EasyChair Preprint

№ 2882

A Survey on Machine Learning Algorithms

T. Senthil Kumar, Sai Sameer Vennam and Sai Saketh Chamarti

March 6, 2020

A SURVEY ON MACHINE LEARNING ALGORITHMS

T. Senthil Kumar

Assistant Professor
Department of Computer Science
SRM Institute of Science and Technology, Chennai, India

Sai Sameer Vennam

Department of Computer Science
SRM Institute of Science and Technology, Chennai, India

Sai Saketh Chamarthi

Department of Computer Science
SRM Institute of Science and Technology, Chennai, India

ABSTRACT

Heart disease can be termed as a significant public health problem that is widely responsible for premature deaths all around the world. In the present decade, there is a need for a system to tackle this problem by using the latest technical developments. Many machine learning techniques have been employed independently to be able to predict the presence of heart diseases in individuals based on structured and unstructured healthcare data. This paper gives a detailed analysis of classification algorithms like Naïve Bayes, KNN, Decision tree, Random Forest, and Support Vector Machine (SVM). Studies have suggested that this accuracy rate is interlinked with the selection of the features from the available dataset. Thus, appropriate feature selection increases the quality of the prediction model. In this research, we take reference from the study of Cleveland heart disease dataset from the UCI-Repository will be used to get the best combination of six features after testing each possible combination from all the thirteen features available against all the five classifiers mentioned above. Each combination will be assigned an individual score, which will correspond to its weighted mean of the accuracy obtained on all the five classifiers. Furthermore, this paper will find out the most prominent feature out of all the thirteen features by counting the number of occurrences of each element in the top hundred combinations giving the maximum score. Thus, this model provides us with an insight into the characteristic feature that has the maximum correlation with the accuracy of the prediction models.

INTRODUCTION

Classification algorithms are generally used to classify the given data into a specified number of classes. There are a number of classification algorithms that are being used today namely Naïve Bayes, Support Vector Machines, Decision Trees, Random Forest, KNN, Logistic Regression etc. The training dataset is used to get better boundary conditions which could be used to determine the list of target classes. Once the boundary conditions are determined, the next task is to predict the target class. The whole process is known as classification. The whole purpose of classification algorithms is to segregate the given dataset depending on the boundary conditions.

In the past decade, prominent research and intriguing developments in the field of these classifiers have broadened the spectrum of the range of applications that the machine learning algorithms have. Classification is now being used as a vital algorithm in many pragmatic problems namely speech recognition, handwriting recognition, bio metric identification, document classification, electronic mail spam classification, facial point detection etc. However, one of its most eminent application is in the healthcare industry. These algorithms have been obtaining significant results in the medical diagnosis and analysis. A few areas problems include drug classification, cancer tumour cells identification, prediction of heart and liver ailments and so on.

Heart disease is one of the leading causes of death in India. It can be termed as a major public health problem that is leading to premature death not just in India but around the world. In India, heart ailments caused more than 2.1 million deaths in 2015 at all ages, which is more than a quarter of all deaths. At ages 30-69 years, of 1.3 million cardiovascular deaths, 0.9 million (68.4%) were caused by coronary heart disease and 0.4 million (28.0%) by stroke. A study has shown that adults born after 1970 are more susceptible to such deaths. People in wealthier and urban places in India and around the world are more vulnerable to heart diseases.

Deaths from heart-related disease among rural Indians have surpassed those among urban Indians, according to a forthcoming study in The Lancet. The study, to be published in the August edition, also suggests that unlike in the West, obesity may not be a big driver of such deaths in India. The study, which provides the first-of-its-kind nationally representative estimates of cardiovascular mortality in India, shows that heart ailments caused more than 2.1 million deaths in India in 2015 at all ages, or more than a quarter of all deaths.

At ages 30-69 years, of 1.3 million cardiovascular deaths, 0.9 million (68.4%) were caused by coronary heart disease and 0.4 million (28.0%) by stroke. The study shows that adults born after the 1970s are much more vulnerable to such deaths than those born earlier. The study is part of the Million Death Study project set up by the Registrar General of India (RGI) in collaboration with global health experts to investigate the causes of deaths in India using nationally representative survey data.

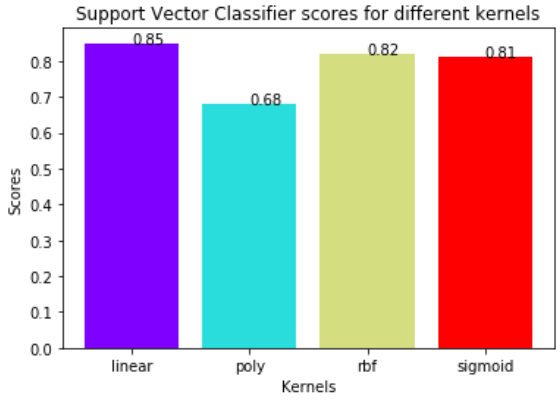
MACHINE LEARNING ALGORITHMS

1. SUPPORT VECTOR MACHINE

A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. SVMs are used in text categorization, image classification, handwriting recognition and in the sciences. A support vector machine is also known as a support vector network (SVN). A support vector machine is a supervised learning algorithm that sorts data into two categories. It is trained with a series of data already classified into two categories, building the model as it is initially trained. The task of an SVM algorithm is to determine which category a new data point belongs in. This makes SVM a kind of non-binary linear classifier.

An SVM algorithm should not only place objects into categories, but have the margins between them on a graph as wide as possible. Applications of a Support Vector Machine include Text and Hypertext Classification, Image Classification, Recognizing Handwritten Symbols, Biological Sciences, Health Care and in Our case Heart disease feature prediction.SVM aims at forming a hyperplane that can separate the classes as much as possible by adjusting the distance between the data points and the hyperplane. There are several kernels based on which the hyperplane is decided.

SVM classifier has been used previously to detect various medical anomalies like diabetes, cancer tumour detection, presence of coronary disease and liver disease. In the prediction of heart disease, the accuracy differed with the usage of each kernel. The kernels used were – linear, sigmoid, poly and rbf. The dataset used was the Cleveland dataset. The dataset is split into two parts namely training and testing. The ratio used is 3:2 for the training and test set respectively. The Cleveland dataset consists of 64 attributes. However, the algorithm uses only 13features with 303 entries.This dataset is tested against all the five kernels. The result is as follows.



As seen, the linear kernel outperforms the rest of the kernels and the main reason so as to why linear kernel performs so well is because of the property of the data set to be linearly separable. This shows that our dataset has points which are linearly separable.

In another case, SVM was used in the detection of coronary disease. The dataset was recruited from five research centres in Beijing and Henan from the same demographic area. The dataset had 80 symptoms and each symptom had four levels namely- none, light, middle and severe. The SVM classifier performed classification tasks by maximizing the margin separating both classes while minimizing the classification

errors. The sequential minimal optimization algorithm was employed to train the SVM here. The SVM achieved an accuracy of 82.5%.

SVM has also been used in the diagnosis of liver diseases. It implements the classification task by maximizing the margin classifies both class while minimizing the classification errors. The accuracy it achieved was close to 80% and it got a better accuracy than Naïve Bayes.

In another prediction model developed to predict heart diseases, a hybrid kernel system was proposed which was then compared to the general kernel. New kernels were created as a result of combinations using the following formula –$k_{mix} = Ek_1 + (1\text{-}E)k_2$ where k_1 and k_2 are two different kernels and E is a real number.Hybrid kernel functions of the documents are constructed between any two traditional kernels. The K-type kernel function which is a new kernel and traditional kernel functions are mixed for constructing a hybrid kernel function. In the same manner, the other hybrid functions are constructed as a cross over between RBF kernel function and traditional kernel function. Simultaneously, PSO algorithm is used to optimize coefficient of linear combination, a penalty parameter C. Compared two models, it can be verify that the first model is much better. The libsvm database of heart data set was used in this experiment. The data set consist of two different classes together with their class labels. The train set and test set consist of 140 and 130 samples respectively. Each of the samples have 17 features, the positive sign +1 and the negative sign -1. The combination of the kernels can't always guarantee a good result. For instance, the combination of K-type kernel with linear and polynomial kernel is worse than K-type kernel. The effect of hybrid kernel between RBF kernel and linear kernel and polynomial kernel is better than RBF kernel function. The experimental results is shown as follows-

Kernel function	Accuracy
Linear	0.8385
Poly	0.8462
K-type	0.8769
RBF	0.8615
ε Linear+ $(1-\varepsilon)$K-type	0.8692
ε Linear+$(1-\varepsilon)$RBF	0.8769
ε Poly+$(1-\varepsilon)$K-type	0.8417
ε Poly+$(1-\varepsilon)$RBF	0.8692

From all the above experiments, it can be inferred that the SVM has been consistently put to use and is an important aspect of healthcare industry.

1.1 Applications of SVM in the Medical Field

Support Vector Machine has been found to have a lot of applications in the medical field. In the field of Bioinformatics, SVM has been used for various different kinds of classification such as the gene classification, biological classification. Recently, Studies have shown that the SVM algorithm can also be used for the protein remote homology detection. It has also been used for the classification of blood cells which are most likely

to turn malignant. Some other applications of SVM in the medical field include Surgery readiness classification, Heart disease classification etc

1.2 Applications of SVM in the Technology Sector

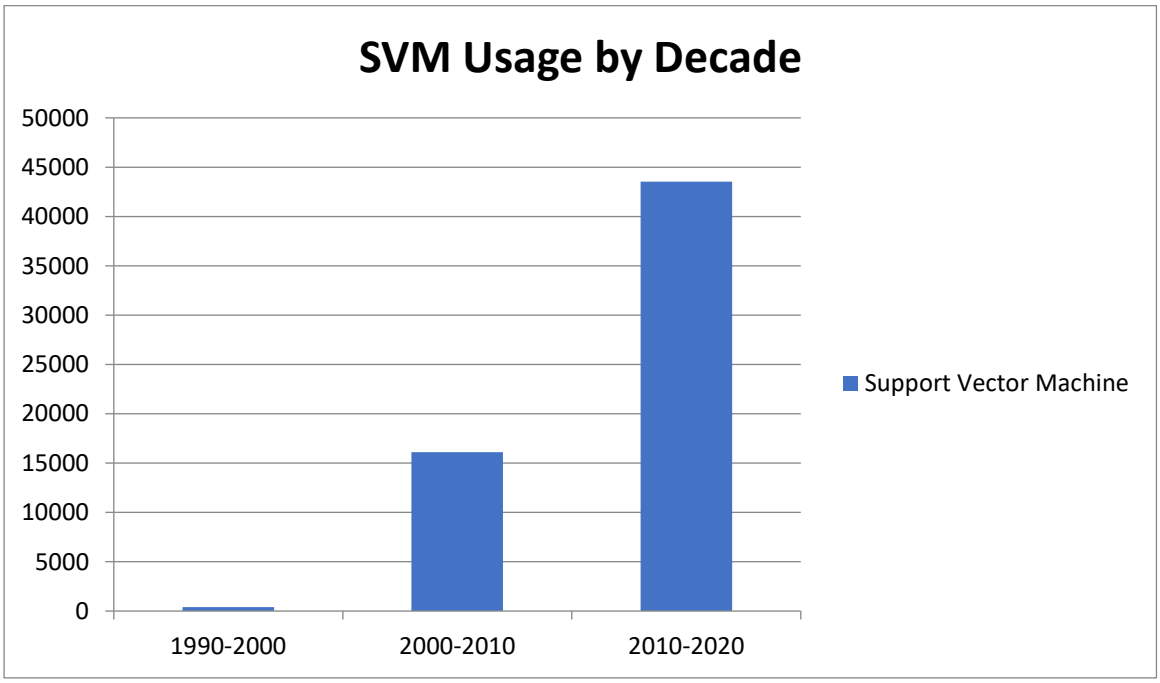
Support Vector Machine has found lots of applications in the technology sector particularly due to the increase in the amount of the data that is available for processing. Some of the applications of the Support Vector Machine in the technology sector are Handwriting Classification, Image classification and facial detection. Nowadays, the support vector machine algorithm has also found usage in the media industry where leading media houses are using it to classify whether the news stories would generate interest in the audience or not.

1.3 Applications of SVM in the Scientific Research

Support Vector Machine has found various applications in both the natural sciences as well as the physical sciences. In the Physical Sciences and chiefly in chemistry. In the various methods of synthesis of a chemical product, chemists are using the Support Vector Machine to determine the better way of synthesis and weed out the other methods. SVM has also been used in Chemometrics which is another important aspect in the chemical synthesis, this is used to describe the various different types of chemical compositions that are present and their uses and applications. In re

1.4 Applications of SVM in the Manufacturing Industry

The Support Vector Machine has also found applications in the manufacturing industry. This has led to the classification of quality of the finished products. In many factories throughout the world, a combination of Computer Vision and SVM are being used to judge the quality of the products and according to the results given by this algorithm the manufacturing methods are also being changed to give the best resultant product.



2. K-NEAREST NEIGHBOURS

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables, the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset. Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value.

Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.This classifier looks for the classes of K nearest neighbors of a given data point andbased on the majority class, it assigns a class to this data point. However, the number of neighbors can be varied.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

The application of KNN has been used in a variety of medical diagnosis before. It is a straightforward classifier where the classification is done based on class of nearest data. Out of all the three formulas that have been shown, the most common type of KNN classification adopted measure is the usage of Euclidean distance. In the prediction of heart disease, KNN was combined with genetic algorithm to analyse a given dataset. The genetic algorithm was used to remove the redundant and unnecessary data attributes. Once the least ranked data attributes have been removed, the classification algorithm is built on the evaluated attributes. The classifier used two classes – one being healthy and the other being sick. The performance of the proposed system was tested with six medical and one non-medical dataset. The experiments were carried out by

assigning various values ofk.As the k value kept increasing, the accuracy of data sets was decreasing. The Accuracies of various data sets using the cross validation with genetic algorithm is shown below. Accuracy of heart disease data decreases by 32% when the cross-validation measure is used. The 5-fold cross validation method was used. Accuracy of the data sets with and without GA is also shown.

Data setname	K=1	K=3	K=5
Weather data	78.57	78.57	78.57
Breast cancer	73.07	71.32	74.47
Heart stalog	80	78.14	81.4
Lympography	85.13	85.13	87.1
Hypothyroid	91.7	93.47	94
Primary tumour	40.7	45.13	47.4
Heart disease A.P	67.5	65	62.5

Knn also found an application in the prediction of alzheimer's disease. MRI (Magnetic resonance imaging) scans were processed by FreeSurfer, a powerful software tool suitable for processing and normalizing brain MRI images. The PSO algorithm was applied to get the best possible set of features that represent the salient features of Alzheimer's condition. The knn algorithm was applied in combination with few other algorithms like the SVM and the naïve bayes algorithm. The results of the experiment are shown below.

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
SVM + KNN	89.22 ± 1.89	82.42 ± 1.26	78.63 ± 1.46	72.31 ± 1.12
BN + SVM + KNN	90.47 ± 1.24	88.62 ± 1.62	90.15 ± 1.84	80.26 ± 2.12
BN + SVM + KNN + PSO	96.31 ± 1.22	91.27 ± 1.44	89.90 ± 1.14	96.05 ± 1.21

Another application of the knn was in the prediction of breast cancer. The algorithm was optimized by clustering and reliability coefficients.Bailey introduced weights to classical KNN to present weighted K-nearest neighbors (WKNN) In WKNN, the weights will be assigned to each calculated value, then the nearest neighbors are computed, and finally, the class is assigned to the processed instance. The Condensed Nearest Neighbor algorithm (CNN) eliminates any record of duplicate data, removes irrelevant instances which do not give additional information, and shows similarity with other training datasets. The Reduced Nearest Neighbor algorithm (RNN) meanwhile includes an additional step: it eliminates patterns which don't affect the result. WKNN algorithm extends classical KNN in two ways- A weighting scheme for nearest neighbors. - A standardization of distances.

The instances of the dataset were replaced by less, but more significant centres of clusters. K-means algorithm is used to form clusters, and the classification will be based on the centres of this new set of clusters. Thus, classifying a new instance into one of the k clusters instead of comparing it to the initial n instances divides the computation time of the

algorithm by k n .As a result, the distance between a given instance and the centre of each cluster is restricted to significant attributes, and then weighted by their reliability coefficients.

The dataset consisted of 355 benign and 210 malignant cancer record entries. The attributes were obtained from computingof a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Each instance contains the ID, the diagnosis (malignant or benign) and 30 attributes such as radius, texture, fractal dimension, etc. The cross-validation approach was considered for evaluation of the dataset. The dataset was divided into 5 subsets of 113 instances each (71 benign and 42 malignant). At each and every iteration, 4 subsets are considered for training and the fifth is taken as the test, the process is thus repeated five times, and the average f-measure is retained.

The knn algorithm gave an f-measure of 91.1% when checked as a result.

2.1 Applications of the K Nearest Neighbours in Healthcare and Medicine

There has been a multitude of applications for K-Nearest Neighbours in recent times especially in the medical field due to its higher accuracy and simplicity in implementation. The KNN algorithm has been used for the detection of breast cancer and has found to increase the chances of detection and cure by 27 percent. Another application is the monitoring of the vital signs in a patient admitted in ICU. Since, there is a large amount of data being observed in ICU and it is difficult to make sense of all that information, the KNN algorithm is used to separate the information into useful and non useful information.

2.2 Applications of the K Nearest Neighbours in Technology

There has been a rapid increase in the applications of the K- Nearest Neighbours algorithm in the technical field. Due to the rise of Virtual Reality in recent time, the KNN algorithm has been used to resolve issues in the Spatial Databases. The study conducted by Gao et al[5] shows that the KNN algorithm can be used to increase the efficiency and provide a better VR experience by reducing the search space of the user and the user would be more adept at overcoming physical obstacles. The other applications of the K nearest neighbours are Image Classification, Social Media Analysis.

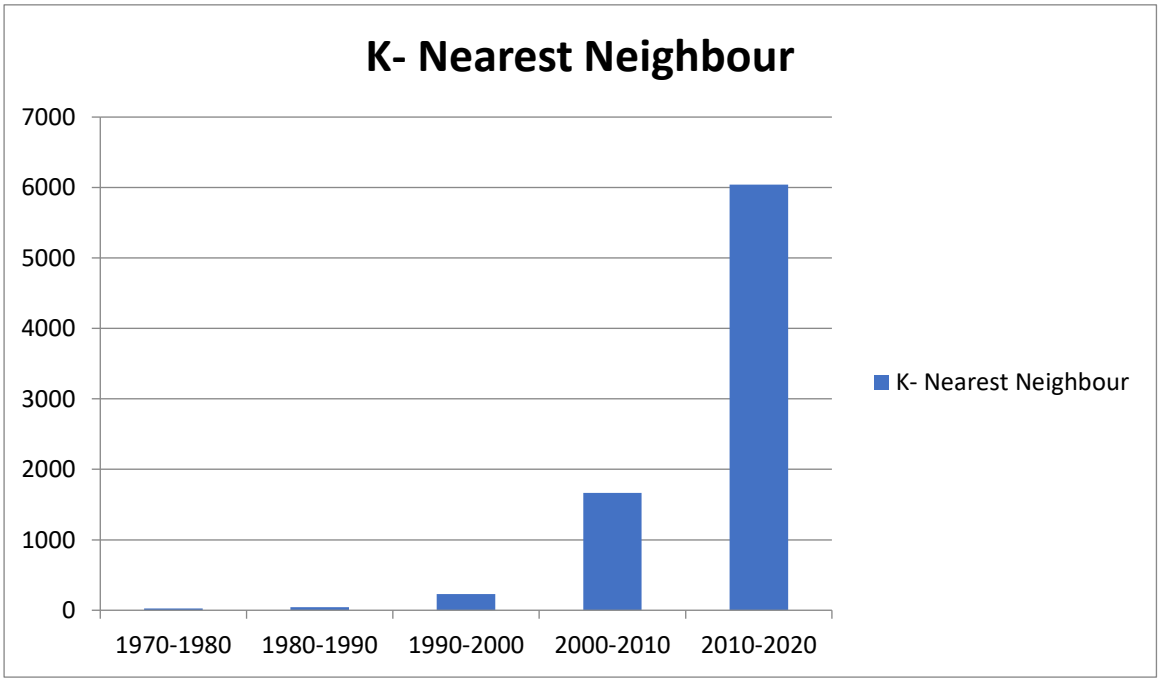
2.3 Applications of the K Nearest Neighbours algorithm in the Financial Sector

One of the most commonly used applications of the K Nearest Neighbours algorithm is the prediction of the value of the stock. Stockbrokers use this to be able to better help their clients with regard to their equity portfolios. Investment Banking sector also uses the KNN algorithm to be able to judge the future performance of the companies and which company has the highest potential for growth and can be invested in. Financial Analysts also use the KNN algorithm to uncover financial trends and to predict the occurrence of recessions.

2.4 Applications of the K Nearest Neighbours algorithm in the Agricultural Sector

There have been some applications of the KNN algorithm in the agricultural sector but they are comparatively lesser compared to the other fields. In the food production, the KNN algorithm has been used for the classification of the seed variety and the seed varieties which provide good yield are separated from the seed varieties which do not. The KNN algorithm coupled with Computer Vision techniques have also been used to analyze satellite images from space of large tracts of land. The algorithm has been accurate in classifying the land according to its arability.

USAGE OF KNN ALGORITHM BY SECTOR OVER DECADES



3. DECISION TREE ALGORITHM

The decision tree has been one of the most popular algorithms for the prediction of heart disease. Especially the concepts of GINI Index and Gain Ration have been used in the process of prediction of heart disease. There have been multiple ways in which decision trees have been used in the prediction of heart disease, and they have generally yielded higher accuracy rates compared to the other classification algorithms. But the usage of the decision trees in the prediction of the most highly correlated feature has been shoddy at best. There have been multiple attempts to use the decision tree algorithm for this usage, but they have yielded deficient F1-Measure Scores. So, It can be concluded that just the usage of the decision tree algorithm alone to be able to predict the most highly correlated feature is not the best method. So, keeping in mind all the

different methods and techniques which can be used for the classification, We can consider that the decision tree algorithm is one of the best in the classification of heart disease. The decision tree algorithm has found several applications in health care such as the early detection of heart disease, reducing the amount of expenditure for both the hospitals and the patients, detection of fraudulent health insurance claims. Studied have also shown that the decision tree algorithm is more accurately working upon the Structured data, and it is obtaining lesser accuracy upon unstructured data. The application of the decision tree algorithm for the heart disease classification has resulted in the accuracies ranging from around 79% to approximately 84%. The various techniques which were used have been as the performance measures of the Decision tree algorithm are Info Gain, GINI Index, and Gain Ratio. The different parameters are Sensitivity, Specificity, Accuracy, which are used to judge the decision tree algorithm from different standpoints.

Algorithms such as a parallel decision tree are also used in the classification of heart disease, which can get accuracies in the higher 80 percent. These algorithms use the parallelism and a multi-threaded approach. The lighter Pthread is used in the parallel decision tree. The main drawback of the parallel decision tree is unable to handle more massive datasets. But this form of the algorithm requires technologies such as Hadoop to be able to handle bigger datasets. Due to these drawbacks, the parallel decision tree algorithm is not used when it comes to more critical data. The decision tree algorithm has also been used for data mining purposes when there have been massive amounts of unstructured data in hospital databases, but to obtain a format that can be fed to an algorithm, the id3 algorithm is used.

Decision trees are also used in other areas such as risk analysis and are highly visual. Decision trees are highly preferred compared to different algorithms because it gets easier to visualize the code that is required to execute the algorithm upon the data that is to be processed. Most of the decision tree algorithms, especially in the application upon the healthcare data, use the UCI Database.

The decision tree classification algorithm is just one method of classification, but the reason that it is preferred is that the algorithm is clear and easy to be understood and implemented. This has led to wide adoption of the decision tree algorithm in the machine learning community. In a research paper authored by professor Yurong Zhong from the Research Institute of electric science and technology, Chengdu, China, The professor uses the improved id3 algorithm to perform the data mining. In this thesis, the professor combines the Taylor formula with the attribute selection of the ID3 Entropy algorithm first. This method yields accuracies that are ranging from mid 70 percent to a maximum of 80 percent. The time taken for this algorithm to run is also much higher, and so this algorithm can be further optimized to be able to decrease the time that it takes to run and also to increase the accuracy.

The decision tree algorithm, along with the ID3 algorithm, is useful, primarily when the target variable that is to be determined is known. If the target variable that is to be understood is not known, then the decision tree algorithm would not be able to predict with high accuracy.

But even the decision tree algorithm alone has not been able to generate the best results in finding the correlation of different features with the occurrence of heart disease. So, this is why an approach that is based upon the combination of different classification algorithms might be able to yield a complete solution

.

3.1 Applications of Decision Tree algorithm in Healthcare and Medical Field

There have been many applications of Decision Tree algorithms coming up in the medical field in recent times. The Study conducted by Vadrevu et al[1] shows that Decision Tree algorithms can be used for the detection and diagnosis of the Dengue fever. There are certain symptoms that are present in the occurrence of Dengue fever. Based upon these symptoms, the Decision Tree algorithm classifies whether the person is suffering from Dengue or not. The study done by Quellec et al[2] shows that the Decision Tree algorithm can also be used for the classification of diseases based upon incomplete medical documents which might contain some crucial missing information.

3.2 Applications of Decision Tree algorithm in Technology Sector

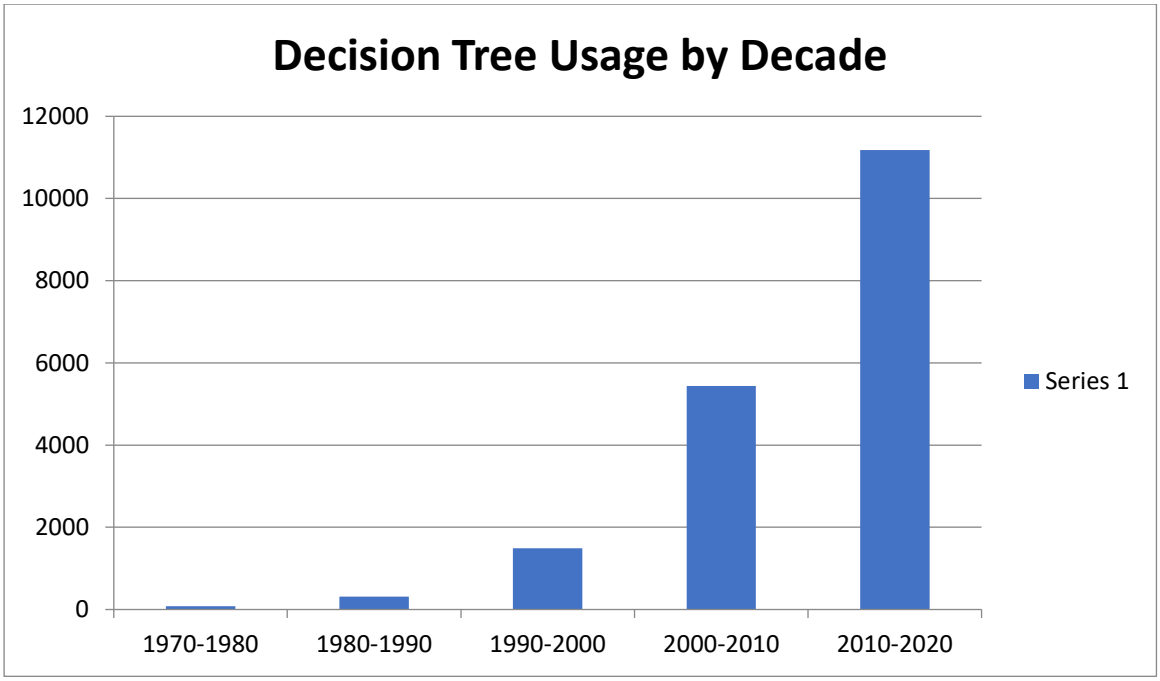
Decision Tree algorithms have been widely used in technology sector since the Later 1990s. Due to the glut of data present today, the applications of Decision Tree algorithms are ever increasing. The Study conducted by Tong et al[3] shows that Decision Tree algorithm can be used in Internet Traffic Management. Based upon the various parameters such as geographical location, data used etc the decision tree algorithm can predict the type of request it is and properly allocate it to the server which is capable of handling it.

3.3 Applications of Decision Tree algorithm in Management Sector

Due to the inherent nature of the Decision Tree algorithm, it has found a multitude of applications in the management sector and has eased the process of decision making. The Study conducted by Nieto et al[4] shows that the Decision Tree algorithms can be used in Educational Institutions to create better learning environments by the evaluation of students. This Evaluation can help in the allocation of the proper amount of resources that are present in the university to be at the required areas. The Decision Tree algorithms are also being used in the management sector for different kinds of companies.

3.4 Application of Decision Tree algorithm in Agricultural Sector

The Decision Tree algorithms are also being used to increase the agricultural efficiency and yield in recent times. By taking into account various environmental factors, climatic conditions along with other factors the Decision Tree algorithms are helping farmers to decide what kind of crops should be placed in the appropriate seasons. Studies have shown that there has been an increase in 36 percent in the yield relative to the mean agricultural yield of the nation when Decision Tree algorithms are being used.



4. RANDOM FOREST CLASSIFICATION ALGORITHM

The Random Forest Classification algorithm is the most popular of the existing classification algorithms which are being used in the classification of heart disease. Random Forest is a special kind of algorithm that can be used not just for classification, but it can also be used for the regression analysis. The Random Forest algorithm is used to assume that there are different trees in a forest. Any forest would be sturdier and more robust if there are several trees in the forest.

Similarly, the random forest algorithm also works well when the number of trees in the forest is more. This gives a higher accuracy to the random forest algorithm. The random forest algorithm could necessarily be considered as consisting of many decision trees. Each decision tree that is present in the random forest can be regarded as just a node in the random forest. Just like the performance metric that is used in the decision tree algorithm, the random forest also uses the same performance metrics, which are the GINI Index and Gain approach.

The main advantage of using a random forest classifier is that this algorithm is capable of handling the missing values. These missing values generally cause a problem in the running of other algorithms. But the random forest algorithm is designed in such a way that it is capable of handling those erroneous values without causing or getting an error in the modelling process. In the random forest algorithm, each decision tree algorithm has one vote, and the output of the random forest algorithm is based upon the maximum outcomes of the decision tree algorithms. For example, if there are about a hundred decision tree algorithms that are present in the random forest algorithm. Then if we obtain the output of fifty-one decision trees as “Yes” and the output of forty-nine decision trees as “No” for a particular scenario, then the final outcome of the random forest algorithm would be “Yes” because the majority of the decision trees that were present in the random forest algorithm have given the output as “Yes”.

The Random Forest algorithm has most applications in the four sectors. These four sectors are Banking, Medicine, Stock Market, and E-Commerce. In the Banking Sector, The Random Forest algorithm is used to find fraudulent customers. Since the usage of the ordinary decision tree algorithm might not always give the correct prediction, the Random Forest Algorithm is used to classify whether the customer is fraudulent or not. The other applications of the Random Forest Algorithm include stock classification into the long term or short term. The Random Forest Algorithm is used by stockbrokers and other merchants who want to know whether the value of the stock would increase in the future or would it decrease in the future. So, Multiple decision trees are used to be able to predict the outcome of the stock, and the prediction which is given by most decision trees would be the final prediction of the random forest.
The Random Forest algorithm is also used in E-Commerce applications. It is used to be able to recommend the products to a particular user. Based upon the search history and the previous purchases of the user, the machine learning model must be able to suggest to the user a specific product. So, The Random forest algorithm is used to be able to generate the final prediction if whether a user is more likely to purchase a particular product that he has been shown or not. The Random Forest algorithm has also been used upon heart disease prediction by the professors and researchers from China Medical University, Taichung. The professors used the Random Forest Algorithm for the classification of whether a person is more likely to suffer from heart disease or not. In the research conducted by these professors, the Random Forest algorithm was able to generate an accuracy of a maximum of 86.4 %, which is more than the accuracy obtained by the decision tree algorithm upon the same dataset. So, We can conclude that the random forest algorithm is one of the best classification algorithms and that the rate of accuracy that it can obtain on a particular dataset would generally be equal to or more than the accuracy that is received upon the usage of the decision tree algorithm upon the same dataset.

4.1 Applications of Random Forest algorithm in Healthcare and Biological warfare

Random Forest is one of the older machine learning algorithms and has been used for a wider variety of applications. The Study conducted by the Sanghir et al[9] shows that the random forest technique can be used to classify the biochemical agents into multiple kinds so that they can be used to synthesize biological weapons and weapons of mass destruction. The random forest algorithm can also be used for the traditional healthcare applications such as classification of tumors into benign and malignant etc. Newer applications of the random forest algorithm are also being found out in animal and drug testing and synthesis.

4.2 Applications of Random Forest algorithm in the Technology sector

Random Forest algorithm has been used for classification of the behaviour of users. Political Analysts have recently started the usage of the Random Forest algorithm to classify the voting population according to whether they would vote for a particular party or not. This has helped them to concentrate their efforts on the non voters and try to convert them into voters. Social media websites have also been using the random

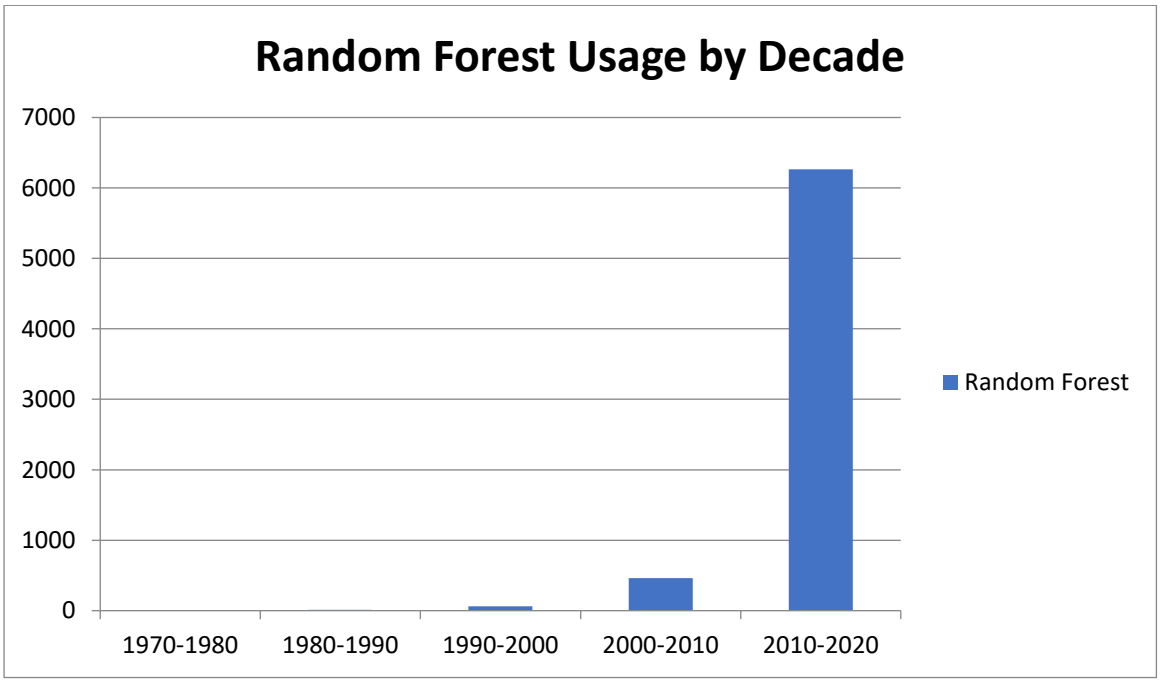
forest algorithm to be able to classify their users according to their interests and be able to show the users advertisements that would potentially interest them.

4.3 Applications of Random Forest algorithm in the News and Media Sector

Due to the recent dwindling numbers of readers of traditional newspapers, newspapers are using Machine learning algorithms such as the Random Forest algorithm to be able to identify the differences between a potential user and a non user. The other application of Random Forest algorithm in the media include classification of a news story. Based upon the results that have been obtained in the Random Forest algorithm, the necessary airtime that a news story requires is decided. This algorithm has also been used for other traditional applications such as weather forecasting etc.

4.4 Applications of Random Forest algorithm in the Sciences

Random Forest algorithm has also found various applications in scientific research and discoveries. The Random Forest algorithm has been used in chemical synthesis to find out the appropriate concentration required to be able to create those required solutions. During Electricity storage and supply, the random forest algorithm has also been used to discover the appropriate voltages which are needed to be able to transmit the electric current with the least power loss and the highest efficiency.



5. NAIVE BAYES ALGORITHM

Naive Bayes Algorithm is a collection of algorithms that are used for the classification. In the Naive Bayes Algorithm, there is not just one algorithm, but a group of algorithms is used to be able to perform the work of classification. The main striking point in the

Naive Bayes Classifier is that there is a family of algorithms. But these family of algorithms which are present in the Naive Bayes algorithm is independent of each other. The formula which is the basis of these Naive Bayes classifier is the Bayes theorem. Based on the Bayes theorem, the Naive Bayes Algorithm tries to classify the data that is present.

The Naive Bayes Classifier is a probabilistic classifier, and all the features that are used are considered to be independent of each other. The most critical application and also the most popular application is spam classification. The Naive Bayes Classifier is used to detect whether a particular piece of mail was from a legitimate source or an illegitimate source. The main advantage with the Naive Bayes Classifiers is that they are very scalable, and the number of features can be added easily without getting any difference in the performance measures such as accuracy. The accuracy of the Naive Bayes algorithm can generally be comparable to the accuracy of the Support Vector Machine.

The Naive Bayes algorithm also finds application in the medical and healthcare field. The Naive Bayes algorithm takes linear time rather than multiple different types such as polynomial etc. There are different types of Naive Bayes algorithms. These different types of Naive Bayes algorithms are Bernoulli Naive Bayes, Multinomial Naive Bayes, Gaussian Naive Bayes. Semi Supervised Parameter estimation etc. Recent studies have also shown that the results that are obtained through the Naive Bayes algorithm are very much similar to the results that are captured in the Logistic Regression algorithm.

5.1 Applications of Naive Bayes Algorithm in Healthcare

In Recent times, the Naive Bayes algorithm has found a lot of usage especially in the healthcare sector. The Study conducted by Altayeva et al(19) shows that Naive Bayes can be used to make medical decision making diagnosis systems. This decision making system contains various parameters such as age, sex, blood pressure, blood sugar levels, chest pain, electrocardiogram. The increased success of the Naive Bayes algorithm in being able to aid healthcare systems has increased its usage tremendously. Newer developments have been taking place in the medical field which has resulted in the increased number of applications for the Naive Bayes algorithm such as Cancer Cell detection, Liver cirrhosis detection and prevention.

5.2 Applications of Naive Bayes Algorithm in the Technology Sector

Naive Bayes algorithm was first employed to be able to solve the problems in the technology such as Spam Filtering, Image Classification etc. The Study conducted by You et al(20) shows that this algorithm can be used for the spam detection and removal. Many of the leading email providers such as Gmail and Yahoo are using the Naive Bayes theorem for their in-house Spam filtering techniques. Some of the other applications of Naive Bayes theorem include Image Classification, Recent Studies have shown that when Naives Bayes theorem is coupled with Computer Vision techniques it can be used for the classification of images into various different categories.

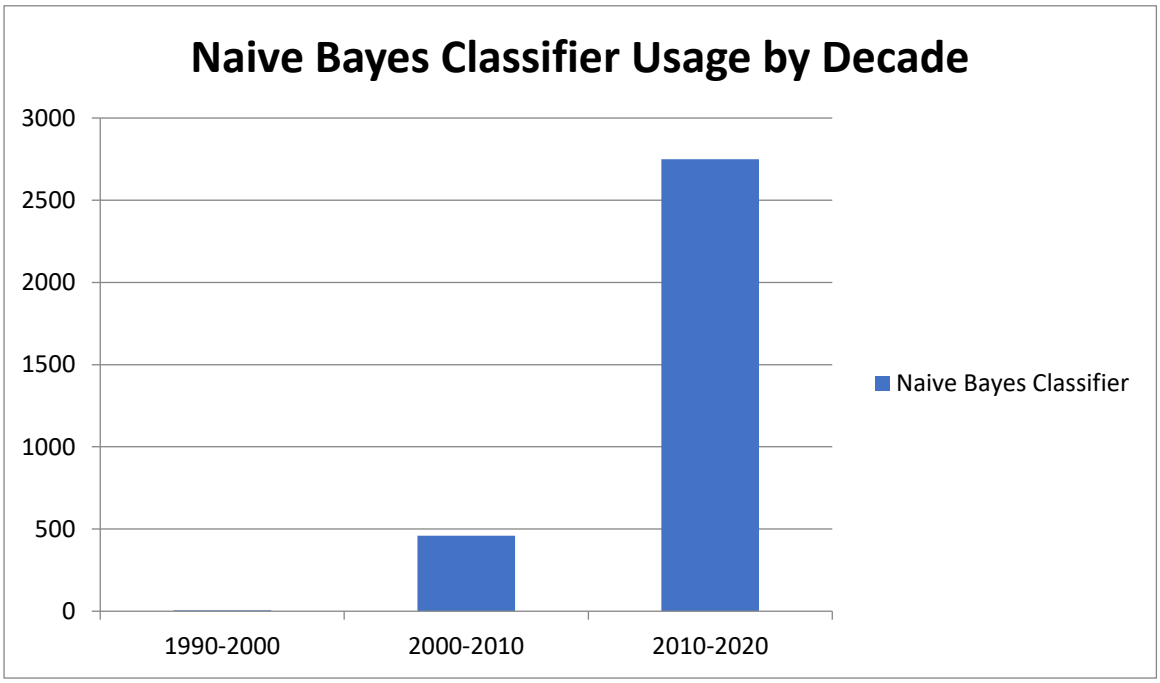
5.3 Applications of Naive Bayes Algorithm in the Manufacturing Sector

There has been an increase in the computerization of the processes in the manufacturing sector. This has led to the machine learning algorithms such as Naive Bayes being used

to increase the industrial output and efficiency. Naive Bayes has been used in conjunction with other algorithms such as the Support Vector Machine to be able to detect the quality of the finished goods and to be able to weed out bad products. Studies shows that the application of the Naive Bayes algorithm has increased efficiency by 30 percent in the Manufacturing Sector.

5.4 Applications of Naive Bayes Algorithm in Media and Sports

Naive Bayes algorithm has also found usage in many different sort of news such as sports and weather. Based upon the average temperatures and climate readings in the particular region, the Naive Bayes algorithm has been able to extrapolate upon the outcomes and predict future weather. Nowadays, sports analysts from leading teams around the world have been using Naive Bayes algorithm to be able to classify the players capabilities in the player auctions and are being able to recruit the best talent for their budget into their teams.



INFERENCE FROM THE SURVEY

The multitude usecases of the various different classification algorithms such as KNN, SVM, Naïve Bayes, Random Forest, Decision Tree have been analyzed and studied. of Machine Learning algorithms such as Logistic Regression and Decision Trees can be used to identify the early phases of the Dengue fever and reduce its prevalence. Decision Trees can also be used to identify the missing information present in the medical documents. The Decision Tree algorithm can also be used for the technological applications such as classifying the Internet

Traffic also. There are a range of applications that are present in the Decision Tree algorithm and also how efficient it is for modelling complex data. There are a plethora of studies which shows the range of application of the Random Forest algorithm. The Random Forest algorithm is used to classify and segregate the features that are necessary for the decision making in higher educational institutions. The K-Nearest Neighbour algorithm to be able to classify the data present in the Virtual Reality databases that is present in the form of Spatial Databases. The variability in the applications of the Random Forest algorithm is also astounding. The Naïve Bayes and Support Vector Machine algorithms can be used for the classification of a person's heart condition based upon various parameters and components that are present in a Electrocardiogram signal. The Support Vector Machine algorithm is being used in the classification of Fetal Heart rate and it helps to diagnose diseases.. SVM can be used for the finding of any abnormal activities that are occurring in the Fetus based upon its heart rate. A combination of the Machine Learning algorithms of Random Forest, Support Vector Machine and Logistic Regression has been able to classify the likelihood of a person being hospitalized due to a chronic disease. The combination of the stated algorithms also shows the value in using a combination of multiple classification algorithms to be able to predict the outcome accurately. Another very unique application of the Random Forest algorithm in which the model is able to predict the likelihood of an organism surviving a catastrophic extinction event based upon the ordering of the DNA structures within its cells. A hybrid Machine Learning technique has been made up of a combination of multiple Machine Learning algorithms such as KNN, Naïve Bayes, Genetic algorithm and Decision Tree. The usage of these combinations of multiple Machine Learning algorithms would be better equipped and more accurate in being able to the predict the outcome where a person is being affected with heart disease or not.

CONCLUSION

Over the past few decades, Heart disease has become the leading cause of death in mankind. Heart disease can be present in a human being over a long period of time without containing any visible symptoms and suddenly exhibit itself. When a person is subjected to a heart attack, time is of crucial importance. Heart disease can be regarded as a complex disease because there are various different factors present in a person, each of which play a role in the condition of the heart. Since, there are multiple forces at action on the human heart and time is of utmost importance for a person subjected to a heart attack, researchers have used Machine Learning models to be able to model the heart disease prediction. Since there are multiple factors that are present in the occurrence of heart disease, it is difficult for the Machine Learning models to predict with a higher amount of accuracy. So, to reduce the number of factors at play, researchers have been using Principal Component Analysis(PCA). When Principal Component Analysis has been implemented on the data, the number of factors are reduced to increase the efficiency of the modelling. But, the drawback that has occurred with Principal Component Analysis is that the accuracy rate has been around fifty to sixty percent. This kind of lower accuracy has led to hospitals and medical institutions not being able to use the algorithm practically. In this Study, the approach that is taken is to use five different classification algorithms which are Naive Bayes Theorem, Random Forest, Decision Tree algorithm, K-

Nearest Neighbours and Support Vector Machine. The number of features are not reduced but are used in the modelling process as it is. To be able to increase the efficiency, we use combinatorics to be able to reduce the number of iterations of modelling. We use each of the classification algorithm on all the data and perform the modelling. Similarly, all the five algorithms are used to model on each and every feature responsible for heart disease. For each combination of features we are going to obtain a different accuracy. In all the combinations that are present for a particular algorithm, we are going to use the first hundred combinations which provide the highest accuracy. The feature which is most prevalent in all these hundred combinations is considered as the feature which is most highly correlated with the occurrence of heart disease. For the generation of the entire accuracy of our model in predicting the heart disease, we can perform the arithmetic mean of all the different accuracies that are obtained for each combination of features. This technique has been generating accuracies between eighty to ninety percent for each algorithm which is a significant improvement compared to the the previous models using the Principal Component Analysis(PCA). Due to the better accuracy rates and also because this model is able to find the feature more highly correlated with the occurrence of heart disease, this model can perform the detection and prevention of heart disease quickly. Since time is a very important factor in the resuscitation of a person undergoing a heart attack, our model reduces the time to identify the occurrence of heart attack and saves lives.

REFERENCES

1. Vadrevu Sree Hari Rao, Mallenahalli Naresh Kumar “ A New Intelligence-Based Approach for Computer-Aided Diagnosis of Dengue Fever ",IEEE, Volume: 16 Issue:1,2012

2. Gwenole Quellec, Mathieu Lamard, Lynda Bekri, Guy Cazuguel, Christian Roux, Beatrice Cochener " Medical Case Retrieval From a Committee of Decision Trees", IEEE, Volume :14 Issue:5,2010

3. Da Tong, Yun R. Qu, Viktor K. Prasanna " Accelerating Decision Tree Based Traffic Classification on FPGA and Multicore Platforms ", IEEE, Volume:10, IEEE, Volume:7,2016

4. Yuri Nieto , Vicente Gacia-Diaz , Carlos Montenegro , Claudio Camilo Gonzalez, Ruben Gonzalez Crespo " Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions ", IEEE, Volume:12 Issue:4, 2019

5. Yunjun Gao, Baihua Zheng, Gencai Chen, Wang-Chien Lee, Ken C.K. Lee, Qing Li " Visible Reverse k-Nearest Neighbor Query Processing in Spatial Databases", IEEE, Volume:21 Issue:9, 2009

6. PengfeiLi ; Yu Wang ; Jiangchun He ; Lihua Wang ; Yu Tian ; Tian-shu Zhou ; Tianchang Li ; Jing-song Li "High-Performance Personalized Heartbeat Classification Model for Long-Term ECG Signal",IEEE, Volume: 64 Issue:1, 2016

7. JiříSpilka ; Jordan Frecon ; Roberto Leonarduzzi ; Nelly Pustelnik ; Patrice Abry ; Muriel Doret "Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification",IEEE, Volume: 21 Issue:3, 2017

8. Theodora S. Brisimi ;Tingting Xu ; Taiyao Wang ; Wuyang Dai ; William G. Adams; Ioannis Ch. Paschalidis "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach",IEEE, Volume:106 Issue:4, 2018

9. Helal Sanghir ; Dalila B. Megherbi "A random forest based efficient comparative machine learning predictive DNA-codon metagenomics binning technique for WMD events & applications" IEEE, Volume:9, Issue:9, 2013

10. SenthilkumarMohan ;ChandrasegarThirumalai ; Gautam Srivastava "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE, Volume:7,2019

11. Amit Krishna Dwivedi ; Syed Anas Imtiaz ; Esther Rodriguez-Villegas,"Algorithms for Automatic Analysis and Classification of Heart Sounds–A Systematic Review", IEEE, Volume:7, 2018

12. Paolo Melillo ; Nicola De Luca ; Marcello Bracale ; Leandro Pecchia "Classification Tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability", Volume:17 Issue:3,IEEE, 2013

13. C. Beulah ChristalinLatha, S. Carolin Jeeva; "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques",ScienceDirect,Volume 16, 2019

14. Mohammad Shafenoor,AminacYinKiaChiama, Kasturi DewiVarathanb; "Identification of significant features and data mining techniques in predicting heart disease", Science Direct,Volume 36, March 2019

15. Deepti Sisodia, Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, Volume 132, 2018.

16. Beunza, J-J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., Hurtado, C., Landecho, M.F., Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease), Journal of Biomedical Informatics (2019)

17. MoloudAbdar , Mariam Zomorodi-Moghadam , Resul Das , I-Hsien Ting , Performance analysis of classification algorithms on early detection of Liver disease, Expert Systems With Applications (2016)

18. Sonam Nikhar, A.M. Karandikar."Prediction of Heart Disease Using Machine Learning Algorithms", *International Journal of Advanced Engineering, Management and Science(ISSN: 2454-1311)*,vol.2,no. 6, pp.617-621,2016

19. Aigerim Altayeva, Suleimenov Zharas, Young Im Cho "Medical decision making diagnosis system integrating k-means and Naïve Bayes algorithms", IEEE 2016

20. Wanqing You**,** kai qian**,** Dan Lo**,** Prabir Bhattacharya**,** Minzhe Guo**,** Ying Qian "Web Service-Enabled Spam Filtering with Naive Bayes Classification" IEE Journal 2015