Analysis of Cardiovascular Risks using Artificial Intelligence and Machine Learning

[1]Pavitra Prakash Bhade, [2] Prof. A. U. Bapat,
[1]Student, [2]Associate Professor,
[1]Computer Engineering Department,
[1]Goa College of Engineering, Farmagudi, Ponda Goa

Abstract : Cardiovascular disease is a disease that is related to heart. It can prove very fatal. The detection of this disease is not only important but is also very complex. This is because, for its detection, a number of parameters have to be considered, or else it may be misinterpreted for some other disease. Due to its complexity, its detection is basically done by highly skilled doctors known as cardiologists. These doctors are not present everywhere and at all the time. Hence some kind of automated techniques have to be invented in order to perform initial risk analysis to overcome delays in this detection. Researchers have been using many data mining techniques so far. This analysis has to be done correctly, efficiently and very minutely. The proposed system uses machine learning technique to do the same. In this method, an ensemble classifier has to be developed, which combines many classifiers working on different algorithms. These classifiers will be trained on a heavy dataset. The dimensionality reduction techniques will also be adopted to reduce the complexity further. Along with this, a multivariate outcome will be provided, which will show severity of risks that helps the patient to accordingly proceed further with diagnosis procedure

IndexTerms **- Cardiovascular, Complexity, Data mining, Dimensionality reduction, Ensemble classifier, Machine learning.**

I. INTRODUCTION

ANY disease that is related to heart or blood vessels is known as the cardiovascular disease. According to the statistics of World Health Organization 2015, heart diseases cause the highest number of morbidity and mortality all over the world. It is one of the main reasons of death around the globe over the last decade. In United States alone, every minute, almost one person dies of this disease.

Due to the current lifestyle, there are a number of diseases in this world. But the count of people suffering from heart diseases is increasing day by day. The major cause for this effect is the current lifestyle itself. The sedentary life coupled with stress and pressure, along with unhealthy eating habits, all lead to this dangerous killer disease. In fact, the beauty of this disease is such that, even though its very dangerous, if its detected in time, major lifestyle changes can be incorporated along with appropriate medication to avoid the fatal effects. These changes can save both the life as well as money of the patients. Timely prediction is the key to overcome dangerous outcome. Various lifestyle changes can be regular exercise, yoga, meditation, healthy eating habits, healthy sleeping habits etc.

Diagnosis of this disease is a very complex task. This is because, a number of parameters have to be considered. For example, age, gender, family history, cholesterol, blood pressure, eating habits, smoking habits, diabetic or not, alcohol consumption etc. All these factors do not have equal weightage. The importance of weightage of these factors is decided by specialized doctors known as cardiologists. These weights are basically linked to the effect or contribution of certain parameter towards the presence of the disease. This is very important because if these factors are not studied properly, some other disease may be misinterpreted to heart disease or vice versa. Hence these skilled doctors are very important. But the problem is that, such highly skilled doctors are not present all the time at all places. And this disease is so fatal that any kind of delay in its diagnosis has to be avoided. Hence automated techniques have to be developed to initiate the diagnosis process to avoid the delay. This automated technique will be based on machine learning (supervised). That is, a meta classifier will be built which will be trained using a dataset prepared by skilled doctors. Now since this dataset is it self prepared by them, it will be considered to be very effective in the training purpose. .

II. LITERATURE SURVEY

A lot of research has been done in the data mining field related to this domain. A brief literature survey is presented here. Intelligent Heart Disease Prediction System (IHDPS) is built using various data mining techniques namely Neural Networks, Naive Bayes and Decision Trees. Results show that each technique has its infrequent strength in realizing the objectives of the defined mining goals. IHDPS can answer complex “what if” queries which conventional decision support systems cannot be proposed by Sellappan Palaniappan et al. [5]. The results illustrated the uncouth strength of each of the methodologies in comprehending the goal of the specified mining objectives. This system could respond the queries that traditional systems could not. It facilitated the installation of crucial knowledge such as patterns, relationships amid medical factors connected with heart disease. IHDPS remains wellbeing web based, user friendly, reliable, scalable and expandable. The diagnosis of Heart Disease, Blood Pressure and diabetes with the aid of neural networks was introduced by Niti Guru et al. [9]. Experiments were carried out on a sampled data set of patient’s records. The Neural Network is trained and tested with 13 input variables such as Blood Pressure, Age, Angiography's report and the like. The supervised network has been advised for diagnosis of heart diseases. Training was carried out with the help of back propagation algorithm. Whenever

unfamiliar data was inserted by the doctor, the system identified the unknown data from comparisons with the trained data and produced a catalog of probable diseases that the patient is vulnerable to.

In 2014, M.A.Nishara BanuB.Gomathy Professor, Department of Computer Science and Engineering has published a research paper "Disease Forecasting System Using Data Mining Methods"[10]. In this article, the preprocessed data is clustered using clustering algorithms as Kmeans to gather relevant data in a database. Maximal Frequent Item set Algorithm (MAFIA) is applied for mining maximal frequent model in heart disease database. The regular patterns can be classified into different classes using the C4.5 algorithm as training algorithm using the concept of information entropy. The result demonstrates that the designed prediction system is capable of predicting the heart attack successfully.

In 2012, T.John Peter and K. Somasundaram Professor, Dept of CSE presented a paper, "An Empirical Study on Prediction of Heart Disease using classification data mining technique"[7].

III. PROPOSED SYSTEM

The proposed system is to use supervised machine learning approach. This is basically a technique of building a classifier that will be working on certain algorithms and trained using a labeled dataset. The algorithms used can be any from the following table.

Table 1
Description of Machine Learning Classification Schemes

Type of Classifier	*Examples*	*Description*
Bayesian	*BayesNet*	*Apply Bayesian Principal of Conditional Probability to classification*
	NaiveBayes(updateable)	*Naïve Bayesian uproaches presume that input factors are independent of one another*
Function Based	*Logistic Regression*	*Creates a seperating surface using linear, logistic or kernel based methods*
	Support Vector Machine	
	Multilayer Perceptron	
	Radial Basis Frequency Network	
	Voted Perceptron	
Lazy	*K nEarest Neighbor*	*Maps outcomes onto a plot and classifies outcomes based upon proximity to neighboring outcomes with known correct classification via majority rules.*
Rule Based	*Decision Tables*	*Creates rules to assign outcomes into correct classification grouping either simply or via iteration.*
	Propositional Rule Learner(JRip)	
	PART Decision list	
	ZeroR	
Decision Trees	*J48*	*Create heirarchy of rules for classification, using an "If this … then that… " paradigm.*

As already mentioned, in this disease detection, a lot of parameters have to be considered and at the same time, these parameters do not have equal weightage. The respective importance of various factors will be decided by highly skilled doctors. This makes this task complex and delicate at the same time. Hence this system will make use of machine learning approach wherein a pre-labeleddataset will be used to train the model. In this approach, all the factors will be considered and the outcome will be studied.

This system will be built in parts. Below is the overall idea of the proposed system

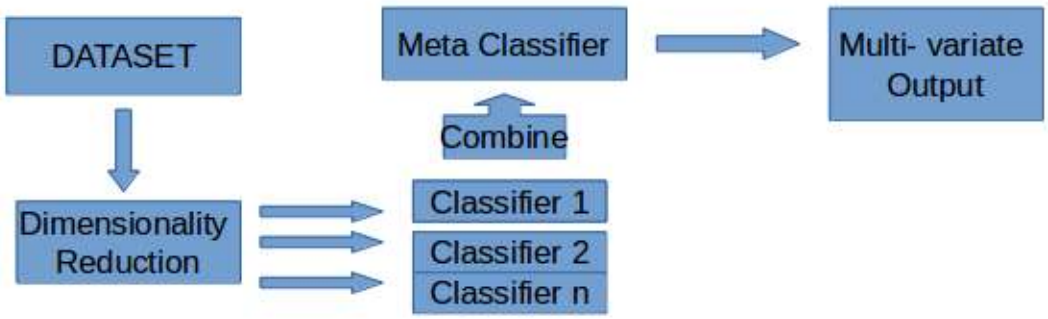

Fig. 1. Flow chart of the proposed system

This system will use a pre-labeled dataset. This dataset will be basically patient records that are attributes like gender, age, blood pressure, diabetes, smoking habits, sleeping habits etc. The dataset to be considered is available at UCI Repository in the name Heart Disease Dataset. It has 76 attributes. First the dimensions of this dataset have to be reduced. We can employ one of the three techniques of feature subset selection: Correlation based feature selection, Information gain based feature selection, Learner based feature selection[13]. Correlation based feature subset selection (CFS) using the greedy stepwise search pattern with 10 fold cross validation. This dimensional reduction strategy identifies those attributes that are highly correlated with the separating class, yet poorly correlated with the other attributes. The resulting "merit" of an attribute results from the following equation:

$$M_s = k * r_{cf} / sqrt(k + k(k-1) r_{ff}) \quad \text{----------(1)}$$

Here M_s is the merit of a subset of feature set S, which contains k number features; r_{cf} is the average correlation between a feature and a class; and r_{ff} is the average intercorrelation between features. Thus, the numerator indicates how predictive a feature set is, and the denominator indicates the redundancy among features.

Another popular feature selection technique is to calculate the information gain. You can calculate the information gain (also called entropy) for each attribute for the output variable. Entry values vary from 0 (no information) to 1 (maximum information). Those attributes that contribute more information will have a higher information gain value and can be selected, whereas those that do not add much information will have a lower score and can be removed. A popular feature selection technique is to use a generic but powerful learning algorithm and evaluate the performance of the algorithm on the dataset with different subsets of attributes selected. The subset that results in the best performance is taken as the selected subset. The algorithm used to evaluate the subsets does not have to be the algorithm that you intend to use to model your problem, but it should be generally quick to train and powerful, like a decision tree method. Once the dimensions are reduced, this dataset will be fed to a number of classifiers. Instead of just one classifier output, more than one classifier will be combined to give more robust output. The techniques by which ensemble classifier can be built are boosting, bagging and stacking.

A. ***Boosting:*** It is an ensemble method that starts out with a base classifier that is prepared on the training data. A second classifier is then created behind it to focus on the instances in the training data that the first classifier got wrong. The process continues to add classifiers until a limit is reached in the number of models or accuracy.

B. ***Bagging:*** It is an ensemble method that creates separate samples of the training dataset and creates a classifier for each sample. The results of these multiple classifiers are then combined (such as averaged or majority voting). The trick is that each sample of the training dataset is different, giving each classifier that is trained, a subtly different focus and perspective on the problem.

C. ***Blending (Stacking):*** Stacked Generalization or Stacking for short is a simple extension to Voting ensembles that can be used for classification and regression problems. In addition to selecting multiple sub-models, stacking allows you to specify another model to learn how to best combine the predictions from the sub-models. Because a meta model is used to best combine the predictions of sub-models, this technique is sometimes called blending, as in blending predictions together. ZeroR meta classifier combines as voted or mean. This model can be seen in figure 2 and 3.

Predictions of the base classifiers with correct class decisions will form the new training set. Based on this training set and a new learning algorithm, a meta classifier will be built. The basic idea behind this technique is that, a meta classifier attempts to learn relationships between predictions and the final decision. It may correct some mistakes of the base classifier. The data set can be used with 10 fold validation to improve the efficiency. Cross validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

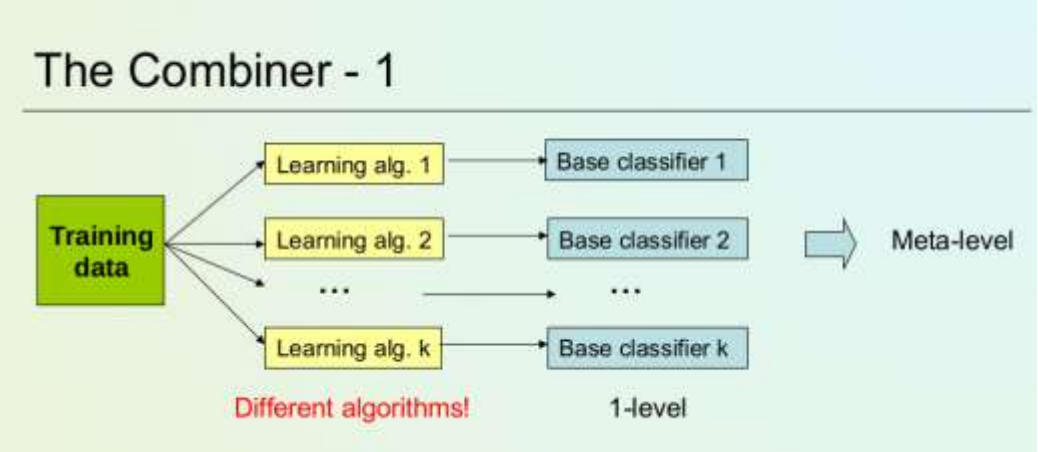

Fig. 2. Stage 1 of Stacking

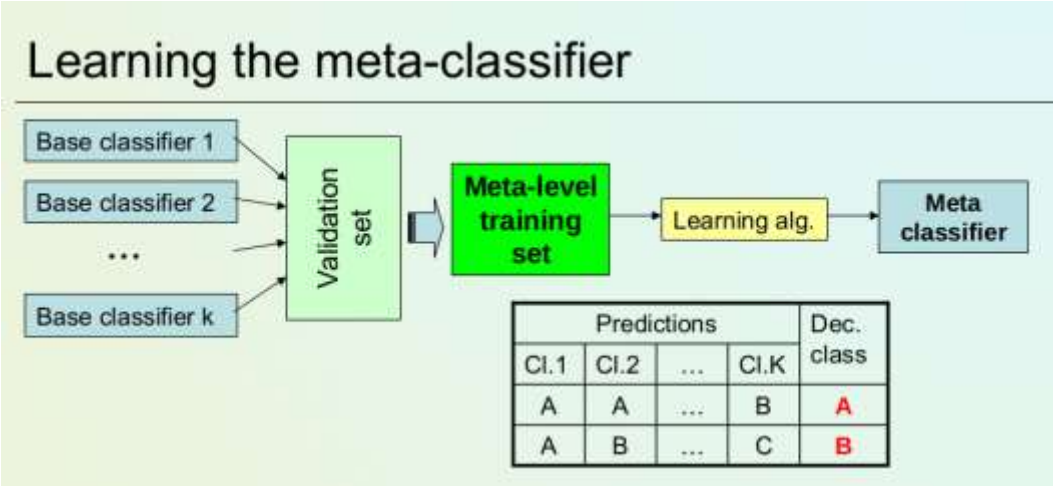

Fig. 3. Stage 2 of Stacking.

In kfold cross-validation, the original sample is randomly partitioned into k equal size sub-samples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. Once the meta classifier is built, it has to output multiple results i.e it should show the severity of risk of the disease instead of the binary presence/absence. This will improve the analysis and diagnosis process and spilt the severity levels between the patients. Hence the output should either mention severity levels as high, moderate, low or no risks, or it should show the probability of risks involved. The dataset that can be used from UCI Repository has an output class label 'num' that has five values from 0 to 4 where 0 means no risk and 1-4 means presence of risk and its severity. This severity levels can be exploited and the output can be accordingly achieved.

IV. CONCLUSION

Using the above mentioned technique, a system can be built which exploits all the factors i.e. it uses most of the attributes linked to the presence of cardiovascular disease and at the same time be less complex as it will be using dimensionality reduction techniques. The building of meta classifier reduces the risk of errors as the weighted average of output from various classifiers will be considered and hence the possibility of false outcomes will be reduced. The multivariate output will in some way be beneficial for the patients to plan their further diagnosis process accordingly.

REFERENCES

[1] Purushottam Prof. (Dr.) Kanak Saxena, Richa Sharma. "Efficient Heart Disease Prediction System using Decision Tree" , IEEE Transaction on International Conference on Computing, Communication and Automation. (2015,May). Noida, India

[2] Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel "Heart disease prediction using Machine Learning Technique". IJCSC Vol 7(2015, Sept), West Bengal, India

[3] Jayshril S. Sonawane,D. R. Patil," Prediction of heart disease using learning vector quantization algorithm," IEEE Transaction on IT in Business. (March 2014),

[4] Stephen F. Weng, Jenna Reps , Joe Kai, Jonathan M. Garibaldi, Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoSONE, (April,2017)

[5] Sellappan Palaniappan and Rafiah Awang, (Intelligent Heart dIsease Prediction System using data mining, IEEE., April 2008),

[6] Abhishek Taneja,, Heart Disease Prediction System using Data mining Techniques, Oriental Journal Of Computer Science and Technology, (December 2013).

[7] T.John Peter , K. Somasundaram, , An Emprical Study on Prediction of Heart Disease using Classification Data Mining Techniques, IEEE, (March 2012)

[8] Tighe PJ, Lucas SD, et. al., Use of machine learning Classifiers to predict requests for pre-operative acute pain service consultation, PMC Journal. (Oct 2012)

[9] Niti Guru, Anil Dahiya, Navin Rajpal, Decision Support System for Heart Disease Diagnosis Using Neural Network, Delhi Business Review (June 2007),

[10] M.A.Nishara Banu, B.Gomathy, Disease Forecasting System Using Data Mining Methods, IEEE, International Conference on (November 2014),

[11] M. A. Hall, "Correlation based Feature Selection for Machine Learing" M.Phill theses, Dept of Computer Science, The University of Waikato, NewZealand, Hemilton, April 1999

[12] Jerzy Stephanowsky, "Multiple Classifiers", Institute of Computing Sciences, Poznan University of Technology, April 2008.