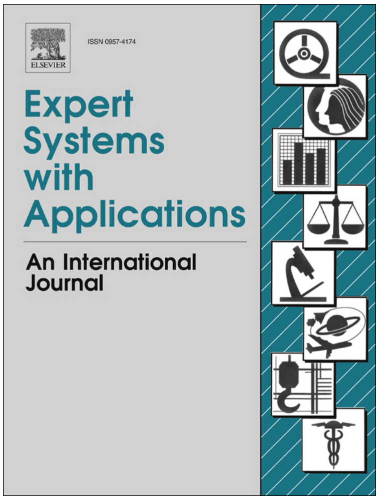

Hybrid Prediction Model with missing value Imputation for medical data

Archana Purwar, Sandeep Kumar Singh

PII: S0957-4174(15)00157-8
DOI: http://dx.doi.org/10.1016/j.eswa.2015.02.050
Reference: ESWA 9900

To appear in: *Expert Systems with Applications*

Please cite this article as: Purwar, A., Singh, S.K., Hybrid Prediction Model with missing value Imputation for medical data, *Expert Systems with Applications* (2015), doi: http://dx.doi.org/10.1016/j.eswa.2015.02.050

Hybrid Prediction Model with missing value Imputation for medical data

Archana Purwar[1] and Sandeep Kumar Singh[2]
[1,2]Department of Computer Science and Information Technology, JIIT Noida, India
[1]archana.purwar@jiit.ac.in
[2] sandeepk.singh@jiit.ac.in

Abstract Accurate prediction in the presence of large number of missing values in the data set has always been a challenging problem. Most of hybrid models to address this challenge have either deleted the missing instances from the data set (popularly known as case deletion) or have used some default way to fill the missing values. This paper, presents a novel Hybrid Prediction Model with missing value imputation (HPM-MI) that analyze various imputation techniques using Simple K-means clustering and apply the best one to a data set. The proposed hybrid model is the first one to use combination of K-means clustering with Multilayer Perceptron. K-means Clustering is also used to validate class labels of given data (incorrectly classified instances are deleted i.e. pattern extracted from original data) before applying classifier. The proposed system has significantly improved data quality by use of best imputation technique after quantitative analysis of eleven imputation approaches. The efficiency of proposed model as predictive classification system is investigated on three benchmark medical data sets namely Pima Indians Diabetes, Wisconsin Breast Cancer, and Hepatitis from the UCI Repository of Machine Learning. In addition to accuracy, sensitivity, specificity; kappa statistics and the area under ROC are also computed. The experimental results show HPM-MI has produced accuracy, sensitivity, specificity, kappa and ROC as 99.82%, 100%, 99.7%, 0.996 and 1.0 respectively for Pima Indian diabetes data set, 99.39%, 99.31%, 99.54%, 0.986, and 1.0 respectively for Breast Cancer data set and 99.98 %, 100%, 96.55%, 0.978 and 0.99 respectively for Hepatitis data set. Results are best in comparison with existing methods. Further, the performance of our model is observed as function of missing rate and train-test ratio using 2D synthetic data set and Wisconsin Diagnostics Breast Cancer data sets. Results are promising and therefore the proposed model will be very useful in prediction for medical domain especially when numbers of missing value are large in the data set.

Keyword *Missing Value Imputation, Multilayer Perceptron (MLP), K-Means Clustering, Data mining.*

1. Introduction

Research in the field of predictive data mining for medical applications is a significant and moving area. Generally, a medical practitioner collects his/her knowledge from patient's symptoms and confirmed diagnosis. Diagnosis is usually made either by evaluating the current test results of the patients or by referring to the previous decisions made on other patients with same test results. The accuracy of diagnosis of patient's disease like diabetes, breast cancer and others is greatly relenting on a experts' experience (Meesad & Yen, 2003). Due to the pace at which numbers of patients are increasing, it has become cumbersome to make diagnostics decisions. On the other side, development of new computational methods and tools makes relatively easy to make decisions from dedicated databases of electronic patient records. As an example, numerous classifiers have been developed for diagnosis and screening of diabetes, cancer and liver disorders (Seera & Lim, 2014). A number of classification systems have been developed in the literature like RBQ, LVQ C4.5, CART, Bayesian Tree, ANN + FNN, HPM, Sim + F2, real coded GA, and FMM-CART-RF to support diagnosis of diabetes as well as breast cancer disease (Seera & Lim, 2014). Various hybrid models (Kahramanli & Allahverdi, 2008; llango & Ramaraj, 2010; Patil, Joshi, & Toshniwal, 2010) namely hybrid system consisting of artificial neural network and fuzzy neural network, HPM consisting of K-means clustering with J48 and HPM with F-score consisting of feature selection using F-score, K-means clustering and SVM are also proposed in the literature.

Researchers have developed a large number of classification systems to improve accuracy of predictive classification. The proposed model uses K-means clustering as a means to analyze 11 MVI techniques under study and selects the best imputation method. This best imputation method is applied on the data set before pattern extraction and subsequently applying prediction. Moreover, to the best of our knowledge, none of the hybrid prediction models have combined use of K-means clustering and MLP for prediction.

This paper makes a novel contribution by first analyzing 11 missing data imputation techniques experimentally and finds the best method for handling the missing values in the data set using K-means clustering. Consequently, it improves the quality of data. Moreover, it also aims at predictive classification using novel model that can classify records in the test data set using training data set. Multi layer Perceptron (MLP) has a capability to learn from examples and can generalize beyond the training data (Carpenter & Markuzon, 1998; Downs, Harrison, Kennedy, & Cross, 1996; Mukhopadhyay, Changhong, Huang, Mulong, & Palakal, 2002). Due to these characteristics of neural networks, MLP with backpropagation is investigated for developing a valid and useful prediction model. K-means clustering is also used to develop proposed Hybrid Prediction Model with Missing Value Imputation (HPM-MI) to extract the patterns from data before applying MLP for classification.

Rest of the paper is grouped in five sections. Section-2 describes the study of data mining methods namely imputation methods, K-means clustering, MLP and review of prediction models. Then, section-3 depicts the proposed model. Evaluation of proposed model is done in section-4. Section-5 shows the results and its discussion. Finally, paper is concluded by section-6.

2. Background Study

This section reviews a few data mining methods and predictive classification models.

2.1 Data Mining Methods

As the amount of data stored in medical databases is increasing, there is growing need for efficient and effective techniques to extract the information. Previous researches have given evidence that medical diagnosis and prognosis is amended by employing data mining techniques on clinical data (Hammer & Bonates, 2006; Saastamoinen & Ketola, 2006; Tsirogiannis, et al., 2004). This has been possible due to extensive availability of data mining techniques and tools for data analysis. Predictive modeling requires that the medical informatics researchers and practitioners to select the most appropriate strategy to cope with clinical prediction problem (Bellazzi & Zupan, 2008). This section discusses mining techniques used to develop the proposed model.

2.1.1 Missing Value Imputation

2.1.1.1 Introduction

In real-life databases, incomplete data or information as shown in Table 1 is frequent owing to the presence of missing values in the attributes. First row in Table 1 shows name of the variables in the data set while other rows show the values of these variables. The values denoted by '?' in Table 1 represent the missing values. Missing values can occur due to large number of reasons such as errors in the manual data entry procedures, equipment errors or incorrect measurements. The presence of missing values (MVs) in data mining produces several problems in the knowledge extraction process such as loss of efficiency, complications in managing and analyzing data. It may also result in bias decisions due to differences between missing and complete data.

Table 1
Sample data showing missing values by '?'.

A	B	C	D	E	F	G	H	class
6	148	72	35	?	33.6	0.62	50	yes
1	85	66	29	?	26.6	0.35	31	no
8	183	64	?	?	23.3	0.67	32	yes
1	89	66	23	94	28.1	0.16	21	no
?	137	40	35	168	43.1	2.28	33	yes

In order to solve these problems, two approaches are found in the literature. First approach consists of missing data toleration techniques which integrate the techniques of missing values handling in specific data mining algorithms such as in classification (David, 2007; Saar-Tsechansky M, 2007), clustering (Hathaway & Bezdek, 2002) and feature selection (Aussem & de Morais, 2008). Second type of approach consists of missing data imputation techniques which fill in missing values before using complete-data methods. One advantage of imputation is that the treatment of missing data is independent of the succeeding mining algorithm, and people can select a suitable learning algorithm after imputation (Qin, Zhang, Zhu, Zhang, & Zhang, 2007).

In our proposed HPM-MI, we have used best proven MVI approach to fill the missing value before pattern extraction and classification on the data set. We have validated our model on three benchmark data sets i.e. Pima Indian diabetes Wisconsin breast cancer and Hepatitis data set from UCI repository (Newman, Hettich, Blake, & Merz, 2007) having 763, 16 and 167 missing values respectively and two complete data sets namely 2D synthetic data set as well as Wisconsin Diagnostics Breast cancer (WDBC) in which missing values were artificially induced.

2.1.1.2 Imputation Techniques

In order to analyse the impact of various MVI techniques (Luengo, García, & Herrera, 2011; Koren, Bell, & Volinsky,2009; Takács,, Pilászy, & Németh, 2008), experimentation is done on the following techniques to choose the best possible one to handle missing values present in the data sets under study.The following approaches have been empirically assessed to find the missing values:

- *Case Deletion:* The examples that have any missing value in their attributes are removed from the data set.
- *Most Common Method (MC):* Missing values present in the data set is substituted by mean value for numerical and mode for nominal attributes.
- *Concept Most Common (CMC):* This method calculates the missing values similar to MC method but it considers only the same class in which MV is missing.
- *K-Nearest Neighbor (KNNI):* Firstly, this method finds the *k* nearest neighbors and then, the most common value among all neighbors is taken for nominal attributes, and the mean value is used for numerical attributes.
- *Weighted Imputation with K-Nearest Neighbor (WKNN):* This method calculates the distance of each missing value instances from its neighbors. This distance is used to calculate the weight. MV is computed by weighted mean for numerical attributes. For nominal attributes, imputed value is the category with highest weight.
- *K-means Clustering Imputation (KMI):* All the instances are clustered using K-means clustering. The instances in each cluster are considered nearest neighbors of each other. The missing value is computed in similar manner as in KNNI method.
- *Imputation with Fuzzy K-means Clustering (FKMI)*: After the fuzzy clustering of the data set, missing values are computed as weighed sum of all centroids, using the membership function of each cluster as the weight.
- *Support Vector Machines Imputation (SVMI)*: SVM model is trained to predict missing attributes from the complete instances which do not have missing values. During testing, missing values are predicted using other attributes by setting missing attribute as a class attribute whose value is intended to be predicted.
- *Singular Value Decomposition Imputation (SVDI)*: In this method, singular value decomposition is used to obtain a set of mutually orthogonal expression patterns that can be linearly combined to approximate the values of all attributes in the data set. In order to do that, first SVDI estimates the MVs within the expected maximization algorithm, and then it computes the singular value decomposition and obtains the eigen values. Now, SVDI can use the eigen values to apply a regression to the complete attributes of the instance and to obtain an estimation of the MV itself.
- *Local Least Squares Imputation (**LLSI):** This method* identifies k similar instances and fits a least squares model between the instances and the known part of the record with missing values
- *Matrix Factorization:* This method aims to get decompositions whose product approximate the values of all attribute in

the data set. Then a missing value for a given feature can be computed by the dot product of these vectors that corresponds to a given instance and feature.

2.1.2 K-means Clustering

Clustering methods partition data set into groups so that the instances in one group are similar to each other, and as dissimilar as possible from the objects in other groups. K-means clustering is a partitioning algorithm that groups the n instances into k clusters (user defined input parameter) so that intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the instances in a cluster. K-means clustering consist of following steps (Kanungo, et al., 2002).

Let X = {x_1,x_2,x_3,……..,x_n} be the set of n instances.

1) Randomly select '*k*' cluster centers as {v_1, v_2,…….,v_k}.

2) Calculate the distance between each instance and cluster centers.

3) Assign instance to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center v_i [$1 \le i \le k$] for i^{th} cluster using:

$$v_i = \frac{1}{c_i}\sum_{j=1}^{c_i} x_j \qquad (1)$$

where, 'c_i' represents the number of data points in i^{th} cluster and x_j is the data point in i^{th} cluster [$1 \le j \le c_i$]

5) Recalculate the distance between each instance and new obtained cluster centers.

6) If no instance was reassigned then stop, otherwise repeat from step 3.

In order to identify the correct patterns, clustering is applied from the given data set before classification.

2.1.3 Multilayer Perceptron (MLP) with backpropogation

MLP (Rumelhart, 1986) has been used earlier for prediction of type-2 diabetes which is known to be strong function approximation for prediction problems. MLP is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable. It utilizes a supervised learning technique called backpropagation for training the network because it learns iteratively by processing data set of training examples, comparing the network's prediction for each example with the actual known targets value known as class labels. For each training example, the weights are modified so as to minimize the mean squared error between the network prediction and the actual target value. These modifications are made in the backward direction.

MLP has been used as a classifier in the proposed HPM-MI model for predictive classification due to its prominent characteristics such as high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained (Han & Kamber, 2006).

2.2 Review of Predictive classification Models

In present years, predictive classification techniques have been applied in medical diagnosis successfully. Over the last few years several researchers have shown the use of predictive data mining to infer clinically relevant models from patient data and to provide decision support in the medical field. Various prediction models have been developed to support various medical decision making tasks such as prediction of breast cancer, diabetes, liver and hepatitis disease (Luukka, 2011; Seera & Lim, 2014; Polat & G¨unes, 2006).

Michie et. al. (Michie, 1994) have applied 22 divergent algorithms to classify the diabetic patients and accuracy of results varies from 67.6% to 77.7%. Adaptive Resonance Theory Map–instance counting (ARTMAP-IC) produced 81% accuracy by testing 576 samples which were used for training dataset, and 192 samples which were used as testing dataset (Carpenter & Markuzon, 1998). Bioach et al. removed the tuples where the attributes Plasma-Glucose level and Body mass index of patients were recorded as having zero value in their study (Bioch, 1996). They got an accuracy of 75.4% and 79.5% using neural network and Bayesian approach respectively. Further, hybrid prediction models (HPM) (Kahramanli & Allahverdi, 2008; llango & Ramaraj, 2010; Patil, et al., 2010) by Patil et. al., Kahramanli and Illango et al gave an accuracy of 84.5%, 92.38% and 98.84%. Kahramanli developed hybrid system consisting of neural network and fuzzy neural network. The model proposed by Patil et al. (Patil, et al., 2010) used the combination of K-means clustering with decision tree classifier. Illango had further increased the accuracy by using F-Score feature selection method, K-means clustering and support vector machine (SVM). Recently Seera et. Al. proposed a hybrid intelligent system (Seera & Lim, 2014) that consists of Fuzzy Min max neural network, the classification & regression tree, and Random Forest and compared their model with Lukka (Luukka, 2011) and Orkcu (Örkcü & Bal, 2011). Table 2 shows that Serra et al got the accuracy rate of 78.39 % and Lukka and Orkcu got accuracy of 75.97 % and 77.60 % in their models respectively.

Table 2
The values of accuracy of classification made on Pima Indian diabetes illness data (Seera & Lim, 2014.

Method	**Accuracy (%)**
Sim	75.29
Sim + F1	75.84
Sim + F2	75.97
Binary-coded GA	74.80
BP	73.80
Real-coded GA	77.60
FMM	69.28
FMM-CART	71.35
FMM-CART-RF	78.39

A lot of classification techniques have also been propounded for the diagnosis of Wisconsin Breast Cancer Data Set. Among these, C4.5 by Quinlan (Quinlan, 1996), LDA by dobinkar (Ster & Dobinkar, 1996) and SVM by (Bennett & Blue, 1998) were proposed with accuracy of 94.74%, 96.8% and 97.2% respectively in the late nineties. Other developed models were NEFCLASS, Fuzzy GA, LVQ, ANFIS, and PSO-SVM (Chen, & et al, 2012). The most recent study has been done by Seera et. al. They have proposed FMM-CART-RF model and compared their model with other methods (Seera & Lim, 2014). Results obtained from their study are listed in Table 3. Accuracy rate shows that none of them have achieved accuracy higher than 98.84 % for breast cancer data set.

Table 3
The values of accuracy of classification made on Breast Cancer data (Seera & Lim, 2014).

Method	**Accuracy (%)**
BC FRPCA1	98.19
BC Original	97.49
BC PCA	97.72
Sim	97.49
Sim + F1	97.10
Sim + F2	97.18
Binary-coded GA	94.00
BP	93.10
Real-coded GA	96.50
AIRS	97.20
Fuzzy-AIRS	98.51
Cooperative coevolution	96.69
Decompositìonal	95.93
Pedagogical	97.07
SVMs	96.50
FMM	95.26
FMM-CART	94.86
FMM-CART-RF	98.84

Many hybrid methods have also been proposed to deal with the automated diagnosis of hepatitis disease problem. Polat and G¨unes (Polat & G¨unes, 2006) proposed a new diagnostic method based on a hybrid feature selection (FS)method and artificial immune recognition system (AIRS) using fuzzy resource allocation mechanism. The obtained classification accuracy of the proposed system was 92.59%. In Polat and Gunes (2007a, 2007b), an artificial immune recognition system (AIRS) based on principal component analysis (PCA) was used for classification, the reported accuracy was up to 94.12%. In Dogantekin et al. (2009), an adaptive network based on fuzzy inference system combining with linear discriminant analysis (LDA-ANFIS) was applied for automatic hepatitis diagnosis, and an accuracy of 94.16% was obtained. Chen et. al. (2011) developed a hybrid model using local fisher discriminant analysis (LFDA) and support vector machine and achieved an accuracy of 96.77%. Recently, Zangooei et. al. have developed Support Vector Regression(SVR) based classification model (Zangooei, Habibi, & Alizadehsani, 2014) where its parameter values were optimized by Non-dominated Sorting Genetic Algorithm-II (NSGA-II) and compared with other models Results obtained from their study are listed in Table 4. Accuracy rate shows that none of them have achieved accuracy higher than 98.52% for hepatitis data set.

Table 4
The values of accuracy of classification made on hepatitis data (Zangooei , Habibi, & lizadehsani, 2014).

Method	**Accuracy (%)**
C4.5	83.60
BNND	90.00
Weighted9NN	92.90
FSM without rotations	88.40
LDA	86.40
FS-AIRS with fuzzy res	92.50
FS-fuzzy-AIRS	94.10
PCA–LSSVM	95.00
GA-SVM	89.60
LDA-ANFIS	94.10
J48	85.67
MLP	90.52
SVR NSGA-II	98.52

Although all these work have demonstrated promising prediction accurately, majority of papers have concentrated on how to increase accuracy of their models .Presence of missing values that affects the data quality and accuracy of mining results have not been addressed properly in their work.Patil, et. al. (2010) used case deletion and Seera & Lim (2014) have taken the default method to handle the missing values in the data set. Research shows that other imputation techniques perform better as compared to case deletion and default methods, if the missing values are large in number. Hence this paper proposes a novel HPM-MI with the goal of improvisation data quality as well as accuracy. Further, it compares results of proposed model with other results reviewed in Table 2, Table 3 and Table 4.

3. Proposed System

In this paper, a Hybrid Prediction Model with Missing value Imputation (HPM-MI) is developed which is shown in Fig. 1 to deal with predictive classification problem of medical patients. This model comprises of three stages namely analysis and selection of Imputation method using K-means

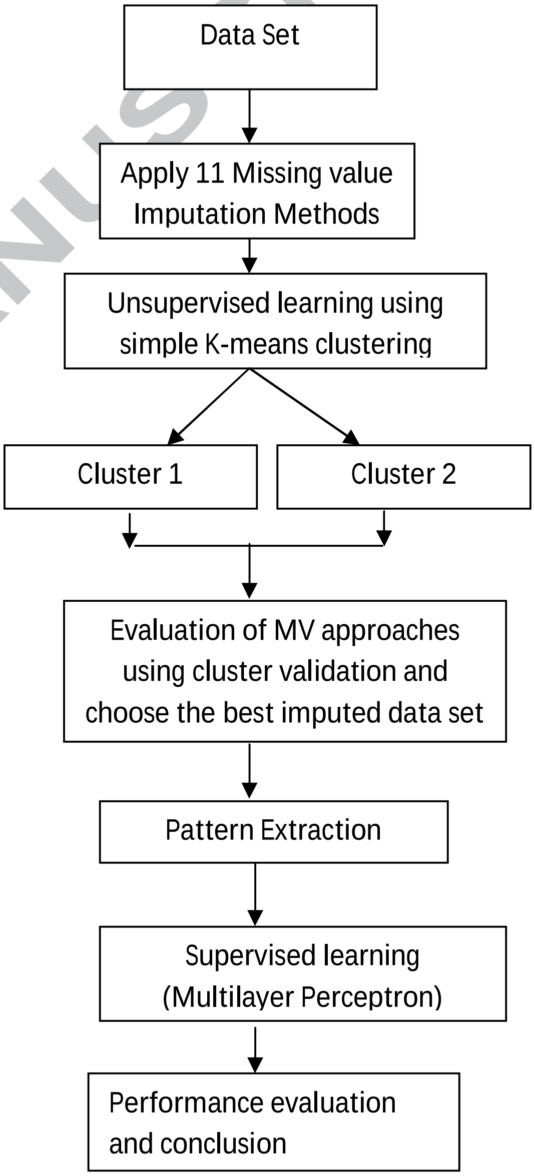
Data Set
Apply 11 Missing value Imputation Methods
Unsupervised learning using simple K-means clustering
Cluster 1
Cluster 2
Evaluation of MV approaches using cluster validation and choose the best imputed data set
Pattern Extraction
Supervised learning (Multilayer Perceptron)
Performance evaluation and conclusion

Fig. 1 Hybrid Prediction Model with Missing Value Imputation

clustering, pattern extraction and Multilayer Perceptron with back propagation as training algorithm. Fig 2 shows procedure of the proposed model and following subsections describe the detailed study.

3.1 Missing Value Imputation

In this paper, we have analyzed 11 approaches to fill missing values in incomplete data set in order to improve the quality of data. Detailed discussion of these approaches is already provided in section 2.1.1

1. For a given data set D (having missing values) under study.
2. Obtain the imputed data sets i.e. $D_{1,}$ $D_{2,}$ $D_{3.....}$, and D_{11} after employing 11 missing values imputation approaches on data set D.
3. Use simple K-means clustering on $D_{1,}$ $D_{2,}$ $D_{3.....}$, and D_{11} obtained from step 2.
4. Validate the clustering results obtained from 11 experiments using their actual classes.
5. Choose imputed data set that has produced least number of incorrectly classified instances and discard other 10 imputed data sets.
6. Extract the instances from chosen imputed data set using correctly classified instances as a result of K-means clustering.
7. Classify the extracted data using Multilayer Perceptron Model.
8. Evaluate the Performance of above model using accuracy, specificity, sensitivity, kappa and ROC.
9. Compare this model with previous models.

Fig. 2 Steps for developing HPM-MI model

MVI is a crucial step incorporated in our proposed (HPM-MI) model .The objective of this step is to find the best imputation technique with respect to incomplete data set under study. As no generalization can be made regarding best imputation method. So we need to analyze empirically all of 11 MVI approaches under study for a given data set. Selection of best MVI method for incomplete data set is based on accuracy achieved (ground truth) after applying mining task. For our proposed model, we have chosen clustering as basis for selection of best imputation technique. Imputation technique which has produced more compact clusters is chosen to impute missing values in the data set. This imputed data set is used for pattern extraction while others are discarded.

Simple K-means clustering algorithm is applied on each imputed data set. Clustering results produced by each imputed data set are validated through their actual classes which are known. As a result, performance of imputation techniques applied on data sets under study is measured using incorrectly classified instances produced by each imputed data set shown in Table 5, Table 6 and Table 7 for diabetes, breast cancer, and hepatitis data sets respectively.

Concept Most Common (CMC) method has given minimum number of incorrectly classified instances in case of diabetes data as well as hepatitis data set. Case Deletion has performed best in case of Wisconsin Breast Cancer data. Therefore, CMC has been chosen to handle the missing values for diabetes data set as well as hepatitis data set and case deletion is selected for breast Cancer data set.

Table 5.
Clustering results with 11 imputation methods for Pima Indian diabetes data.

S.No	Imputation Method	Incorrectly classified instances (%)
1	**CMC**	**25.78**

2	FKMI	31.77
3	KMI	26.56
4	KNNI	27.99
5	LSSI	31.64
6	MC	32.42
7	SVDI	44.79
8	SVMI	31.51
9	WKNN	27.73
10	Case Deletion	30.72
11	Matrix Factorization	27.60

Table 6
Clustering results with 11 imputation methods for Wisconsin Breast Cancer Data.

S.No	Imputation Method	Incorrectly classified instances (%)
1	CMC	4.00
2	FKMI	4.00
3	KMI	4.00
4	KNNI	4.00
5	LSSI	4.00
6	MC	4.15
7	SVDI	4.29
8	SVMI	4.00
9	WKNN	4.00
10	**Case Deletion**	**3.95**
11	Matrix Factorization	4.15

Table 7
Clustering results with 11 imputation methods for hepatitis data.

S.No	Imputation Method	Incorrectly classified instances (%)
1	**CMC**	**29.68**
2	FKMI	30.32
3	KMI	30.32
4	KNNI	29.68
5	LSSI	29.68
6	MC	49.03
7	SVDI	30.97
8	SVMI	29.68
9	WKNN	29.68
10	Case Deletion	37.50
11	Matrix Factorization	33.55

3.2 Pattern Extraction

We have obtained wrongly classified instances, as shown in Table 8 for the data sets under study after cluster validation done in previous step. These instances were eliminated from best imputed data set chosen for diabetes, breast Cancer, and hepatitis data sets before applying classification.

3.3 Prediction using MLP (Multilayer Perceptron)

It has been proved that MLP with single hidden layer is able to accurately approximate continuous functions (Cibenko, 1989). MLP network architecture built by Weka as shown in Fig. 3 is used in my proposed model as a classifier. It consists of an input layer, one hidden layer, and an output layer. Each layer is made up of neurons. The number of neurons in input layers is same as the number of attributes or features present in each training example. The attributes are fed simultaneously into the neurons making up the input layer. The output of input layer is calculated using sigmoidal function and in turn, fed simultaneously to a second layer known as hidden layer. The outputs of the hidden layer are the input to output layer. This output layer shows the network's prediction for given instances of data set as "yes" and "no" for binary classification problem MLP uses backpropagation algorithm to train the network.

Table 8
Clustered Instances

Pima diabetes data set			Breast Cancer data set			Hepatitis data set		
Cluster attribute (clusters)	**Samples**	**Incorrectly classified**	**Cluster attribute (clusters)**	**Samples**	**Incorrectly classified**	**Cluster attribute (clusters)**	**Samples**	**Incorrectly classified**
Cluster 1 (yes) Cluster 0 (no)	768	198	Cluster 1 (benign) Cluster 0 (malignant)	683	27	Cluster 1 (die) Cluster 0 (live)	155	46

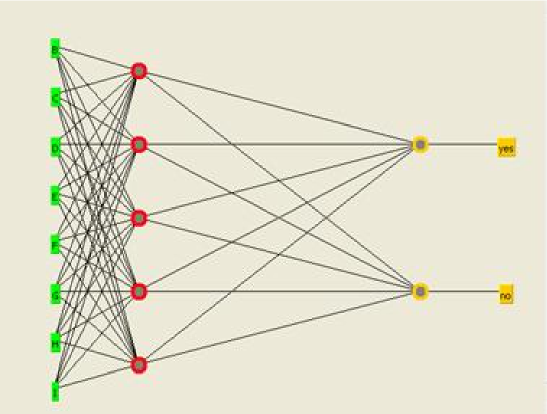

Fig. 3 MLP Architecture built in Weka

4. Evaluation of Proposed System

4.1 Experiments

In this section, proposed model is validated on three publicly available data sets namely Pima Indians Diabetes, Wisconsin Breast Cancer Data Set and Hepatitis data set from the UCI machine learning data repository (Newman, Hettich, Blake, & Merz, 2007).

We have taken two more data sets which do not have missing values namely 2D synthetic data set and Wisconsin Diagnostic Breast Cancer Data Set (WDBC) from the UCI machine learning data repository (Newman, Hettich, Blake, & Merz, 2007). We have artificially induced missing values in these data sets to observe the robustness of our proposed model as a function of missing rate.

4.1.1 Pima Indian Diabetes data Set

This data set includes a total of 768 instances depicted by 8 attributes and a predictive class. Out of 768 instances , 268 instances belong to class ‘1’ which indicate that diabetic cases and 500 instances belong to class ‘0’ means non diabetic cases i.e. they are healthy persons. Most of the cases contain missing values. Number of missing values corresponding to each attribute in the data set is shown in Table 9.

4.1.2 Wisconsin Breast Cancer Data Set

The data set contains total 699 instances described by 9 attributes and a predictive class. The attribute values for all 9 attributes lie from 1 to 10. The class attribute has only two categories namely benign and malignant. Class ‘malignant’ has 241 instances and class ‘benign’ consist of 458 instances. This data set consists of missing values .The number of missing values with their attribute name is mentioned in Table 9.

Table 9
Numbers of missing values in each variable of data sets.

Pima diabetes data set (763 missing values in 768 instances)		**Breast Cancer Data Set (16 missing values in 699 instances)**	
Variable	**Number of missing values**	**Variable**	**Number of missing values**
Pregnant	111	Clump thickness	Nil*
Plasma glucose	5	Uniformity of cell size	Nil*
Diastolic BP	35	Uniformity of cell shape	Nil*
TricepsSFT	227	Marginal adhesion	Nil*
Serum-Insulin	374	Single epithetical size	Nil*
BMI	11	Bare nuclei	16
DPF	Nil*	Bland chromatin	Nil*
Age	Nil*	Normal nuclei	Nil*
Class	Nil*	Mitoses	Nil*
		Class	Nil*

Nil* indicates no missing value with respect to a given feature.

4.1.3 Hepatitis Data Set

The data set contains total 155 instances described by 19 attributes and a predictive class. The class attribute has only two categories namely live and die. Class 'live' has 123 instances and class 'die' consist of 32 instances. Many instances in the data set consist of missing values .The number of missing values with their attribute name is mentioned in Table 10.

Table10
Numbers of missing values in each variable of the Hepatitis data set (167 missing values in 155 instances).

Variable	**Number of missing values**
Age	Nil*
Sex	Nil*
Steroid	1
Antivirals	Nil*
Fatigue	1
Malaise	1
Anorexia	1
Liver Big	10
Liver Firm	11
Spleen Palpable	5
Spiders	5
Ascites	5
Varices	5
Bilirubin	6
Alk Phosphate	29
Sgot	4
Albumin	16
Protime	67
Histology	Nil*
Class	Nil*

Nil* indicates no missing value with respect to a given feature.

4.1.4 2D Synthetic Data Set

We created 2D data set shown in Fig. 4. It is described by three attributes namely x, y and a class. Both the attributes x and y have numerical values and a class attribute has two categories namely "yes" and "no" Out of 200 instances, 97 instances belong to class "yes" and 103 instances belong to class "no. This data set does not have any missing values. We have artificially created missing values to observe the performance of our model at four missing rate (the amount of data missing) such as 20%,40 %, 60% and 80 %.

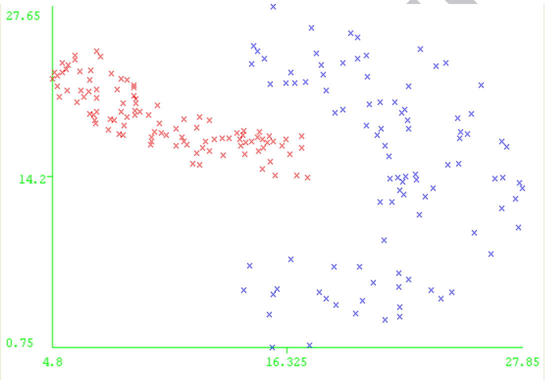

Fig. 4. 2D data set

4.1.5 Wisconsin Diagnostic Breast Cancer Data Set

Wisconsin Diagnostic Breast Cancer Data Set (WDBC) contains total 569 instances described by 30 attributes and a predictive class. The class attribute has only two categories namely B (benign) and M (malignant). Class 'B' has 357 instances. This data set does not have any missing values. We have artificially created missing values to observe the performance of our model as a function of missing rate (the amount of data missing) such as 25%, 50% and 75 %.

4.1.6 Experimental Framework

MV techniques except matrix factorization (implemented in Python) under study are implemented using Knowledge extraction based on evolutionary learning (KEEL) (Alcalá-Fdez, et al., 2009). Then we have applied simple K-mean clustering using Waikato Environment for Knowledge Analysis (Weka) toolkit (Witten & Frank, 2000) on 11 imputed data sets. Clusters produced by K-means are evaluated using their classes (

ground truth) for 11 experiments .CMC and case deletion are chosen as the best imputation approaches for diabetes dataset as well as hepatitis data sets and case deletion for breast cancer dataset respectively. Correctly classified instances were extracted from the respective data sets under study. These resulting data sets were classified using MLP classifier. Results were collected and evaluated using performance metrics discussed in section 4.2.

4.2 Performance metrics

This section describes measures that were used to evaluate the performance of classifiers.

(a) Accuracy, sensitivity and specificity. Accuracy is popular metric that refers to the ability of the model to correctly predict the class label of new or unseen data (Han & Kamber, 2006). In addition to this, sensitivity and specificity are also used to assess how well classifier can recognize true examples as well as false examples. These measures are calculated as.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

$$sensitivity = \frac{TP}{TP+FP} \qquad (3)$$

$$specificity = \frac{TN}{TN+FP} \qquad (4)$$

where,

True positives (TP) = no. of correct classifications predicted as yes (or positive).

True negatives (TN) = no. of correct classifications predicted as no (or negative)

False positive (FP) = no. of examples that are incorrectly predicted as yes (or positive) when it is actually no (negative).

False negative (FN) = no. of examples that are incorrectly predicted as no when it is actually yes.

(b) Kappa statistics. This measure evaluates the pair wise agreement between two different observers, corrected for an expected chance agreement (Thora, Ebba, Helgi, & Sven, 2008). Kappa value 0 indicates chance agreement and 1 shows prefect agreement between classifier and the ground truth (true classes). Kappa value using proposed model is 0.996, 0.986 and 0.978 for Pima Indians Diabetes, Wisconsin Breast Cancer and Hepatitis data set respectively. It is calculated as

$$k = \frac{[p(a)-p(e)]}{[1-p(e)]} \qquad (5)$$

where, p(a) and p(e) can been found from Eq.(6) and Eq.(7) respectively.

$$p(a) = \frac{TP+TN}{N} \qquad (6)$$

$$p(e) = \frac{[(TP+FN)*(TP+FP)*(TN+FN)]}{N^2} \qquad (7)$$

where N is the total number of instances used.

(c) Area under ROC Curve. ROC curve is the graph between True positive rate (TPR) and False positive rate (FPR). Accuracy of a classifier can also be measured by the area under the ROC curve (AUC). Most of the classifiers have AUC between 0.5 and 1. AUC is 1 for perfect classifiers. An area under the ROC curve of 0.8, for example, means that a randomly selected case from the group with the target equals 1 has a score larger than that for a randomly chosen case from the group with the target equals 0 in 80% of the time. When a classifier cannot distinguish between the two groups, the area will be equal to 0.5.

(d) Confusion matrix. A confusion matrix is calculated for the classifier to interpret the results as shown in Table 11, Table 12 and Table 13. Table11 shows 187 and 382 correctly classified instances in class ‘yes’ and class ‘no ’respectively for diabetes data set. Similarly, 432 are correctly classified instances in class “benign” and 220 are correct instances in class “malignant” for breast cancer data se, as shown in Table 12. Table 13 shows 80, correctly classified instances in class “live” and 28, correct instances in class “die” for Hepatitis data set. The diagonal elements of the matrix show the incorrect classification made by classifier.

Table 11
Confusion matrix. (570 instances) after pattern extraction for Pima diabetes data set.

a	b	Classified as
187(True Positive)	0(False Negative)	a=yes
1(False Positive)	382(True Negative)	b=no

Table 12
Confusion matrix. (656 instances) after pattern extraction for Wisconsin Breast cancer data set.

a	b	Classified as
432(True Positive)	3(False Negative)	a=benign
1(False Positive)	220(True Negative)	b=malignant

Table 13
Confusion matrix. (109 instances) after pattern extraction for Hepatitis data set.

a	b	Classified as
80(True Positive)	0(False Negative)	a=live
1(False Positive)	28(True Negative)	b=die

4.3 k-fold cross-validation

Cross-Validation (Delen, Walker & Kadam, 2005) is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments. One is used to learn or train a model and the other is used to validate the model. The basic form of cross-validation is k-fold cross-validation. In this form, data set is divided into k sub sets. Each sub set is tested via classifier constructed from the remaining (k-1) sub sets. Thus the k different test results are obtained for each train–test pair. The average result gives the test accuracy of the algorithm. We used 10 fold cross-validations in our proposed HPM-MI since it reduces the bias associated with random sampling.

5. Results and Analysis

The results obtained from the proposed model are tabulated in Table 14 for diabetes, Wisconsin breast cancer and Hepatitis data sets. The following subsections give the detailed discussion of results obtained for data sets used under experimental work.

Table14
Obtained classification accuracy, sensitivity and specificity, Kappa value and ROC .

Performance Metric	Pima Indian diabetes Data Set	Wisconsin Breast Cancer Data Set	Hepatitis data set
Accuracy (%)	99.82	99.39	99.08
Sensitivity (%)	100	99.31	100
Specificity (%)	99.7	99.54	96.55
Kappa	0.996	0.986	0.978
ROC	1	1	0.99

5.1 Discussion for Pima Indian Diabetes Data Set

The accuracy of proposed HPM-MI is computed as 99.82% .We have also compared with existing methods (Seera & Lim, 2014; Kahramanli & Allahverdi, 2008; N., 2010; Patil, Joshi, & Toshniwal, 2010) in the literature shown in Table 15. It is clearly evident from Table 15 that none of the studies had the success rates higher than 98.92 % for the mentioned algorithms on Indian Pima Diabetes data set.

Table 15
Classification accuracies of proposed model and other classifiers for the Pima Indian diabetes.

Method	Accuracy (%)	Reference
HPM -MI	**99.82**	**Our Proposed Model**
FMM-CART-RF	78.39	(Seera & Lim, 2014)
FMM-CART	71.35	(Seera & Lim, 2014)
FMM	69.28	(Seera & Lim, 2014)
Sim + F2	75.29	(Luukka, 2011)
Sim + F1	75.84	(Luukka, 2011)
Sim	75.29	(Luukka, 2011)

Binary-coded GA	74.80	(Örkcü & Bal, 2011)
BP	73.80	(Örkcü & Bal, 2011)
Real-coded GA	77.60	(Örkcü & Bal, 2011)
Hybrid Prediction Model with F-score	98.92	(llango & Ramaraj, 2010)
Hybrid Prediction Model	92.38	(Patil et al, 2010)
Hybrid model	84.5	(Kahramanli & Allahverdi, 2008)

In addition to accuracy, specificity as well as sensitivity is also computed and compared with three publications shown in Fig. 5. HPM-MI produced best results of sensitivity & specificity as 100% and 99.74%. (Kahramanli & Allahverdi, 2008) has reported sensitivity & specificity of 80.3% & 87.2%. HPM proposed by (Patil et al, 2010) used resulted sensitivity & specificity of 90.4% & 93.4%, (llango & Ramaraj, 2010) produced gave sensitivity & specificity as 99.3% & 98.7%.

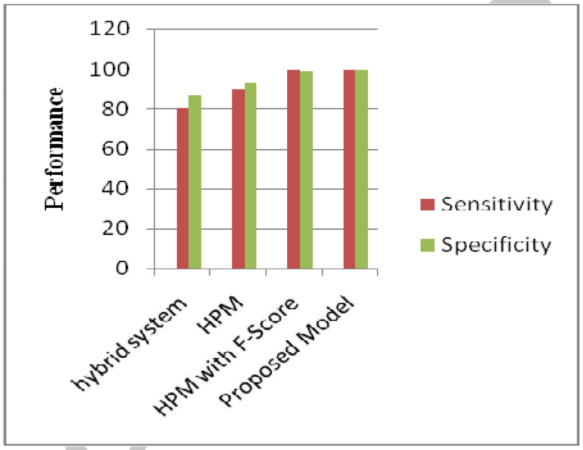

Fig. 5 Comparison of sensitivity & specificity with hybrid system , HPM , HPM with F- Score and Proposed Model for Pima diabetes data set

Another important metric ROC is also calculated. (Lukka, 2011) and (Seera & Lim, 2014) has reported ROC score for their prediction models in their publications. Hence, Fig. 6 shows the comparisons of the areas under ROC curve of HPM-MI with their models and HPM-MI has ranked highest as compared to recent methods proposed by (Luukka, 2011) as well as (Seera & Lim, 2014). Comparing the results produced by proposed model, we conclude that the proposed Hybrid Prediction Model (HPM) using missing value imputation obtains very promising results in classifying the possible diabetes patients. The proposed system can be very helpful to the physicians for their final decision on their patients as by using such an efficient model they can make very accurate decisions.

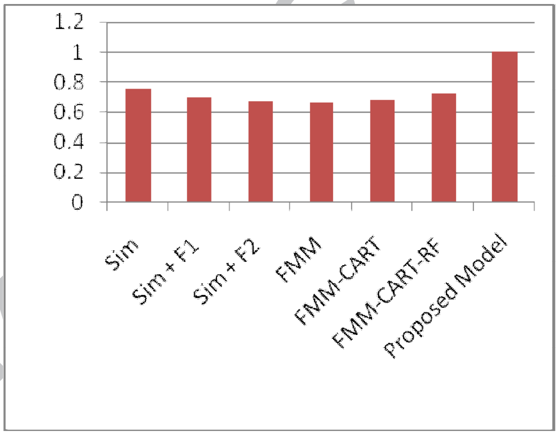

Fig. 6 ROC score for Pima Indian diabetes data set

5.2 Discussion for Wisconsin Breast Cancer Data Set

The accuracy of HPM-MI is computed as 99.39%. We have also compared with existing methods as shown in Table 16. It is clearly evident from Table 16 that none of the studies had the success rates higher than 98.84% for the mentioned algorithms on Wisconsin Brest Cancer Diabetes data set.

In addition to accuracy, specificity as well as sensitivity is also computed .The values of sensitivity & specificity are 99.39% and 99.54% respectively. The publications shown in Table 16 have not mentioned sensitivity & specificity rates; hence no comparison is made for these metrics.

Another important metric ROC was also calculated. Area under ROC for HPM-MI is 1 of proposed model. (Lukka, 2011) and (Seera & Lim, 2014) has reported ROC score for their prediction models in their publications. Hence, Fig. 7 shows the comparisons of the areas under ROC curve of HPM-MI with their models and HPM-MI has ranked highest as compared to recent methods proposed by (Lukka, 2011) as well as (Seera and Lim, 2014). Comparing the results produced by proposed model, we conclude that the proposed Hybrid Prediction Model (HPM) using missing value imputation obtains very promising results in classifying the possible diabetes patients. The proposed system can be very helpful to the physicians for their final decision on their patients as by using such an efficient model they can make very accurate decisions.

Table 16
The values of accuracy of classification made on Wisconsin Breast Cancer data

Method	Accuracy (%)	Reference
HPM-MI	**99.39**	**Proposed Model**
FMM-CART-RF	98.84	(Seera & Lim,2014)
FMM-CART	94.86	(Seera & Lim,2014)
FMM	95.26	(Seera & Lim,2014)
Pedagogical	97.07	(Stoean & Stoean,2013)
Cooperative coevolution	96.69	(Stoean & Stoean, 2013)
SVMs	96.50	(Stoean & Stoean, 2013)
Decompositìonal	95.93	(Stoean & Stoean, 2013)
Sim	97.49	(Luukka, 2011)
Sim + F1	97.10	(Luukka, 2011)
Sim + F2	97.18	(Luukka, 2011)
Binary-coded GA	94.00	(Örkcü & Bal, 2011)
BP	93.10	(Örkcü & Bal, 2011)
BC FRPCA1	98.19	(Luukka, 2009)
BC FRPCA2	98.13	(Luukka, 2009)
BC FRPCA3	98.16	(Luukka, 2009)
BC Original	97.49	(Luukka, 2009)
BC PCA	97.72	(Örkcü & Bal, 2011)
Fuzzy-AIRS	98.51	(Polat et al., 2007)
AIRS	97.20	(Polat et al., 2007)

5.3 Discussion for Hepatitis Data Set

The accuracy of HPM-MI is computed as 99.08%. We have also compared with existing methods as shown in Table 17. It is clearly evident from Table 18 that none of the studies had the success rates higher than 98.52 % for the mentioned algorithms on Hepatitis data set.

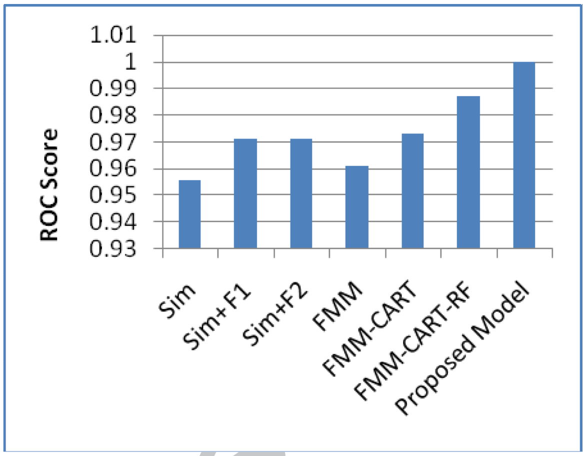

Fig. 7. ROC score for Wisconsin breast cancer data set.

Table 17
The values of accuracy of classification made on Hepatitis dataset.

Method	Accuracy (%)	Reference
HPM-MI	**99.08**	**Proposed Model**
SVR NSGA II	98.52	(Zangooei , Habibi, & lizadehsani, 2014)
LFDM SVM	96.77	(Chen et al , 2011)
LDA-ANFIS	94.16	(Dogantekin et al. , 2009)
PCA AIRS	94.12	(Polat and Gunes , 2007 b)
FS –AIRS with fuzzy resource	92.59	(Polat and Gunes , 2006)

5. 4 Robustness testing of proposed model

To show the robustness of our proposed model at different missing rate, we made the experiments on 2D Synthetic Data Set at different missing rate as 20 %, 40 %, 60 % and 80 %. Results are produced in Table 18. Table 18 shows the incorrectly classified instances produced by 11 imputation techniques. At highest missing rate SVM was chosen as best imputation method and for other missing rate, CMC was selected to impute the missing values in the data set. Moreover, proposed model has produced good results at different missing rates. Table 19 summarizes accuracy, specificity, sensitivity, Kappa and ROC value achieved at different missing rates.

Table 18
Clustering results with 11 imputation methods for 2D synthetic Data Set at different missing rate.

Imputati Methoc	**Incorrectly classified instances(%)**			
	20 % missing	**40 % missing**	**60 % missing**	**80 % missing**
CMC	**8**	**10**	**8.5**	8
FKMI	16	22.5	19.5	19
KMI	14	17	21.5	16.5
KNNI	16	18.5	16.5	15
LSSI	12.5	22.5	29.5	12.5
MC	16.5	27.5	36	41
SVDI	15.5	26	31	31
SVMI	8	10.5	9.5	**5**
WKNN	16	18.5	16.5	15
Case Deletion	11.38	16.68	24.77	18.60
Matrix Factoriz ation	16.5	24.5	34	29

Table 19
Obtained classification accuracy, sensitivity and specificity in percentage for 2D synthetic data set.

Missing Rate (%)	**20**	**40**	**60**	**80**
Accuracy (%)	100	100	99.45	100
Sensitivity (%)	100	100	100	100
Specificity (%)	100	100	98.86	100
Kappa	1	1	0.98	1
ROC	1	1	1	1

We have also analyzed the performance of our model as a function of missing rate and train-test ratio. We experimented HPM-MI on WDBC data set similar to a missing tolerant algorithm namely logistic regression (David, 2007) on different missing rate at 25 %, 50 % and 75 % .At each missing rate, accuracy as well as ROC for HPM-MI was calculated by different train-test ratio as shown in Table 20.Accuracy varies between 98% to 100 % in all the experiments and area under ROC varies from 0.99 to 1 which is better than missing tolerant approach proposed by David et al. (David, 2007) as AUC produced by their approach varies from 0.94 to 0.99 for similar experiments.

Table 20
Obtained classification accuracy and ROC on different missing rate and different train test ratio for WDBC Data Set.

Train –Test Ratio	**25 % Missing rate**		**50 % Missing rate**		**75 % Missing rate**	
	Accuracy (%)	**ROC**	**Accuracy (%)**	**ROC**	**Accuracy (%)**	**ROC**
10-90	99.58	1	99.37	1	99.16	1
20-80	98.82	1	98.82	1	99.06	1
30-70	99.19	1	99.19	1	99.19	.99
40-60	99.35	1	99.37	1	99.06	0.99
50-50	99.24	1	99.24	1	98.86	1
60-40	98.59	1	98.58	1	99.53	1
70-30	98.74	1	98.11	1	99.37	1
80-20	98.11	.99	98.11	.99	100	1
90-10	96.23	.99	96.23	.99	100	1

Table 21
Obtained classification accuracy of HPM-MI, Naivebayes,NaiveBayes (MVI) Bayesnet and J48 in percentage for WDBC synthetic data set at different missing rate.

Missing Rate (%)	**25**	**50**	**75**
Naïve Bayes	93.14	92.97	93.14
Naïve Bayes(MVI)	93.14	93.49	94.02
BayesNet	94.55	94.37	94.20
J48	94.72	93.32	93.32
Proposed Model	**99.43**	**99.24**	**99.06**

Moreover, we have also compared classification accuracy achieved by HPM-MI on WDBC data set with missing value tolerant techniques such as naïve bayes, bayesnet and J48 using WEKA. Table 21 shows that accuracy obtained from proposed model is 99.43%, 99.24% and 99.06 at missing rate of 25%, 50% and 75% respectively that is best in comparison to others. Table 21 also shows that comparison between naïve bayes classifier (that handles the values by omitting the probability while computing the likelihoods of membership of each class) and naïve bayes with best missing value imputation approach selected from our analysis and results are better for naïve bayes with imputation at higher missing rates as 50% and 75%.

6. Conclusion and Future Work

This has paper has proposed a hybrid prediction model with missing value imputation, i.e., HPM-MI to support medical decisions. This model comprises of analysis and selection of imputation method using K-mean clustering, pattern extraction and MLP. Proposed model has emphasized on the analysis and selection of missing value imputation techniques before making any prediction as no generalization can be made for the best imputation method relative to incomplete data set. K-means clustering is also incorporated to extract the correct patterns before applying MLP for classification. A lot of experiments have been performed to validate the proposed model using three medical data sets namely Pima Indian diabetes data set, breast cancer data set, and hepatitis data set from UCI repository. Results obtained from the experiments shows that none of the studies have achieved accuracy of 99.82% for Pima Indian diabetes data set. Further, accuracy obtained for Wisconsin breast cancer as well as hepatitis data set are 99.30% and 99.08% respectively and comparable with other models reported in the literature. Moreover, proposed model is validated on two other data sets namely 2D synthetic and WDBC data sets at different missing rates. Results achieved for these two data sets are also consistent with previous work in the literature.

Future work will focus on the testing and improving the model for multi-class imbalanced classification problems. Imbalanced classification problem arises due to imbalanced distribution of instances among multiple classes present in the data set. Further we would also like to incorporate more MVI approaches such as BPPCA, non-linear PCA in the set of MVI approaches under study (Scholz, Kaplan, Guy, Kopka, & Selbig, 2005). On the other hand, real world data may consist of noise in addition to missing values that has been neglected in our proposed model. Therefore, it is advantageous to equip HPM-MI to deal with noisy data. Finally, it would be nice to examine the efficiency of HPM-MI in other domains in addition to medical domain.

References

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M. J., Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., Fernández, J. C., & Herrera, F. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing, 13*, 307-318.

Aussem, A., & de Morais, S. R. (2008). A Conservative Feature Subset Selection Algorithm with Missing Data. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on* (pp. 725-730).

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics, 77*, 81-97.

Bennett, K. P., & Blue, J. A. (1998). A support vector machine approach to decision trees. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on* (Vol. 3, pp. 2396-2401 vol.2393).

Bioch, J. C., Meer, O., & Potharst, R. . (1996). Classification using Bayesian neural nets. In *International conference on neural networks* (pp. 1488–1493).

Carpenter, G. A., & Markuzon, N. (1998). ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. *Neural Networks, 11*, 323-336.

Chen, H.,Liu D., Yang B., Liu J., & Wang G.(2011), A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis, Expert Systems with Applications, 38(9), 11796-11803.

Chen, H., Yang, B., Wang, G., Wang, S., Liu, J., & Liu, D. (2012) , Support Vector Machine Based Diagnostic System for Breast Cancer Using Swarm Intelligence. Journal of Medical Systems, 2505-2519

Cibenko, G.(1989),Approximation by superpositions of a sigmoidal function," Mathematics of Control, Signal and systems, 2(4), 303-314.

David, W. (2007). On Classification with Incomplete Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*, 427-436.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34(2), 113–127.

Dogantekin, E., Dogantekin, A., & Avci, D. (2009). Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system. Expert Systems with Applications, 36(8), 11282–11286.

Downs, J., Harrison, R. F., Kennedy, R. L., & Cross, S. S. (1996). Application of the fuzzy ARTMAP neural network model to medical pattern classification tasks. *Artificial Intelligence in Medicine, 8,* 403-428.

Hammer, P., & Bonates, T. (2006). Logical analysis of data—An overview: From combinatorial optimization to medical applications. *Annals of Operations Research*, *148*, 203-225.

Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed.). Morgan Kaufmann Publishers.

Hathaway, R. J., & Bezdek, J. C. (2002). Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm. *Pattern Recognition Letters, 23*, 151-160.

llango, S. B., & Ramaraj, N. (2010). A hybrid prediction model with F-score feature selection for type II Diabetes databases. In *1st Amrita ACM-W Celebration on Women in Computing in India* (pp. 1-4). Coimbatore, India.

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*, *35*, 82-89.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 881-892.

Koren, Y. Bell, R, & Volinsky, C. (2009) Matrix factorization techniques for recommender systems. Computer 42.8 30-37

Luengo, J., García, S., & Herrera, F. (2011) On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems, 32, 77–108

Luukka, P. (2009). Classification based on fuzzy robust PCA algorithms and similarity classifier. Expert Systems with Applications, 36(4), 7463–7468.

Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications*, *38*, 4600-4607.

Meesad, P., & Yen, G. G. (2003). Combined numerical and linguistic knowledge representation and its application to medical diagnosis. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 33*, 206-222.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine learning, neural and statistical classification*: Ellis Horwoo.

Mukhopadhyay, S., Changhong, T., Huang, J., Mulong, Y., & Palakal, M. (2002). A comparative study of genetic sequence classification algorithms. In *Neural Networks for Signal Processing,* 57-66.

Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (2007). UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science

Örkcü, H. H., & Bal, H. (2011). Comparing performances of backpropagation and genetic algorithms in the data classification. *Expert Systems with Applications*, *38,* 3703-3709.

Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for Type-2 diabetic patients. *Expert Systems with Applications, 37*, 8102-8108.

Polat, K., & G¨unes, S(2006). "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation. *Digital Signal Processing*, 16(6), pp. 889– 901

Polat, K., & Gunes, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 17(4), 702–710.

Polat, K., & Gunes, S. (2007). Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system. Applied Mathematics and Computation, 189(2), 1282–1291.

Polat, K., & Gunes, S. (2008). Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm.

Expert Systems with Applications, 34(1), 773–779.

Polat, K., Sahan, S., Kodaz, H., & Günes, S. (2007). Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism. Expert Systems with Applications, 32(1), 172–183.

Qin, Y., Zhang, S., Zhu, X., Zhang, J., & Zhang, C. (2007). Semi-parametric optimization for missing data imputation. *Applied Intelligence*, *27*, 79-88.

Quinlan, J. (1996). Improved use of continuous attribute in C4.5. *Journal of Artificial Intelligence Research, 4*, 77-90.

Rumelhart, D. E., Geoffrey E. Hinton, and R. J. Williams. (1986). *Learning Internal Representations by Error Propagation* (Vol. 1: Foundations.): MIT Press.

Saar-Tsechansky M, P. F. (2007). Handling missing values when applying classification models. *The Journal of Machine Learning Research, 8*, 1625–1657

Saastamoinen, K., & Ketola, J. (2006). Medical Data Classification using Logical Similarity Based Measures. *IEEE Conference on Cybernetics and Intelligent Systems,* (1-5).

Scholz, M. Kaplan, F. Guy, C. L. Kopka, J. ,and Selbig, J. (2005)Non-linear PCA: a missing data approach, In Bioinformatics, Vol. 21, Number 20, pp. 3887-3895, Oxford University Press, 2005

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications, 41*, 2239-2249.

Ster, B., & Dobinkar, A (1996), Neural Networks in medical diagnosis: Comparison with others methods. In proceeding of international conference on engineering applications of neural networks, pp 427-430.

Stoean, R., & Stoean, C. (2013). Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. Expert Systems with Applications, 40(7), 2677–2686.

Takács, G., Pilászy, I., & Németh, B. (2008). Matrix factorization and neighbor based algorithms for the Netflix prize problem. ACM Conference on Recommender Systems, 267-274.

Thora, J., Ebba, T., Helgi, S., & Sven, S. (2008). The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. Expert Systems with Applications, 34, 108–118

Tsirogiannis, G. L., Frossyniotis, D., Stoitsis, J., Golemati, S., Stafylopatis, A., & Nikita, K. S. (2004). Classification of medical data with a robust multi-level combination scheme., In *Neural Networks Proceedings. IEEE International Joint Conference on* (Vol. 3, pp. 2483-2487 vol.2483).

Witten, I. H., & Frank, E. (2000). Data mining: practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufman

Zangooei M. H., Habibi J., & Alizadehsani R. (2014). Disease Diagnosis with a hybrid method SVR using NSGA-II, Neurocomputing, 136, 14-29.

Highlights

Proposed novel Hybrid prediction model with missing value imputation.

HPM-MI has improved accuracy, sensitivity, specificity, kappa & ROC on 2 datasets

The best accuracy is achieved for diabetes and breast cancer datasets

MVI is one of the important step of proposed model