values in the A_incomplete' subset. The same process is performed over the testing set A_te. That is, the M dimensional testing set A_te is reduced to the N dimensional testing set, denoted by A_te'. The next step involves missing value imputation performed by the learned model trained by A_complete'. Finally, the imputed dataset, which is denoted as reduced data A_tr', is used to train a classifier, and its classification performance is examined by the reduced testing set (i.e., A_te').

Without feature selection being performed, the baseline computation imputation process uses A_complete directly with the model to produce estimations for the missing values of A_incomplete. The aim of this study is to examine differences in performance between the combined feature selection imputation method and the baseline imputation method.

Feature Selection

Feature selection is defined as a process of selecting a subset of relevant features (or variables) from a given dataset. Since real-word datasets generally have some features that are either redundant or irrelevant, they can be discarded without witnessing much loss of information [35,36]. In simpler terms, feature selection can be regarded as a case of dimensionality reduction, which tries to decrease the number of random variables under consideration by obtaining a set of principal variables. The difference between feature selection and dimensionality reduction is that the set made by dimensionality reduction does not have to be a subset of the original set of features. For principal component analysis, linear combination of the original ones creates new synthetic features and the less important ones are omitted.

In general, feature selection algorithms can be classified into three types of methods—filter, wrapper, and embedded methods [35]. One major type of filter method used to select important features is based on ranking techniques. Essentially the input features are scored via a suitable ranking criterion and features that are below a certain threshold are omitted. Many statistical

techniques belong to the filter type of method, including information gain and stepwise regression.

The wrapper methods are based on using a predictor (or learning model) as the objective and their function is to evaluate different feature subsets. The subset that can make the predictor produce the highest accuracy rate is chosen as the best feature subset. Evolutionary compaction techniques, such as the genetic algorithm and particle swarm optimization methods, have gained attention in recent times and shown some success [37,38]. The representative wrapper methods are the genetic algorithm and particle swarm optimization methods. However, the wrapper methods involve a large computational cost for model training and searching for the best subset. During the process of model learning the embedded methods perform feature selection during the model learning process [38–40]. In other words, feature selection is incorporated into the classifier training process. Embedded methods not only measure the relations between the input features and the output features, but also search for features that allow better classification accuracy. The decision tree model is one such embedded method, where the constructed tree contains a number of selected features (i.e., decision nodes) that can differentiate between different classes (i.e., leaf nodes).

In this paper, we conduct our research based on these two characteristics and the specific methods will then be introduced in the next section.

Random Forests

In machine learning Ensemble or Network Committees are algorithms which essentially combine individual paradigms to form different combinations which are often more accurate than the individual classifier alone [41]. In the classification case, overall predictions can be arrived at from such a network using a weighted or an un-weighted voting system. Whereas in the regression case through an averaging technique overall predictions can be chosen. Obtaining a general understanding of why such methods succeed is an active area of research [41, 42].

A *Decision Tree* is a tree with nodes which contains information related to attributes in the input vectors.

For a given set of input attributes this information is used to follow a certain decision path, which depends on either thresholding nodes (as in the case of a continuous variable) or categorical nodes (as in the case of categorical data) [43]. Even though decision trees have appeal for being straightforward and fast, they are prone to being overly adapted to the training data or to a loss in accuracy for generalization through tree pruning [44].“

Through bagging (Bootstrap Aggregation) which combines multiple random predictors in order to aggregate predictions [46], Random Forest” (RF) is an algorithm which generalizes ensembles of Decision Trees [45]. They allow for complexity without over-generalizing the training data [44]. RF can be used for both regression and classification, and has been used with success in the context of missing data [43]. Random Forests were first introduced in 2000 by Breiman, and “Random Forests” is a trademark of Cutler and Breiman [41]. Each tree in the

RF is grown according to algorithm 1, and each tree forms an independent member of the forest [43].

Algorithm about growth of a tree in Random Forest

1) The Splitting Criterion is selected i.e. is constant for the Random Forest, n to be less than the number of the input variables N(n<<N).
2) For A training Samples, Sample A cases at random with replacement.
3) With the A sampled cases, growth of each tree according to A sampled case is performed
    a) n variables are selected at random from N and the best split on these n is used to split at each node.
    b) Each tree is grown as much as possible (with no pruning)

Thus, each tree in the forest contains an individually selected subset of the overall collection of features. The error rate of the RF is shown to be dependent on two properties [45]-1) ***correlation*** *between different tree in the forest and 2)* ***strength*** of an individual tree in the forest.

The *correlation is related to the similarity between one tree to another tree. If the correlation increases between trees increases* the Forest Error Rate (FER). The *strength relates* to how powerful a classifier the tree is, and increasing the power of each trees decreases the FER. The parameter *n* is directly related to both *correlation* and strength, so there is an optimized range of *n* in which the *correlation* made lowest and the *strength made highest*. Sample with replacement results in some of the training set not being used in training (approximately a third of the training data) [45]. These data are referred to as "out-of-bag" (oob) data that are used to get an unbiased estimate of the performance of the RF, which is unlike *cross-validation* which may be biased [47]. Furthermore, oob data are used in predicting variable importance, which is discussed further in section VI. Information regarding strength and correlation can also be obtained from the oob methods, allowing one to gain insight into the forest [47]. The *proximity* is an *N* X *N* matrix obtained by running all the data down the tree, and if two cases are in the same terminal node, their proximity is increased by one [45].

This is a useful property which can be used in locating outliers or estimating missing data. RFs have been an area of active research in the last few years for their numerous advantageous features and high success [41]. RFs are said to work fast, have excellent accuracy offering improvements over single classification and regression trees (CART).It is impervious to overfitting the data, runs efficiently on thousands variable numbers (no dimensionality problems), give an unbiased self-assessment as well as a variable importance assessment [41], [45]. These properties make the RF algorithm a logical and suitable candidate for this missing data study.

3.2 ANFIS Classification

Neuro-Fuzzy:

With knowledge expressible in linguistic rules, a fuzzy inference system (FIS) can be developed .Fuzzy systems involve interpretation of the rules through a process of fuzzification (resolving the antecedent to a degree of membership), fuzzy operation and implication (the

consequent assigns a fuzzy set to the output) [48]. Fuzzy inference is the entire process of mapping from a given input to a given output using fuzzy logic. While a fuzzy system makes use of natural language, a NN can be used if we have data for training. Drawbacks of each of the systems are seen to be complementary, and thus the integration of the two systems is logical. The FIS offers an advantage in terms of learning capability, while the extraction and learning of rules is a problem well suited to ANNs [49].

Neuro-Fuzzy systems consist of rule sets and inference systems combined with or governed by a connectionist structure for optimisation and adaptation to given data. Adaptive Neuro Fuzzy Inference System (ANFIS) implements a Takagi Sugeno (TS) FIS and consists of five layers, the first of which is for fuzzification of the input variables [49]. The second layer uses a T-norms operation which computes the rules. The third layer normalizes the rule strength, where as the fourth layer helps determine the consequent of the rule.The final layer is the output layer, which combines all the inward flowing information to compute the weighted global output. It is important to note that in TS FIS, the consequent part of the rule is mathematically zero order or first order [48].

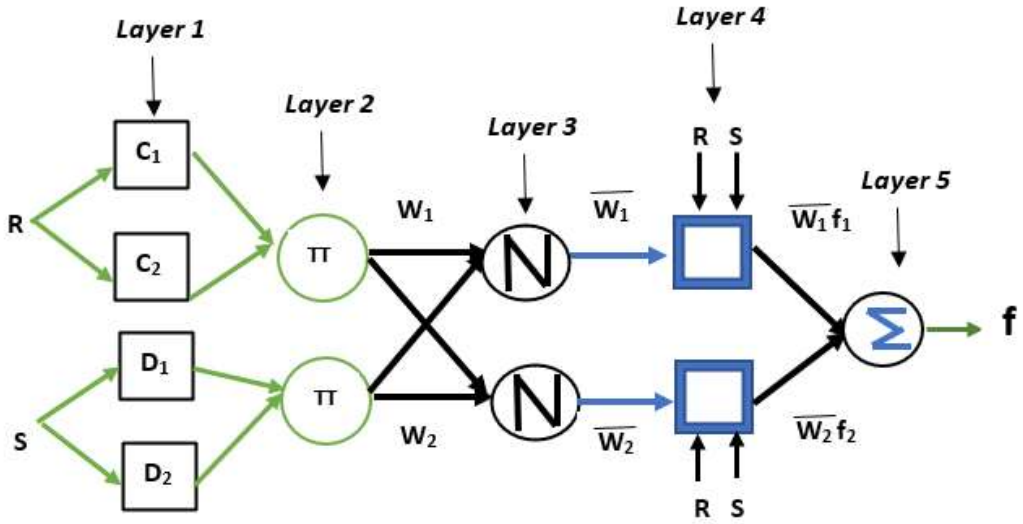

Figure 3 represents ANFIS Structure

The schematic architecture of this system is presented in figure 3 as in [50], [51]. The system can employ grid partitioning or subtractive clustering techniques [48]. In the learning process, the parameters associated with the membership functions change – this change is an optimisation essentially facilitated by a gradient vector [51]. Using a
combination of back-propagation and with the use of a least squares method [48], the fuzzy inference system is able to learn from the model data. A TS system is suited for modelling of non-linear systems by interpolating multiple linear models [20].

4. Experimental Results

In this section, we perform comparison experiments with respect to missing value imputation methods and classification methods for arrhythmia datasets with different proportions of missing values (e.g., 10%~70%). All the experiments are carried out with Intel(R) Core (TM)

i5-8250U CPU (1.6 GHz) and 8 GB RAM under the MATLAB 2015 a programming environment. The dataset comes from the UCI machine learning repository described in Section 2, and the different proportions of missing values are generated by computer at random based on the original UCI arrhythmia dataset.

4.1. Experimental Procedure.

The experimental flow on the UCI arrhythmia dataset is delineated in the flow block diagram (Figure 3), which comprises of the following four steps:

(1) The first step is to create different proportions of the missing value data sets by deleting different proportion values on the original UCI arrhythmia dataset at random .

(2) The second step essentially requires imputing these missing values using different methods for all created datasets.In our experiment, we apply four methods, i.e., Zero, Mean, PCA, RPCA and Random Forest imputation methods.

"Zero method" means to input the missing values with zero which is used for comparison; "Mean method" refers to replace each missing value with the average value of the corresponding attribute which is commonly applied in arrhythmia classification;

The Random forest, PCA and RPCA method are the inductive learning-based methods we introduced in Section 3.1.

(3) The third step involves classification of these datasets using different classifiers (KNN, ANN, SVM, ANFIS). Arrhythmia feature with reduced length is used for the training of the ANFIS for 10 to 300 epochs as shown in table 2. In our experiment, the results have been calculated on the basis of 10 trials of epoch on an average. The usage of Gbellmf MF for doing fuzzification. Due to which the Gbellmf MF provides the very realistic real world to fuzzy conversion is fulfilled from these experiments. The consideration of ten such MFs which give optimum results is performed. The root mean squared error (RMSE) is the measure with which calculation of the cost function parameter showed for training the ANFIS. The equation of Root Mean Squared Error (RMSE) shown below as follows:

$$RMSE = \sqrt{\frac{1}{U}\sum_{V=1}^{U} (A_V - F_V)^2}$$

where A_V and F_V are actual and fitted values respectively, and the number of training or testing sample is U. The RMSE is usually is the primary measure to make sure that the extent of learning performed by the ANFIS.

Table 2 Specifications of ANFIS

Input data size	**For arrhythmia - features of length 452**
SNR	0-6 Db
ANFIS type	ANFIS with five layers
ANFIS training method	Least-squares is included with the back propagation gradient descent method
Average training epochs	10-300
Total no. of membership functions	10

Type of membership function	Gbellmf, Gaussmf
RMSE with 6 Gbellmfs at 300 epochs	1.903×10^{-3}

(4) As a final step we compare the performance of these methods for arrhythmia classification and visualize the experiment result.

4.2. Classification Performance. In this section, we show the results of the experiment that were implemented using an accuracy indicator to examine the performance of missing value imputation methods and five classifiers for classifying cardiac arrhythmia. The experiments compare the performance on different proportions of missing value datasets generated from the UCI arrhythmia database. And the whole result is shown in Table 2.
Performance is measured on the basis of the accuracy of the classification of a model.. From Table 3, we can infer that out of four different classifiers the Adaptive Neuro fuzzy inference system (ANFIS) model gives very attractive classification results in terms of classification accuracy through the four missing value imputation methods of 77%, 78%, 79%, and 82%, respectively. To further research on the influence. of different missing value imputation methods on the accuracy of certain classifiers and observe the trends obviously, we take the ANFIS classifier as an example and visualize the last four rows of Table 2 in form of line chart (Figure 4).

Table 3 represents classification accuracy% achieved and comparison was done with respect to imputation methods and classifier.

Proportion %		10%	20%	30%	40%	50%	60%	70%
KNN	Mean	58	56	54	54	56	53	54
	Pca	57	56	55	55	54	53	54
	Rpca	57	57	55	56	55	54	53
	Random forest	59	57	56	57	54	54	55
ANN	Mean	61	59	55	54	52	53	52
	Pca	62	59	58	56	55	54	53
	Rpca	64	61	58	58	55	55	51
	Random forest	63	59	59	59	54	55	57
SVM	Mean	71	66	61	59	58	59	56
	Pca	70	69	65	60	61	57	55
	Rpca	71	68	65	61	61	58	55
	Random forest	71	67	66	59	62	59	57
ANFIS	Mean	77	76	73	72	69	70	67
	Pca	78	77	74	73	70	69	68

	Rpca	79	78	75	73	71	66	68
	Random forest	82	78	74	72	68	69	71

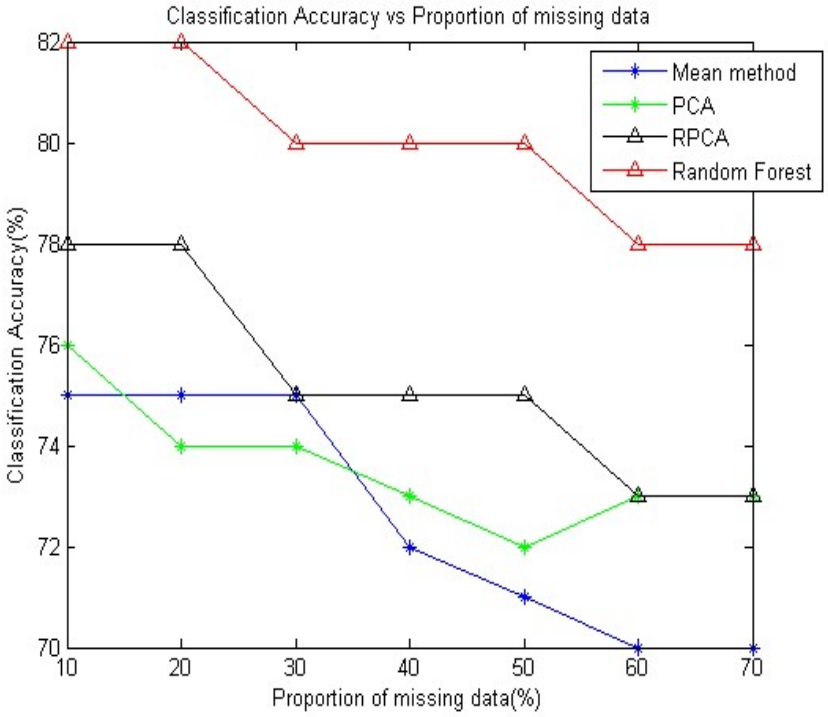

Figure 4 represents graphically the classification accuracy of ANFIS classifier with four other imputation methods.

We empirically tested 7 simulations based on the dataset with the percentage of missing values range from 10% to 70% using four methods to estimate the missing values, and the accuracy is obtained by the classification on the data after imputation.

From Figure 4, as we can see, the accuracy becomes lower and lower with the increase of the percentage of missing values for all the four broken lines.When the proportion of missing data ranges from 0% to 30%, the last three imputation methods are similar and perform better than the "Mean method". For 30% to 70% missing data range the "RF-based method" outperforms the other three; however, the "PCA-based method and RPCA based method" sharply gets worse, which indicates the lack of stability of the "PCA-based method". Throughout the whole picture, the "RF-based method" is more stable and accurate particularly for the data including a larger number of missing values.This indicates that the "RF-based method" can successfully handle missing values in an arrhythmia dataset irrespective of the number of values missing. And the "Mean method", a most commonly used imputation method in arrhythmia classification, is second only to it.

Based on the experimental results, we can conclude that for arrhythmia classification, when the small part (0% to 30%) of data is missing, we can apply any of the other three imputation

methods except the “Mean method”. On the other hand when the missing values are large (30% to 70%), we can use the “RF-based method” to replace the classical method, i.e., the “Mean method”. Figure 4 illustrates the mean classification accuracy of four classifiers, i.e., KNN, SVN, ANN, and ANFIS with four missing value estimation methods. Obviously, we can see that whatever method is chosen to fill the missing data, ANFIS outperforms the others in terms of classification accuracy.This indicates that the modification is effective for the uneven dataset like the UCI arrhythmia dataset in which the major class labels are “Normal”. Additionally, we can infer that when using our ANFIS classifier the mean accuracy is more than 70%, whereas the mean accuracy of the traditional KNN is lower than 60%.This implies that our proposed algorithm is reliable for the classification of different arrhythmia types so that the problem that has been stated in Section 1 can be solved.

While these methods seem to show good work, the computational time trade-off for the use of these methods (due to the need to cascade NNs) does not guarantee good performance or improvement. This is especially so in via of the relative computation time taken, as indicated in Table 3 given below. It is to be noted that the study to obtain this table was performed in MATLAB, using the tools specified in section IV. Thus, the programming is not standardised, and this result should be treated as a basic evaluation. That said, it is to be noted that RF is generally documented as being relatively fast machine learning tools.

Table 3 Computation time taken for the different methods for propagation through 245 cardiac instances with missing data points

Method	Training time (s)	Propagation Time (s)
Mean +KNN	32	41
PCA + KNN	33	43
RPCA +KNN	31	39
Random Forest + KNN	25	28
Mean +SVM	41	45
PCA + SVM	39	43
RPCA +SVM	36	39
Random Forest + SVM	35	37
Mean +ANN	21	26
PCA + ANN	23	27
RPCA +ANN	18	21
Random Forest + ANN	13	16
Mean +ANFIS	11	14
PCA + ANFIS	6	9
RPCA +ANFIS	3	11
Random Forest + ANFIS	0.039	0.49

5. Conclusion

Missing value is an acute problem that can deteriorate the quality of data, so missing value estimation is a significant pre-processing step for further experiments. Missing value imputation is a solution for the incomplete dataset problem.The quality of the observed data is critical considering that the imputation process requires a set of observed data, regardless of whether statistical or machine learning techniques are used to produce estimates to replace the missing values. In this paper, we delve on the problem from the feature selection perspective ,assuming that some of the collected features may be unrepresentative and affect the imputation results, resulting in degradation of the final performance of the classifiers.In this paper, we compare the main methods for estimating the missing values in electrocardiogram data like the "Mean method", "PCA-based method", "RPCA-based method" and RF-based method. In our comparative study, the "RF-based method" can successfully handle missing values in the arrhythmia dataset no matter how many values in it are missing.This indicates that when a large number of values in the dataset are missing ,higher classification accuracy can be expected in the practical application of the RF-based method. As for the imbalance data classification problem, we also presented an ANFIS classification algorithm, which is modified by a correction factor to handle the imbalance datasets problem to get better performance. In the future, we will further study the Five layer of the ANFIS model and improve the robustness of the method of selecting better parameters using Random forest method.

References

[1] S. M. Jadhav, S. L. Nalbalwar, and A. A. Ghatol, "Artificial neural network models based cardiac arrhythmia disease diagnosis from ECG signal data," International Journal of Computer Applications, vol. 44, no. 15, pp. 8–13, 2012.

[2] R. G. Kumar and Y. S. Kumaraswamy, "A neural network approach for cardiac arrhythmia classification," IUP Journal of Computer Sciences, vol. 7, no. 1, p. 62, 2013.

[3] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "Classification of ECG signals using machine learning techniques: a survey," in 2015 International Conference on Advances in Computer Engineering and Applications, pp. 714–721, Ghaziabad, India, March 2015.

[4] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," Current Bioinformatics, vol. 13, no. 4, pp. 352–359, 2018.

[5] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine," Frontiers in Bioengineering and Biotechnology, vol. 7, p. 224, 2019.

[6] L. Jiang, J. Zhang, P. Xuan, and Q. Zou, "BP neural network could help improve pre-miRNA identification in various species," BioMed Research International, vol. 2016, Article ID 9565689, 11 pages, 2016.

[7] X. Zeng, W. Wang, G. Deng, J. Bing, and Q. Zou, "Prediction of potential disease-associated microRNAs by using neural networks," Molecular Therapy-Nucleic Acids., vol. 16, pp. 566–575, 2019.

[8] Q. Kaiyang, L. Wei, and Q. Zou, “A review of DNA-binding proteins prediction methods,” Current Bioinformatics,
vol. 14, no. 3, pp. 246–254, 2019.
[9] J. S. Wang, C. W. Lin, and Y. T. C. Yang, “A k -nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition,” Neurocomputing, vol. 116, no. 10, pp. 136–143, 2013.
[10] S. A. Dudani, “The distance-weighted k-nearest-neighbor rule,” IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-6, no. 4, pp. 325–327, 1976.
[11] R.J.A. Little and D.B. Rubin, Statistical "Analysis with Missing Data," 2nd ed, United States of America: Wiley-Interscience, 2002, pp. 200-220.
[12] F.Z. Poleto, J.M. Singer and C.D. Paulino, "Missing data mechanisms and their implications on the analysis of categorical data," Statistics and Computing, vol. 21, no. 1, pp. 31-43 Jan. 2011
[13] N. V. Thakor and Y.-S. Zhu, “Applications of adaptive filtering to ECG analysis: noise cancellation and arrhythmia detection,” IEEE Transactions on Biomedical Engineering, vol. 38, no. 8, pp. 785–794, 1991.
[14] D. A. Coast, R. M. Stern, G. G. Cano, and S. A. Briller, “An approach to cardiac arrhythmia analysis using hidden Markov models,” IEEE Transactions on Biomedical Engineering, vol. 37, no. 9, pp. 826–836, 1990.
[15] J.-B. Pan, S.-C. Hu, H. Wang, Q. Zou, and Z.-L. Ji, “PaGeFinder: quantitative identification of spatiotemporal pattern
genes,” Bioinformatics, vol. 28, no. 11, pp. 1544-1545, 2012.
[16] L. Wei and Q. Zou, “Recent progress in machine learning based methods for protein fold recognition,” International Journal of Molecular Sciences, vol. 17, no. 12, p. 2118, 2016.
[17] Q. Zou, J. Zeng, L. Cao, and R. Ji, “A novel features ranking metric with application to scalable visual and bioinformatics data classification,” Neurocomputing, vol. 173, pp. 346–354, 2016.
[18] L. Chen, L. Lu, K. Feng et al., “Multiple classifier integration for the prediction of protein structural classes,” Journal of Computational Chemistry, vol. 30, no. 14, 2009.
[19] M. Diana and C. Deisy, “A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data,” Kuwait Journal of Science, vol. 42, no. 2, pp. 206–235, 2015.
[20] R. J. Little and D. B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, 2014.
[21] S. K. Pati and A. K. Das, “Missing value estimation for microarray data through cluster analysis,” Knowledge and Information Systems, vol. 52, no. 3, pp. 709–750, 2017.
[22] O. Troyanskaya, M. Cantor, G. Sherlock et al., “Missing value estimation methods for DNA microarrays,” Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.
[23] C. Meng, L. Wei, and Q. Zou, “SecProMTB: a SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data set,” Proteomics, vol. 19, article e1900007, 2019.

[24] P. Komal and C. Usha, “Relevance of machine learning techniques and various protein features in protein fold classification: a review,” Current Bioinformatics, vol. 14, no. 8, pp. 688–697, 2019.
[25] M. S. Islam, M. A. Hoque, M. S. Islam et al., “Mining gene expression profile with missing values: a integration of kernel PCA and robust singular values decomposition,” Current Bioinformatics, vol. 14, no. 1, pp. 78–89, 2019.
[26] B. Ran, H. Tan, J. Feng, W. Wang, Y. Cheng, and P. Jin, “Estimating missing traffic volume using low multilinear rank tensor completion,” Journal of Intelligent Transportation Systems, vol. 20, no. 2, pp. 152–161, 2015.
[27] H. Tan, Y. Wu, B. Cheng, W. Wang, and B. Ran, “Robust missing traffic flow imputation considering nonnegativity and road capacity,” Mathematical Problems in Engineering, vol. 2014, no. 1, pp. 12–25, 2014.
[28] M. Mitra and R. Samanta, “Cardiac arrhythmia classification using neural networks with selected features,” Procedia Technology, vol. 10, pp. 76–84, 2013.
[29] S. Khare, A. Bhandari, S. Singh, and A. Arora, “ECG arrhythmia classification using spearman rank correlation and support vector machine,” in Advances in Intelligent and Soft Computing, pp. 591–598, Springer, 2011.
[30] W. Zuo, W. Lu, K. Wang, and H. Zhang, “Diagnosis of cardiac arrhythmia using kernel difference weighted KNN classifier, computers in cardiology,” in Computers in Cardiology, pp. 253–256, IEEE, 2008.
[31] Z. Lin, M. Chen, and Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, Eprint, 2013, https://arxiv.org/abs/1009.5055.
[32] E. Candes and B. Recht, “Exact matrix completion via convex optimization,” Communications of the ACM, vol. 55, no. 6, pp. 111–119, 2012.
[33] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.
[34] C. Blake and C. J. Merz, UCI repository of machine learning databases, University of California. Department of Information and Computer Science, Irvine, CA, 1998, http://archive.ics.uci.edu/ml/datasets/Arrhythmia.
35. Tang, J.; Alelyani, S.; Liu, H. Feature selection for classification—A review. In Data Classification Algorithms and Applications; Aggarwal, C.C., Ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2014.
36. Li, Y.; Li, T.; Liu, H. Recent advances in feature selection and its applications. Knowl. Inf. Syst. **2017**, 53, 551–577.
37. De la lglesia, B. Evolutionary computation for feature selection in classification problems. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **2013**, 3, 381–407.
38. Xue, B.; Zhang, M.; Browne, W.N.; Yao, X. A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. **2016**, 20, 606–626.
39. Zhao, Z.; Liu, H. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 1151–1157.

40. Zhu, X.; Zhang, S.; Hu, R.; Zhu, Y.; Song, J. Local and global structure preservation for robust unsupervised spectral feature selection. IEEE Trans. Knowl. Data Eng. **2018**, 30, 517–529.
[41] G. Biau, L. Devroye, G. Lugosi. "Consistency of random forests and other averaging classifiers." Journal of Machine Learning Research, to appear, 2008.
[42] L. Masisi, F. V. Nelwamondo, T. Marwala. "The effect of structural diversity of an ensemble of classifiers on classification accuracy." IASTED International Conference on Modelling and Simulation (Africa-MS), 2008.
[43] Y. Qi, J. Klein-Seetharaman, Z. Bar-Joseph. "Random forest similarity for protein-protein interaction prediction from multiple sources," in Pacific Symposium on Biocomputing 10, pp. 531 - 542, 2005.
[44] T. K. Ho. "Random decision forests." ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition, Vol. 1, 1995.
[45] L. Breiman, A. Cutler. "Random forests." Department of Statistics, University of California Berkeley. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home. htm. Last accessed 12 August, 2008.
[46] J. R. Brence, D. E. Brown. "Improving the robust random forest regression algorithm." Systems and Information Engineering Technical Papers, Department of Systems and Information Engineering, University of Virginia, 2006. http://www.sys.virginia.edu/techreps/2006/sie06_0004.pdf. Last accessed 10 August, 2008.
[47] L. Breiman. "Random forests." Machine Learning, 45:5–32, Kluwer Academic Publishers, 2001.
[48] J-S. R. Jang, N. Gulley, "Fuzzy logic toolbox," The MathWorks Inc., 1997.
[49] A. Abraham, "Neuro-fuzzy systems: state-of-the-art modelling techniques," in Lecture Notes in Computer Science, Vol. 2084, pp. 269 - 276, Springer Verlag Germany, 2001.
[50] J-S. R. Jang, "Neurofuzzy modelling and control." Proceedings of the IEEE, Vol. 83, Issue 3, 1995.
[51] J-S. R. Jang, "Input selection for ANFIS learning." Proceedings of IEEE International Conference on Fuzzy Systems, 1998.