# A SURVEY ON MACHINE LEARNING APPROACHES FOR DISEASE PREDICTING SYSTEM

**[1]Akhila C S, [2]Vidya M**
[1]PG Student, [2]Assistant Professor
[1]Department of Computer Science & Engineering,
[1]Vidya Academy of Science & Technology,
Thalakkottukara P. O, Thrissur, Kerala, India

***Abstract— Nowadays, the prediction of various diseases by using an automated system is a significant alternative for medical decision by doctors. Such automated systems are able to assist the patients, hospital enquiry personals and doctors to acquire the correct information about the disease. Machine learning approaches are used intensively to predict the diseases such as heart disease, lung cancer, liver diseases etc. In this paper, we focus on the various machine learning approaches used in disease predicting systems.***

***Index Terms—Disease Predicting System, Naïve Bayes Algorithm, K-Means Clustering Algorithm, Support Vector Machine.***

## I. INTRODUCTION

Modern era marks the development of disease diagnosis systems in the medicinal trade. A disease predicting system is an automated system refers to the process of attempting to determine or identify possible diseases or disorder. Such attempts provide the users with information about the disease. A normal patient doesn't have the facility to predict the disease he is suffering from until he consults a doctor. Every so often, patients disregard their symptoms in the early stages of the disease, which can be harmful. If the patients will be provided with the automated system, they can know about the disease they are suffering from.

Disease predicting systems are developed using several types of machine learning techniques. Machine learning is a method used to devise complex models and algorithms that lend themselves to prediction purposes. Predictive analytics encompass a variety of statistical techniques that analyze current and historical facts to make predictions about future or unknown events. Some of the common approaches are Naïve Bayes Classifier Algorithm, K Means Clustering Algorithm, Support Vector Machine Algorithm, etc. The final goal of this research is to survey different machine learning approaches used in the development of disease prediction systems.

## II. RELATED WORKS

Basic working of a disease prediction system consists of two parts: training and testing phase. Once the medical data is divided into training and testing data, an appropriate classifier model is build. Then using this model, the test data is classified accordingly to the relevant class. The study carried out here, concentrates on various machine learning approaches used in disease predicting systems. This will help to the enhanced understanding of the tools and technologies used in such systems. Figure 1 shows the working of a typical disease prediction system.

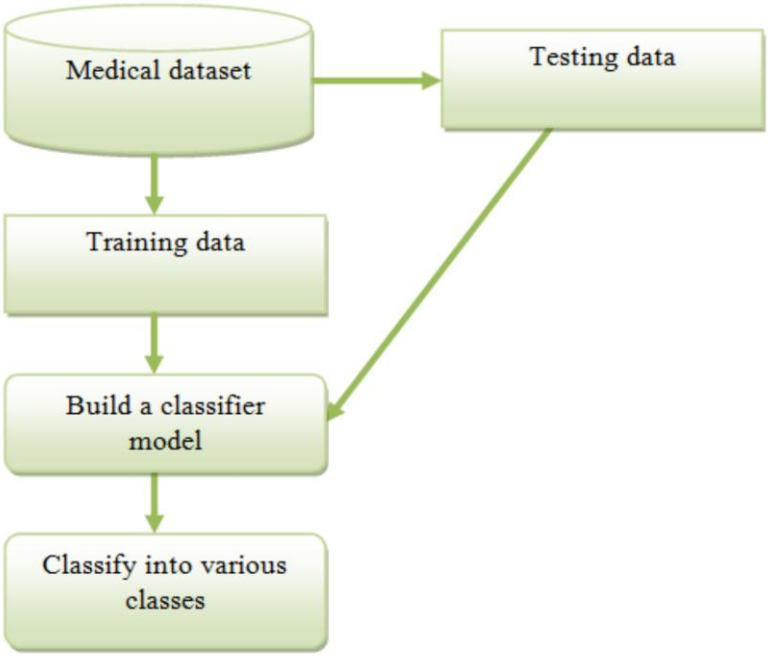


**Fig.1.** Working of disease prediction system

### *Naïve Bayes Classifier*

Medical data can be transformed into valuable data and for this purpose different data mining techniques are used. One such technique is Naïve Bayes Classifier, where classification represents a supervised learning method as well as a statistical method. It allows capturing uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

Naïve Bayes classifiers can be trained very efficiently in a supervised learning setting, for some types of probability models. Parameter estimation for these models uses the method of maximum likelihood. Naïve Bayes algorithm is based on Bayesian Theorem.
Bayesian Theorem: Equation 1 follows the Bayes theorem, given training data X and the posterior probability of a hypothesis H, P(H|X) is;

$$P(H|X)=P(X|H)P(H)/P(X). \quad (1)$$

Dhanashree S. Medhekar, Mayur P. Bote and Shruti D. Deshmukh[1], proposed a classifier approach for detection of heart disease, which shows how Naïve Bayes can be used for classification purpose. Mainly, the system proposed by them works in two phases: Training phase (classification) & Testing phase (prediction). Dataset contains information concerning heart disease diagnosis with 14 parameters like age, gender, chest pain, cholesterol, etc. In this method, medical data was categorized into five categories of prediction, namely: no, low, average, high and very high. The training dataset was given as input to the classifier. This classified data was further used for testing purpose. If unknown sample were given, then the system predicted the class label of that sample.

Naïve Bayes is simple and fast to train besides handling real and discrete data. A benefit of Naïve Bayes is that it only requires a small number of training data to estimate the parameters essential for classification. A main disadvantage can be coined as; it assumes the independence of features. Moreover the accuracy of the system proposed depends on the database used.

***K-Means Clustering Algorithm***

Clustering algorithm finds clusters of data objects that are similar to one another. The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high. One of the most commonly used clustering algorithm is the distance based K-Means clustering algorithm. Its objective is to divide *n* observations into *k* clusters in which each observation belongs to the cluster having closest mean.

Priyanka D and S Shahar Banu [2], presented a paper on lung disease prediction using K-means algorithm. The database used here is the Wisconsin lung cancer database. Here the objective is to determine whether a patient has a benign cancer or malignant using several descriptors. Beginning with a decision on the value of k = number of clusters, it determines the centroid coordinates. Then each sample in sequence was taken and its distance from the centroid of each of the clusters was computed. If a sample was not currently in the cluster with the closest centroid, this sample was switched to that cluster and the centroid of the cluster was updated. Repeat this until convergence, i.e., until a pass through the training sample causes no new assignments. The compactness and connectedness of this method provided better efficiency and effectiveness for predicting lung disease.

Shilna S and Navya E K [3], proposed k-means clustering algorithm for heart disease forecasting system. The heart disease database was preprocessed efficiently by necessary steps of cleaning and filtering. This heart disease dataset was then processed by K-means algorithm and Particle Swarm Optimization (PSO) clustering algorithm with the K value of 2. At the initial stage, the PSO was executed to search for the location of cluster's centroid. These locations were used as initial centroid for K-means clustering algorithm for generating the optimal clustering output. Then the maximal frequent forms were mined efficiently from the dataset, using the Maximal Frequent Itemset Algorithm (MAFIA), which uses the strategy of integrating a depth-first traversal of the item set lattice with efficient pruning mechanisms. These frequent patterns were classified using C4.5 algorithm as training algorithm.

***Support Vector Machine***

Support Vector Machines (SVM), also called support vector networks, are supervised learning models that can analyze data used for classification and regression analysis. Given a set of training example, SVM training algorithm builds a model that assigns new examples to one category or the other.

In addition to the linear classification, SVMs can efficiently execute a non-linear classification using what is called the kernel trick which replacing its features by a kernel function. It implicitly maps their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space. Since support vector machines employing the kernel trick do not scale well to large numbers of training samples or large numbers of features in the input space, several approximations to the Radial Basis Function (RBF) kernels have been introduced. SVMs are helpful in text and hypertext categorization, classification of images and recognition of hand-written characters also.

Anuja Kumari and R.Chitra [4], proposed support vector machine as the classifier for diagnosis of diabetes. The dataset used in this work was a collection of medical diagnostic reports from 768 records of female patients at least 21 years old. In SVM the binary target variable takes the values '0' or '1' while '1' means a positive test for diabetes, '0' means a negative test. Here, 268 cases are given to the class '1' while 500 cases to the class '0'. The importance of the involuntarily selected set of variables was further manually evaluated by fine tuning of parameters. There are eight numeric variables, like number of times pregnant, diastolic blood pressure, body mass index, age, etc. After the deletion there were 460 cases with no missing values. The training data set was first partitioned into 10 equal-sized subsets. Each subset was used as a test data set for a model trained on all cases and an equal number of non-cases randomly selected from the remaining nine datasets. This cross-validation process was continued for 10 times, and each subset will act once as the test data set. Test data sets assess the performance of the models.

The machine learning method focuses on classifying diabetes disease from high dimensional medical dataset. The SVM classifier with RBF kernel was used for classification. The diabetes dataset contains 460 data, 200 data are used for training and 260 data for testing. The entire patient's data were trained by using SVM. The choice of best value of parameters for particular kernel is critical for a given amount of data. SVM approach with Radial basis function kernel can be successfully used to detect a common disease with simple clinical measurements, without laboratory tests. The performance parameters such as the classification accuracy, sensitivity, and specificity of the SVM and RBF were found to be high thus making it a good option for the classification process.

## III. CONCLUSION

A study of various machine learning approaches which are used in the development of disease prediction systems is done. Each classifier has different level of performance. The performance level largely depends on the dataset collection. More refined and meaningful dataset along with a good classifier can lead to an efficient disease predicting system. Such an attempt can be done using SVM for some of the lifestyle disease prediction. The accuracy of system can be increased by using a query answering interface with users. The answers given by users will produce an efficient test data input to the SVM classifier. Also such systems can be enhanced by providing remedies for predicted

disease, details about doctor for treatment, etc. This advancement will provide better working of machine learning approach- SVM, for a more accurate disease predicting system.

## IV. ACKNOWLEDGMENT

We wish to record our indebtedness and thankfulness to all those who helped us to prepare this paper and present it in a satisfactory way. Our sincere thanks to Dr. Sudha Balagopalan, our Principal, for providing us all the necessary facilities. We are also thankful to Ms. Sunitha C, Head of Department of Computer Science and Engineering, for encouragement. Last but not the least, we wish to thank our family and friends for supporting and encouraging us throughout the work of this paper.

## REFERENCES

**[1]** Dhanashree S. Medhekar, Mayur P. Bote and Shruti D. Deshmukh, "Heart disease prediction system using Naive Bayes", International Journal Of Enhanced Research In Science Technology & Engineering, Vol 2 Issue 3, ISSN NO: 2319-7463, March 2013.

**[2]** Priyanka D and Ms.S Shahar Banu, "Prediction on Lung disease using K means Algotithm", International Journal of Innovative Research in Technology, Volume 1 Issue 11, ISSN: 2349-6002, 2014.

**[3]** Shilna S and Navya E K, "Heart disease forecasting system using K-Mean Clustering Algorithm with PSO and other Data Mining methods", International Journal On Engineering Technology and Sciences, ISSN(P): 2349-3968, ISSN (O): 2349-3976, Volume III, Issue IV, April- 2016.

**[4]** V.Anuja Kumari and R.Chitra, "Classification of Diabetes disease using Support Vector Machine", International Journal of Engineering Research and Applications, Vol. 3, Issue 2, pp.1797-1801, ISSN: 2248-9622, March -April 2013.

**[5]** M.A.Nishara Banu and B Gomathy, "Disease predicting system using Data Mining techniques", International Journal of Technical Research and Applicationse-ISSN: 2320-8163, Volume 1, Issue 5, PP. 41-45, Nov-Dec 2013.

**[6]** G.Subbalakshmi, K. Ramesh, M. Chinna Rao, "Decision support in Heart disease prediction system using Naive Bayes", Indian Journal of Computer Science and Engineering, ISSN : 0976-5166, Vol. 2 No. 2 Apr-May 2011.

**[7]** Hian Chye Koh and Gerald Tan, "Data Mining applications in Healthcare", Journal of Healthcare Information Management - Vol. 19, No. 2.

**[8]** Sridevi Radhakrishnan and Dr. D. Shanmuga Priyaa, "A critical study on Data Mining Techniques in HealthCare Dataset", International Research Journal of Engineering and Technology, Volume: 02, Issue: 05, Aug-2015.

**[9]** K. Rajalakshmi and Dr. S. S. Dhenakaran, "Analysis of Data Mining Prediction Techniques in Healthcare Management System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, ISSN: 2277 128X, April 2015.