Journal of Computational Science 13 (2016) 10–25

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

Saba Bashir[a], Usman Qamar[a,*], Farhan Hassan Khan[a], Lubna Naseem[b]

[a] Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan
[b] Shaheed Zulfiqar Ali Bhutto Medical University, PIMS, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form 2 January 2016
Accepted 4 January 2016
Available online 6 January 2016

Keywords:
Data mining
Prediction
Majority voting
Disease classification
Multi-layer
Ensemble technique

ABSTRACT

Decision support is a crucial function for decision makers in many industries. Typically, Decision Support Systems (DSS) help decision-makers to gather and interpret information and build a foundation for decision-making. Medical Decision Support Systems (MDSS) play an increasingly important role in medical practice. By assisting doctors with making clinical decisions, DSS are expected to improve the quality of medical care. Conventional clinical decision support systems are based on individual classifiers or a simple combination of these classifiers which tend to show moderate performance. In this research, a multi-layer classifier ensemble framework is proposed based on the optimal combination of heterogeneous classifiers. The proposed model named "HMV" overcomes the limitations of conventional performance bottlenecks by utilizing an ensemble of seven heterogeneous classifiers. The framework is evaluated on two different heart disease datasets, two breast cancer datasets, two diabetes datasets, two liver disease datasets, one Parkinson's disease dataset and one hepatitis dataset obtained from public repositories. Effectiveness of the proposed ensemble is investigated by comparison of results with several well-known classifiers as well as ensemble techniques. The experimental evaluation shows that the proposed framework dealt with all types of attributes and achieved high diagnosis accuracy. A case study is also presented based on a real time medical dataset in order to show the high performance and effectiveness of the proposed model.

1. Introduction

Data mining in medical domain, is a process of discovering hidden patterns and information from large medical datasets; analyzes them and uses them for disease prediction [1]. The basic goal of data mining process is to extract hidden information from medical datasets and transform it into an understandable structure for future use [2]. A large number of predictive models can be developed from data mining techniques which enable classification and prediction tasks. After discovering knowledge from data, learning phase starts; where a scientific model is built. This learning method evolves the concept of machine learning and can be formally defined as "the complex computation process of automatic pattern recognition and intelligent decision making based on training sample data" [3]. The machine learning classifiers are further categorized into supervised learning and unsupervised learning depending on the availability of data. In supervised learning, labeled training data is available and a learning model is trained. Some examples include Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Trees (DT). In unsupervised learning, there is no class label field in sample data. Examples include, *K*-mean clustering and Self-Organization Map (SOM). An ensemble approach performs better than individual machine learning techniques by combining the results of individual classifiers [4,5]. There are multiple techniques that can be utilized for constructing the ensemble model and each result in different diagnosis accuracy. Most common ensemble approaches are bagging [6], boosting [7] and stacking [8].

1.1. Contribution

Significant amount of work has already been done on disease classification and prediction. However, there is no single methodology which shows highest performance for all datasets or diseases,

* Corresponding author.
E-mail addresses: saba.bashir@ceme.nust.edu.pk (S. Bashir), usmanq@ceme.nust.edu.pk (U. Qamar), farhan.hassan@ceme.nust.edu.pk (F.H. Khan), doctorlubna@hotmail.com (L. Naseem).

http://dx.doi.org/10.1016/j.jocs.2016.01.001
1877-7503/

while one classifier shows good performance in a given dataset, another approach outperforms the others for other dataset or disease. The proposed research focuses on a novel combination of heterogeneous classifiers for disease classification and prediction, thus overcoming the limitations of individual classifiers. The novel combination of heterogeneous classifiers is presented which is Naïve Bayes, Linear Regression, Quadratic Discriminant Analysis, *K*-Nearest Neighbor, Support Vector Machine, Decision tree using Information Gain and Decision tree using the Gini Index. The multiple classifiers are used at multiple layers to further enhance disease prediction accuracy. An application has also been developed for disease prediction. It is based on the proposed HMV ensemble framework. The proposed application can help both doctors and patients in terms of data management and disease prediction.

The rest of the paper is organized as follows: Section 2 relates to literature review. Section 3 presents the proposed ensemble framework. Section 4 provides the results and discussion from the experiments carried out. Section 5 provides a case study about proposed ensemble model, whereas Section 6 is related to the discussion. Medical application for disease diagnosis is detailed in Section 7 and finally the conclusion is provided in Section 8.

2. Literature review

Extensive amount of work has already been done on disease classification and prediction. However, most of the literature has focused on using a single classifier for a specific disease.

Pattekari and Parveen [9] presented a heart disease prediction system based on a Naïve Bayes algorithm to predict the hidden patterns in a given dataset. The proposed technique limits the use of only categorical data and uses only single classifier. Other data mining techniques such as ensembles, time series, clustering and association mining can be incorporated to improve the results. Similarly Ghumbre, Patil, and Ghatol [10] presented a heart disease prediction system using radial based function network structure and support vector machine. Again a single classifier is being utilized. Prashanth et al. [11] proposed automatic classification and prediction of Parkinson's disease. SVM and logistic regression are used for model construction. SVM classifier with RBF kernel produced high classification accuracy. Improvements can be made by incorporating ensemble classifier instead of a single classifier. Übeyli [12] used different classifiers for disease diagnosis and analyzed that support vector machine achieved the highest performance. The method limits the use of a single classifier for diagnosis and prediction. Ba-Alw et al. [13] presented a survey on data mining approached for hepatitis classification and prediction. The comparison of results indicates that Naïve Bayes attained high classification and prediction accuracy. However, again a single machine learning technique is considered.

Ensemble techniques have been for disease prediction. Zolfaghari [14] proposed a framework for diagnosis of diabetes in female patients. The proposed framework uses an ensemble classifier which is based on neural network and support vector machine. Sapna and Tamilarasi [15] proposed an algorithm that uses fuzzy systems and neural networks for learning membership functions. However, both frameworks use a single layer approach. Multiple layers of classifiers can be incorporated to further increase accuracy. Temurtas [16] introduced a neural network ensemble method for thyroid disease diagnosis in medical datasets. The proposed research focuses on using multilayer, probabilistic and learning vector quantization methods for implementing the neural networks. However the framework is tested only for thyroid disease.

The literature review shows that multiple techniques that have been utilized for disease classification and prediction. However, there is no single methodology which shows highest performance for all datasets or diseases Therefore, the proposed research focuses on multi-classifier and multi-layer ensemble framework for disease classification and prediction with high accuracy for all diseases and datasets.

3. HMV ensemble framework

The proposed ensemble framework consists of three modules, namely data acquisition and preprocessing, classifier training and HMV (Hierarchical Majority Voting) ensemble model for disease classification and prediction with three layered approach.

3.1. Data acquisition and pre-processing module

Data acquisition and pre-processing module includes feature selection, missing value imputation, noise removal and outlier detection. There are multiple methods for feature selection and some of them are given as follows. The HMV ensemble framework utilizes F-score feature selection method.

3.1.1. Feature extraction using principal component analysis (PCA)

The principal component analysis technique assumes that most interesting and useful feature in the dataset is one which has the largest variance and spread. This theory is based on the fact that the dimension with the largest variance represents the dimension which has the largest value of entropy and thus corresponds to maximum information. Eigen vector represents x and y coordinates for a given data. The smallest eigenvectors will often simply represent noise components, whereas the largest eigenvectors often correspond to the principal components that define the data. Dimensionality reduction by means of PCA is then accomplished simply by projecting the data onto the largest eigenvectors of its covariance matrix. Therefore, we obtain a linear M-dimensional subspace of the original N-dimensional data, where $M \leq N$. The Singular Value Decomposition (SVD) is a way to perform PCA analysis and is given by [17]:

$$[U, S, V] = \mathrm{SVD}(A) \tag{1}$$

Therefore,

$$A = USV^T \tag{2}$$

where A is covariance input, U and S hold the eigenvectors of A. The advantages of feature extraction using PCA comprise stability, robustness and fancy extension.

3.1.2. Feature extraction using particle swarm optimization (PSO)

The particle swarm optimization is an evolutionary computational technique where a population is termed as swarm which consists of candidate solutions that are encoded as particles in the search space. The random initialization of a population of particles is used as starting criteria having its own objective function value. The feature extraction method uses PSO as a filter and Correlation-based Feature Selection (CFS) as fitness function. A search algorithm is used by CFS for evaluation of feature subsets. The usefulness of each feature is then evaluated for predicting the class label along with level of inter-correlation between features. Highly correlated features with the class and uncorrelated with each other are considered as good feature subsets. PSO searches for the optimal solution by updating the velocity and the position of each particle because each particle flies in the search space with a velocity adjusted by its own flying memory [18].

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{3}$$

where t denotes the tth iteration and d represents the dth dimension in search space. The best position of particle i is calculated as follows:

$$y_i(t+1) = \begin{cases} y_i(t) & \text{if } f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) & \text{if } f(x_i(t+1)) \langle f(y_i(t)) \end{cases} \quad (4)$$

3.1.3. Feature extraction using forward selection and backward elimination

The feature selection process involves reduction of attributes by selecting only those features which contribute toward final prediction of disease and setting others as rejected. These rejected features will not be used for subsequent modules and analysis. There are multiple steps involved in the process of feature selection and identification [19]. The generation procedure and selection procedure are two of the most important steps. The generation procedure involves generation of subset of features whereas selection procedure will evaluate these features on the basis of different evaluation criteria. The generation procedure can result in empty set, subset based on randomly selected attribute set or a set based on all attributes set. Forward selection method is used in case of empty set which iteratively adds the attributes in feature set and backward elimination method is used in case of all attributes set which iteratively eliminates the irrelevant attributes from feature set. The relevancy of attributes is measured based on wrapper approaches. The main focus of wrapper approaches is classification accuracy. The estimation accuracy of each feature set is calculated that is candidate of adding or removing from the dataset. We have used the cross validation for the accuracy estimation of each feature set for training set. The feature selection process continuous until pre-specified number of features are achieved or some threshold criteria is attained [19].

3.1.4. F-score feature selection

The HMV ensemble framework utilizes F-score feature selection method in order to select most appropriate and relevant features from medical datasets. F-score method can distinguish between two classes having real values. For a given dataset, F-score of a particular feature is calculated by following formula [20]:

$$F(i) = \frac{(X_i'^{(+)} - X_i')^2 + (X_i'^{(-)} - X_i')^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(X_{k,i}^{(+)} - X_i'^{(+)})^2 + \frac{1}{n_--1}\sum_{k=1}^{n_-}(X_{k,i}^{(-)} - X_i'^{(-)})^2} \quad (5)$$

where X_i', $X_i'^{(+)}$ and $X_i'^{(-)}$ is the average of ith feature of the whole, positive and negative datasets respectively.

$X_{k,i}^{(+)}$ is used to represent the ith feature of kth positive instance whereas $X_{k,i}^{(-)}$ is the ith feature of kth negative instance. The discrimination between positive and negative sets is shown in numerator whereas the denominator represents one within each of the two sets. A threshold value is used to select the appropriate features from feature set. If the F-score value of a given feature is greater than threshold value then the feature is added to selected feature space, otherwise it is removed from feature space. The threshold value is obtained by calculating the average of F-scores of all features. Eq. (1) is used to calculate the F-score of ith feature where n_+ is number of positive instances and n_- is number of negative instances. X_k is a given vector where $k = 1, 2, 3, \ldots, n$. A feature is more discriminative indicating large value of F-score.

3.1.5. Missing data imputation using kNN approach

The kNN approach is used for missing data imputation. The proposed procedure is named as "kNNimpute". It is defined as, given a set of instances with incomplete pattern, the K closest cases with known attribute values are identified from the training cases for which the attribute values need to be determined. Once K-nearest neighbors are identified, the missing attribute values are then identified. Heterogeneous Euclidean-Overlap Metric (HEOM) is used for distance measure in order to determine the K-nearest neighbors and then to impute the missing value. HEOM has a benefit that it can calculate different distance measures for different types of attributes. For instance, x_a and x_b are two variables and distance between them is denoted by $d(x_a,x_b)$. Then following formula is used to calculate the distance between them [21]:

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} d_j(x_{aj}, x_{bj})^2} \quad (6)$$

where distance is determined for jth attribute. d_0 is an overlap distance function which assigns a value of 0 if qualitative features are same other $d = 1$. For example:

$$d_0(x_{aj}, x_{bj}) = \begin{cases} 0 & x_{aj} = x_{bj} \\ 1 & x_{aj} \neq x_{bj} \end{cases} \quad (7)$$

Consider jth attribute is missing for an instance x i.e. $m_j = 1$. The distance of x from all training instances is calculated and K-nearest neighbors are identified using the given notation:

$$V_x = \{V_k\}_{k=1}^{K} \quad (8)$$

where K-nearest neighbors of x are arranged in increasing distance order. For instance, v_1 is the closest neighbor of x. Once K-nearest neighbors are identified, the unknown value is determined by an estimate from jth feature values of V_x.

If the jth feature value is continuous value, then missing value is replaced by mean value of its K nearest neighbors. If the jth input feature is qualitative value then the missing value is imputed by determining the category of value using V_x. The mode of $\{V_k\}_{k=1}^{K}$ is imputed where same importance is given to all neighbors. A weighted method assigns a weight α_k to each V_k and closer neighbors are assigned greater weight. The grouping is performed using discrete value in the jth input feature. The distance weighted method for kNN classifier where α_k is calculated by [22]:

$$\alpha_k(x) = \frac{d(V_k, x) - d(V_k, x)}{d(V_k, x) - d(V1, x)} \quad (9)$$

For imputation, consider the jth feature is qualitative having S possible discrete values. The imputed value is given by

$$S = arg\max\{\alpha_x^s\} \quad (10)$$

where s is the possible category that is determined as imputed value. The missing value imputation method can be defined by flowchart given in Fig. 1.

3.1.6. Outlier detection and elimination

The proposed HMV method uses Grubb's test for outlier detection and elimination from medical datasets. The outliers in medical data can exist due to several reasons such as abnormal condition of patient, equipment malfunction or recording error. Grubb's test is also termed as ESD (Extreme Studentized Deviate) method. The advantage of this method is that the complete dataset is considered for outlier detection and elimination.

The test continues as long as all outliers are removed from the dataset. Two step process is applied in order to execute the Grubb's test: calculation of how far the outliers are from others and calculation of ratio G which is the difference between dataset value and mean divided by standard deviation. The outliers are determined using the following formula [23]:

$$G = \frac{\max_{i=1\ldots n}|x_i - \bar{x}|}{\sigma} \quad (11)$$

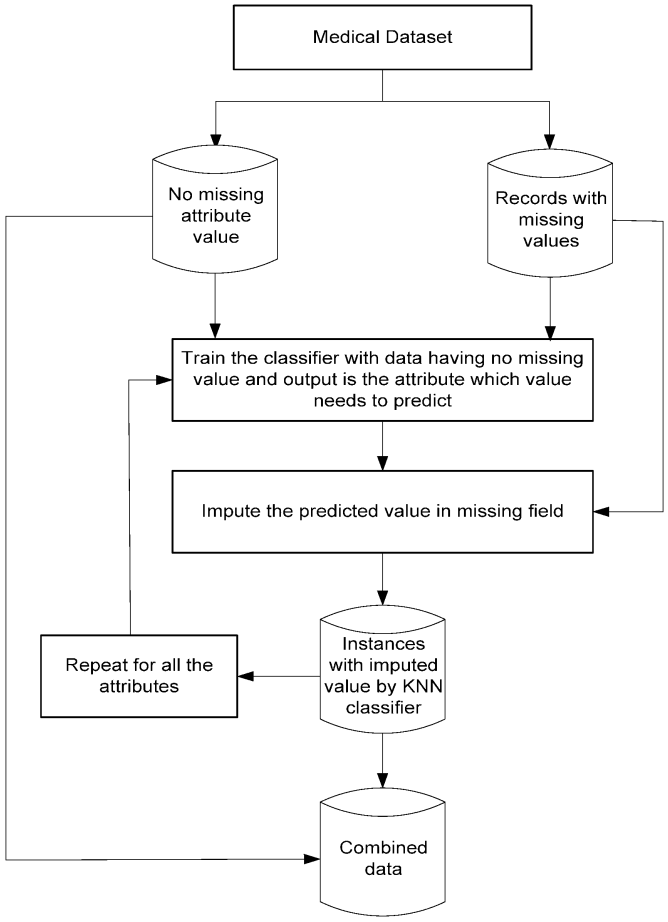

Fig. 1. Proposed missing data imputation method.

where x_i represents the elements of dataset and $\bar{x}$ shows mean of the dataset. The standard deviation of dataset is denoted by σ. The threshold value is termed as statistical significance (α) level. Each Grubb's test (G) value is compared with α level threshold value. If G is higher or lower than significance level, then this value is considered as outlier [23].

3.1.7. Noise removal

The proposed HMV framework eliminates noise from medical dataset using clustering approach. The dataset is also reduced by noise removal method as irrelevant and noisy data is removed. *K*-mean clustering approach is utilized for data clustering. Initially, all dataset is clustered into *K* number of clusters or groups having similar characteristics. Then, centroid of each cluster is calculated and then distance of each point from the particular centroid is calculated. A threshold value is used to remove the noisy data from clusters. If the distance is greater than threshold than the particular point is declared as noise and it is removed. Following rule of thumb is used to set the value of *K* to

$$K \approx \sqrt{\frac{n}{2}} \tag{12}$$

where n denotes the number of objects (data points) and K stands for number of clusters for each dataset [24]. The threshold value is provided by user and mean of all values within a cluster are used to set the threshold. The Euclidean distance between pair of objects is calculated using pair-wise distance and is denoted by n^*p data matrix. The rows of x show instances or observations whereas attributes against each instance are given by matrix columns. The Euclidean distance between pair of attributes for two instances is calculated by following formula:

$$d_{rS}^2 = (x_r - x_S)(x_r - x_S)' \tag{13}$$

where

$$\bar{x}_r = \frac{1}{n}\sum_j x_{rj} \tag{14}$$

and

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj} \tag{15}$$

The noise from disease dataset is removed by following steps: the pair-wise distance using Euclidean distance is calculated between pair of objects; calculate square distance of maximum value is taken from square distance values; threshold value for each cluster is then calculates; if distance > threshold, then the value is considered as noise and is removed from the dataset. The proposed ensemble framework performs noise removal for each dataset individually. We have used benchmark medical datasets in the research and they do not contain any irrelevant features as the respective publishers have already processed them. Therefore, the entire feature set of each dataset will be used for clustering and subsequent analysis. However, the proposed noise removal method can work for any type of features such as binary, numerical and categorical etc. depending upon the type of medical data.

3.2. Classifier training module

The further computations are then performed on preprocessed data by classifier training module. Training set is a labeled dataset used for ensemble training. The instances of training set are described as attribute-value vectors. For instance, let A denotes the set of input attributes containing n attributes: $A = \{a_1,\ldots,a_i,\ldots,a_n\}$ and y represents the class variable or the target attribute. The base classifiers are trained using training dataset and then they can be further utilized for disease classification and prediction.

3.3. HMV ensemble

The construction of HMV ensemble is explained in this section. The selection of optimal set of classifiers for constructing an ensemble is a crucial step for most of the ensemble methods. The HMV ensemble satisfied two conditions that are accuracy and prediction diversity in order to achieve high quality.

3.3.1. Majority voting ensemble scheme

The two of most common methods for combining models are Unweighted [25] and Weighted voting [26]. The unweighted voting is also called majority voting method where each model's output is combined using voting method. The ensemble classifier will output the class with the most votes by base classifiers/models. Weighted voting scheme considers the weight of each classifier. It combines the results of base classifiers and the ensemble model will output the class which has highest weight associated with it. The weights to each classifier can be assigned on the basis of classification accuracy. The highest weight will be assigned to the classifier which has highest accuracy and vice versa. The final prediction will be done based on highest weighted vote. In case of unbalanced classes, the biasness of accuracy results can be generated due to biased dataset. In order to avoid such situation, a multi-layer ensemble framework is proposed based on majority voting ensemble technique. The private judgment of each classifier is turned into collective decision by aggregating their results.

The Majority Voting (MV) ensemble scheme classifies an unlabeled instance into a class that obtains the high number of votes or the most frequent vote. The MV ensemble method is also termed as Plurality Vote (PV) method. Most frequently, the MV method is utilized for comparing the performance of different models [27]. Mathematically:

$$\text{class}(x) = arg\max_{c_i \in dom(y)} \sum_k g(y_k(x), c_i) \quad (16)$$

where the classification of kth classifier is denoted by $y_k(x)$ and $g(y,c)$ represents the indicator function which can be defined as follows:

$$g(y, c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases} \quad (17)$$

In case of probabilistic classifiers, the crisp classification $y_k(x)$ is obtained by following formula:

$$\text{class}(x) = arg\max_{c_i \in dom(y)} \hat{P}M_k(y = c_i|x) \quad (18)$$

where M_k is used to represent the classifier k and $\hat{P}M_k(y = c_i|x)$ denotes the probability of class c for an instance x.

3.3.2. Selection of classifiers

There are two basic categories of ensemble frameworks [28,29]: homogeneous ensemble frameworks and heterogeneous ensemble frameworks. In homogeneous ensemble framework, the base classifiers of same type are used whereas heterogeneous ensemble framework is composed of base classifiers that belong to different types. The basic idea behind ensemble classifies is that they outperforms their base classifiers when the components are not identical. A necessary condition for the ensemble approach is that the components or members of ensemble classifier should have a substantial level of disagreement such as their error level must be independent with respect to each other. The heterogeneous ensemble classifiers overcome the limitations of homogeneous ensemble frameworks. The ensemble criterion for heterogeneous ensemble classifiers is usually a two phase process: overproduce and select. The training dataset is used to train many different base models and then these models are aggregated in order to make a heterogeneous ensemble framework. Multiple studies show that the strength of heterogeneous ensemble is related to the performance of the base classifiers and the lack of correlation between them (model diversity) [30]. Furthermore, the HMV ensemble framework utilizes multi-layer classification scheme in order to further enhance the prediction. The computational complexity of HMV ensemble framework is reduced by dividing it into three layer approach. Fig. 2 shows the detailed architecture of proposed ensemble framework.

3.3.2.1. Layer-1 classifiers. The HMV ensemble framework attains diversity by selecting entirely different set of classifiers. Naïve Bayes (NB) is a probabilistic classifier and shows high classification and prediction accuracy [31]. Linear regression (LR) is a statistical procedure which shows the relationship between different variables and has high interpretability [32]. Quadratic Discriminant Analysis (QDA) is mostly used for classification task and no parameters are required to train the model [33]. It has generally good performance and is inherently multiclass. In layer-1, no decision tree based classifier is used. The NB classifier considers each attribute independently without considering the relationship between them. The correlation analysis between set of attributes is performed using LR and QDA which overcomes the limitation of NB classifier.

3.3.2.2. Layer-2 classifiers. The output of layer-1 classifiers is obtained at layer-2 using majority voting ensemble approach. The MV ensemble will output either class 0 or 1. Moreover, two more classifiers are added at this layer. The classifiers are named as Support vector machine (SVM) and k-Nearest Neighbor (kNN) classifier. kNN is a distance based classifier which does not build a model explicitly [34].

It simply stores the training instances and classifies a new instance based on training dataset by generating the model at run time. It overcomes the risk of overfitting to noise in training set. SVM classifier has the capabilities of outlier detection, classification and regression analysis [35]. The different kernel functions of SVM classifier makes it useful for performing in high dimensional feature space. The limitations of kNN include: computationally expensive and requires a lot of storage space. The SVM works only on subset of feature space and selects the relevant features based on information gain. Thus, the requirement of lot of storage space by kNN is resolved by utilizing SVM classifier. Moreover, the problem of overfitting is also resolved by SVM which results in high classification and prediction accuracy. In any scenario where one classifier has some limitation, the other classifier performs well, consequently giving better performance.

3.3.2.3. Layer-3 classifiers. In layer 3, the output of layer-2 classifiers is combined with Decision Tree using Information Gain (DT-IG) and Decision Tree using Gini Index (DT-GI). Majority voting ensemble approach is used to combine the results of layer-2 classifiers. Decision trees work well with numerical and categorical data. Moreover, the models generated are reliable and robust. Large amount of data is handled well in reasonable time. Each tree will generate an output in form of either 0 or 1. Therefore, there will be three outputs at layer-3 i.e. one output result of layer-2 classifiers and two outputs of decision trees. Majority voting is then applied at these outputs and final prediction is obtained. The final output of hierarchical majority voting ensemble will be either 0 or 1 and that will be the class label of test instance.

The HMV ensemble framework utilized diverse set of base classifiers. Each classifier has some strength and weaknesses that are resolved by combination of heterogeneous classifiers. Hence, each classifier in HMV ensemble framework has diverse set of qualities which complement each other to make a framework which has high classification and prediction accuracy. Moreover, the diversity parameter is also determined by the extent to which each base classifier disagrees about the prediction result for test datasets. Thus all classifiers in HMV ensemble model complement each other very well.

4. Experimental results

4.1. Datasets description

The experimental evaluation of HMV ensemble framework is performed on two heart disease datasets, two breast cancer datasets, two diabetes datasets, two liver disease datasets, one hepatitis dataset and one Parkinson's disease dataset. Each dataset contains diverse set of attributes that are ultimately used to determine the disease classification and prediction such healthy or sick. The two heart disease datasets (Cleveland heart disease, Statlog) are taken from UCI data repository [36]. One breast cancer dataset named Wisconsin Breast Cancer (WBC) dataset is obtained from UCI data repository whereas other dataset (UMC dataset) is retrieved from Wisconsin clinical sciences center [37] repository. Two diabetes datasets are named as Pima Indian Diabetes Dataset (PIDD) and Biostat Diabetes Dataset (BDD). PIDD dataset is taken from UCI data repository whereas BDD dataset is obtained from Biostat data repository [38]. Both liver disease datasets are obtained from UCI data repository and they are named as BUPA liver disease dataset and Indian Liver Patient Dataset (ILPD). Hepatitis dataset is also

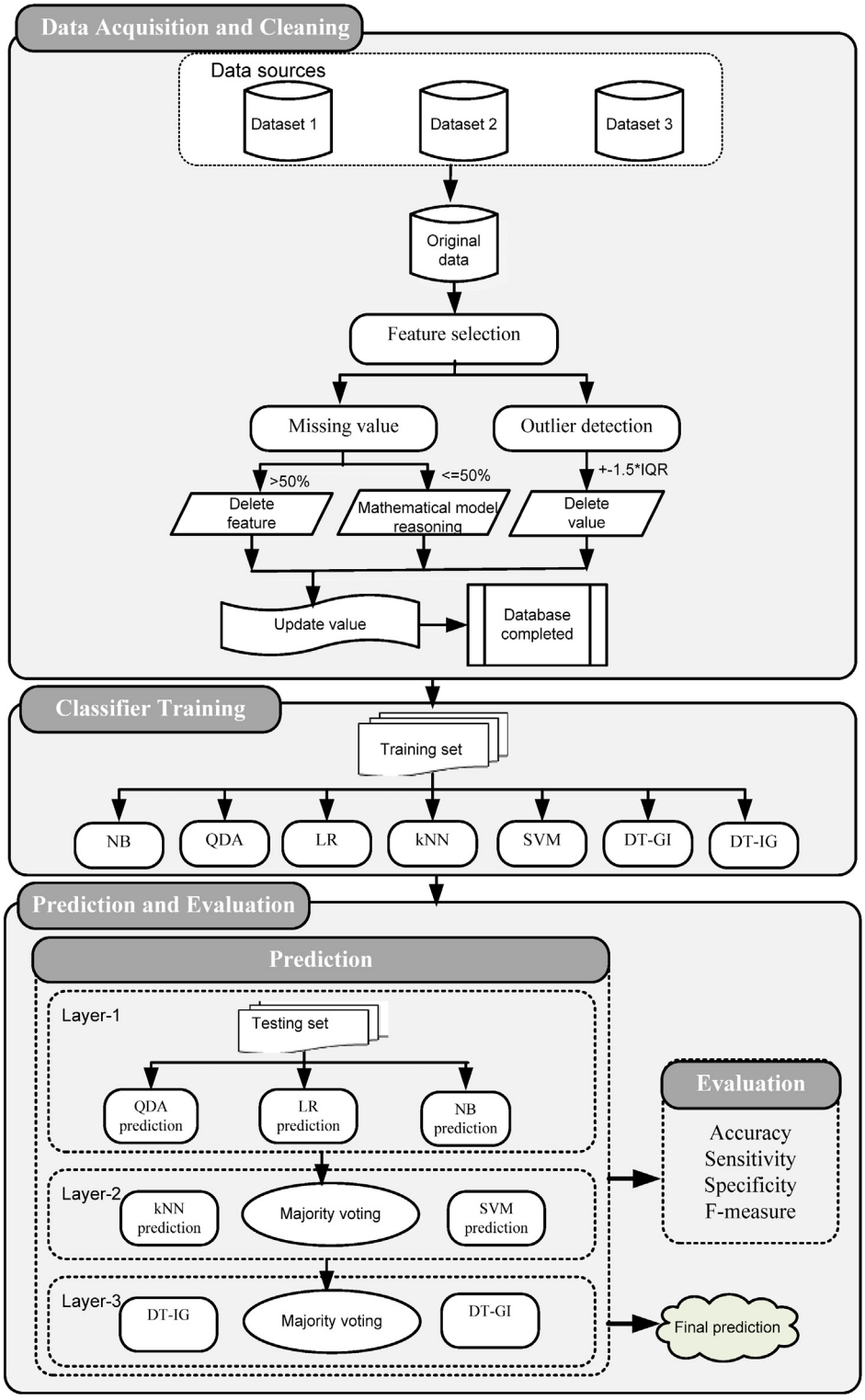

Fig. 2. Detailed architecture of proposed ensemble framework.

taken from UCI data repository and it is named as hepatitis disease dataset. Parkinson's dataset is termed as Parkinson's disease dataset and it is obtained from UCI data repository. The class labels of each dataset are replaced with 0 and 1 in order to maintain consistency where 0 represents absence of disease or healthy and 1 indicates the presence of disease or sick.

Each dataset is divided into training set and test set. The HMV ensemble framework is applied on each test set. Ten-fold cross validation is applied and confusion matrices are obtained. The average prediction result of all confusion matrices is then calculated and analyzed.

4.2. Heart disease prediction

Table 1 shows the comparison of Accuracy (Acc), Sensitivity (Sen), Specificity (Spec) and F-measure (F-M) results of the HMV

Table 1
Comparison of HMV ensemble framework with other classifiers for heart disease.

Classifiers	Acc	Sen	Spec	F-M	Acc	Sen	Spec	F-M
	Cleveland dataset				Statlog dataset			
NB	78.22%	78.49%	82.32%	80.36%	76.67%	77.71%	81.33%	79.48%
QDA	65.68%	68.29%	68.29%	68.29%	68.15%	75.00%	64.00%	69.06%
LR	83.50%	82.39%	88.41%	85.29%	82.59%	82.39%	87.33%	84.79%
kNN	58.42%	60.92%	64.63%	62.72%	57.41%	61.01%	64.67%	62.78%
SVM	80.86%	76.24%	93.90%	84.15%	81.85%	77.60%	94.67%	85.29%
DT-GI	74.92%	76.83%	76.83%	76.83%	75.56%	79.17%	76.00%	77.55%
DT-IG	73.27%	75.78%	74.39%	75.08%	73.33%	76.00%	76.00%	76.00%
RF	69.64%	84.76%	51.80%	64.30%	71.11%	81.33%	58.33%	67.94%
AdaBoost	79.21%	83.54%	74.10%	78.54%	78.89%	83.33%	73.33%	78.01%
HMV	84.49%	83.82%	88.41%	86.05%	84.44%	86.00%	86.00%	86.00%

Table 2
Comparison of HMV ensemble framework with state of art techniques for Cleveland dataset.

Reference	Year	Accuracy	Sensitivity	Specificity
Yang et al. [39]	2015	73.74%	–	–
Peter et al. [40]	2015	78.2%	–	–
Kiruthika et al. [41]	2014	78%	–	–
Shouman et al. [42]	2013	80%	–	–
Bashir et al. [43]	2015	84.16%	92.68%	74.10%
HMV	2015	85%	83.82%	88.41%

for all datasets with individual and ensemble classifiers such as random forest (RF) and AdaBoost. It can be seen from Table 1 that, HMV ensemble framework produces the highest accuracy level for all heart disease datasets when compared to other classifiers. HMV ensemble framework shows a consistent accuracy level of around 85%, whereas other classifiers are not stable as observed in the results. HMV ensemble framework achieved best accuracy of 84.49%, 86% sensitivity, 88.41% specificity, and 86.05% F-Measure. Table 2 shows a comparison of accuracy, sensitivity, specificity and F-measure for HMV ensemble framework with other state of art classification techniques for the Cleveland heart disease dataset. The comparison of the results shows that HMV ensemble framework performed much better than state of the art classification techniques.

Table 3 shows comparison of HMV ensemble with other classification methods with less number of classifiers for heart disease dataset. It is clear from the comparison that HMV ensemble has high classification and prediction accuracy when compared with other classification methods.

4.3. Breast cancer prediction

The evaluation of HMV ensemble framework is also performed for breast cancer datasets as shown in Table 4. The highest accuracy of 97% and 95.78% sensitivity is achieved for WBC dataset whereas high specificity of 96.7% with 98.01% sensitivity, 96.94% specificity and 97.48% F-measure is achieved for WBC dataset by HMV ensemble framework.

Table 5 shows accuracy, sensitivity, specificity and F-measure comparison of HMV ensemble framework with state of the art techniques for breast cancer datasets. HMV ensemble framework produces the highest accuracy level for all the datasets when compared to other state of the art techniques. It has achieved an accuracy of 97% for WBC dataset whereas 98.01% sensitivity and 96.94% specificity are achieved when compared with the state of art techniques for WBC dataset.

4.4. Diabetes prediction

The comparison of HMV ensemble framework is also performed with other classifiers for diabetes datasets as shown in Table 6. We have used two diabetes datasets i.e. Pima Indian Diabetes Dataset (PIDD) and Biostat Diabetes Dataset (BDD) for comparison of results. Accuracy, sensitivity, specificity and F-measure comparison is performed for HMV ensemble framework with other classifiers. The analysis of the results indicates that HMV has

Table 3
Comparison of HMV ensemble with other classification methods for heart disease.

Classification method	Classifiers	Acc	Sen	Spec	F-M
3 classifiers based ensemble	Naïve Bayes, Decision tree, Support vector machine	78.79%	68.42%	85.71%	76.09%
5 classifiers based ensemble	Naive Bayes, DT-IG, DT-GI, *K* Nearest Neighbor, Support vector machine	81.82%	73.68%	92.86%	82.17%
7 classifiers based ensemble (HMV ensemble)	Naive Bayes, Decision Tree-Info Gain, Decision Tree-Gini Index, *K* Nearest Neighbor, Support vector machine, Linear regression, Quadratic discriminant analysis	84.49%	83.82%	88.41%	86.05%

Table 4
Comparison of HMV ensemble framework with other classifiers for breast cancer.

Classifiers	Acc	Sen	Spec	F-M	Acc	Sen	Spec	F-M
	WBC dataset				UMC dataset			
NB	95.99%	98.64%	95.20%	96.89%	70.98%	51.47%	41.18%	45.75%
QDA	67.10%	66.57%	100.00%	79.93%	31.82%	29.48%	92.94%	44.76%
LR	95.85%	95.74%	98.03%	96.87%	71.33%	53.85%	24.71%	33.87%
kNN	95.85%	96.53%	97.16%	96.84%	62.24%	33.33%	27.06%	29.87%
SVM	95.28%	96.30%	96.51%	96.40%	70.28%	50.00%	24.71%	33.07%
DT-GI	94.28%	95.83%	95.41%	95.62%	69.58%	47.37%	21.18%	29.27%
DT-IG	93.28%	94.38%	95.41%	94.90%	69.58%	25.00%	1.18%	2.25%
RF	92.85%	87.14%	95.85%	91.29%	70.28%	11.76%	95.02%	20.94%
AdaBoost	95.85%	92.95%	97.38%	95.11%	64.34%	28.24%	79.60%	41.68%
HMV	96.71%	98.01%	96.94%	97.48%	72.38%	65.00%	15.29%	24.76%

Table 5
Comparison of HMV ensemble framework with state of art techniques for WBC dataset.

Reference	Year	Accuracy	Sensitivity	Specificity
Sørensen et al. [44]	2015	92%	90%	65%
Zand [45]	2015	84.5%	70%	57%
Chaurasia et al. [46]	2014	74.47%	76.2%	92.5%
Chaurasia et al. [47]	2014	–	95.8%	98%
Aljahdali et al. [48]	2013	75%		
HMV	2015	97%	98.01%	96.94%

achieved the highest prediction accuracy for both the datasets. It has achieved 93.05% accuracy for BDD dataset and 77.08% for PIDD dataset respectively.

Table 7 shows accuracy, sensitivity, specificity and F-measure comparison of HMV ensemble framework with other state of art classification techniques. The analysis of the results indicates that HMV ensemble framework has achieved the highest accuracy of 77.08%, when compared with the state of the art techniques for PIDD dataset.

4.5. Liver disease prediction

The comparison of HMV ensemble framework is also performed for liver disease datasets with other classifiers. Table 8 shows accuracy, sensitivity, specificity and F-measure comparison of different classifiers for BUPA liver disease dataset and ILPD dataset. The analysis of the results indicates that HMV ensemble framework has achieved highest accuracy in both liver disease datasets. It has achieved 71.53% accuracy for ILPD dataset and 67.54% accuracy for Bupa liver disease dataset when compared with other classifiers.

Table 9 shows accuracy, sensitivity, specificity and F-measure comparison of HMV ensemble framework with the state of the art classification techniques for liver disease datasets. The analysis of results indicates that HMV has achieved the highest accuracy of 71.53% for ILPD dataset when compared with other classifiers.

4.6. Hepatitis prediction

Table 10 shows accuracy, sensitivity, specificity and F-measure comparison of HMV ensemble framework with other classifiers. The analysis of the results indicates that HMV ensemble framework has achieved the highest accuracy, sensitivity and F-measure of 86.45%, 90.48%, 92.68% and 91.57% respectively when compared with other classifiers.

Table 11 shows accuracy, sensitivity, specificity and F-measure comparison of HMV ensemble framework with the state of the art techniques for hepatitis disease dataset. The analysis of the results indicates that HMV ensemble framework has achieved the highest accuracy when compared with the state of the art techniques.

Table 6
Comparison of HMV ensemble framework with other classifiers for diabetes.

Classifiers	Acc	Sen	Spec	F-M	Acc	Sen	Spec	F-M
	Pima Indian diabetes dataset				Biostat diabetes dataset			
NB	75.52%	79.77%	83.60%	81.64%	91.32%	94.77%	95.04%	94.91%
QDA	57.94%	80.41%	46.80%	59.17%	14.89%	#DIV/0!	0.00%	#DIV/0!
LR	75.78%	79.07%	85.40%	82.12%	91.81%	92.35%	98.54%	95.35%
Knn	68.23%	75.40%	76.00%	75.70%	87.84%	92.24%	93.59%	92.91%
SVM	76.95%	78.99%	88.00%	83.25%	91.56%	94.02%	96.21%	95.10%
DT-GI	74.61%	78.08%	84.80%	81.30%	89.58%	92.88%	95.04%	93.95%
DT-IG	74.35%	78.42%	83.60%	80.93%	89.33%	93.35%	94.17%	93.76%
RF	65.10%	3.73%	98.00%	7.19%	85.11%	100.00%	0.00%	0.00%
AdaBoost	76.43%	52.99%	89.00%	66.42%	88.83%	96.79%	43.33%	59.87%
HMV	77.08%	78.93%	88.40%	83.40%	93.05%	94.12%	97.96%	96.00%

Table 7
Comparison of HMV ensemble framework with state of art techniques for PIDD dataset.

Reference	Year	Accuracy	Sensitivity	Specificity
Kandhasamy et al. [49]	2015	71.74%	53.81%	80.4%
Bashir et al. [50]	2014	74.48%	81.4%	61.5%
Gandhi et al. [51]	2014	75%	–	–
Tapak et al. [52]	2013	75.3%	13.3%	99.9%
Karthikeyani et al. [53]	2012	74.8%	–	–
HMV	2015	77.08%	78.93%	88.40%

Table 8
Comparison of HMV ensemble framework with other classifiers for liver disease.

Classifiers	Acc	Sen	Spec	F-M	Acc	Sen	Spec	F-M
	BUPA liver disease dataset				ILPD dataset			
NB	56.52%	48.89%	75.86%	59.46%	56.43%	96.02%	40.63%	57.09%
QDA	42.03%	42.03%	100.00%	59.18%	29.16%	100.00%	0.72%	1.43%
LR	67.54%	64.10%	51.72%	57.25%	71.01%	71.78%	97.84%	82.81%
kNN	62.03%	54.67%	56.55%	55.59%	63.98%	75.62%	73.08%	74.33%
SVM	57.97%	#DIV/0!	0.00%	#DIV/0!	71.36%	71.36%	100.00%	83.28%
DT-GI	57.97%	#DIV/0!	0.00%	#DIV/0!	71.36%	71.36%	100.00%	83.28%
DT-IG	57.97%	#DIV/0!	0.00%	#DIV/0!	71.53%	71.70%	99.28%	83.27%
RF	59.42%	7.59%	97.00%	14.07%	71.70%	99.76%	1.80%	3.53%
AdaBoost	68.41%	53.79%	79.00%	64.00%	66.55%	81.49%	29.34%	43.15%
HMV	67.54%	68.54%	42.07%	52.14%	71.53%	71.48%	100.00%	83.37%

Table 9
Comparison of HMV ensemble framework with state of art techniques for ILPD dataset.

Reference	Year	Accuracy	Sensitivity	Specificity
Gulia et al. [54]	2014	67.2%	–	–
Jin et al. [55]	2014	65.3%	72.7%	46.7%
Sug [56]	2012	67.82%	–	–
Ramana et al. [57]	2011	62.6%	55.86%	67.5%
Karthik et al. [58]	2011	55%	–	–
HMV	2015	71.53%	71.48%	100.00%

Table 10
Comparison of HMV ensemble framework with other classifiers for hepatitis.

Classifiers	Acc	Sen	Spec	F-M
NB	74.84%	92.86%	73.98%	82.35%
QDA	85.16%	87.88%	94.31%	90.98%
LR	84.52%	89.60%	91.06%	90.32%
kNN	66.45%	79.34%	78.05%	78.69%
SVM	84.52%	89.60%	91.06%	90.32%
DT-GI	80.00%	87.70%	86.99%	87.35%
DT-IG	81.29%	89.17%	86.99%	88.07%
RF	83.12%	100.00%	18.75%	31.58%
AdaBoost	83.12%	95.08%	37.50%	53.79%
HMV	86.45%	90.48%	92.68%	91.57%

Table 11
Comparison of HMV ensemble framework with state of art techniques for hepatitis dataset.

Reference	Year	Accuracy	Sensitivity	Specificity
Houby [59]	2014	69%	87.3%	50.6%
Karthikeyan et al. [60]	2013	84%	–	–
Neshat et al. [61]	2012	70.29%	–	–
Kumar et al. [62]	2011	83.12%	–	–
Khan et al. [63]	2008	72%	83%	66%
HMV	2015	86.45%	90.48%	92.68%

Table 12
Comparison of HMV ensemble framework with other classifiers for Parkinson's disease.

Classifiers	Acc	Sen	Spec	F-M
NB	70.77%	95.92%	63.95%	76.73%
QDA	25.64%	100.00%	1.36%	2.68%
LR	88.72%	89.81%	95.92%	92.76%
kNN	83.59%	88.59%	89.80%	89.19%
SVM	87.69%	86.83%	98.64%	92.36%
DT-GI	88.72%	91.95%	93.20%	92.57%
DT-IG	85.64%	89.40%	91.84%	90.60%
RF	87.92%	87.66%	69.9%	77.78%
AdaBoost	88.90%	89.91%	78.87%	84.02%
HMV	89.23%	91.45%	94.56%	92.98%

4.7. Parkinson's disease prediction

The accuracy, sensitivity, specificity and F-measure comparison of proposed HMV ensemble is also performed with individual as well as ensemble classifiers for Parkinson's dataset as shown in Table 12. HMV has achieved high performance when compared with other classifiers. It has achieved highest accuracy of 89.23%, sensitivity of 91.45%, 94.56% specificity and 92.98% F-measure.

Table 13 shows accuracy, sensitivity, specificity and F-measure comparison of HMV ensemble framework with the state of the art techniques for Parkinson's dataset. The analysis of the results indicates that HMV ensemble framework has achieved the highest accuracy when compared with the state of the art techniques.

The graphical comparison of HMV ensemble approach is also performed with other classifiers as shown in Figs. 3–7. It is clear from the analysis that HMV ensemble has high classification and prediction accuracy as compared to other classifiers for all medical datasets.

4.8. Analysis of HMV ensemble with other techniques

The proposed HMV ensemble method outperforms the other methods for all medical datasets. HMV comprises of heterogeneous base classifiers and layered architecture. The strength of heterogeneous ensemble is related to the performance of the base classifiers and the lack of correlation between them (model diversity). We have introduced multi-layer classification to further enhance the prediction. In order to keep the ensemble computationally less expensive we have used a three layer approach. The HMV ensemble framework attains diversity by selecting entirely different set of classifiers. Naïve Bayes (NB) is probabilistic classifier and shows high classification and prediction accuracy. Linear regression (LR) is a statistical procedure which shows the relationship between different variables and has high interpretability. Quadratic Discriminant Analysis (QDA) is mostly used for classification task and no parameters are required to train the model. kNN is a distance based classifier which does not build a model explicitly. SVM classifier has the capabilities of outlier detection, classification and regression analysis. Decision trees work well with numerical and categorical data. Moreover, the models generated are reliable and robust.

Table 13
Comparison of HMV ensemble framework with state of art techniques for Parkinson's dataset.

Reference	Year	Accuracy	Sensitivity	Specificity
Prashanth et al. [11]	2014	96%	–	–
Willis et al. [64]	2012	69.6%	–	–
Hariganesh et al. [65]	2014	74%	–	–
Yadav et al. [66]	2011	76%	97%	15%
HMV	2015	89.23%	91.45%	94.56%

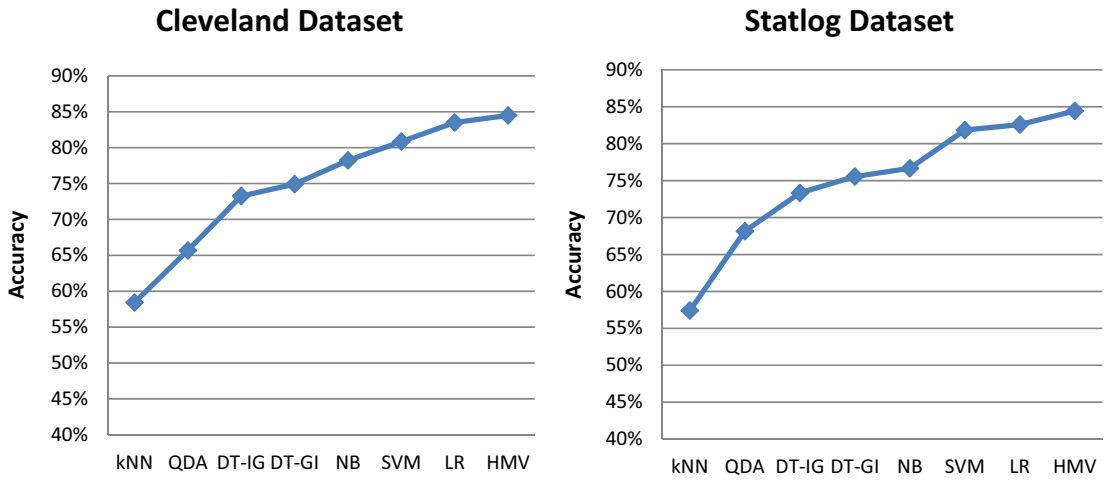

Fig. 3. Accuracy comparison of HMV ensemble with other classifiers for heart disease datasets.

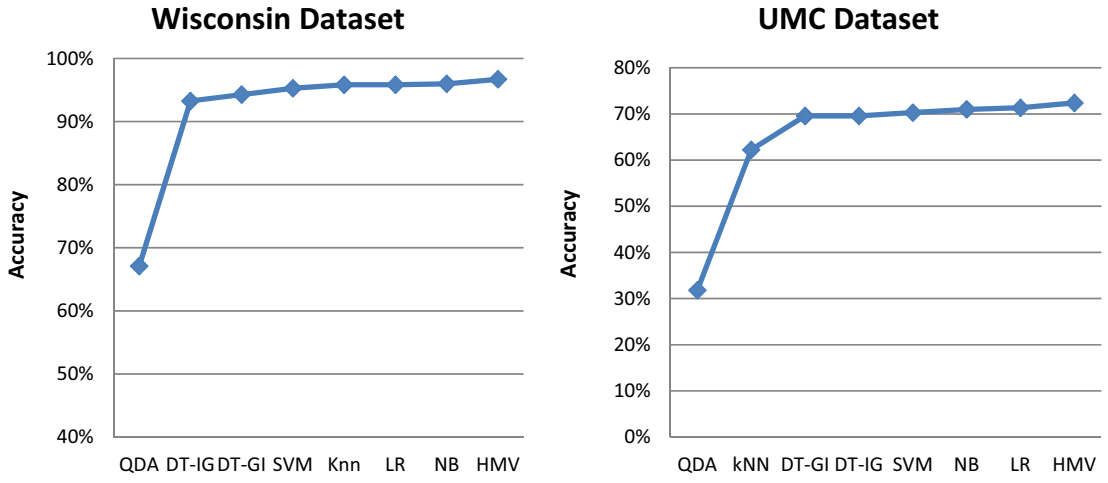

Fig. 4. Accuracy comparison of HMV ensemble with other classifiers for breast cancer datasets.

The computation cost of classification models is as important as accuracy or other parameters in classification methods [67]. The computational cost is usually determined by the cost of computing implicit features for raw data. The computational cost of running a classifier has two major contributors: (i) information extraction from the unseen test instance, e.g. lab tests, parsing, remote database lookups, etc. and (ii) running the decision function or algorithm using the test and training information. Feature extraction plays a vital role in computation cost of any model since the unseen instances do not yield to pre-computation as much. If the number of selected features gets smaller, the computation cost gets lower and in return we get an efficient model. Therefore, the

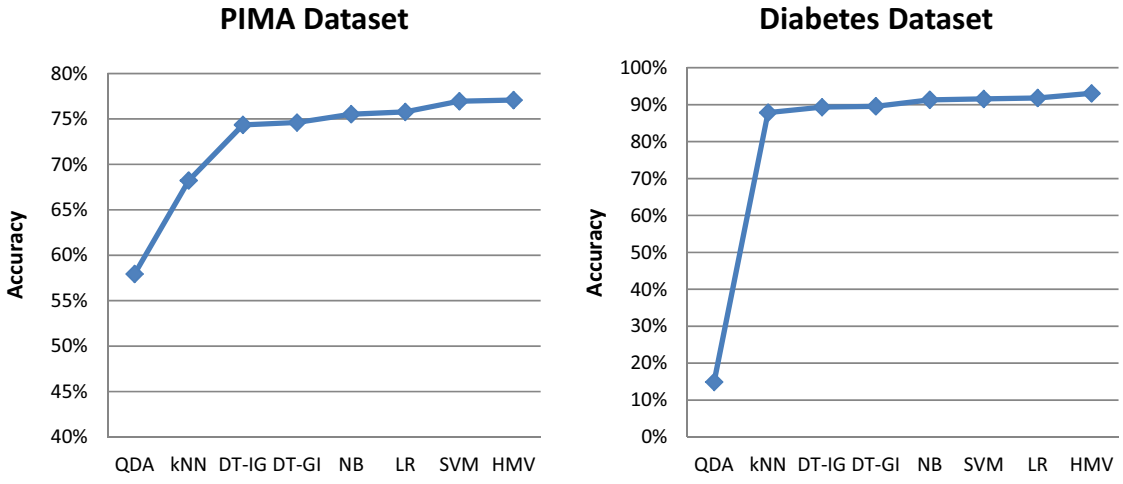

Fig. 5. Accuracy comparison of HMV ensemble with other classifiers for diabetes datasets.

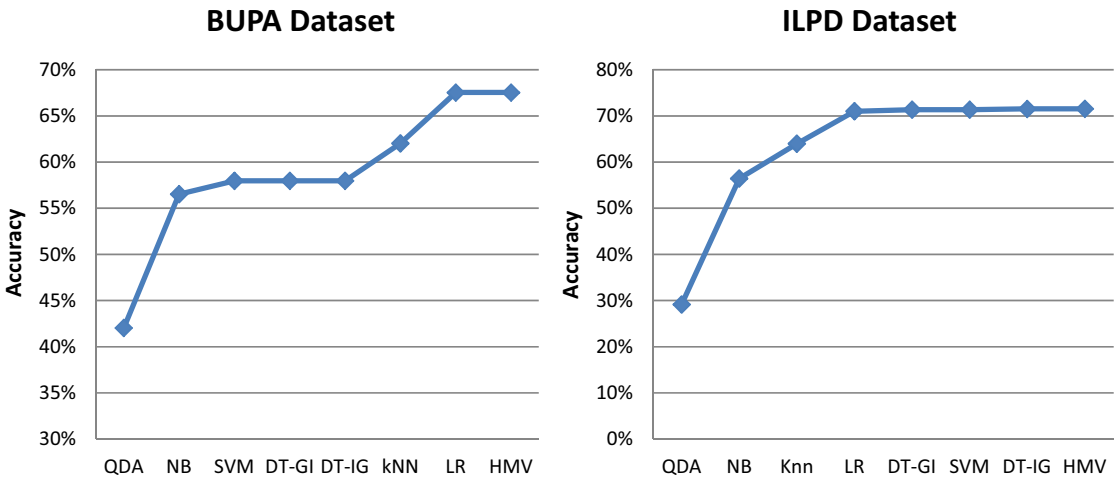

Fig. 6. Accuracy comparison of HMV ensemble with other classifiers for liver disease datasets.

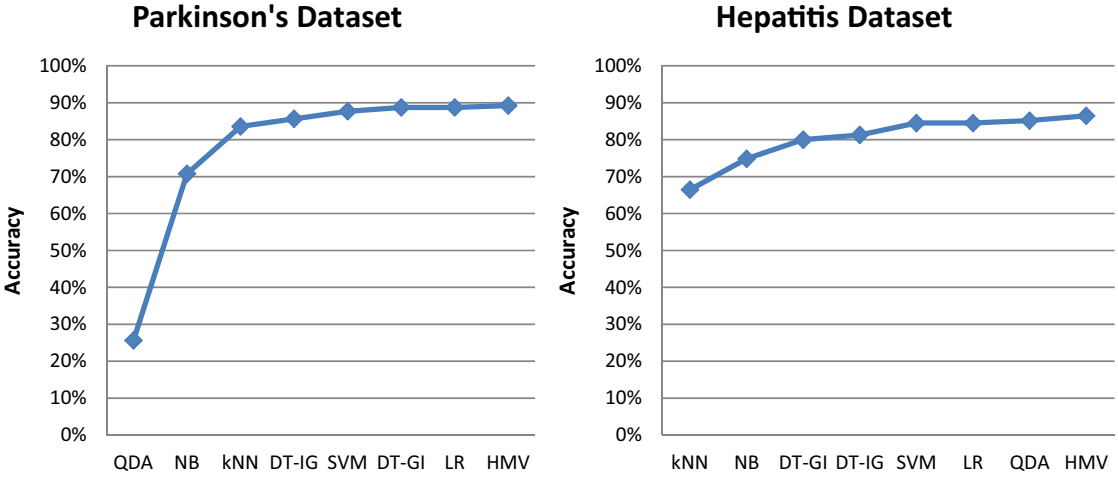

Fig. 7. Accuracy comparison of HMV ensemble with other classifiers for Parkinson's and hepatitis datasets.

proposed ensemble utilizes reduced set of feature space in order to make the proposed ensemble computationally effective.

The computation cost comparison of HMV ensemble is now performed with other techniques for all medical datasets as shown in Table 14. It is clear from the computation time that HMV ensemble is not much expensive in terms of computation time.

Looking at the performance of HMV ensemble, it is analyzed that HMV ensemble framework has achieved high classification and prediction accuracy for all medical datasets. While some single as well as ensemble classifiers have produced good results for one medical dataset, they have produced poor results when applied to other medical datasets; whereas, HMV ensemble has consistently good performance and highest accuracy for all diseases datasets. This is contributed to our ensemble framework, which utilizes an optimal model of class diversity. Each classifier in HMV ensemble has diverse set of qualities which complement each other and overcome the limitations of individual classifiers and results in a framework which has high classification and prediction accuracy for all diseases.

5. Case study: real-time implementation of the proposed framework

The proposed DSS is evaluated on real time dataset of blood CP taken from Pakistan Institute of Medical Science (PIMS) hospital. PIMS is located in Islamabad, Pakistan. It is opening in 1985. P.I.M.S. hospital provides patient health care, medical facilities and as medical appointment hospital also to ways and training of doctors and other health workers in the field of medication and surgery. It consists of multiple departments such as Cardiology, Dental, Urology, Dermatology, Blood bank, Radiology, Plastic surgery, Pulmonology, Oncology and Pathology etc.

Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST) has signed a Memorandum of Understanding (MoU) with Shaheed Zulfiqar Ali Bhutto Medical University which is based at PIMS. The MoU allows for facilitation and sharing of data, however the following guidelines must be met:

1. Handling of data can only be done in presence of a medical doctor and its usage will also be supervised by the medical doctor.
2. Any data that is used will be anonymized by the concerned members of the hospital.

Therefore there is no direct key or identifier type variable in the blood dataset used in the proposed research. The dataset is not individually identifiable. The usage of the data and its monitoring was carried out by Dr. Lubna Naseem.

The first step was to build a knowledge from which the classifiers maybe trained for the prediction of disease. A team of five

Table 14
Computation cost comparison of HMV ensemble with other techniques.

Classifiers	Time (ms) per instance				
	Heart disease datasets		Breast cancer datasets		Hepatitis dataset
	Cleveland	Statlog	UMC	WBC	Hepatitis
NB	0.1	.03	.06	.09	0.14
QDA	.03	.03	.02	.03	.06
LR	.01	.01	.02	.01	.03
Knn	.11	.11	0.1	.24	.09
SVM	.03	.01	.02	.01	.03
DT-GI	.05	.04	.03	.05	.06
DT-IG	.05	.04	.02	.06	.04
HMV	25.24	22.3	15.46	20.9	24.8
Classifiers	Diabetes datasets		Liver disease datasets		Parkinson's dataset
	PIMA	Diabetes	ILPD	BUPA	Parkinson's
NB	.09	.13	0.1	0.1	0.2
QDA	.03	.03	.02	.01	.05
LR	.01	.01	.01	.01	.01
Knn	.19	.19	.24	0.1	.21
SVM	.02	.02	.02	.02	.02
DT-GI	.06	.07	.03	.05	.07
DT-IG	.07	.05	.04	.04	.07
HMV	48.6	32.5	35.6	29.9	27.8

based on medical practitioners and doctors helped us to define the medical knowledge in order to classify the healthy and diseased patients. A patient will come to the doctor to be diagnosed, the medical knowledge was stated in natural language and was written as follows:

A person is having disease if a person has high T.L.C of 20 unit, platelets count of 80 unit, hemoglobin of 6 unit and red cell count of 3 unit, then there are strong chances of having any disease. On the other hand if a person has normal T.L.C of 10 unit, platelets count of 300 unit, hemoglobin of 10 unit and red cell count of 4.8 unit then there are less chances of disease.

According to the defined knowledge of doctor, the features of blood CP are entered into the database along with reference values and diagnosis performed by doctor. The personal information of each patient and other identifying tags are kept anonymyzed for privacy purposes. We have used 495 patients' data consisting of 13 attributes. Following blood CP features are used for real time evaluation of data.

The tests attribute in Table 15 shows feature set that is used for experimentation and analysis, Unit attribute shows measurement of unit or each feature and reference value indicates unit range for healthy person. The proposed DSS write rules based on feature sets and values entered into the database. The dataset is divided into training set and test ser. The training of base classifiers is then performed using training dataset. Table 16 shows a sample of rules generated from decision trees. The training of base classifiers is performed on monthly basis.

Table 15
PIMS blood CP dataset features for disease prediction.

Tests	Unit	Reference value
T.L.C	*1000/μL	4.5–11.5
Red cell count	million/μL	3.5–5.5
Hemoglobin	g/dL	12–15
PCV/HCT	Fl	35–55
MCV	Fl	75–100
MCH	Pg	25.0–35.0
MCHC	g/dL	31.0–38.0
Platelet count	*1000/μL	100–400
RDW-CV	%	11.6–15.0
Neutrophils	%	60–70
Lymphocytes	%	30–40

Table 16
An example of if-then rules generated by proposed ensemble framework.

Rules	Diagnosis
If T.L.C < 4.5 and red cell count < 3.5 and hemoglobin < 12	Class = 1
If T.L.C < 4.5 and red cell count < 3.5 and platelets count < 100	Class = 1
If T.L.C > 4.5 and hemoglobin > 12 and platelets count > 100	Class = 0
If T.L.C > 4.5 and hemoglobin > 12 and platelets count < 100	Class = 1

After classifiers training, the proposed DSS is then used by medical practitioners for disease classification and prediction. A patient showing certain disease symptoms goes to the doctor. The patient's data is fed into the system and then the prediction performed by DSS is discussed with panel of doctors in order to verify the accuracy of disease prediction. Moreover at the end of each discussion, the recommendations provided by the proposed DSS were compared with panel's decisions in order determine whether the two recommendations matched.

5.1. Analysis of results

The prediction performed by proposed DSS is matched with prediction performed by panel of doctors and then accuracy is calculated. The process of first 10 patients is shown in Table 17.

Table 17
Diagnosis comparison of individual patients.

Patient_ID	By doctor	Prediction by multi-layer ensemble
1	1	1
2	1	1
3	1	1
4	1	1
5	1	0
6	0	0
7	1	1
8	0	0
9	1	1
10	1	1

Table 18
Comparison with other classifiers for PIMS patients.

		Class 1	Class 0	Accuracy	Sensitivity	Specificity	F-Measure
HMV	Class 1	366	2	99.19%	99.46%	99.46%	99.46%
	Class 0	2	125				
NB	Class 1	302	9	84.85%	97.11%	82.07%	88.95%
	Class 0	66	118				
QDA	Class 1	0	0	25.66%	0.00%	0.00%	0.00%
	Class 0	368	127				
LR	Class 1	347	121	71.31%	74.15%	94.29%	83.01%
	Class 0	21	6				
kNN	Class 1	294	52	74.55%	84.97%	79.89%	82.35%
	Class 0	74	75				
SVM	Class 1	366	127	73.94%	74.24%	99.46%	85.02%
	Class 0	2	0				
DT-GI	Class 1	366	2	96.19%	99.46%	99.46%	99.46%
	Class 0	2	125				
DT-IG	Class 1	366	2	96.19%	99.46%	99.46%	99.46%
	Class 0	2	125				

Table 18 shows accuracy, sensitivity, specificity and F-measure of proposed ensemble framework for real-time blood CP dataset against other classifiers. It is clear from the analysis of results that proposed DSS (HMV) has achieved high classification and prediction accuracy when compared with other classifiers.

These results again reflect the effectiveness of the proposed DSS. For each patient, the proposed ensemble framework can also store the state of care process such as recommendation done by doctors, patients history related to disease and diagnosis of disease type.

6. Discussion

We have used heterogeneous classifier ensemble model by combining entirely different type of classifiers and achieved a higher level of diversity. Decision Trees, QDA, LR, SVM and Bayesian classification are eager evaluation methods, whereas kNN is a lazy evaluation method. Combining lazy and eager classification algorithms (hybrid approach) overcomes the limitations of both eager and lazy methods. Eager method suffers from missing rule problem in case when there is no matching exists for given test instance. Eager algorithm adopts a solution of default prediction in this scenario. Since choosing a pre-determined class does not take into account any characteristic of the test instance, therefore the solution is too restrictive and lacks diversity. The hybrid approach works similar to eager method, the only difference is that hybrid algorithm generates a customized rule set for each test instance that is not covered by the original rule set and prediction is performed with high accuracy.

Moreover the diversity parameter can be determined by the extent to which each individual classifier disagrees about the probability distributions for the test datasets. The individual Naïve Bayes classifier considers each attribute independently without taking into account the relation between them, whereas HMV model can handle dependency and relations between given attribute set. The linear regression classifier and quadratic discriminant analysis are used in HMV model in order to perform correlation analysis between attribute sets. Thus HMV model resolves the limitation of individual Naïve Bayes classifier by handling correlation. The individual kNN algorithm has various limitations such as it is computationally intensive and requires lots of storage space. The HMV model has resolved the storage problem by selecting only necessary and useful features for disease analysis and prediction. The SVM algorithm performs the feature selection by using only subset of data chosen based on information gain. Moreover SVM algorithm decreases the overfitting issue and increases the prediction and classification accuracy. Thus, all of the selected classifiers complement each other very well. In any scenario where one classifier has some limitation, the other classifier overcomes it and as a result higher ensemble performance is achieved.

Whilst human decision-making performance can be suboptimal and deteriorate as the complexity of the problem increases, the proposed framework can help healthcare professionals to make correct decisions. It should be noted, that an intelligent DSS is not a replacement for doctor or practitioner but it can help them to gather and interpret information and build a foundation for decision-making related to particular disease. The doctor can compare the prediction made by him and the proposed DSS. If these results are same, this adds weight to the diagnosis performed by the doctor but if they are different, further investigations can be performed by doctor.

Following are some advantages of the proposed clinical decision support system.

- The proposed clinical DSS can reduce the margin of diagnostic error. Medical practitioners can use diagnosis support tool as an aid to identify diagnostic possibilities.
- The proposed framework improves efficiency and patient throughput. If clinicians can rapidly determine the diagnosis, they can order relevant tests and make appropriate referrals, saving time and eliminating unnecessary costs for the patients
- The access to all information is available at one place with the help of proposed clinical decision support system. The ability to access the most current medical resources in a central location eliminates the need for multiple logins.

6.1. Limitations

Currently, the proposed ensemble model predicts healthy and sick individuals based on their vital signs. It predicts either class 0 or class 1, representing either absence or presence of a disease. However, the proposed system can be extended to predict the levels and types of particular disease such as for heart disease, it can predict different levels of disease like early, acute, etc. and prediction of type-1 diabetes or prediction of type-2 diabetes in addition of predicting diabetes.

There are multiple data mining techniques that can be used for disease classification and prediction into different levels. Genetic programming (GP) has been vastly used in research in the past

10 years to solve data mining classification problems. The reason genetic programming is so widely used is the fact that prediction rules are very naturally represented in GP. Additionally, GP has proven to produce good results with global search problems like classification. The search space for classification can be described as having several 'peaks', this causes local search algorithms, such as simulated annealing, to perform badly. GP consists of stochastic search algorithms based on abstractions of the processes of Darwinian evolution. Each candidate solution is represented by an individual in GP. The solution is coded into chromosome like structures that can be mutated and/or combined with any other individual's chromosome. Each individual contains a fitness value, which measures the quality of the individual, in other words, how close the candidate solution is from being optimal. Based on the fitness value, individuals are selected to mate. This process creates a new individual by combining two or more chromosomes, this process is called crossover. They are combined with each other in the hope that these new individuals will evolve and become better than their parents. Additionally to matting, chromosomes can be mutated at random.

7. Application for disease diagnosis

A medical application is also developed with the purpose of assisting the physician in diagnosing diseases. It is based on HMV framework and is divided into different modules in order to maintain simplicity and consistency.

There are three main users of the application: Admin staff, Doctor and Patient. Each user has its login id and password in order to interact with the system. There are four main modules of the proposed application:

1. Data Acquisition and Preprocessing Module
2. Classifier Training and Model Generation
3. Disease Prediction Module
4. Report Generation

The data acquisition and preprocessing module is handled by the admin staff member. This module involves entering the patient's data; and preprocess it to remove any anomaly or inconsistency. After preprocessing, classifier training is performed and a model

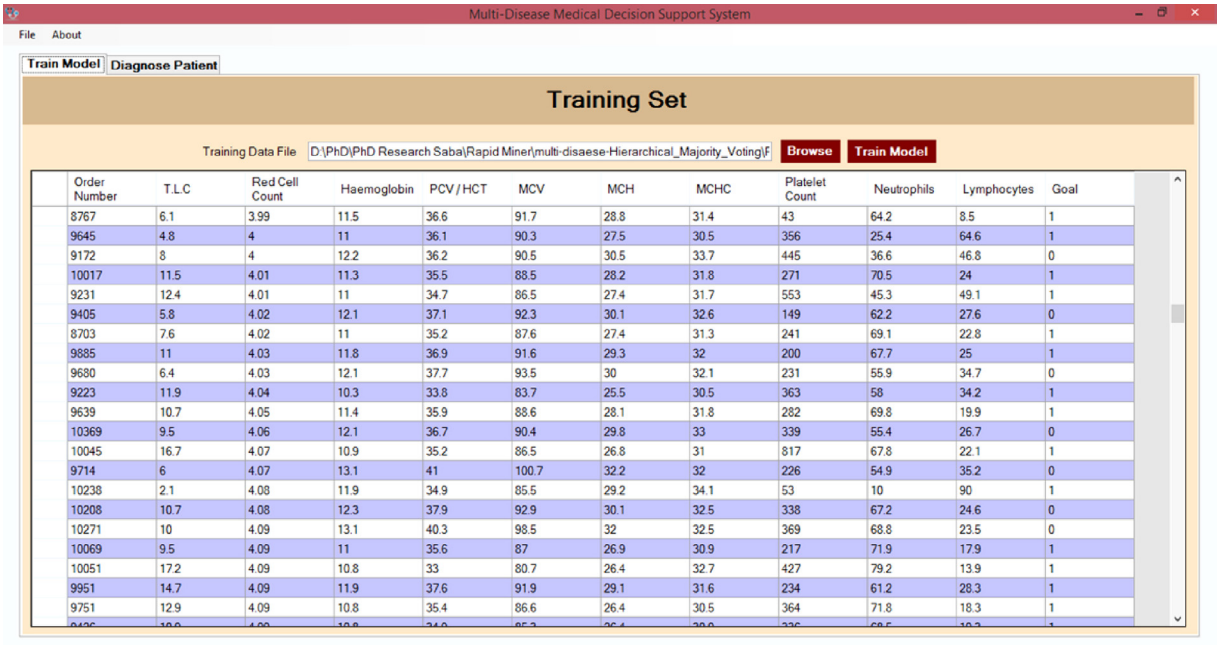

Fig. 8. Load training set and generate model screen.

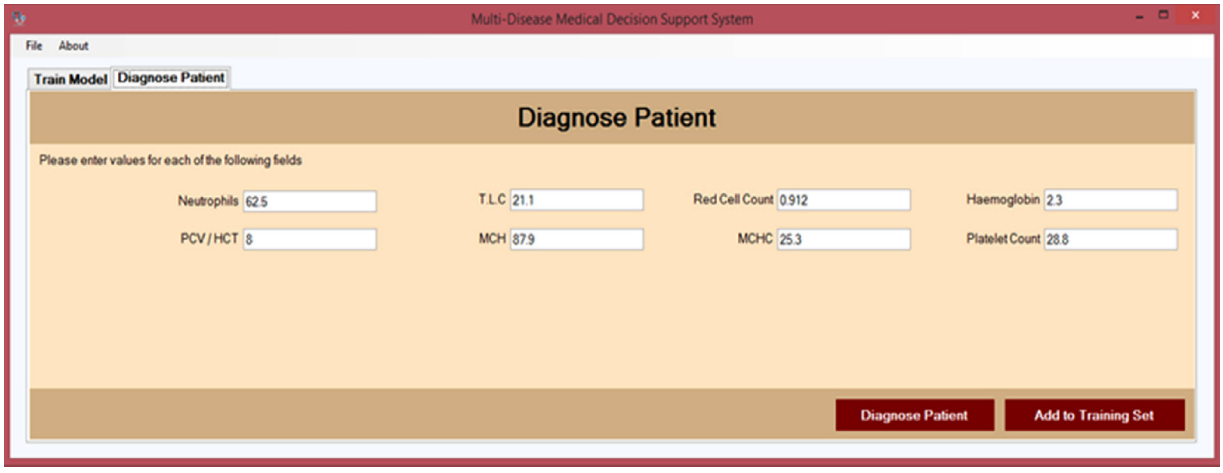

Fig. 9. Data entry screen for disease diagnosis.

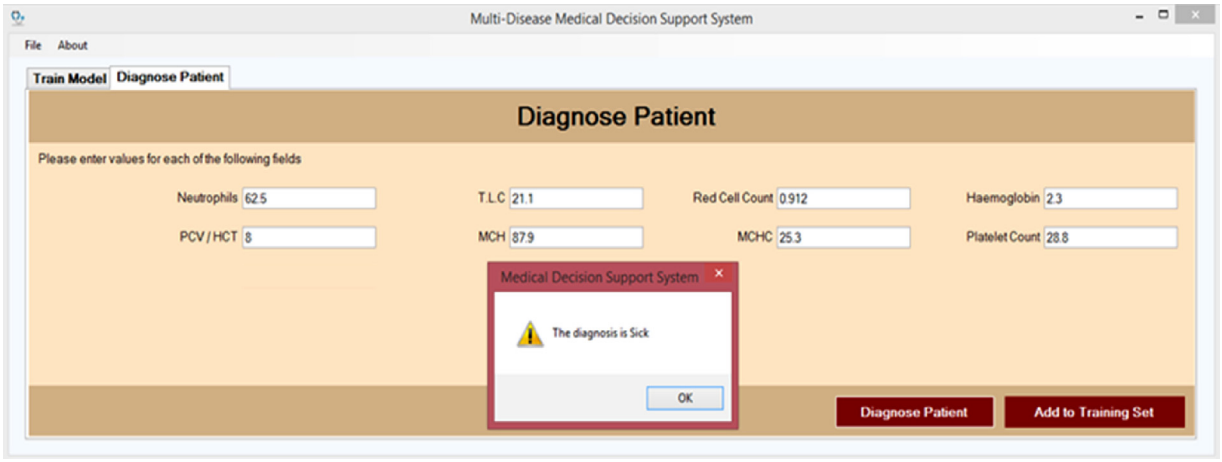

Fig. 10. Disease diagnoses screen.

is generated which performs disease classification and prediction. Disease prediction module shows the prediction whether patient is healthy or sick based on disease symptoms and historical data. Reports generation module generates reports in computerized as well as printed format.

Fig. 8 shows the screen of training set for blood CP dataset that can be used for classifier's training and model generation. When user of the application clicks on "Train Model" tab, following screen will appear showing the data that is loaded for training. When user will click on "Train Model" button, the classifiers will be trained and a training model is generated.

Fig. 9 shows the screen for entering patient's data for disease diagnosis. When user will click on "Diagnose Patient" tab, following screen will appear showing fields where a new patient's information needs to be entered.

After entering patient's information, when user will click on "Diagnose Patient" button, the diagnosis result will be displayed showing either patient is healthy or sick. Fig. 10 shows the screen that will appear when user will click on "Diagnose Patient" button after entering the data. When user will click on "Add to Training Set" button, the information will be added to the training set and a new instance will be created.

8. Conclusion

Accuracy plays a vital role in the medical field as it concerns with the life of an individual. Data mining in the medical domain works on the past experiences and analyzes them to identify the general trends and probable solutions to the present situations. This research paper presents an ensemble framework using hierarchical majority voting and multi-layer classification for disease classification and prediction using data mining techniques. The proposed model overcomes the limitations of conventional performance by utilizing an ensemble of seven heterogeneous classifiers: Naïve Bayes (NB), Linear Regression (LR), Quadratic Discriminant Analysis (QDA), *K* Nearest Neighbor (kNN), Support Vector Machine (SVM), Decision tree using Information Gain (DT-IG) and Decision tree using Gini Index (DT-GI). The proposed framework is based on three modules. The first module is data acquisition and preprocessing which obtains data from different data repositories and preprocess them. Each classifier's training is then performed on the training set in second module and then they are used to predict unknown class labels for test set instances. The prediction and evaluation is the third module of the proposed ensemble framework which is comprised of three classification layers. The evaluation of HMV framework is performed on two different heart disease datasets, two breast cancer datasets, two diabetes datasets, two liver disease datasets, one Parkinson's disease dataset and one hepatitis dataset obtained from public repositories. The analysis of results indicates that proposed HMV ensemble framework has achieved highest accuracy of disease classification and prediction for all medical datasets. Moreover, a real-time implementation of proposed ensemble framework is also performed on blood CP dataset obtained from PIMS hospital in order to determine healthy and diseased patients. The analysis of results again shows high accuracy of disease prediction for real -time patients' data and also it can help practitioners and patients for disease prediction based on the disease symptoms.

Acknowledgements

This research uses real time data from PIMS (Pakistan Institute of Medical Sciences) hospital, Islamabad, Pakistan. We thank Doctor Lubna Naseem and PIMS hospital staff for the data collection and analysis files.

References

[1] C. Fernández-Llatas, J.M. García-Gómez (Eds.), Data Mining in Clinical Medicine, Humana Press, 2015.
[2] J.H. Chen, T. Podchiyska, R.B. Altman, OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records, J. Am. Med. Inform. Assoc. (2015), Ocv091.
[3] S. Dua, X. Du, Data Mining and Machine Learning in Cyber Security, CRC Press, 2011.
[4] A. Ahmad, G. Brown, Random ordinality ensembles: ensemble methods for multi-valued categorical data, Inf. Sci. 296 (2015) 75–94.
[5] B. Sluban, N. Lavrač, Relating ensemble diversity and performance: a study in class noise detection, Neurocomputing 160 (2015) 120–131.
[6] F. Moretti, S. Pizzuti, S. Panzieri, M. Annunziato, Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling, Neurocomputing (2015).
[7] M.J. Kim, D.K. Kang, H.B. Kim, Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, Expert Syst. Appl. 42 (3) (2015) 1074–1082.
[8] S. Kang, S. Cho, P. Kang, Multi-class classification via heterogeneous ensemble of one-class classifiers, Eng. Appl. Artif. Intell. 43 (2015) 35–43.
[9] S.A. Pattekari, A. Parveen, Prediction system for heart disease using Naïve Bayes, Int. J. Adv. Comput. Math. Sci. 3 (3) (2012) 290–294.
[10] S. Ghumbre, C. Patil, A. Ghatol, Heart disease diagnosis using support vector machine, in: International Conference on Computer Science and Information Technology (ICCSIT), Pattaya, 2011.
[11] R. Prashanth, S.D. Roy, P.K. Mandal, S. Ghosh, Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging, Expert Syst. Appl. 41 (7) (2014) 3333–3342.
[12] E.D. Übeyli, Implementing automated diagnostic systems for breast cancer detection, Expert Syst. Appl. 33 (4) (2007) 1054–1062.
[13] F.M. Ba-Alw, H.M. Hintaya, Comparative study for analysis the prognostic in hepatitis data: data mining approach, Int. J. Sci. Eng. Res. 4 (8) (2013) 680–685.

[14] R. Zolfaghari, Diagnosis of diabetes in female population of PIMA Indian heritage with ensemble of BP neural network and SVM, IJCEM 15 (2012).
[15] S. Sapna, a. Tamilarasi, Data mining-fuzzy neural genetic algorithm in predicting diabetes, Res. J. Comput. Eng. (2008).
[16] F. Temurtas, A comparative study on thyroid disease diagnosis using neural networks, Expert Syst. Appl. 36 (1) (2009) 944–949.
[17] D.C. Li, C.W. Liu, S.C. Hu, A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets, Artif. Intell. Med. 52 (1) (2011) 45–52.
[18] S.W. Lin, K.C. Ying, S.C. Chen, Z.J. Lee, Particle swarm optimization for parameter determination and feature selection of support vector machines, Expert Syst. Appl. 35 (4) (2008) 1817–1824.
[19] An introduction to feature extraction, in: I. Guyon, A. Elisseeff (Eds.), Feature Extraction, Springer, Berlin/Heidelberg, 2006, pp. 1–25.
[20] Combining SVMs with various feature selection strategies, in: Y.W. Chen, C.J. Lin (Eds.), Feature Extraction, Springer, Berlin/Heidelberg, 2006, pp. 315–324.
[21] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artif. Intell. Res. (1997) 1–34.
[22] P.J. García-Laencina, J.L. Sancho-Gómez, A.R. Figueiras-Vidal, M. Verleysen, K nearest neighbours with mutual information for simultaneous classification and missing data imputation, Neurocomputing 72 (7) (2009) 1483–1493.
[23] K. CHROMIŃSKI, M. Tkacz, Comparison of outlier detection methods in biomedical data, J. Med. Inform. Technol. 16 (2010) 89–94.
[24] K. Mardia, et al., Multivariate Analysis, Academic Press, 1979.
[25] J.F. Díez-Pastor, J.J. Rodríguez, C. García-Osorio, L.I. Kuncheva, Random balance: ensembles of variable priors classifiers for imbalanced data, Knowl. Based Syst. (2015).
[26] M.A. King, A.S. Abrahams, C.T. Ragsdale, Ensemble learning methods for pay-per-click campaign management, Expert Syst. Appl. 42 (10) (2015) 4818–4829.
[27] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (1–2) (2010) 1–39.
[28] H. Parvin, M. MirnabiBaboli, H. Alinejad-Rokny, Proposing a classifier ensemble framework based on classifier selection and decision tree, Eng. Appl. Artif. Intell. 37 (2015) 34–42.
[29] J. Mendes-Moreira, A.M. Jorge, J.F. de Sousa, C. Soares, Improving the accuracy of long-term travel time prediction using heterogeneous ensembles, Neurocomputing 150 (2015) 428–439.
[30] S. Whalen, G.K. Pandey, A comparative analysis of ensemble classifiers: case studies in genomics, in: 2013 IEEE 13th International Conference on Data Mining (ICDM), IEEE, 2013, pp. 807–816.
[31] S.H. Park, J. Fürnkranz, Efficient implementation of class-based decomposition schemes for Naïve Bayes, Mach. Learn. 96 (3) (2014) 295–309.
[32] H. Hino, K. Koshijima, N. Murata, Non-parametric entropy estimators based on simple linear regression, Comput. Stat. Data Anal. 89 (2014) 72–84.
[33] S. Bose, A. Pal, R. SahaRay, J. Nayak, Generalized quadratic discriminant analysis, Pattern Recognit. 48 (8) (2015) 2676–2684.
[34] D. Lin, X. An, J. Zhang, Double-bootstrapping source data selection for instance-based transfer learning, Pattern Recognit. Lett. 34 (11) (2013) 1279–1285.
[35] S. Datta, S. Das, Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs, Neural Netw. 70 (2015) 39–52.
[36] http://archive.ics.uci.edu/ml/datasets.html (accessed 25.09.2014).
[37] https://www.lri.fr/~antoine/Courses/Master-ISI/TD-TP/breast-cancer.arff (accessed 02.04.15).
[38] http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.html (accessed 25.09.14).
[39] J. Yang, Y. Lee, U. Kang, Comparison of prediction models for coronary heart diseases in depression patients, Int. J. Multimed. Ubiquitous Eng. 10 (3) (2015) 257–268.
[40] P.J. Somasundaram, Probabilistic classification for prediction of heart disease, Aust. J. Basic Appl. Sci. 9 (7) (2015) 639–643.
[41] C. Kiruthika, S.N.S. Rajini, An ill-identified classification to predict cardiac disease using data clustering, Int. J. Data Min. Tech. Appl. 03 (2014) 321–325.
[42] M. Shouman, T. Turner, R. Stocker, Integrating clustering with different data mining techniques in the diagnosis of heart disease, J. Comput. Sci. Eng. 20 (1) (2013).
[43] S. Bashir, U. Qamar, F.H. Khan, BagMOOV: a novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting, Australas. Phys. Eng. Sci. Med. (2015) 1–19.
[44] K.P. Sørensen, M. Thomassen, Q. Tan, M. Bak, S. Cold, M. Burton, M.J. Larsen, T.A. Kruse, Long non-coding RNA expression profiles predict metastasis in lymph node-negative breast cancer independently of traditional prognostic markers, Breast Cancer Res. (2015).
[45] H.K.K. Zand, A comparitive survey on data mining techniques for breast cancer diagnosis and prediction, Indian J. Fundam. Appl. Life Sci. 5 (S1) (2015) 4330–4339.
[46] V. Chaurasia, S. Pal, Data mining techniques: to predict and resolve breast cancer survivability, Int. J. Comput. Sci. Mob. Comput. 3 (1) (2014) 10–22.
[47] V. Chaurasia, S. Pal, A novel approach for breast cancer detection using data mining techniques, Int. J. Innov. Res. Comput. Commun. Eng. 2 (1) (2014).
[48] A.A. K. Aljahdali, S.N. S. Hussain, Comparative prediction performance with support vector machine and random forest classification techniques, Int. J. Comput. Appl. 69 (11) (2013).
[49] J.P. Kandhasamy, S. Balamurali, Performance analysis of classifier models to predict diabetes mellitus, Procedia Comput. Sci. 47 (2015).
[50] S. Bashir, U. Qamar, F.H. Khan, An efficient rule based classification of diabetes using ID3, C4.5 and CART ensemble, in: IEEE Frontier Information Technology, Islamabad, Pakistan, 2015.
[51] K.K. Gandhi, N.B. Prajapati, Diabetes prediction using feature selection and classification, Int. J. Adv. Eng. Res. Dev. (2014).
[52] L. Tapak, H. Mahjub, O. Hamidi, J. Poorolajal, Real-data comparison of data mining methods in prediction of diabetes in Iran, Healthc. Inf. Res. 19 (3) (2013) 177–185.
[53] V. Karthikeyani, I.P. Begum, K. Tajudin, I.S. Begam, Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction, Int. J. Comput. Appl. 60 (12) (2012).
[54] A. Gulia, R. Vohra, P. Rani, Liver patient classification using intelligent techniques, Int. J. Comput. Sci. Inf. Technol. 5 (4) (2014) 5110–5511.
[55] H. Jin, S. Kim, J. Kim, Decision factors on effective liver patient data prediction, Int. J. BioSci. BioTechnol. 6 (4) (2014) 167–178.
[56] H. Sug, Improving the prediction accuracy of liver disorder disease with oversampling, Appl. Math. Electr. Comput. Eng. (2012) 331–335.
[57] B.V. Ramana, M.P. Babu, N.B. Venkateswarlu, A critical study of selected classification algorithms for liver disease diagnosis, Int. J. Database Manag. Syst. 3 (2) (2011).
[58] S. Karthik, A. Priyadarishini, J. Anuradha, B.K. Tripathy, Classification and rule extraction using rough set for diagnosis of liver disease and its types, Adv. Appl. Sci. Res. 2 (3) (2011) 334–345.
[59] E.M.F. Houby, A framework for prediction of response to HCV therapy using different data mining techniques, Adv. Bioinform. 2014 (2014).
[60] T. Karthikeyan, P. Thangaraju, Analysis of classification algorithms applied to hepatitis patients, Int. J. Comput. Appl. 62 (Januray (15)) (2013).
[61] M. Neshat, M. Sargolzaei, A.N. Toosi, A. Masoumi, Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization, Artif. Intell. (2012).
[62] M.V. Kumar, V.V. Sharathi, B.R.G. Devi, Hepatitis prediction model based on data mining algorithm and optimal feature selection to improve predictive accuracy, Int. J. Comput. Appl. 51 (19) (2012) 13–16.
[63] D.A. Khan, F.T. Zuhra, F.A. Khan, A. Mubarak, Evaluation of diagnostic accuracy of APRI for prediction of fibros is in hepatitis C patients, J. Ayub. Med. Coll. Abbottabad (2008).
[64] W. Willis, A. Schootman, M. Kung, N. Evanoff, B.A. Perlmutter, J.S.B.A. Racette, Predictors of survival in Parkinson disease, Arch. Neurol. 69 (5) (2012) 601–607.
[65] S. Hariganesh, S.G. Annamary, A survey of Parkinson's disease using data mining algorithms, Int. J. Comput. Sci. Inf. Technol. 5 (4) (2014) 4943–4944.
[66] G. Yadav, Y. Kumar, G. Sahoo, Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical, and support vector machine classifiers, Indian J. Med. Sci. 64 (6) (2011).
[67] L. Li, U. Topkara, B. Coskun, N. Memon, CoCoST: a computational cost efficient classifier, in: Ninth IEEE International Conference on Data Mining, ICDM'09. IEEE, 2009, pp. 268–277.

Saba Bashir is an Assistant Professor in Computer Science Department at Federal Urdu University of Arts, Science and Technology, Pakistan. She is also a Ph.D. research scholar at NUST, Pakistan. Her research interest lies in predictive systems, web services and object oriented computing. She has published more than 8 research papers in international conferences and journals.

Usman Qamar is the head of Knowledge and Data Engineering Research Centre (www.kdegroup.wordpress.com) at Department of Computer Engineering, College of Electrical and Mechanical Engineering, NUST, Pakistan. He has done his MS in Computer Systems from UMIST, UK whereas his M.Phil., Ph.D. and Post-Doc are from University of Manchester, UK in Data Engineering. His expertise are in Data and Text Mining, Expert Systems, Knowledge Discovery and Feature Selection.

Farhan Hassan Khan has been working as a Project Manager in a software development organization in Pakistan since 2005. He is also a Ph.D. research scholar at NUST, Pakistan. His research interest lies in text mining, web service computing and VoIP billing products. He has published many research papers in international conferences and journals.