Health and Technology (2019) 9:533–545
https://doi.org/10.1007/s12553-019-00299-3

ORIGINAL PAPER

Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach

Dharavath Ramesh [1] · Yogendra Singh Katheria [1]

Received: 21 September 2018 / Accepted: 28 January 2019 / Published online: 5 February 2019

Abstract
Medical datasets have attracted the research community for possible analysis and suitable prediction, which helps the human to take proper precautions in preventing future diseases. To perform related operations, data mining techniques have been widely used in developing decision support systems for disease prediction through a set of medical datasets. This work proposes a new predictive model for disease prediction using pre-processing techniques for various disease datasets. The proposed model not only analyses the datasets also improves the performance by using ensemble methods. To process the datasets, pre-processing techniques such as discretization, resampling, principal component, and decision tree have been used. To classify the datasets, classification techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF) have been used. The algorithms are applied with 10 fold validation technique. A predictive analysis has also been performed on various disease datasets, where every dataset results in significant improvement for various performance measures. We perform a predictive analysis on the datasets such as CKD (Chronic Kidney Disease), Cardiovascular Disease (CVD) or heart, Diabetes, Hepatitis disease, Cancer disease and ILPD (Indian Liver Patient disease). Experimental results show that the proposed predictive model outperforms in terms of better accuracy.

Keywords Disease prediction · Ensemble methods · Machine learning

1 Introduction

Machine learning plays a major role in prediction of diseases from the reacted healthcare datasets [1]. It analyses different attributes and patient lab records. On the basis of a suitable learning strategy, it predicts whether the patient has a certain type of disease or not. It can also predict the severity of the disease by analysing the outcomes of the various attributes or features, which are considered from different health issues or diseases. For example, it predicts by examining the dataset of the patients whether they have cancer or not. And by gaining the knowledge from different features of cancer dataset, it can also predict the type of cancer the patient is having that is whether it is benign or malignant [2]. The machine learning techniques are broadly divided into two categories; supervised and unsupervised [3]. It has different applications and usefulness in predicting the diseases and also analyses the disease datasets. Supervised learning provides the facility to have the result of interest from the related information, whereas unsupervised learning works in a different manner.

Only patients seem abnormal because they have unusual combinations of labs and comorbidity. So we look for interesting structures within the data, not categorized, but related to the properties of the data themselves. This is called learning by without supervision. Automatic learning can help the healthcare analysers with the things such as precision medicine. In fact, automatic learning plays an important role in promoting these efforts to achieve important goals such as helping healthcare evolves. To build a suitable and efficient learning model, we can utilize the information gathered from ponders completed, socioeconomic of patients, medical records, and other sources. The distinction between the customary approach and the

✉ Dharavath Ramesh
drramesh@iitism.ac.in

Yogendra Singh Katheria
y2singh.katheria@gmail.com

[1] Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand 826004, India

programmed learning technique for illness prediction is depending on the number of factors. In a conventional approach, many important factors have not been considered for proper prediction, whereas the substantial number of factors, which brings more prominent precision of wellbeing information has been considered in the machine learning approach. Machine learning has tremendous capability of analysing, visualizing and predicting different kinds of data. Due to its wide applicability in healthcare sector, one can build a machine learning model which can analyse, visualize and predict various kinds of diseases [4]. To accommodate the disease prediction instances mentioned above, in the form of a proposed predictive model, we use different classification algorithms of machine learning for the analysis and prediction of suitable results. For better understanding of results, various performance measures have been considered. We also obtain statistical analysis of various algorithms used in the model.

The rest of the paper is organized as follows. Section 2 describes about the different work done in the past on various disease datasets. In section 3, the methodologies used in our new predictive model for classifying disease datasets are discussed. Section 4 deals with experimental and its results and gives prediction of various performance measures and section 5 conclude the paper.

2 Related works and disease dataset characteristics

2.1 Disease dataset

We look into some of the most vulnerable diseases which are commonly found in patients, so we analyse the datasets for those disease. We collect the datasets for different diseases from UCI repository [5]. We also compare different aspects of ML algorithms on disease datasets for the better prediction of various causes of that diseases in which circumstances lead to particular disease by analysing the attributes of the datasets and finally predicting the severity of the disease in the patient and predicting the patient who are diagnosed with the disease as per the outcome of the performance analysis of different disease dataset using our new predictive model. Diseases we have chosen for our experiment setup are as follows.

- CKD (Chronic Kidney Disease)
- Heart Disease (Cardio Vascular Disease)
- Diabetes
- Wisconsin Cancer dataset
- Hepatitis
- ILPD (Indian Liver Patient dataset)

2.2 CKD (chronic kidney disease)

Chronic Kidney disease (CKD) [6] is a very famous worldwide common health problem, with a hazardous risk to the predictable life of >50%, greater than that of invasive cancer, diabetes, and coronary heart disease. CKD is defined as the presence of renal impairment, revealed by abnormal excretion of albumin or diminished renal function. The disease is measured or estimated by the glomerular filtration rate (FG) which persists for more than 3 months for patients suffering from CKD [7]. The glomerular filtration rate (FG) is the best indicator of how the kidneys work. CKD is analyzed [8, 9] with various machine learning techniques for a better prediction in the literature [10]. Dataset we used for CKD is from the UCI repository [5]. Dataset features, it has 400 numbers of instances and the number of attributes is 25 and there is some missing value and the dataset is analyzed to predict whether the patient has CKD or not CKD.

2.3 Heart disease (cardiovascular disease)

Cardiovascular disease (CVD) [3] is a class of diseases involving the heart or blood vessels. Cardiovascular diseases include coronary artery disease (CAD), such as angina and myocardial infarction (commonly known as a heart attack). Other CVD includes stroke, heart failure, hypertension, rheumatic heart disease, cardiomyopathy, cardiac arrhythmia, a congenital heart defect, valvular cardiac disorder, aortic aneurysms, peripheral artery diseases, disease Thromboembolism and venous thrombosis [11]. In the literature, many heart disease datasets are analyzed by different machine learning techniques [11, 12]. The cardiovascular dataset is the Cleveland Heart dataset that has 303 numbers of instances and the number of features 14 and the number of classes 5 [12]. The goal field refers to the presence of cardiac disease in the patient is an integer evaluated from 0 (no presence) to 4. Experiments with the Cleveland database focused on the simple attempt to distinguish the presence (values 1, 2, 3, 4) from the absence (value 0).

2.4 Hepatitis

Hepatitis is inflammation of the liver tissue [2, 13]. Some people have no symptoms, while others develop yellowing of the skin and the whites of the eyes, lack of appetite, vomiting, fatigue, abdominal pain or diarrhoea. Hepatitis may be temporary (acute) or long-term (chronic) depending on whether it lasts less than six months or more. Acute hepatitis can sometimes be resolved by itself, either by progressing to chronic

hepatitis, or by rarely giving rise to acute hepatic failure. Over time, chronic form can progress to liver scarring, liver failure, or liver cancer. Machine learning plays a vital role in the prediction of hepatitis by its data analysis and there is work done in the literature for the same [14–16]. The Dataset used for analysis is taken from the UCI repository [5]. The features of the datasets are: The number of instances is 155, and the number of features is 20 and the number of classes is 2, and they are categorized as Live or die and there are some missing values that will be frequented using the Dataset's preprocessing.

2.5 Diabetes

Diabetes mellitus (DM), commonly known as diabetes [17], it is a group of metabolic disorders in which there are high levels of blood sugar for an extended period of time. High blood sugar symptoms include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. The classic symptoms of untreated diabetes are weight loss, polyuria (increased urination), polydipsia (increased thirst) and polyphagia (increased hunger). Machine learning plays a vital role in the prediction of diabetes by its data analysis and there is work done in the literature for the same [18, 19]. The dataset is used by the UCI Pima Diabetes Dataset Repository. The number of instances is 768 and the number of attributes are 9 and the class has values such as tested_positive and tested_negative. Some major findings of diabetes prediction methodologies are summarized in Table 1.

2.6 Wisconsin breast Cancer dataset

Breast cancer is a disease that occurs when the cells of the mammary tissue change (or mutate) and continue to reproduce [26]. These abnormal cells are usually grouped together forming a tumor. A tumor is cancerous (or malignant) when these abnormal cells invade other parts of the breast or when they spread (or reproduce) to other areas of the body through the bloodstream or lymphatic system, a network of vessels and knots in the body that carries a PA in the fight Against the infection. The Dataset used here for the Wisconsin breast cancer dataset taken from the UCI repository. Machine learning plays a vital role in the prediction of cancer by its data analysis and there is work done in the literature for the same [27–30]. The number of instances is 699, and the number of features is 10, and the class represents 2 values if the type is benign or malignant.

2.7 ILPD (Indian liver patient dataset)

Liver disease (also known as hepatic diseases) is a type of liver injury or disease [31]. A number of liver (hepatic) function tests are available to test the correct function of the liver. These tests for the presence of enzymes in the blood that is usually more abundant in liver tissue, metabolites or products. Machine learning plays a vital role in the prediction of diabetes by its data analysis and there is work done in the literature for the same [32, 33]. The dataset used in this study is the ILPD dataset [34], taken from the UCI machine learning repository [5]. This Dataset contains 583 records with 11 attributes. It contains 416 records of liver patients and 167 records of non-hepatic patients. The characteristics of the datasets considered are described in the Table 2.

Table 1 Findings of diabetes prediction methodologies

Related works	Methodologies applied	Major findings
Zheng et al. [20]	Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, and SVM	Performed the accuracy related strategy to opt better result. The criteria which is applied in terms of filtering can be further improved.
Pradeep & Naveen [21]	Variant of Decision Tree i.e. J48	The criteria of feature selection has increased the time variant for prediction in a particular area.
Bashir et al. [22]	ID3, C4. 5, & CART Ensembles	Rule-Based Classification strategy has been applied on two datasets.
Guo et al. [23]	Naïve Bayes with its network model	Accuracy of prediction has been measured with network which can be further improved.
Lee et al. [24]	Logistic Regression, Naïve Bayes, and Anthropometry	The accuracy has been measured by Anthropometry based on Glucose levels.
Nai & Moungmai [19]	Logistic Regression, Naïve Bayes, Boosting, Bagging, Decision Tree and Artificial Neural Networks	The accuracy level is recorded with the variant of RF as a maximum value.
Meng et al. [25]	Comparison of three data mining models	Different rules were generated for better prediction

Table 2 Characteristics of disease dataset

Name of the dataset	Number of Instances	Number of Attributes	Number of Missing Values	Number of Class Values
Chronic kidney disease (CKD)	400	25	Yes	2
Heart Disease (CVD)	303	14	Yes	5
Hepatitis	155	20	Yes	2
Diabetes	768	9	Yes	2
ILPD	583	11	Yes	2
Wisconsin Breast Cancer Dataset	699	10	Yes	2

3 New predictive model for analyzing disease datasets

3.1 Predictive model

In the literature, various works have been described with different classification algorithms on few disease datasets [1, 2]. But, no combined efforts have been made, where we can find multiple disease datasets analysed under a common entity. On the other hand, using as many different algorithms with ensemble methods to predict and analyse for obtaining better performance measures have not been made. To accommodate this instance, we consider various disease dataset as previously described into consideration and analysed them by our proposed model.

1In this section, we describe various pre-processing methods, different classification algorithms and ensemble methods which are used in our new predictive model for better prediction of the class in the disease dataset. The disease datasets are taken from the open source UCI repository and are first pre-processed using different pre-processing methods such as: Discretization, Resampling and Principal Component Analysis [5]. On the other hand, the misclassified instances were removed by using Decision Tree algorithm. The traditional classification algorithms such as SVM, Naïve Bayes, KNN, Decision Tree, Random Forest [3, 35] are then again processed with ensemble methods such as Bagging and Boosting [36] to obtain better results and improved performance metrics for different disease datasets. Different techniques for pre-processing, various classification algorithms and ensemble techniques we used for our predictive model are described as follows.

3.2 Pre-processing techniques

3.2.1 Discretization

Discretization is a method to transform the datasets from numeric to nominal value. We use discretization mechanism to transform the disease dataset from numerical values to nominal values, which represents the class labels of the classification problem.

3.2.2 Resampling

It forms a new dataset by producing a subsample of the previous dataset using sampling with replacement. It also used to handle missing values in the datasets.

3.2.3 Principal component

Principal components are used for reducing the number of features of the data. Generally, it is desirable for the set of features to describe a large amount of "information". It helps in reducing features and improving prediction.

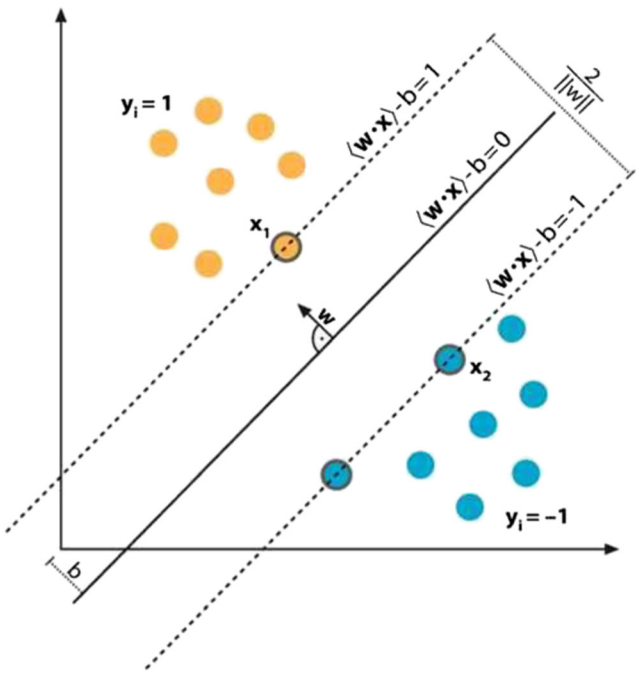

Fig. 1 SVM Classifier

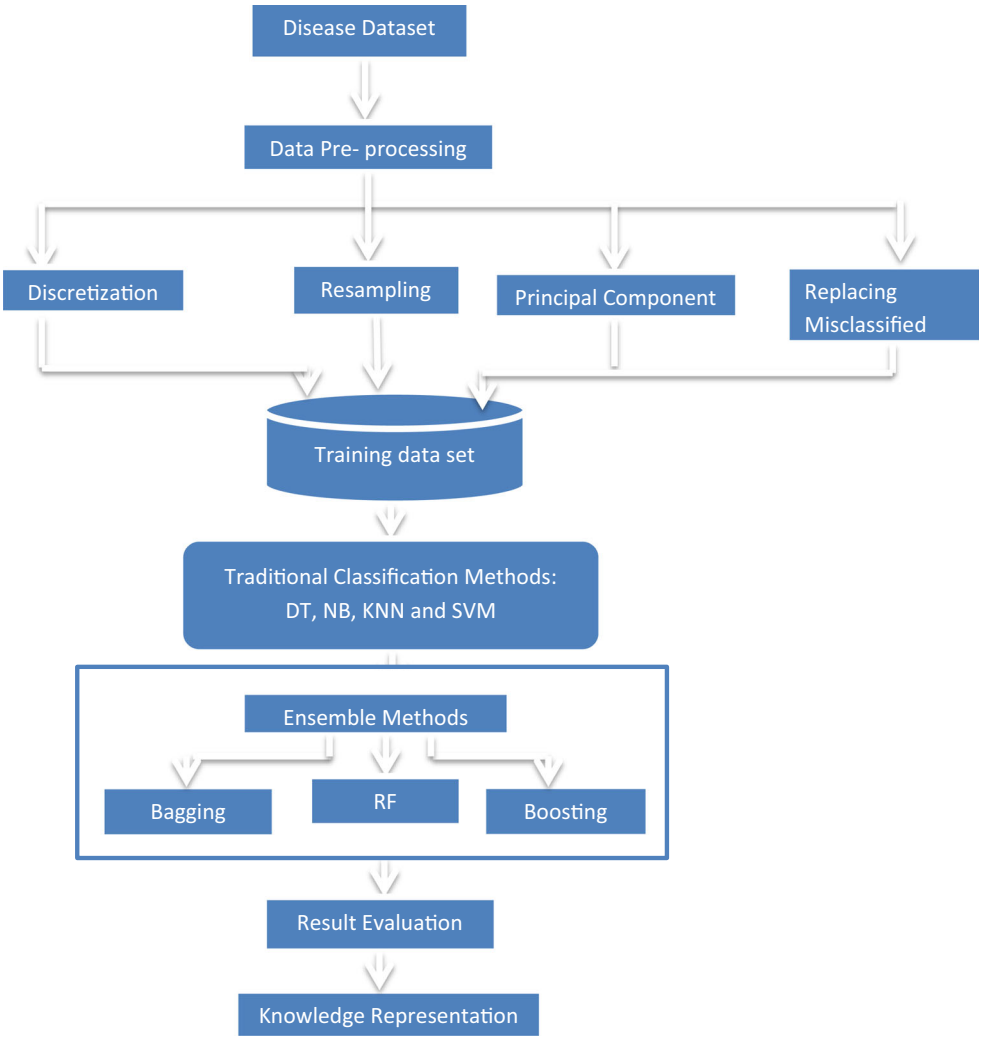

Fig. 2 Flow diagram of predictive model for analysis of disease datasets

3.2.4 Replacing misclassified with decision tree

In this technique, the misclassified instances are firstly identified by using decision tree and then they are removed. This improves the accuracy of prediction of the class labels.

3.3 Classification algorithms

3.3.1 Support vector machine (SVM)

A Support Vector Machine (SVM) is a classifier that tries to **maximize the margin** between training data and the classification boundary (i.e. the plane defined by $X\beta = 0$). The idea is that maximizing the margin **maximizes the chance that classification will be correct on new data**. We assume the new data of each class is near the training data of that type. The instance of the classifier is shown in Fig. 1.

It is formulated as follows:

- **w**: decision hyperplane normal vector
- $\mathbf{x}_i$: data point i
- y_i: class of data point i (+1 or − 1)
- Classifier is:

$$f(\mathbf{x}_i) = \text{sign}(\mathbf{w}^T\mathbf{x}_i + b) \quad (1)$$

Table 3 Predictive analysis of CKD dataset with Traditional algorithms

Evaluation Measure	Traditional Algorithms				
Classifier Type	DT	NB	KNN	SVM	RF
Accuracy	0.97	0.965	0.967	0.983	0.99
Recall	0.98	0.965	0.968	0.983	0.99
Precision	0.98	0.965	0.968	0.983	0.99
F-measure	0.98	0.965	0.968	0.983	0.99

Table 4 Predictive analysis of CKD dataset with Bagging and Boosting

Evaluation Measure	Bagging					Boosting				
Classifier Type	DT	NB	KNN	SVM	RF	DT	NB	KNN	SVM	RF
Accuracy	0.973	0.965	0.96	0.987	0.993	0.983	0.99	0.967	0.987	0.99
Recall	0.973	0.965	0.96	0.988	0.992	0.985	0.99	0.968	0.988	0.99
Precision	0.973	0.968	0.961	0.988	0.993	0.985	0.99	0.968	0.988	0.99
F-measure	0.973	0.965	0.968	0.995	0.992	0.985	0.99	0.968	0.988	0.99

- Functional margin of $\mathbf{x}_i$ is:

$$y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right) \tag{2}$$

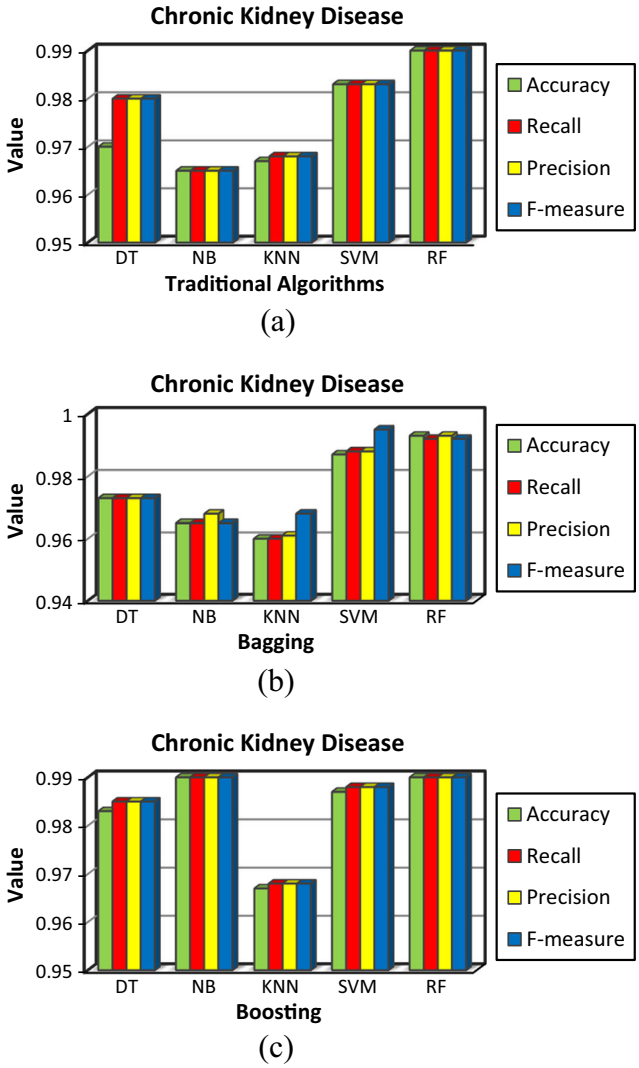

Fig. 3 CKD **a** Traditional Algorithms, **b** Algorithms with Bagging, **c** Algorithms with Boosting

3.3.2 N B (Naïve Bayes) classifier

Bayes Theorem:

P (A|B) = probability of A given that B is true.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{3}$$

B = Data, A = some event.

Naive Bayes classifiers [37–39] are the statistical classifiers designed based on the bayes theorem. These classifiers predict the likelihood of membership in the class, such as the probability that a particular tuple belongs to a particular class. Bayes classifier assumes that a class label is independent of the attributes. It first calculates the frequency of the vaious instances on the basis of class labels and then calculates the probability of the same where the class lablel having highest probabililty value is predicted.

3.3.3 K – Nearest neighbour (KNN)

K-nearest neighbors of a record x are the data points that have the k smallest distance to x [4, 38]. To classify an unknown record, the distance is measured to compute with other training records. The Euclidean distance between two points or tuples, say, x = (×1, ×2,..., xn) and y = (y1, y2,...., yn) is calculated as;

$$Dist(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

Table 5 Predictive analysis of heart (CVD) dataset with traditional algorithms

Evaluation Measure	Traditional Algorithms				
Classifier Type	DT	NB	KNN	SVM	RF
Accuracy	0.885	0.692	0.814	0.877	0.915
Recall	0.885	0.693	0.815	0.878	0.915
Precision	0.887	0.706	0.818	0.894	0.918
F-measure	0.855	0.693	0.815	0.875	0.914

Table 6 Predictive analysis of heart (CVD) dataset with bagging and boosting

Evaluation Measure	Bagging					Boosting				
Classifier Type	DT	NB	KNN	SVM	RF	DT	NB	KNN	SVM	RF
Accuracy	0.911	0.692	0.788	0.833	0.862	0.918	0.862	0.814	0.862	0.915
Recall	0.911	0.693	0.789	0.833	0.863	0.919	0.863	0.815	0.863	0.915
Precision	0.911	0.699	0.790	0.864	0.863	0.918	0.863	0.818	0.866	0.918
F-measure	0.911	0.694	0.789	0.826	0.863	0.918	0.863	0.815	0.862	0.914

Class labels of nearest neighbors is used to determine the class label of unknown record (e.g., by taking majority vote).

3.3.4 Decision tree

J48 [39, 40] is the Weka [21] implementation of the C4.5 software paradigm. This paradigm is used to induce the classification rules in the form of decision trees from a set of given instances. This is a software extension of the basic ID3 algorithm designed by Quinlan which is used to construct a tree. It works on categorical as well as continuous values. The nodes of tree denote different attributes. The branches between nodes represent the possible value of attributes and the terminal node represents the final values of the dependent variables. In literature, it is widely used for disease prediction [40].

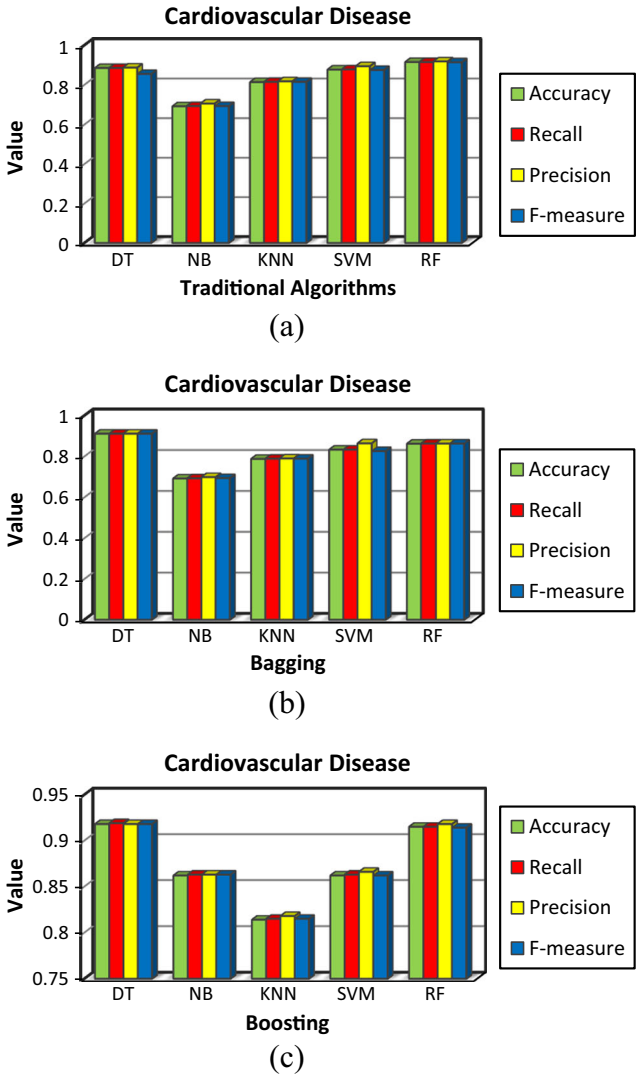

Fig. 4 CVD analysis **a** traditional Algorithms, **b** Algorithms with Bagging, **c** Algorithms with boosting

3.4 Ensemble methods

Ensemble methods construct a set of classifiers from the training data and predicts a class label of previously unseen records by aggregating predictions made by multiple classifiers [36]. These methods use a combination of models to increase accuracy and combine a series of k learned models, M_1, M_2, ..., M_k, with the aim of creating an improved model M*.

3.4.1 Random Forest, bagging, and boosting

Random forest [41, 42] is an extension of bagging which is used for classification or regression. Decision trees are built using a greedy algorithm that selects the best starting point in each step in the process of construction of the tree. Thus, the resulting trees end up looking very similar to reduce the variance of the estimates of all bags which in turn harms the robustness of the predictions. Random forest is an improvement on

Table 7 Predictive analysis of Diabetes (PIMA) dataset with traditional algorithms

Evaluation Measure	Traditional Algorithms				
Classifier Type	DT	NB	KNN	SVM	RF
Accuracy	0.83	0.761	0.889	0.764	0.897
Recall	0.89	0.762	0.889	0.764	0.897
Precision	0.89	0.765	0.89	0.765	0.897
F-measure	0.89	0.763	0.887	0.749	0.897

Table 8 Predictive analysis of Diabetes (PIMA) dataset with bagging and boosting

Evaluation Measure / Classifier Type	Bagging					Boosting				
	DT	NB	KNN	SVM	RF	DT	NB	KNN	SVM	RF
Accuracy	0.869	0.750	0.889	0.776	0.895	0.885	0.750	0.889	0.767	0.901
Recall	0.871	0.762	0.889	0.760	0.896	0.885	0.760	0.889	0.768	0.901
Precision	0.870	0.765	0.890	0.790	0.896	0.887	0.775	0.89	0.768	0.901
F-measure	0.870	0.763	0.887	0.780	0.895	0.883	0.775	0.887	0.765	0.900

bags of decision trees that stops the greedy algorithm division when building the tree.

The ensemble bagging has been constructed by forming a sequence of classifiers which runs a specific algorithm repeatedly on different versions of the training dataset. In other words, bagging is the combination of predictions which exactly provides the same type of occurrences. In the same fashion, boosting has been constructed by framing a sequence of classifiers which runs a learning algorithm repeatedly by changing the distribution of the training set. In other words, this is same as bagging, where the performance of the previous classifier has an effect on new classifier [43].

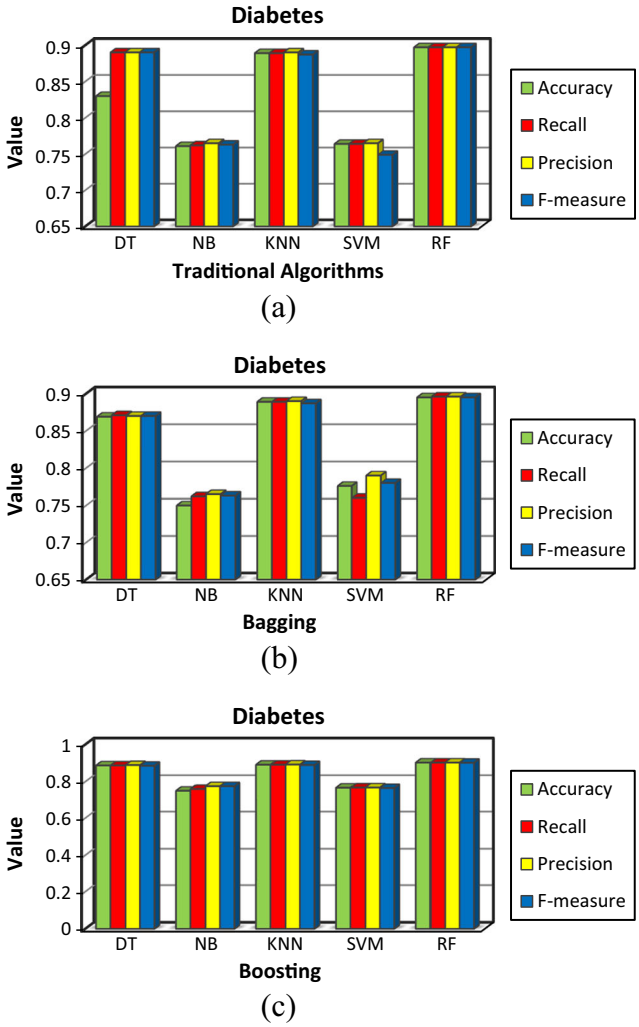

Fig. 5 Diabetes analysis **a** Traditional Algorithms, **b** Algorithms with Bagging, **c** Algorithms with Boosting

3.5 Performance measures

We have used four different measures for the evaluation of the classification quality such as accuracy, precision, recall and F-Measure [41, 42]. These measures can be calculated using confusion matrix given below.

		Labels being Predicted	
		Postive	Negative
True Label	Positive	TP(True Positive)	FN(False Negative)
	Negative	FP(False Positive)	TN (True Negative)

3.5.1 Accuracy

It is a measure to identify the total number of correctly classified instances. It is calculated in the following manner.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

Table 9 Predictive analysis of Hepatitis dataset with traditional algorithms

Evaluation Measure	Traditional Algorithms				
Classifier Type	DT	NB	KNN	SVM	RF
Accuracy	0.957	0.914	0.878	0.842	0.914
Recall	0.957	0.914	0.879	0.843	0.914
Precision	0.957	0.918	0.876	0.843	0.909
F-measure	0.957	0.916	0.877	0.915	0.909

Table 10 Predictive analysis of Hepatitis dataset with bagging and boosting

Evaluation Measure	Bagging					Boosting				
Classifier Type	DT	NB	KNN	SVM	RF	DT	NB	KNN	SVM	RF
Accuracy	0.907	0.935	0.90	0.842	0.886	0.942	0.914	0.878	0.842	0.928
Recall	0.907	0.936	0.90	0.843	0.902	0.943	0.914	0.879	0.843	0.929
Precision	0.903	0.935	0.894	0.843	0.887	0.941	0.912	0.876	0.843	0.929
F-measure	0.904	0.935	0.896	0.915	0.877	0.942	0.913	0.877	0.915	0.922

3.5.2 Precision

It is a measure that describes what portion of the instances are true when they are predicted true.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

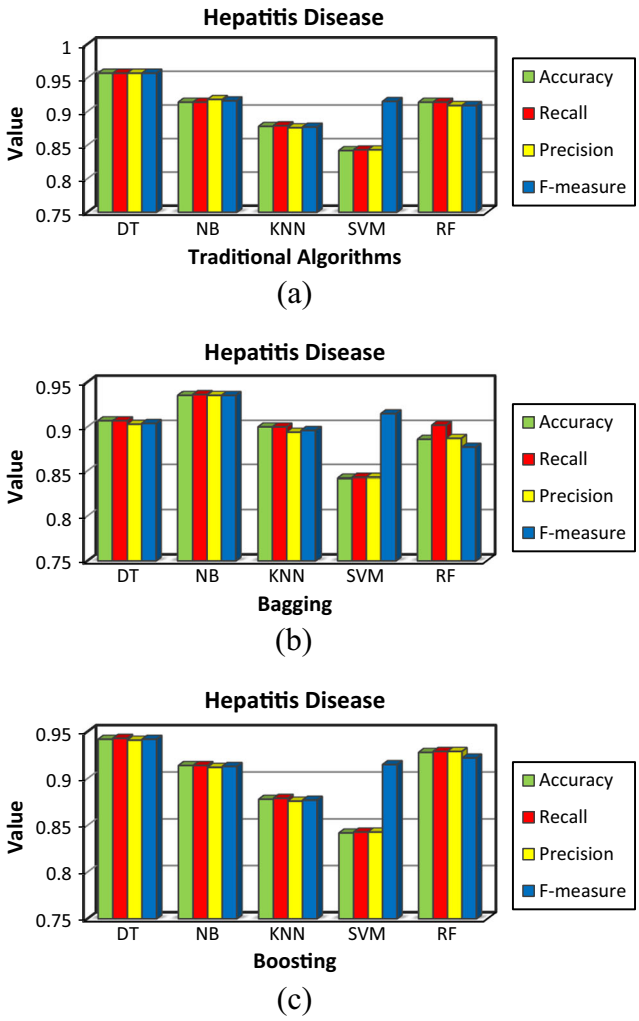

Fig. 6 Hepatitis analysis a traditional Algorithms, b Algorithms with Bagging, c Algorithms with boosting

3.5.3 Recall

It is a measure that describes the number of correctly classified cases to the number of positive cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

3.5.4 F-measure

It is a measure which combines the previous two defined measures Precision and Recall to produce a combined model. It is formulated as follows:

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Fig. 2 illustrates the flow diagram of our proposed ensemble method based predictive model. After the completion of analysis by the ensemble methods, result evaluation is carried out on the basis performance metrics and result obtained is visualized through the graphical representation of the metrics obtained after the prediction.

4 Experimental analysis of predictive model

In this segment, we focus on the experimental analysis of the predictive model of different disease datasets. The whole experiment is carried out in the well-known machine learning tool known as Weka [21]. The disease

Table 11 Predictive analysis of Wisconsin Breast Cancer dataset with traditional algorithms

Evaluation Measure	Traditional Algorithms				
Classifier Type	DT	NB	KNN	SVM	RF
Accuracy	0.972	0.969	0.973	0.967	0.982
Recall	0.972	0.969	0.974	0.968	0.983
Precision	0.972	0.97	0.974	0.970	0.983
F-measure	0.972	0.971	0.974	0.968	0.982

Table 12 Predictive analysis of Wisconsin Breast Cancer dataset with bagging and boosting

Evaluation Measure / Classifier Type	Bagging					Boosting				
	DT	NB	KNN	SVM	RF	DT	NB	KNN	SVM	RF
Accuracy	0.983	0.970	0.973	0.967	0.983	0.985	0.975	0.973	0.969	0.983
Recall	0.984	0.971	0.974	0.968	0.984	0.985	0.975	0.974	0.974	0.984
Precision	0.984	0.972	0.974	0.970	0.984	0.985	0.975	0.975	0.975	0.984
F-measure	0.984	0.971	0.974	0.968	0.984	0.985	0.974	0.975	0.975	0.984

dataset contains nominal as well as numeric data. All the datasets were analysed through the pre-processing phase to ensure the data is converted into the nominal data which in turn useful for the prediction of class value through the proposed model. The traditional algorithms [44, 45] were applied with 10-fold validation technique and the performance is improved by applying ensemble methods with classification algorithms.

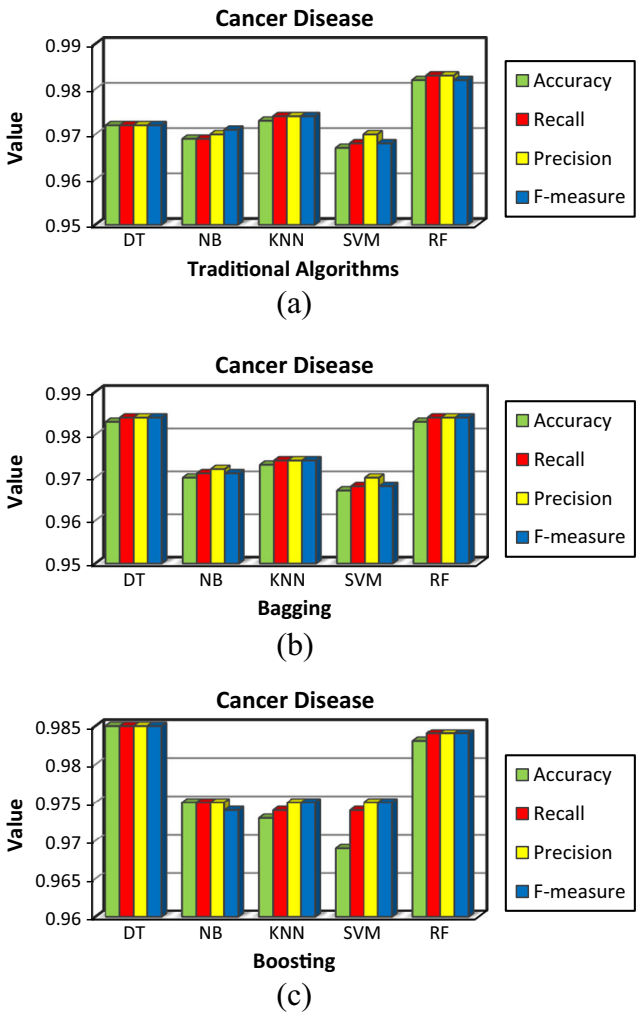

Fig. 7 Cancer disease analysis **a** traditional Algorithms, **b** Algorithms with Bagging, **c** Algorithms with boosting

4.1 CKD (chronic kidney disease) analysis

The CKD dataset contains numerical and as well as nominal data. We applied our model to it and Discretization; Resampling, Principal Component are used as pre-processing methods. The result that we obtained from Bagging and Boosting improves the performance in the case of Naïve Bayes, KNN, SVM and performance is same in the case of Random Forest and Decision Tree but we can also observe there's some margin improvement visible from the result as given in the Table 3 and Table 4.

The table analysis is shown as the graphical representation in Fig. 3. From this, it is observed that the proposed combined model predicts better in comparison to the previous models in the literature [46].

4.2 Heart (cardiovascular disease) analysis

The Heart dataset, also known as Cleveland Heart dataset [12] contains numerical and as well as nominal data. We initially apply our model and perform pre-processing using discretization, resampling and principal component analysis. After applying ensemble methods, it is found that Bagging and Boosting improves the performance in the case of Decision Tree, Naïve Bayes and behaves almost same in case Random Forest. The predicted values are shown Table 5 and Table 6.

The table analysis is shown as the graphical representation in Fig. 4. From this, it is observed that the proposed model predicts better in comparison to the previous models in the literature [11, 47].

Table 13 Predictive analysis of ILPD dataset with traditional algorithms

Evaluation Measure	Traditional Algorithms				
Classifier Type	DT	NB	KNN	SVM	RF
Accuracy	0.807	0.678	0.885	0.891	0.886
Recall	0.808	0.570	0.885	0.892	0.887
Precision	0.812	0.710	0.883	0.896	0.902
F-measure	0.810	0.590	0.883	0.886	0.877

Table 14 Predictive analysis of ILPD dataset with bagging and boosting

Evaluation Measure / Classifier Type	Bagging					Boosting				
	DT	NB	KNN	SVM	RF	DT	NB	KNN	SVM	RF
Accuracy	0.881	0.59	0.855	0.885	0.89	0.862	0.732	0.885	0.891	0.89
Recall	0.882	0.59	0.855	0.885	0.91	0.86	0.732	0.885	0.892	0.89
Precision	0.880	0.716	0.852	0.863	0.89	0.863	0.794	0.883	0.893	0.898
F-measure	0.878	0.608	0.851	0.860	0.91	0.860	0.74	0.883	0.887	0.883

4.3 Diabetes (PIMA) dataset

The Diabetes dataset, also known as PIMA (Diabetes) dataset contains numerical as well as nominal data. We apply our model and perform pre-processing using discretization, re-sampling and principal component analysis. After applying ensemble methods, it is found that Bagging and Boosting improves the performance in the case of Decision Tree, KNN, and SVM. On the other hand, in case of Boosting with Random forest there is significant improvement but in case of SVM there are marginal changes have been obtained. The obtained values are shown Table 7 and Table 8.

The table analysis is shown as the graphical representation in Fig. 5 where, analysis of tradition algorithms is presented with algorithm analysis with Bagging and the algorithms analyzed with the Boosting methods are presented. From this it is concluded that our model predicts better results in comparison to the previous models [17, 48].

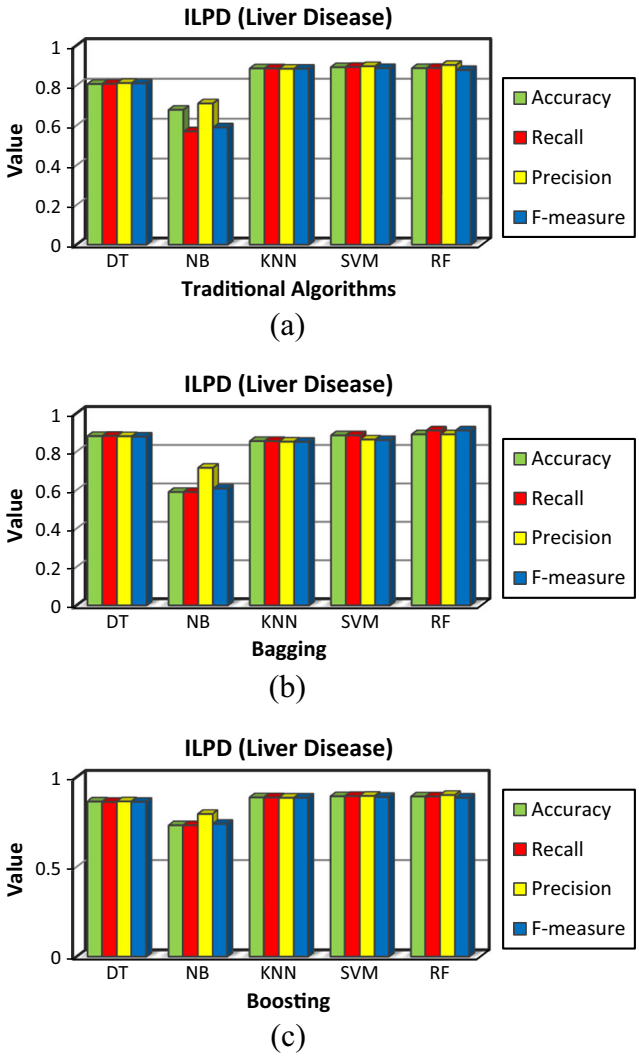

Fig. 8 ILPD analysis **a** traditional Algorithms, **b** Algorithms with Bagging, **c** Algorithms with boosting

4.4 Hepatitis disease dataset analysis

The Hepatitis disease dataset contains numerical nominal data. Discretization, Resampling and Replacing misclassified with Decision Tree is used as pre-processing methods and traditional algorithms were applied which then again used with ensemble methods to improve the performance. From the results, it is found that Bagging improves the performance in the case of Naïve Bayes, KNN and behaves almost same in case SVM and there is some margin improvement seen in case of Random Forest with Boosting. On the other hand, there is no visible improvement seen in case of Decision Tree. The result is presented in Table 9 and Table 10.

The table analysis is depicted as the graphical representation in Fig. 6, where analysis of tradition algorithms is presented with algorithm analysis with Bagging and the algorithms analyzed with the Boosting methods are presented. From this it is found that the proposed model predicts better results in comparison with other existing models [31, 49].

4.5 Wisconsin breast cancer dataset analysis

The Breast cancer dataset, also known as Wisconsin Breast Cancer dataset contains numerical as well as nominal data. We apply our model to perform pre-processing with discretization, resampling and principal component analysis methods and traditional algorithms were applied, which are again used with ensemble methods to improve the performance. From the results, it is found that Bagging and Boosting improves the performance in the case of Decision Tree, Naïve Bayes, and SVM. In the case of RF with Bagging

and boosting there found a marginal improvement. The obtained results are depicted in Table 11 and Table 12.

The table analysis is shown as the graphical representation in Fig. 7, where analysis of tradition algorithms is presented with algorithm analysis with Bagging and the algorithms analyzed with the Boosting methods are presented. From this it is concluded that our model predicts better results in comparison to the previous methodologies existing in the literature [27–29, 50].

4.6 ILPD (Indian liver patient disease) dataset analysis

The Liver dataset also known as Indian Liver Patient dataset [34] contains both numerical and nominal data. The methods such as discretization, resampling and principal component are used as pre-processing methods and traditional algorithms were applied which then again used with ensemble methods to improve the performance. From the results, we found that Bagging and Boosting improves the performance in the case of Decision Tree and behaves almost same in case of Boosting with SVM and KNN. On the other hand, there is some marginal improvement seen in case of Naïve Bayes and Random Forest with Boosting but no significant improvement found in case of Bagging. The related experimental observations are given Table 13 and Table 14.

The table analysis is shown as the graphical representation in Fig. 8, where analysis of tradition algorithms is presented with algorithm analysis with Bagging and the algorithms analyzed with the Boosting methods are presented. From this it is concluded that the proposed model predicts better results in comparison to the previous models [32, 51, 52].

The experiments performed on six disease datasets proves that our new predictive model obtained significant improvement in almost every case as compared with other existing methodologies [13, 41, 42].

5 Conclusion and future work

In this paper, we developed a new predictive model for the disease datasets to improve their performance measures for various traditional classification algorithms with ensemble methods where different pre-processing methods use to handle the numerical as well as nominal data. The performance of classical algorithms with boosting found better where in some cases bagging performed significantly well and in few cases both performed marginally well as compared to the traditional algorithms. And for some cases, they performed almost same and for very few cases only there was no significant changes. Overall, the new predictive model obtains better results as compared to the existing methods. The future work can be compiled by adding new algorithms and using this other different kind of datasets can be further improved.

Compliance with ethical standards

Conflict of interest The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Magoulas GD, Prentza A. Machine learning in medical applications. Advanced course on artificial intelligence. Berlin, Heidelberg: Springer; 1999. p. 300–7.
2. World Health Organization. The Top 10 Causes of Death, 2018. https://www.who.int/news-room/factsheets/detail/the-top-10-causes-of-death.
3. World Health Organization, Cardiovascular, 2017. http://www.mediacentre/mediacentre/factsheets/fs317/en/. Accessed 15 January 2009
4. Godara S, Singh R. Evaluation of predictive machine learning techniques as expert systems in medical diagnosis. Indian J Sci Technol. 2016;9(10):1–14.
5. Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn. 1999;36(1–2):105–39.
6. UCI Machine learning Repository: http://www.archive.ics.uci.edu/mVabout.html.
7. John R, Webb M, Young A, Stevens PE. Unreferred chronic kidney disease: a longitudinal study. Am J Kidney Dis. 2004;5(3):825–35.
8. de Lusignan S, Chan T, Stevens P, O'donoghue D, Hague N, Dzregah B, et al. Identifying patients with chronic kidney disease from general practice computer records. Fam Pract. 2005;22(3):234–41.
9. Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, et al. Definition and classification of chronic kidney disease: a position statement from kidney disease: improving global outcomes (KDIGO). Kidney Int. 2005;67(6):2089–100.
10. Ribeiro RT, Marinho RT, Miguel Sanches J. Classification and staging of chronic liver disease from multimodal data. IEEE Trans Biomed Eng. 2013;60(5):1336–134.
11. Bhatla N, Jyoti K. An analysis of heart disease prediction using different data mining techniques. IJERT. 2012; 1(8).
12. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. Int J Comput Sci Netw Sec. 2008;8(8):1–8.
13. Ho C, Pai T, Peng Y, Lee C, Chen Y, Chen Y. Ultrasonography image analysis for detection and classification of chronic kidney disease. IEEE Complex Intell Softw Intens Syst. 2012; 624–629.
14. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D Top 10 algorithms in data mining. Knowl Inf Syst 14, 1–37, 2008.
15. Kim MJ, Suh DJ. Profiles of serum bile acids in liver diseases. Korean J Intern Med. 1986;1(1):37–43.
16. Adekanle O, Ndububa DA, Olowookere SA, Ijarotimi O, Ijadunola KT. Knowledge of hepatitis B virus infection, immunization with hepatitis B vaccine, risk perception, and challenges to control hepatitis among hospital workers in a Nigerian tertiary hospital. Hepatitis Res Treat. 2015, 1:6.

17. Sharma P, Kaur M. Classification in pattern recognition: a review. Int J Adv Res Comput Sci Softw Eng. 2013;3:298.
18. Kumar Dewangan A, Agrawal P. Classification of diabetes mellitus using machine learning techniques. Int J Eng Appl Sci. 2015;2(5): 145–8.
19. Nai-arun N, Moungmai R. Comparison of classifiers for the risk of diabetes prediction. Proc Comput Sci. 2015;69:132–42.
20. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120–7.
21. Pradeep KR, Naveen NC. Predictive analysis of diabetes using J48 algorithm of classification techniques. Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on. 2016; 347–352). IEEE.
22. Bashir S, Qamar U, Khan FH, Javed MY. An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. 2014 12th International Conference on Frontiers of Information Technology (FIT). 2014; 226–231. IEEE.
23. Guo Y, Bai G, Hu Y. Using bayes network for prediction of type-2 diabetes. Internet Technology Secured Transactions, 2012 International Conf. 2012; 471–472. IEEE.
24. Lee BJ, Ku B, Nam J, Pham DD, Kim JY. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE J Biomed Health Inform. 2014;18(2): 555–61.
25. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013;29(2):93–9.
26. Übeyli ED. Implementing automated diagnostic systems for breast cancer detection. Expert Syst Appl. 2007;33(4):1054–62.
27. Gerson SL, Jensen RA. Patient access to academic cancer centers. J Med Syst. 2018;42(5):86.
28. Gupte A, Joshi S, Gadgul P, Kadam A. Comparative study of classification algorithms used in sentiment analysis. Int J Comput Sci Inform Technol. 2014;5(5):1–4.
29. Polat K, Günes S. Breast cancer diagnosis using least square support vectormachine. Digit Sign Process. 2007;17(4):694–701.
30. Weka https://www.cs.waikato.ac.nz/ml/weka/g.
31. Cleveland Clinic Foundation. Heart disease dataset. http://archive.ics.uci.edu/ml/datasets/Heart+Disease. Date accessed: 22/07/1988.
32. Kirubha V, Priya SM. Survey on data mining algorithms in disease prediction. Int J Comput Trends Technol. 2016;38(3):24–128.
33. Pakhale H, Xaxa DK. A survey on diagnosis of liver disease classification. Int J Eng Techn. 2016;2:2395–1303.
34. Sen SK, Dash S. Application of Meta learning algorithms for the prediction of diabetes disease. Int J Adv Res Comput Sci Manag Stud. 2014;2:396–401.
35. World Health Organization. Diabetes, 2018. https://www.who.int/news-room/fact-sheets/detail/diabetes1.
36. Patil TR, Sherekar SS. Performance analysis of naive Bayes and J48 classification algorithm for data classification. Int J Comput Sci Appl. 2013;6(2):256–61.
37. Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. Healthcare Inform Res. 2016;22(3):196–205.
38. Teli S, Kanikar P. A survey on decision tree based approaches in data mining. Int J Adv Res Comput Sci Softw Eng. 2015;5(4):1–5.
39. Sindhuja D, Priyadarsini RJ. A survey on classification techniques in data mining for analyzing liver disease disorder. Int J Comput Sci Mobile Comput. 2016;5(5):483–8.
40. Kaur R. Using some data mining techniques to predict the survival year of lung cancer patient. Int J Comput Sci Mobile Comput. 2013;2(4):1–6.
41. Romani S, Hosseini SM, Mohebbi SR, Kazemian S, Derakhshani S, Khanyaghma M, et al. Interleukin-16 gene polymorphisms are considerable host genetic factors for patients' susceptibility to chronic hepatitis B infection. Hepatitis research and treatment. 2014, 1:5.
42. Sira MM, Behairy BE, Abd-Elaziz AM, Abd Elnaby SA, Eltahan EE. Serum inter-alpha-trypsin inhibitor heavy chain 4 (ITIH4) in children with chronic hepatitis C: relation to liver fibrosis and viremia. Hepatitis Res Treat. 2014, 1:7.
43. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. Comput Commun (ISCC), 2017 IEEE Symposium. 2017; 204–207. IEEE.
44. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9(01):1–16.
45. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerg Artif Intell Applic Comput Eng. 2007;160:3–24.
46. Mythili MS, Shanavas ARM. An analysis of students' performance using classification algorithms. IOSR J Comput Eng. 2014;16(1): 63–9.
47. Elsayad A, Fakr M. Diagnosis of cardiovascular diseases with Bayesian classifiers. J Comput Sci. 2015;11(2):274–82.
48. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. J Artif Intell Med. 2001;1:89–109.
49. Karabatak M. A new classifier for breast cancer detection based on Naïve Bayesian. Measurement. 2015;72:32–6.
50. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial meta plasticity neural network. Expert Syst Appl. 2011;38(8):9573–9.
51. Ba-Alwi FM, Hintaya HM. Comparative study for analysis the prognostic in hepatitis data: data mining approach. Int J Sci Eng Res. 2013;4:680–5.
52. Singh Y, Bhatia PK, Sangwan O. A review of studies on machine learning techniques. Int J Comput Sci Secur. 2007;1:70–84.