IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

KIDNEY DISEASE PREDICTION USING DIFFERENT CLASSIFICATION TECHNIQUES OF MACHINE LEARNING

Prof. Deepali Joshi
Information Technology Dept.
Vishwakarma Institute of Technology, Pune
Pune, India
deepali.joshi@vit.edu

Harsh Said
Information Technology Dept.
Vishwakarma Institute of Technology, Pune
Pune, India
harsh.said21@vit.edu

OmkarBhosale
Information Technology Dept.
Vishwakarma Institute of Technology, Pune
Pune, India
omkar.bhosale21@vit.edu

Ritika Garad
Information Technology Dept.
Vishwakarma Institute of Technology, Pune
Pune, India
ritika.garad21@vit.edu

Rakeshkumar Visave
Information Technology Dept.
Vishwakarma Institute of Technology, Pune
Pune, India
rakeshkumar.visave21@vit.edu

Samarth Usture
Information Technology Dept.
Vishwakarma Institute of Technology, Pune
Pune, India
samarth.usture20@vit.edu

Abstract — Based on the rise in chronic kidney disease (CKD) incidence in recent years, a more accurate early prediction model is required to identify high-risk individuals before they develop end-stage renal failure. To date, it has been determined that diabetes mellitus6, obesity5, and female sex4 are all significant risk factors for chronic renal disease. Recently, several biomarkers connected to CKD have been identified. Treatment for renal failure and chronic kidney disease is both expensive and inefficient. Only around 5% of those with early CKD are aware of their illness, though20. Once CKD is detected, glomerular damage has typically reached over 50% and is irreversible. An accurate chronic renal illness prediction can be very helpful in this regard. This study aims to forecast renal failure. Static Vector Machine (SVM)

Keywords — disease, prediction, algorithm, SVM,k-nearest neighbor algorithm ; Latent Dirichlet Allocation

I. INTRODUCTION

Chronic renal disease impact on the healthcare system is increasing. It is a significant burden on the medical field and other healthcare centers or system due to its increased prevalence, high risk of occurrence of end-stage renal disease, and poor prognosis for morbidity and mortality. A global health issue is progressively becoming more serious.

The damage to your kidneys can be significantly slowed down by adopting healthy lifestyle changes, though, if CKD is identified and treated early on. We must gain a deeper awareness of a few signs brought on by renal illness if we are to prevent serious harm. The primary goal of this project is to apply various algorithms to predict renal illness by evaluating data sets, flow diagrams, and block diagrams. Correction, which is not always possible, may help with early detection of renal illness.We need a better knowledge and information of a few kidney illness symptoms in order to prevent irreversible injury.By performing some analysis on the data of disease and patients health parameters and comparing it on indices, using three machine learning classification algorithms to forecast the sickness, and then selecting the schema which has the highest accuracy value, the main reason for implementation of this system and making research on it is to anticipate renal illness. The three classification techniques employed are Random Forests, K-nearest Neighbor, and SVM. The class, target, labels, and categories of a data point are predicted using machine learning classifiers.

This is the issue that worries us. A typical kidney disfunctioning and progressive illness causing due to some renal diseases or progressive renal failure over time define chronic kidney disease (CKD), also called

as chronic renal disease. As we see the health parameters the people who suffers from this renal or kidney failure are those who have high blood pressure or diabetes,or who have a blood relative who has the ailment, are regularly screened for chronic kidney disease. Because the decline in kidney function must extend longer than three months, it differs from acute renal disease. The primary goal of this research was to predict chronic renal illness.

Chronic kidney disease or renal diseases are predicted using data mining classification methods by applying various classification method on the data record of patients health condition. In this study, classifiers used are such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Latent Dirichlet Allocation (LDA) were used. The performance of each is then assessed for calculation of accuracy, precision, and F-measure for each classifier separately and according to the classifiers value and degree of accurateness of each classifier.

II. LITERATURE SURVEY

In 2015, A research has been gone on the topic of application of machine learning on the application in cancer patients health detection on basis of that problem statement they defined a system by which their may be chances of developing a system which predicts the diseases id causes or not to that patients.that is mentioned in the paper A research of Machine learning applications in cancer prognosis and prediction was proposed by Konstantina Kourou et al [1].

In this paper , they provided an overview of numerous contemporary machine-learning algorithms techniques for cancer detection prediction. They have provided an analysis of recently published materials for cancer detection studies to date.

In 2015 another researcher P.Swathi Baby et al [2] proposed as a system to develop a predictive data mining-based analysis and prediction system. The data collection for renal disease is analyzed using the Weka and Orange software. Machine learning algorithms that give statistical analysis and renal disease prediction, such as Nave Bayes, J48, AD Trees,K Star, and Random forest, are used for the performance calculation of each algorithm. According to the research they hav into conclusion that K-star and random forest are the best algorithms for the dataset utilized, since they create models faster (0 sec and 0.6 sec, respectively), and their ROC values are 1 and 2, respectively.

In 2015,Mr. S Dayanand [3]proposed using Support Vector Machine (SVM) and Artificial Neural Networks to predict kidney disorders (ANN). The comparison of these two algorithms' precision and execution times is the main goal of this research. According on the experimental data, the ANN performs better than the other algorithm.

IN 2020 NV researchers like Ganapathi Raju, K Gayathri Praharshitha, K Prasanna Lakshmi,[4] This study's primary goal is to diagnose chronic renal disease by applying several classification algorithms to the patient's medical file. Finding the most appropriate classification algorithm to utilise for the diagnosis of CKD based on the classification report and performance variables is the main focus of this research project.

In 2017 Alessia Sarica, Antonio Cerasa, Aldo Quattrone[5] addressed the benefits of RF, taking into account any potential drawbacks, and urging additional research on comparisons of this algorithm with other widely employed classification systems, notably in the early prediction of the development from MCI to AD.

III.OVERVIEW

The objective of this research is to conduct a comparative examination of kidney disease prediction and analysis of patients medical history data utilizing this intelligent machine learning approaches are made according to the end results of this learning approaches the decision has been that the occurrence of disease have chances or not.. According to the study's findings, the decision tree made by the data values and logistic regression can be utilized to better correctly forecast chronic renal disease. we developed a convolutional neural network based model which is based on multimodal sickness risk prediction algorithm using structured and unstructured patients previous medical history data. The use of data mining techniques to disclose and extract hidden information from clinical and laboratory medical history of patient data is discussed in this paper. Multilayer Perceptron (MLP),Probabilistic Neural Networks (PNN), and other algorithms are used. The results demonstrate that the PNN algorithm performs better in terms of classification and prediction when it comes to determining the severity stage of chronic renal disease. In this study, the Random Forest (RF) technique is used to decrease high-dimensional and multi-source data in a variety of scientific fields. The goal of the paper was to look at the current state of the art in using RF on single and multi-modal neuroimaging data to predict renal / kidney disease. Existing machine learning methods do not deliver the level of prediction accuracy that is required. To close the gap, this research suggests a new classification technique for predicting chronic kidney disease from a medical dataset that includes environmental elements. The Optimal Fuzzy-K Nearest Neighbor Technique is the methodology presented in this article.

IV. METHODOLOGY/EXPERIMENTAL

A) Flowchart

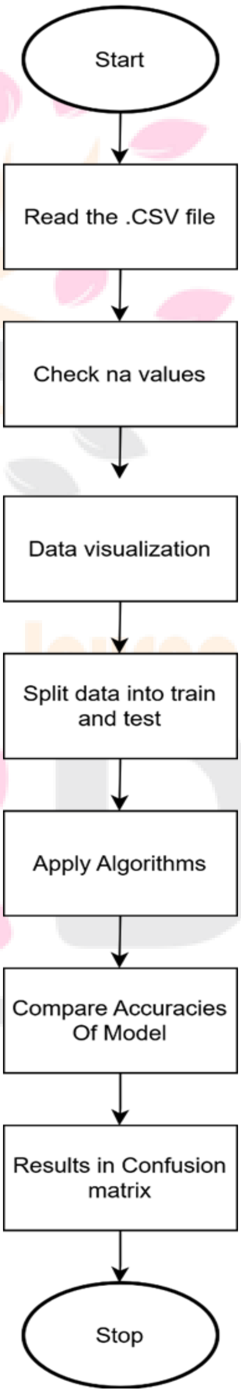
Start
Read the .CSV file
Check na values
Data visualization
Split data into train and test
Apply Algorithms
Compare Accuracies Of Model
Results in Confusion matrix
Stop

Figure 1: Flowchar

Three categorization algorithms are used in this study to predict the existence of chronic renal disease in adults. Support vector machine and KNN classifiers were employed as classifiers. The data set for chronic kidney/renal disease was gathered and applied to each classifier to predict the disease, with the higher accuracy, precision, and F measure used to evaluate the classifier's performance. Proposed Approach to Predictive Data Mining Architecture

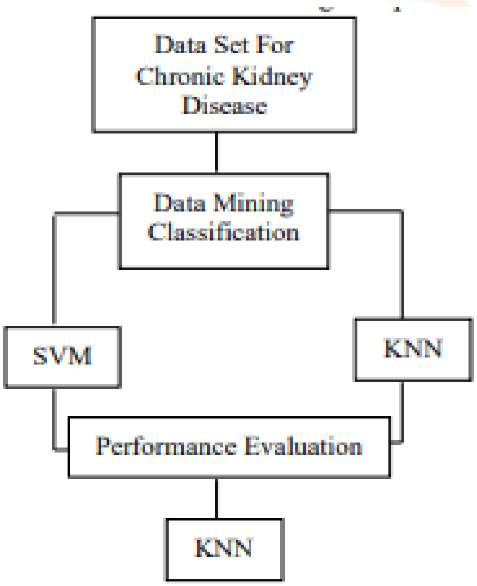
Data Set For Chronic Kidney Disease
Data Mining Classification
SVM
KNN
Performance Evaluation
KNN

Figure 2:-Algorithms selection procedure

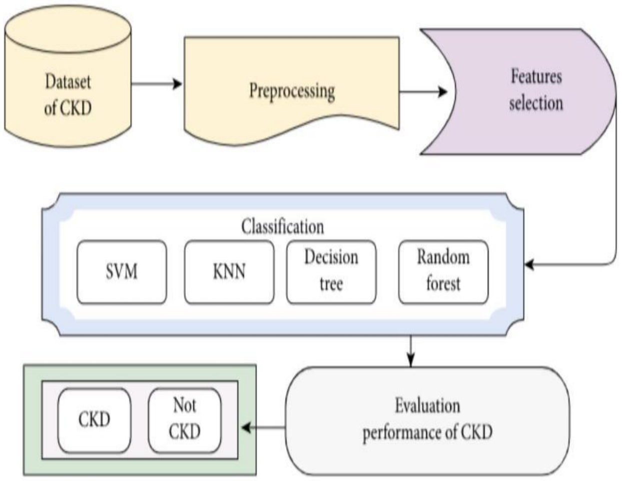

Figure 3:-prediction of chronic kidney disease using and Chronic Kidney Disease; SVM; KNN

Dataset

The data was gathered from a variety of medical labs, centers, and hospitals. The synthetic kidney function test (KFT) dataset was developed from this for renal disease analysis.

This comparative study is based on a dataset with 401 cases and six attributes. age, gender, urea, creatinine, and glomerular filtration rate are the attributes in this KFT dataset (GFR). This dataset contains information about renal diseases that are affected.

B) Method

Many classification techniques are applied on patients med details dataset and comparison made according to the performance of each algorithm on testing dataset the prediction has been done accordingly. Training dataset contains 320 rows which are used to analyze and train the model. Testing dataset contains 80 rows which is used to measure the performance of the model on unseen dataset. The proposed system here uses three classification The algorithms used are KNN classifier, Random Forest and Support vector machine. The data set information for chronic kidney disease was gathered and applied on each classifier to predict the disease occurrence result and the performance of the classifier is evaluated based on rate of accuracy, precision and F measure values collected from whole analysis . The description of the architecture is as follows: The main reason to use an SVM is because the problem that has been given may or may not be linearly separable. At that point , we have to use an SVM algorithm which has a non linear kernel. KNN is robust to noisy training data and is effective in a large number of training examples. But to use this algorithm , we have to determine the value of the K parameter which is the value of the number of nearest neighbors and the type of distance to be used. The computation time is also large as we need to calculate the distance of each query instance to all training samples. Random Forest is nothing more than a group of Decision Trees bunched together. They can be used to handle categorical features of data set values . This algorithm has the ability to handle high dimensional spaces not only this but also the large number of training examples.

The following are some parameters to define are.

Data Mining Techniques

Support Vector Machines (SVMs)

Support Vector Machines (SVM) is can be said as a strong, cutting-edge linear and nonlinear regression technique. The application of this algorithms is Oracle Data Mining uses SVM for binary and multiclass classification.. The SVM's advantage is that it can estimate the separation between a molecule and the hyper plane in a modified (nonlinear) feature space without specifically changing the initial descriptors., thanks to the so-called "kernel trick." The most generally utilized radial basis function kernel (Gaussian kernel) was used in this investigation.

K-nearest neighbor Classification

For classification and regression, the K-Nearest Neighbor algorithm (K-NN) is a nonparametric pattern recognition model which takes the input with K closest training instances in the feature space in both the situations. Instance-based learning is a form of K-NN. The output of KNN Classification is a class membership. A 75% vote of neighbors is required for classification. If K = 1, the class has only one nearest neighbor. In a conventional weighting system, each neighbor is assigned with a value of a weight is 1/d, where d is the distance between the data points .The shortest distance between any two neighbors is always a straight line, which is called the Euclidean distance . The K-NN algorithm's drawback is that it is insensitive to the data's local configuration value. Feature extraction refers to the process of converting raw data into a set of features. Before using the K-NN method in Feature space, the raw data is extracted.

Random Forest Technique for classification

Random Forest is an algorithm used for supervised classification. To efficiently compute accuracy, . The accuracy of the classifier is measured by value which is directly proportional to the number of no. of trees. Random forest results are reliable even without hyperparameter tuning due to their flexibility. It's simple and works very efficiently, especially for large dataset sizes. Maintain accuracy by detecting outliers and anomalies. However, it is not very easy to implement and computationally intensive.

V. LIMITATIONS

With certain shortcomings, the current chronic renal disease prediction method is adequate. The table below shows the research that has been done to predict and identify various kidney disorders. There is still a need for a new chronic kidney disease prediction system. There is still a requirement of a decision support system for early detection of chronic renal disease, as little research has been done in this area.

VI. ALGORITHMS AND EVALUATION

The below specifies results analyze the accuracy for three algorithms and displays the accurate values on the graph as shown below.

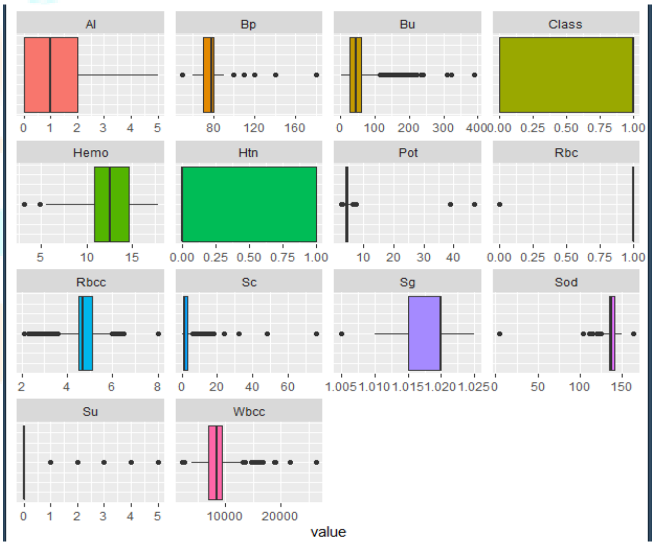

fig:- box plots of attributes

Total time for KNN is : 16.13 sec elapsed

Confusion Matrix and Statistics

```
 Reference
 Prediction
                   0        1
               0 117        9
               1   5      189
```

<u>Accuracy : 0.9562</u>

95% CI : (0.9277, 0.9759)

No Information Rate : 0.6188

P-Value [Acc > NIR] : <2e-16

Kappa : 0.9078

Mcnemar's Test P-Value : 0.4227

Sensitivity : 0.9590

Specificity : 0.9545

Pos Pred Value : 0.9286

Neg Pred Value : 0.9742

Prevalence : 0.3812

Detection Rate : 0.3656

Detection Prevalence : 0.3937

Balanced Accuracy : 0.9568

'Positive' Class : 0

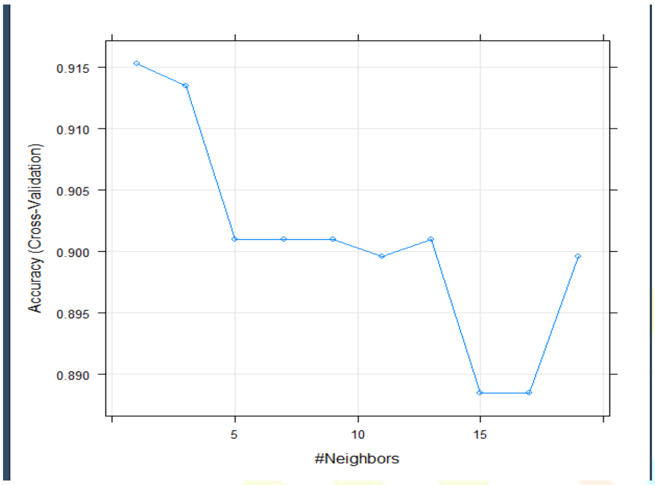

fig2:- accuracy prediction using knn algorithm

Total time for RF is : 1.56 sec elapsed

Confusion Matrix and Statistics

Reference

Prediction

```
        0    1
     0  116  11
     1  6    187
```

Accuracy : 0.9469

95% CI : (0.9163, 0.9688)

No Information Rate : 0.6188

P-Value [Acc > NIR] : <2e-16

Kappa : 0.8883

Mcnemar's Test P-Value : 0.332

Sensitivity : 0.9508

Specificity : 0.9444

Pos Pred Value : 0.9134

Neg Pred Value : 0.9689

Prevalence : 0.3812

Detection Rate : 0.3625

Detection Prevalence : 0.3969

Balanced Accuracy : 0.9476

'Positive' Class : 0

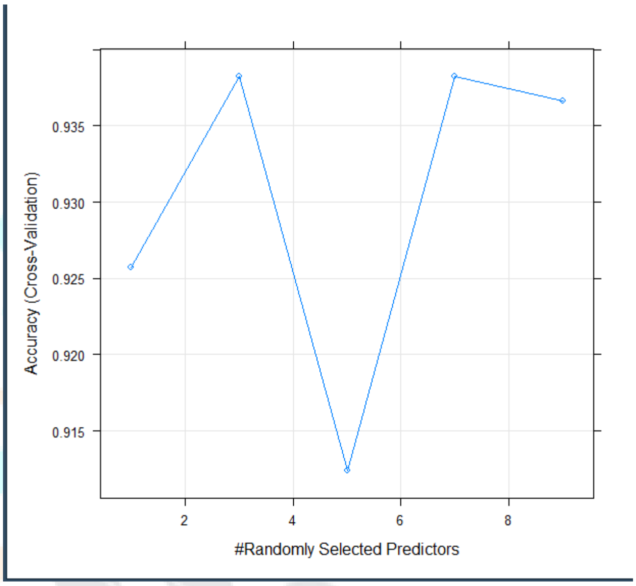

fig3:Randomly selected predicators

Total time for SVM :: : 3.9 sec elapsed

Confusion Matrix and Statistics

Reference

Prediction

```
        0    1
     0  122  5
     1  0.   193
```

Accuracy : 0.9844

95% CI : (0.9639, 0.9949)

No Information Rate : 0.6188

P-Value [Acc > NIR] : < 2e-16

Kappa : 0.9671

Mcnemar's Test P-Value : 0.07364

Sensitivity : 1.0000

Specificity : 0.9747

Pos Pred Value : 0.9606

Neg Pred Value : 1.0000

Prevalence : 0.3812

Detection Rate : 0.3812

Detection Prevalence : 0.3969

Balanced Accuracy : 0.9874

'Positive' Class : 0

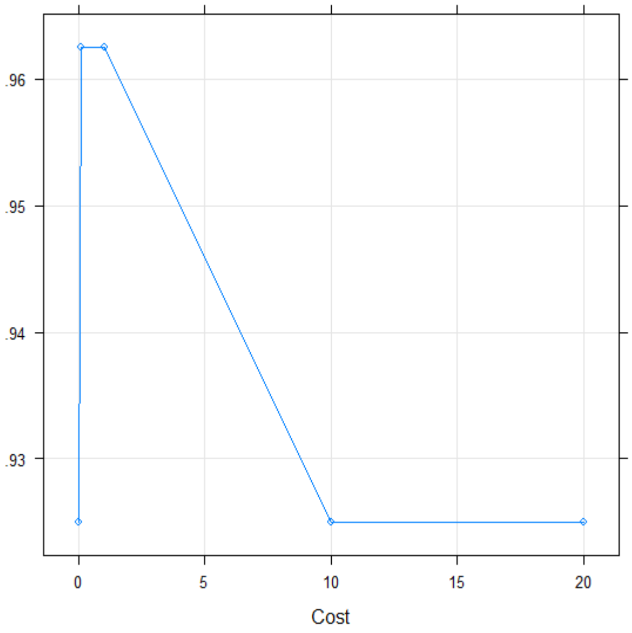

fig4:- SVM accuracy matrix plot

LDA

```
Confusion Matrix and Statistics

Reference
Prediction
        0    1
   0   121   9
   1    1    189

 Accuracy : 0.9688
 95% CI : (0.9433, 0.9849)
 No Information Rate : 0.6188
  P-Value [Acc > NIR] : < 2e-16
  Kappa : 0.9346

 Mcnemar's Test P-Value : 0.02686
 Sensitivity : 0.9918
 Specificity : 0.9545
 Pos Pred Value : 0.9308
 Neg Pred Value : 0.9947
 Prevalence : 0.3812
 Detection Rate : 0.3781
 Detection Prevalence : 0.4062
 Balanced Accuracy : 0.9732

     'Positive' Class : 0
```

VII. Future Scope

To detect CKD, this work uses SVM and KNN and Random forest classification techniques. The performance of the utilized classifiers can also be evaluated and compared to that of other classifiers. Early identification of CKD aids in the timely treatment of patients with the disease, as well as the prevention of the disease worsening. The medical sector has to be able to predict diseases early and treat them promptly. In future work, new classifiers can be utilized and their performance evaluated in order to find better objective function solutions.

VIII. Conclusion

As we have already seen, data mining and machine learning are being used in the medical industry. In order to forecast CKD, a new decision support system is used in this work. Even so, the classifiers were effective at predicting other illnesses. In this study, three distinct classifiers are used to predict the presence of chronic kidney disease, and the effectiveness of each is compared. As a result of the investigation, we discovered that SVM classifier outperformed RF and KNN classifiers. The likelihood of CKD prediction has increased.

IX. References

[1] Gazi Mohammed Ifraz, Muhammad Hasnath Rashid, Tahia Tazin, Sami Bourouis and Mohammad Monirujjaman Khan, “Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods”. Published on 3 Dec 2021.

[2] El- Houssainy A. Ready, Ayman S. Anwar, “Prediction of Kidney Disease Stages Using Data Mining”. Published on 1 April 2019..

[3] Suraiya Aktar, Abhijit Pathak, Abrar Hossain Tasin, “Chronic Kidney Disease (CKD) Prediction Using Data Mining Techniques”. Published on Feb 2021.

[4] Publisher IEEE, NV Ganapathi Raju, K Prasanna Lakshmi, K Gayathri Praharshitha, “Prediction Of Chronic Kidney Disease (CKD) Using Data Science.

[5] Institute of Bioimaging and Molecular Physiology, National Research Council, Catanzaro, Italy, Alessia Sarica, Antonio Cerasa,

[6] Aldo Quattrone, “Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review”.

[7] R. Subhashini and 2M.K. Jeyakumar 1Noorul Islam Center for Higher Education, Kanyakumari, Tamil Nadu, India, “OF-KNN Technique: An Approach for Chronic Kidney Disease Prediction.”

[8] Dr. S. Vijayarani1 , Mr.S.Dayanand Assistant Professor1 , M.Phil Research Scholar Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamil Nadu, India : “Kidney Disease Prediction Using SVM and ANN algorithm”.

[9] https://static.javatpoint.com/tutorial/machine-learning/images/random-forest-algorithm.png

[10] https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_21.png

[11] https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_3.