Disease prediction tool: an integrated hybrid data mining approach for healthcare

Megha Rathi
Computer science & engineering
Jaypee institute of information technology,
Noida, India

Vikas Pareek
Computer science & IT
Banasthali University
Banasthali, India

Abstract— **In recent hybrid Data Mining techniques are gaining popularity for disease prediction. In this paper we develop a software tool which will detect any disease at an early stage. An approach is proposed for disease prediction that combines Support Vector machine and Bootstrap bagging using REP tree. We tested the proposed tool on different disease datasets and outcome shows that proposed hybrid technique is proven to be very informative and utilizable in healthcare domain. Disease datasets are first clean and features are extracted using MRMR feature selection algorithm. After that data is classified using SVM classifier and evaluated. If predicted outcome is positive then it will pass for validation otherwise we again classify remaining data using Bagging classifier. Results are then passed for validation along with previous results. Accurate results have been obtained which proves that proposed tool is effective in disease diagnosis within no time.**

Keywords-disease diagnosis; SVM; bootstrap bagging; hybrid approach.

I. INTRODUCTION

In today's fast growing world people are becoming more and more prone to diseases, whether they live in developed or third world countries. The tremendous advancement in technology has led medical information systems in hospitals and medical institutions become larger and larger and in turn the process of extracting useful information is becoming more and more difficult and time consuming. Now a day's more number of institutions are changing from keeping the paper based records of the patient to Computerized Patient Records (CPR) [1] as relying on Manual data analysis has become costly and inefficient. We now require a technology advancement that is not only efficient but also easy to use hence we require computer based analysis which can effectively diagnose the patient's problem.

Data mining and knowledge discovery from data (KDD) is the process of extracting knowledge from large amounts of data [2] and have been successfully applied to different classification tasks including, but not limited to, decision making, fault detection, pattern recognition, weather forecasting and image processing. Extracting knowledge from data aims at building a model from the data to predict the future behavior.

Before Data Mining the CPR was known as Electronic Patient Records and was just a centralized data stored which provides very limited possibility to analyse and process the data and more overly the data was just capable of diagnosing only simple cases. Data mining is the key to evaluate, interpret and the huge amount of data and processes the query with great accuracy and increases the CRM value [3].

Data Mining uses algorithms and tools to convert lifecycles of knowledge, and formalizations to extract patterns, information and knowledge extracted from the data stored in the CPR. Thus we can say that DM is useful in transforming the transactional data in the CPR from a mare tacit knowledge into more useful and efficient explicit knowledge [3].

DM is not only knowledge creation but also a set of techniques for data analysis, it is the key for extracting information out of the huge data set. Without Data Mining, having data as CPR is not necessary as it makes no difference to diagnosis. What Data Mining does is to build a group of heterogeneous tools and techniques to different purposes along the process to create the knowledge. They use both descriptive and predictive models. The descriptive model help in identifying similar patterns in the analyzed data by using classification, association rules, and visualization on the other hand Predictive model uses classification, regression and time series analysis to show the impact of a treatment to a patient based on the data of the past. We can summaries that data mining techniques distinguishes the by using model building and clustering techniques [4].

We know the fact that medical decisions must always be supported by explanation, arguments for predicting the correctness of making a particular decision/diagnose, and hence Data Mining has become area of a great interest for clinical practice and research. Data Mining is one of the most important technologies which enable Evidence Based Medicine which proposes the strategy to apply evidence gained from scientific study of patients.

These inconsistency and consideration has prompted us to look into the case and finding out a way to predict faster with an increased accuracy. Hybrid approach was used to classify these instances. The results obtained from the analysis help physicians to treat the patients with more meaningful and resulting way.

II. LITERATURE REVIEW

Data mining is the process of analyzing huge or voluminous amount of data in various perspectives in order to bring about trends or patterns leading to business intelligence [5]. Data mining also plays an important role in IT as it can discover knowledge from the historical data of various domains or fields. Data mining can also be used to mine medical data as we are aware that Healthcare domain produces huge amount of data about patients, diseases, diagnosis and so on. The knowledge which is discovered by the application of data mining techniques can be used by the healthcare administrators to improve the quality of service [3]. Data mining is also useful in many medical applications including medications, medical tests, prediction of surgical procedures, and discovery of relationships between pathological data and clinical data [6].

MS Uzer et al. [7] had developed a hybrid breast cancer detection system via neural network by using feature selection based on SBS, SFS and PCA. Authors have opted for hybrid feature selection method, artificial neural network was used as a classifier in order to improve its performance and accuracy. They also have used ten-fold cross validation technique to validate their results with 98.57% accuracy.

In the study author [8] presented to use local linear wavelet neural network for recognizing breast cancer by training its parameters using Recursive Least Square (RLS) approach in order to improve its performance. The average correct classification rate of proposed system is 97.2 %.

In the research [9] author proposed a modified Support Vector Machine (SVM) which is based on rough set attribute reduction and using RS, as an anterior preprocessor of SVM, to find out these relevant features which influence the medical disease, and make great use of the advantages of RS in eliminating redundant information. The technique was practiced on three datasets which shows an increase in accuracy if compared to greedy method to choose classifier.

Research [10] suggested a decision making support system which is based on Fuzzy Logic and Fuzzy Decision Trees to increase performance of classification by the introduction of new type of FDT-stable FDT which is both simple to understand and apply. The split ratio of 70:30 was used to test the results on two breast cancer datasets (Breast Cancer and Breast Cancer-Wisconsin) and the error rates were 0.3661 and 0.1414 respectively.

Paper [11] presented a system which was developed for remote health monitoring of patients suffering from Heart Failure. It also included advanced functionalities of Data Mining for continuous patient monitoring. The aim of the system was the early detection of any worsening in patient's condition, using automatic "Heart Failure severity assessment" which used Data Mining via CART classifiers, with the assumption that during worsening, patients will gradually show characteristic of a more severe Heart Failure. The results were obtained testing the system with biomedical signals from public databases. The system had achieved accuracy and a precision of 96.39% and 100.00% in detecting HF and of 79.31% and 82.35% in distinguishing severe versus mild HF, respectively.

In the paper [12] author put forward the use of boosting in order to increase the accuracy of CAD-based techniques so as to solve the breast cancer characterization problem. The authors have also proposed a hybrid boosting algorithm which combines the advantages of several boosting techniques. With the application of the different boosting techniques investigated on real breast cancer benchmarks show that the hybrid boosting algorithm outperforms the other boosting techniques on average by 48%.

Research [13] has presented a hybrid system which combines the genetic algorithm and random forest for diagnosing lymphatic diseases. The genetic algorithm which was used as a feature selection technique for reducing the dimension of the lymphatic diseases dataset and the technique of random forest was used as a classifier. The proposed system performance was compared with that of other feature selection algorithms combined with RF classifier such as principal component analysis (PCA), Relief, Fisher, sequential forward floating search (SFFS), and the sequential backward floating search (SBFS). The result of the experiments performed has shown that GA–RF has achieved a high classification accuracy which is 92.2%.

This paper [14] presented a model to predict the risk score of heart disease in the state of Andhra Pradesh with reduced no. of attributes. It also used feature selection measures such as SU, IG and genetic search to determine the specific attributes that contribute more towards the prediction of heart disease and hence indirectly reduces the no. of diagnosis tests which the patient needs to take. In addition they have also used associative classification in order to improve the accuracy of classification and have this very approach on public datasets and have compared it with other existing solutions.

In this study [15] author had used supervised learning algorithms of Artificial Neural Network for predicting diabetes. The network was trained to use the data of 250 patients suffering from diabetes. The use of Scaled Conjugate Gradient algorithm with value of R=0.88026, had produced the best prediction performance in comparison to the other algorithms.

Ö. Akin [16] had selected a resampling strategy which is based upon RF ensemble classifier to improve diagnosis of cardiac arrhythmia. The strategy was found to be 90.0% accurate, and the experiments were able to demonstrate the efficiency of random sampling strategy in training RF ensemble classification algorithm.

III. METHODOLOGY

The proposed framework consists of four main modules namely: Data Collection, Data Pre-processing, Feature Selection, and Classification. First step is the collection of data set. One data set is collected from one of the known hospital of India and rest is collected from UCI ML Repository. Data Pre-processing is the second step in which we process the data set so that high quality data is available which is error free. Data should be non ambiguous, correct, and complete because classification accuracy depends on the quality of data. Data Pre-processing is applied to remove inconsistencies from the data set, also to fill missing values. Data set obtained from different sources contains redundant and irrelevant data. In order to remove such kind of inconsistencies from the data set we apply data cleaning techniques. In third module we select the relevant features out of the given data set so that dimensionality of data set is reduced. Feature Selection is the process of selecting the subset of features or attributes that is inputted to the system. Accuracy of classifier depends upon the features of data set which contribute to the prediction of breast cancer. In this study features are selected using MRMR (Maximum Relevance and Minimum Redundancy) [17] algorithm. Minimum redundancy is a feature selection algorithm which is used for the identification of some very important characteristics of phenotype and genes and reduces their relevancy [17]. This algorithms works by selecting those features which were mutually far away and having "high" correlation to the classification variable. In last module we apply classification algorithm, we combined two algorithm SVM and Bootstrap aggregation using REP tree and applied on disease database for disease detection for better results.

A. *Data Preprocessing and Feature Selection*

Data Preprocessing is the important step as quality result always depends on the quality of data. Data preprocessing is used to remove inconsistency from data set, also to fill missing values. As we gathered data from various different sources it may contain erroneous entries also data is having different file format with attributes consists of irrelevant and redundant data. For removing all such anomalies we apply data cleaning techniques. In this research we fill attribute values with the mean value for unknown instances and extract relevant features using MRMR feature selection algorithm. We have done this process by selecting combination of different attributes then observed that we are obtaining as classification rate as the initial set of variables.

B. *Algorithms Used*

- Support Vector Machine: SVM (Support Vector Machine) is a kind of supervised learning algorithm and widely used in medical diagnosis for classification and regression [18]. It is the nature of this algorithm that it reduces the classification error by maximizing geometric margin.SVM that is the reason why it is also known as Maximum Margin Classifiers.SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory i.e. the so called structural risk minimization principle.

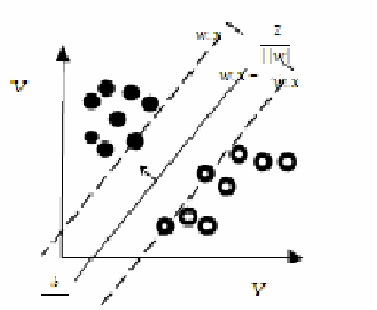


Figure 1. SVM Hyperplane

SVM represents example data as points in space, then mapped data into that feature space so that example data which belongs to different categories are divided by a gap that is as wide as possible [18,19]. Consider an instance, given a set of points belonging to either one of the two classes, an SVM finds a hyperplane having the largest possible fraction of points of the same class on the same plane. This separating hyperplane is called the optimal separating hyperplane (OSH) that maximizes the distance between the two parallel hyper planes and can minimize the risk of misclassifying examples of the test dataset.

Given some training data ‘D’, a set of *n* points of the form

$$\mathcal{D} = \left\{(x_i, y_i) \mid x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\right\}_{i=1}^{n}$$

Value of is 1 or -1 which indicates the class to which point belongs. Basically is a real vector of dimension ‘p’. We are interested in a hyperplane that maximize margin and separate points with value = -1 from those having =1. Hyperplane with set of points ‘X’ is defined s: “W.X-b=0”.
where ‘w’ is the normal vector to hyperplane. Offset of hyperplane with respect to normal vector ‘w’ is:

$$\frac{b}{||W||}$$

In case of linearly separable training data we can choose two hyper planes in a manner that they can separate data with no points between them and then try to maximize their distance. The region bounded by them is called "the margin". These hyper planes can be described by the equations:

$$W.X-b=1,$$
and
$$W.X-b=-1,$$

From the above given two equations distance between two hyper planes is described by , so main objective is to minimize ||W|| so that we can prevent data points from falling into margin. Thus we add the following constraint: for each 'i' either

$W.X_i$-b ≥1 for X_i of the first

OR

$W.X_i$-b ≤ -1 for X_i of the second.

Rewrite the equation as:

$$Y_i(W.X_i - b) \geq 1, for\ a \quad (1)$$

We can put this together to get the optimization problem:Minimize (W,b) ||W|| subject to (for any i=1,2….n) Yi. (W.Xi-b) ≥ 1.

- Bagging: Bootstrap aggregating also known as bagging, is one of the most widely used Machine Learning algorithm which is basically used for improving accuracy and stability of algorithms and is deployed in statistical classification and regression. It works by reducing variance for avoiding Overfitting. Mostly deployed with decision tree methods but can be applied with any machine learning techniques. It is one of the special cases of model averaging approach.

  Suppose there is Dataset (training dataset) 'D' of size 'n' , bagging approach creates 'm' new training datasets each of size 'n', suppose newly created datasets are designated as 'D_i ' and generated after uniform sampling from D with replacement. When we generate new training examples by sampling with replacement some observations may be repeated in D_i. In case n' = n, for large value of 'n' D_i is equivalent to have fraction (1-1/e) (≈ 64%) that is only 64% contains unique examples rest being duplicates. This type of sampling is known as Bootstrap Sampling. Total 'm' bootstrap samples are used to fit 'm' models and combined by averaging the output (for regression) or voting (for classification) [20,21].

- REP (Reduced Error Pruning) Tree: A Decision tree is a tree like structure of decisions along with the consequence of each and every decision. It is a kind of flowchart model in which internal nodes of tree represents test condition on an attribute , each branch represents outcome of that test condition and leaf nodes represents class label which is the decision obtained after computation of all attributes. Classification rule in decision tree is the path from root to leaf.

  Pruning is a machine learning technique which is used for size reduction of decision tree. Size of decision tree is reduced by eliminating sections of the tree that add little weight age to classify instances. Main objective is to reduce complexity of classifier as well as to provide better predictive accuracy by reducing Overfitting and removal of sections of classifier that may based on noisy data.

  Reduced error pruning is one of the simplest type of pruning technique [22]. It works as: Starting from leaf nodes, each node of tree is replaced with its most popular class. If accuracy of prediction is not affected then this change is saved. REP has advantage of simplicity and speed.

- MRMR: Minimum Redundancy Maximum Relevance (MRMR) [17] is a feature selection algorithm which is used to identify the characteristics of medical datasets and minimize their relevance accordingly described with its pairing with relevant feature selection algorithm. This algorithms works by selecting those features which were mutually far away and having "high" correlation to the classification variable. In machine learning feature extraction is the important subfield which selects subset of data that are relevant to particular problem domain and is known as Maximum Relevance. Sometimes these subsets of data contain redundant entries and MRMR aims to remove those redundant entries. MRMR has variety of applications like speech recognition and Cancer Diagnosis. Using this algorithm we can extract features in many ways. One method of feature extraction is to choose features that correlate strongest to classification variable and is known as Maximum-relevance Selection. Other scheme is to select those features that are mutually far away from each other having high correlation to classification variable. This technique is known as "Minimum Redundancy Maximum Relevance" and is found those features subsets are more influential than Maximum relevance technique. In few specialized cases correlation can be replaced by statistical dependency between variables. In such case, MRMR works as to maximize the dependency between mutual distribution of selected features and classification variable.

C. *Data Description*

We have used three different datasets to validate our approach: Diabetes Dataset (A): We have used a dataset of 1171 tuples which was gathered from one of the known hospital in India, but due to the sensitive nature of the database it cannot be made public. Each record in the database has seven attributes. In this database, eight hundred and forty-three samples of the dataset belong to non-diabetic class, and three hundred and twenty-eight samples of the dataset are of diabetic class. The seven attributes are detailed in Table 1. Our dataset has 7 attributes excluding class label which are as follows:

TABLE I. DIABETES DATASET DESCRIPTION

Attribute No.	Attribute Description	Values of attributes
1	Gender of a patient	M,F.
2	Plasma glucose concentration in an oral glucose tolerance test	[44-385].
3	Diastolic blood pressure	[24-124].
4	Skin fold thickness	[7-99].
5	Serum insulin	[14-846].
6	Body mass Index	[15.171-67.100].
7	Age of patient	[19-92].

Wisconsin Breast Cancer Dataset (B): We have used Wisconsin Breast Cancer Database (taken from UCI machine learning repository) which was obtained from University of Wisconsin Hospitals, Madison from Dr.William H. Wolberg in our experiments. This Dataset is commonly used among researchers who use machine learning methods for breast cancer classification. There are 699 tuples in this database. Each record in the dataset has nine attributes. In this database, four hundred and fifty-eight samples of the dataset contain benign class and two hundred and forty-one samples are malignant class. The nine attributes are detailed in Table 2. [23].

TABLE II. BREAST CANCER DATASET DESCRIPTION

Attribute No.	Attribute Description	Values of attributes
1	Clump Thickness	1-10.
2	Uniformity of Cell Size	1-10.
3	Uniformity of Cell Shape	1-10.
4	Marginal Adhesion	1-10.
5	Single Epithelial Cell Size	1-10.
6	Bare Nuclei	1-10.
7	Bland Chromatin	1-10.
8	Normal Nucleoli	1-10.
9	Mitoses	1-10.

Hypothyroid Dataset (C): We have used the hypothyroid database (taken from the UCI machine learning repository). This dataset is common among researchers who use machine learning methods for thyroid classification. There are 3163 records in this database. Each record in the database has twenty-five attributes. In this database, three thousand twelve instances of the dataset belong to negative class those who are not suffering from hypothyroid and one hundred and sixty-one instances of the dataset are of hypothyroid class. The twenty five attributes are detailed in Table 3. [24].

TABLE III. HYPOTHYROID DATASET DESCRIPTION

Attribute No	Attribute Description	Values of Attributes
1	Age	[1-98].
2	Sex	M,F.
3	On_thyroxine	f,t.
4	Query_on_thyroxine	f,t.
5	On_antithyroid_medication	f,t.
6	Thyroid_surgery	f,t.
7	Query_hypothyroid	f,t.
8	Query hyperthyroid	f,t.
9	Pregnant	f.t.
10	Sick	f,t.
11	Tumor	f,t.
12	Lithium	f,t.
13	Goitre	f,t.
14	TSH_measured	n,y.
15	TSH	[0-530].
16	T3_measured	n,y.
17	T3	[0-10.2]
18	TT4_measured	n,y.
19	TT4	[2-450].
20	T4U_measured	n,y.
21	T4U	[0-2.21].
22	FTI_measured	n,y.
23	FTI	[0-881]
24	TBG_measured	n,y.
25	TBG	[0-122].

IV. PROPOSED FRAMEWORK

In this research we presented an integrated framework for disease prediction generic to all types of disease using data mining techniques. We developed a software tool known as “Disease Prediction tool” with the help of Java Netbeans Interface which is able to diagnose any disease at initial stage. First step is the collection of disease datasets. We have collected datasets form two sources: One from the known hospital of India (Diabetes Dataset) while rest is collected from UCI ML Repository. After collection of datasets Data Preprocessing is the second step in which we clean data for classification purpose as quality outcome depends on quality data it is essential to remove anomalies from all datasets. For Data preprocessing unknown or miss instances are filled with mean values. And by feature selection removed some attributes. We have done this process by selecting combination of different attributes then observed that we are obtaining as classification rate as the initial set of variables. For Feature Extraction MRMR algorithm is used to extract features which account more in classification. After data preprocessing data is split into two: Training data set and test dataset with split ratio of 75 : 25 that is 75% train data set and 25% test dataset. Training examples are trained by SVM and Bagging classification algorithm. Test data is first classified by SVM classifier and classification performance is evaluated. If predicted value comes out to be positive then it will be passed for validation else test data is again classified using Bagging classifier. We observed significant increase in accuracy using proposed hybrid approach. Figure 2 shows the overall architecture of proposed system. Proposed hybrid approach is prove to be very useful in healthcare domain as it diagnose disease at initial stage and prediction accuracy also enhances because of the use of two Data mining algorithms together. If a tool is available in healthcare industry which diagnose disease and is generic to all types of disease, just after inputting some values one can easily predict life threatening disease like Cancer at initial stage and without doctor’s intervention. As disease is diagnosed at initial stage, right treatment started at right time and chances of death also decrease in case of disease like Cancer and Heart related problems. Various phases of proposed approach are listed below:

A. *Data Collection*

We collect data from two sources on data set from the known Hospital of India and rest are collected from UCI ML Repository.

B. *Data Cleaning and Data Preprocessing*

As we collect data from different sources it is essential to integrate all the data and convert in a same file format. Data cleaning techniques are applied on all data sets to remove inconsistent, missing, erroneous, incorrect entries from the datasets. We fill all unknown instances by putting mean values.

C. *Feature Selection*

Once data is clean and error free we can select relevant features form the entire data set to improve classifier accuracy. Irrelevant features are those which does not play any significant role while classification. We have used MRMR algorithm for selecting relevant features form the database.

D. *Classification*

Data is divided into training (75%) and test (25%) data sets. For the classification purpose we have two algorithms: SVM and bagging. We combined both the algorithm to improve the prediction accuracy of proposed system. Classifiers are trained to predict any disease with in no time.

E. *Hybrid Approach*

Training examples are passed for classification learning by bagging and SVM. While predicting test examples, they first are classified by using SVM classifier and then evaluated. If predicted value is comes out as positive then it will be passed for validation else the remaining portion of test examples will be classified again by bagging classifier. The results are then passed for validation along with the previous results.

F. *Validation*

We have tested our approach on different disease datasets and it is found that our approach works well than other existing hybrid approaches in disease detection. Moreover the tool is not specific to any disease type it is able to detect all types of diseases. In next section we have presented our result which shows that proposed approach achieves desired result and proven to be very fruitful in healthcare domain.

V. PROPOSED FRAMEWORK

A. *Feature Extraction using MRMR Algorithm*

Table 4, 5, and 6 presents the snapshot of features selected for three different datasets. After relevant features are selected we apply Classification algorithm for checking the performance of hybrid classifier.

TABLE IV. DIABETES DATASET AFTER FEATURE SELECTION

Selected Features of Diabetes Dataset
Feature Selection Method: MRMR Total Number of instances: 569 with 8 attribute(one class attribute) Selected Attribute: Attribute Id:2,3,5,7 Attribute Name: plasma_glucose, diastolic blood_pressure, serum insulin, age.

TABLE V. BREAST CANCER DATASET AFTER FEATURE SELECTION

Selected Features of Breast Cancer Dataset
Feature Selection Method: MRMR Total Number of instances: 699 with 10 attribute(one class attribute) Selected Attribute: Attribute Id:1,2,3,4,5,6,7,8,9 Attribute Name: Clump_Thickness, Cell_Size_Uniformity, Cell_shape_Uniformity, Marginal Adhesion, Single_Epi_Cell_size, Bare_Nuclei, Normal_Nuclei, Mitoses.

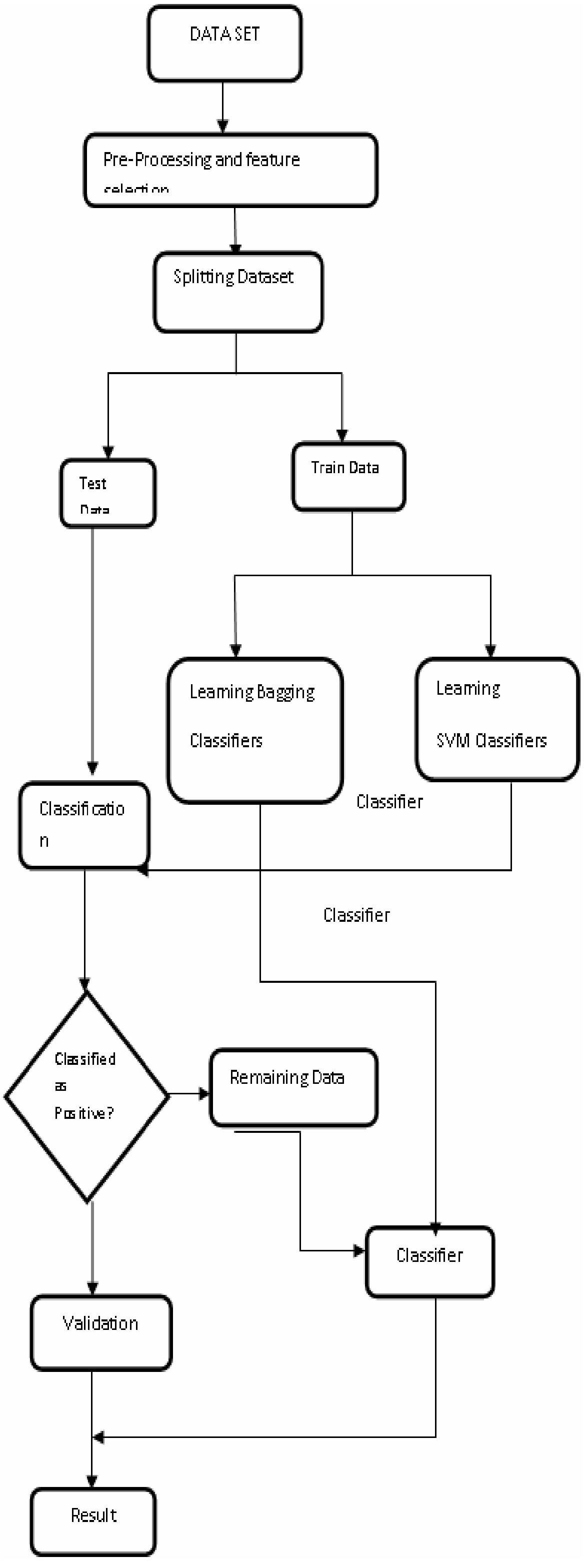


Figure 2.Overall Architecture of Proposed System

TABLE VI. HYPOTHYROID DATASET AFTER FEATURE SELECTION

Selected Features of Hypothyroid Dataset
Feature Selection Method: MRMR Total Number of instances: 106 with 26 attribute(one class attribute) Selected Attribute: Attribute Id:7,9,15,17,23,24 Attribute Name: q_hypo, preg, tsh, t3, fti, tbg_m.

B. *Evaluation Methodology*

The In classifying an unknown case, depending on the class predicted by the classifier and the true class of the patient (Control or HCC), four possible types of results can be observed for the prediction as follows:

- True positive—the result of the patient has been predicted as positive (disease type) and the patient has particular disease.
- False positive—the result of the patient has been predicted as positive (disease type) but the patient does not have disease.
- True negative—the result of the patient has been predicted as negative (Control), and indeed, the patient does not have particular disease.
- False negative—the result of the patient has been predicted as negative (Control) but the patient has disease.

Let TP, FP, TN, and FN, respectively, denote the number of true positives, false positives, true negatives, and false negatives. For each learning and evaluation experiment: Accuracy, Sensitivity, and Specificity defined below are used as the fitness or performance indicators of the classification [25],[26],[27].

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN),$$

$$\text{Sensitivity} = TP / (TP + FN),$$

$$\text{Specificity} = TN / (TN + FP).$$

Table 7 presents the matrix for Performance Evaluation:

TABLE VII. PERFORMANCE EVALUATION METRICS

	Predicted Class		
		Class=Yes	**Class=No**
Actual Class	**Class=Yes**	a:	b:
	Class=No	c:	d:

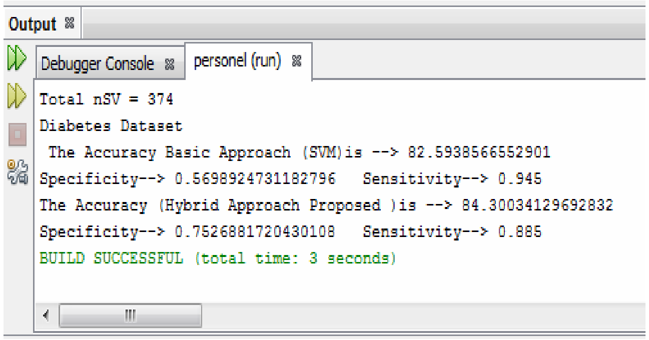
C. *Results (Proposed hybrid Approach + Other Statistical)*

About Dataset: We experimented with three different data sets. All the experiments are implemented in Java NetBeans interface, mention of split ratio 75:25 necessary in result/conclusion and this process was repeated various times, and average estimations were produced. Table 8 shows the results of our proposed hybrid approach and snapshots of "Disease Prediction tool" are presented in figure 3, 4, 5 for three different dataset.

From this study it has been found that our proposed hybrid classifier generates better results in terms of accuracy, sensitivity and specificity. As shown in table 8 SVM when combined with Bagging using REP tree produce better output in terms of correctly identified instances. Hybrid classifier achieves accuracy ≈99% . Above table also shows results when we combine SVM with J48, REP tree only but prediction accuracy is not satisfactory for both the cases.

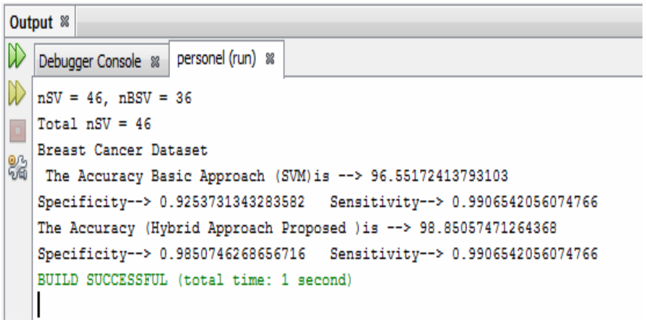
TABLE VIII. PROPOSED HYBRID APPROACH RESULTS

Data	**Statistical Measure**	**Proposed Approach**	**SVM + J48**	**SVM + REP Tree**	**SVM**
A Testing	Acc	.843	.825	.825	.825
	Sen	.885	.945	.945	.945
	Spec	.752	.569	.569	.569
B Testing	Acc	.988	.971	.971	.965
	Sen	.990	.972	.962	.990
	Spec	.985	.970	.985	.925
C Testing	Acc	.992	0.992	.992	.972
	Sen	.995	0.997	.995	.998
	Spec	.937	0.937	.937	.562



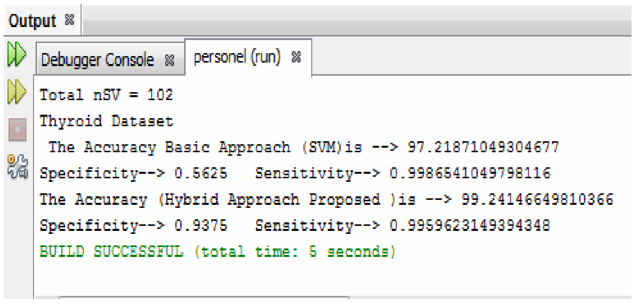
```
Total nSV = 374
Diabetes Dataset
 The Accuracy Basic Approach (SVM)is --> 82.5938566552901
Specificity--> 0.5698924731182796   Sensitivity--> 0.945
The Accuracy (Hybrid Approach Proposed )is --> 84.30034129692832
Specificity--> 0.7526881720430108   Sensitivity--> 0.885
BUILD SUCCESSFUL (total time: 3 seconds)
```

Figure 3. Tool Performance on first dataset



```
nSV = 46, nBSV = 36
Total nSV = 46
Breast Cancer Dataset
 The Accuracy Basic Approach (SVM)is --> 96.55172413793103
Specificity--> 0.9253731343283582   Sensitivity--> 0.9906542056074766
The Accuracy (Hybrid Approach Proposed )is --> 98.85057471264368
Specificity--> 0.9850746268656716   Sensitivity--> 0.9906542056074766
BUILD SUCCESSFUL (total time: 1 second)
```

Figure 4. Tool Performance on second dataset



```
Total nSV = 102
Thyroid Dataset
 The Accuracy Basic Approach (SVM)is --> 97.21871049304677
Specificity--> 0.5625   Sensitivity--> 0.9986541049798116
The Accuracy (Hybrid Approach Proposed )is --> 99.24146649810366
Specificity--> 0.9375   Sensitivity--> 0.9959623149394348
BUILD SUCCESSFUL (total time: 5 seconds)
```

Figure 5. Tool Performance on third dataset

Table 9 present the comparison of our proposed hybrid data mining approach with other existing hybrid approaches in disease detection and it is seen that our approach works well for the diagnosis of disease when compared with other existing approaches. Our approach achieves almost 99% accuracy for disease detection.

TABLE IX. COMPARATIVE ANALYSIS OF OUR APPROACH WITH OTHER EXISTING HYBRID APPROACHES

Comparison of our and other approaches	**Algorithm I**	**Algorithm II**	**Accuracy**
Our Approach for Disease Detection	SVM	Bagging with REP Tree	99.5%.
Heart Disease Prediction using hybrid data mining techniques [28].	Neural Network	Genetic Algorithm	98%.
Hybrid Breast Cancer Detection system [7].	Neural Network	Sequential Forward Selection and Sequential Backward Selection	SFSP+NN (97.5%) and SBSP+ NN (98.5%).
Breast Cancer Classification [29].	Neural Network	Particle Swarm Optimization	96.47%.
Data mining in healthcare using hybrid approach [30].	Naïve Bayes	Decision Table	97.4%.
Breast Cancer Diagnosis [31].	Least Square SVM	SVM	98.5% .
Breast Cancer Diagnosis [32].	Particle Swarm Optimization	Statistical Method	98.7 %.
Prediction of Coronary Heart Disease using Hybrid Approach [33].	Neural Network	Fuzzy Rules	Error rate is very low.
Breast Cancer Diagnosis [34].	Feature Selection Artificial Immune System	C4.5 Decision Tree	98.5%.
Intelligent Hybrid Method for Breast Cancer diagnosis [35].	Fuzzy Clustering	SVM	97.34%.
Novel Algorithm for Breast Cancer Detection [36].	Constrained search sequential floating forward search(CSS FFS)	SVM	98%.
Breast Cancer Detection in peripheral Blood [37].	Recursive Feature elimination and cross validation	SVM	98.4%.

VI. CONCLUSION

To summarize, in this paper we develop a software tool for the prediction of disease within no time and thus helps in decision making for the treatment method.This paper implemented using data mining techniques can be helpful in diagnosing the disease type and to assist for decision support in healthcare domain. In this study hybrid classification technique is proposed for classification of medical data sets and is applicable in healthcare domain. Overall classification accuracy is improved by combining two classification techniques together: SVM and bootstrap aggregation using REP tree. To evaluate the performance of proposed system we tested our approach on different datasets. Experimental results illustrate that our approach is more efficient that other hybrid techniques in classification of disease. In conclusion this study shows that data mining techniques can be a useful for medical diagnosis and applications particularly at disease prediction and treatment decision statement. This tool helps doctors or patient to decide in a short time whether the person is suffering from disease and is generic to all types of disease.

REFERENCES

[1] Homayounfar, P. and Owoc, M.L.,"Data mining research trends in computerized patient records", Proceedings of the Federated Conference on Computer Science and Information Systems, 2011 pp. 133–139.

[2] Han J. and Kamber M., "Data Mining Concepts & Techniques". ,CA: Elsevier: Morgan Kaufmann Publisher, 2006.

[3] Cios K.J. and Moore G.W., "Medical data mining and knowledge discovery," Berlin Heidelberg: Spinger, 2001, pp. 1-67.

[4] H. Wasan S.K., Bhatnagar V., Kaur H., "The impacts of data mining techniques on medical diagnostics. Data Science Journal. Vol.5, 2008 pp.119-126.

[5] Pujari A.K."Data Mining Techniques", Edition , 2001.

[6] Prather J.C., Lobach D.F., Goodwin L.K.,Hales J.W. , Hage M.L., Hammond W.E., "MedicalData Mining: Knowledge Discovery in a Clinical DataWarehouse", 1997.

[7] Uzer S. Mustafa, Inan O., Yilmaz N.," A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS, and PCA. Springer Journal of Neural Computing and application, 2012.

[8] Senapati MR, Mohanty AK, Dash S, Dash PK , "Local linear wavelet neural network for breast cancer recognition",Neural Computing and Applications,Volume 22, Issue 1, 2013, pp. 125-131.

[9] Zhang G., "A Modified SVM Classifier Based on RS in Medical Disease Prediction", 2009.

[10] Levashenko V. Zaitseva E.,"Fuzzy Decission Trees in Medical Decision Making Support System",Proceedings of the Federated Conference onComputer Science and Information Systems, 2012, pp. 213–219.

[11] Pecchia L., Meilillo P., and Bracale M.,"Remote Health Monitoring of Heart Failure with Data Mining via CART Method on HRV Feature". IEEE Transaction on Biomedical Engineering, Vol.58, 2011.

[12] Hilal A,R., Basir O., "Combination of enhanced AdaBoosting techniques for the characterization of breast cancer tumors", International Conference on Future BioMedical Information Engineering, 2009.

[13] Elshazly H., Azar A.T., El-korany A., Hassanien A.E., "Hybrid System for Lymphatic Diseases Diagnosis", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013.

[14] Jabbar M.A., Chandra P. Deekshatulu B.L., "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection", 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012.

[15] Sapon M.A., Ismail K., Zainudin S. and Ping C.S., "Diabetes Prediction with Supervised Leaming Algorithms of Artificial Neural Network," Intemational Conference on Software and Computer Applications, Kathmandu, Nepal, 2011.

[16] Akin O., "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," Computers in Biology and Medicine, vol. 41, 2011, pp. 265-271.

[17] Peng H., Long F., and Ding C., “"Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy ,"IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, 2005, pp.1226-1238.

[18] Vapnik C., “Support Vector networks”, Machine Learning, 20(2), 1995, pp.273-297.

[19] Burges J.C., “A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery”, Springer, 2(2), 1998, pp.121-167.

[20] Breiman, “Bagging predictors,” Machine Learning, Vol. 24, 1996, . pp. 123-140.

[21] Efron B., and Tibshirani R., “An Introduction to the Bootstrap,” Chapman & Hall, 1993.

[22] Esposito F., Malerba D., Semeraro G., and Tamma V., “The Effects of Pruning Methods on the Predictive Accuracy of Induced Decision Trees,” Applied Stochastic Models in Business and Industry, 1999 , pp. 277-299.

[23] UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science Wisconsin Breast Cancer Database Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data.

[24] UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science. Thyroid Database Available:http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/hypothyroid.data.

[25] Vapnik V., Statistical Learning theory. New York: Wiley, 1998.

[26] Yugal K. Sahoo G.,”Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers” International Journal of Information Technology & Computer Science , Vol. 5 Issue 6, 2013, pp. 57-64.

[27] Altman D.G., J.M. Bland,” Diagnostic tests Sensitivity and Specificity”, BMJ 308(6943):1552, 1994.

AUTHORS PROFILE

Megha Rathi: She is Assistant Professor at Jaypee Institute of Information Technology, Noida, India. She holds Masters of Technology and Bachelor of Engineering degree in Computer Science and engineering. Currently she is pursuing her PhD in computer Science and engineering. Her areas of Interest are Data Mining, Database, Software Engineering, Software Testing, Data Structure, Data warehouse and web mining.

Vikas Pareek: He is Associate Professor at Mahatma Gandhi Central University, Patna, India. He obtained his doctorate in the area of Cryptography. He also holds a Bachelor of engineering Degree in Computer Science and Engineering. His areas of interest are Cryptography, Algorithms, Data Structures and Electronic Commerce. He has many publications in International journals and conferences to his credit.

[28] Dewan A., Sharma M., “Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification”,2nd International Conference on Computing for Sustainable Global Development, 2015.

[29] Zhang L., wang L., Wang X., Liu K., and Abraham A., "Research of Neural Network Classifier Based on FCM and PSO for Breast Cancer Classification”. Springer, 2012.

[30] Sharma M., Kaur R., “Data Mining in Healthcare using Hybrid Approach”, International Journal of Computer Applications, Vol.128. no.4, 2015.

[31] Polat K., Gunes S. “Breast Cancer diagnosis using least square support vector machine” ,J. Digital Signal Processing, vol.17,2007, pp. 694-701.

[32] Yeh WC, Chang WW, Chung YY," A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method". Expert System application, 2009.

[33] Sen et.al., “A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level”, International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 2 Issue 9, 2013, pp. 2663-267.

[34] Polat K, Sahan S, Halife, and Gunes S,” A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS)”, Lecture Notes in Computer Science Volume 3611, Springer, 2005.

[35] Addeh J. and Ebrahimzadeh A., “Breast Cancer Recognition using a Novel Intelligent Hybrid Method”, J Med Signals Sens. Apr-Jun; 2(2), 2012 pp. 95–102.

[36] S. Aruna, S.P. Rajagopalan,” A Novel SVM Based CSSFFS Feature Selection Algorithm for Detecting Breast Cancer”, International Journal of computer Applications, 2011.

[37] Zhang F, Kaufman H L, Deng Y, and Drabier R, “Recursive SVM Biomarker selection for early detection of breast cancer in peripheral blood”, International Conference on Bioinformatics and Computational Biology, Vol.6. , 2011.