Quest Journals
Journal of Software Engineering and Simulation
Volume 10 ~ Issue 4 (2024) pp: 11-19
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Disease Prediction, Analysis and Visualization using Machine Learning and Tableau

[1)]Jaidi Srikanth
[2)] Vadlamudi Immaniyel
[3)] Peddi Vara Naga Koteswararao
[4)]Kagga Venkata Naga Gopi
[5)]Dr.Sanjay Kumar Sahu(Mentor)
School of Electrical and Electronics Engineering (SEEE)
Lovely Professional University (LPU)
Phagwara-(144411), INDIA nn

Abstract:
The focus of the study is on the use of computer programs by doctors to predict patients' outcomes in various diseases. We collected a lot of information about individuals' health such as having fever, coughing or feeling tired and other things like age and blood pressure. Moreover, we trained a special computer program called random forest classifier using part of this information. Thereafter, we evaluated its predictive ability in another sample. This was done by comparing the predictions made against what actually occurred.
Also, we applied Tableau tool to interpret all our collected data. By this way; we understood which disease some health factors are more prevalent than others. According to our findings, the computer program was fairly successful at predicting results for other cases. Thus, researchers propose that doctors can utilize such software applications alongside tools like Tableau in order to make better choices concerning different ailments treatment plans for their patients with medical conditions.

Keywords: *Disease Prediction, Machine Learning, Predictive Analytics, R Programming, Tableau, Visualization.*

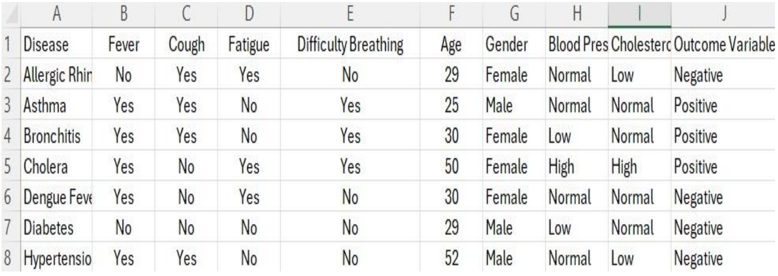
R*eceived 12 Apr., 2024;* ***R****evised 21 Apr., 2024;* ***A****ccepted 23 Apr., 2024 © The author(s) 2024.*
P*ublished with open access at* ***www.questjournals.org***

I. Introduction:

The integration of technology into healthcare has gained unprecedented momentum, offering promising ways to improve patient care and medical outcomes One such innovation is the use of machine learning techniques, where computers can analyze vast amounts of data to identify patterns, make predictions and inform medical decisions Department delivery revolution has great potential.

Among the machine learning approaches, random forest segmentation has emerged as a versatile and powerful tool for predictive modeling. Unlike traditional statistical methods, random forests can handle complex data with many predictors and correlations, making them ideally suited for multidimensional medical data analysis Using a variety of decision trees a, random forest segmentation is able to capture nonlinear relationships well and provide robust predictions for a variety of sectors including health

The aim of this study is to use the predictive power of random forest segmentation to predict outcomes in 20 different diseases The focus of our research is on health-related issues health-related, demographic variables, and a wide variety of physical factors including markers. These include symptoms such as fever, cough, fatigue and shortness of breath, as well as demographics such as age and physical parameters including blood pressure and lipid levels.The essence of this is to come up with a predictive framework that can be used to look into the future as far as various medical problems are concerned through intense analysis and training models with data subset. The performance of this model on an independent test set will be assessed by metrics like sensitivity, specificity and accuracy for prediction of disease outcomes. Additionally, it also contains exploratory data analysis using advanced visualization tools such as Tableau. By visuallyexploring the dataset's characteristics, relationships, and trends, we try to get more understanding about what makes disease outcomes happen. This is a multidimensional approach that improves our knowledge ondiseases' dynamics but also presents clinical decision-making insights and patient management opportunities.



	A	B	C	D	E	F	G	H	I	J
1	Disease	Fever	Cough	Fatigue	Difficulty Breathing	Age	Gender	Blood Pres	Cholesterc	Outcome Variable
2	Allergic Rhir	No	Yes	Yes	No	29	Female	Normal	Low	Negative
3	Asthma	Yes	Yes	No	Yes	25	Male	Normal	Normal	Positive
4	Bronchitis	Yes	Yes	No	Yes	30	Female	Low	Normal	Positive
5	Cholera	Yes	No	Yes	Yes	50	Female	High	High	Positive
6	Dengue Feve	Yes	No	Yes	No	30	Female	Normal	Normal	Negative
7	Diabetes	No	No	No	No	29	Male	Low	Normal	Negative
8	Hypertensio	Yes	Yes	No	No	52	Male	Normal	Low	Negative

Fig 1: Sample Data set for Prediction

Random forest is integrated with R Studio and used to predict solutions when dealing with various diseases and their symptoms. The proposed scheme involves three phases: clustering, prediction, and visualization. The clustering phase deals with cleaning and grouping data based on the different diseases. The prediction phase finds the dependent variable, which includes Positive and Nagative, using the Random Forest technique. Here, we had tried various models for the prediction, and we had chosen the most accurate prediction model. We found that the Randomforest model has the best accuracy for our dataset, so we use the Randomforest model for our further prediction process. As a result, we are able to develop a predictive analytics model for disease prediction using machine learning techniques for the future prediction of patients.

Related Works

Prasanna Kumar (1) probabilistic data collection method was proposed. conducted a correlational analysis of the data collected. Finally, a stochastic forecasting model is developed.To predict future highly correlated health status
It depends on the patients current health status. The bandwidth used in this way reduces the analysis time.
Sudha Ram(2) proposed an alternative multi-data approach sources for determining rates of asthma-related emergency department visits. The system used Twitter data and environmental sensors to quantify asthma emergencies
Departmental visits.
Dequan Chenet (3) proposed Mayo Clinic health care machine to accumulate and keep business enterprise facts.
It plays analytics on scientific data for analysis, remedy,
prevention, or clinical reporting and nonclinical facts medical research health informatics.
Alexandra(4) discussed the challenges associated with device gaining knowledge of tactics in Big Data platform. Finally, they concluded that

the gadget getting to know strategies are nice appropriate
for diagnosed challenges in the generation of Big
statistics analytics.
Min Chenet (5) proposed a CNN-based multiplex diagnosis algorithm. Risk prediction systems for health information. Analyzes medical data to diagnose diseases in health caremedical research health informatics.
Alexandra(4) discussed the challenges associated with device gaining knowledge of
tactics in Big Data platform. Finally, they concluded that
the gadget getting to know strategies are nice appropriate for diagnosed challenges in the generation of Big statistics analytics.
Min Chenet (5) proposed a CNN-based multiplex diagnosis algorithm.Risk prediction systems for health information. Analyzes medical data to diagnose diseases in health care
community.
Abdulsalam Yassin(6) suggested models using smart home big data as a tool Identifying and recognizing human activity patterns for Health care programs. It is suggested to use the constant Pattern mining, cluster analysis, and predictive measurement and adapt to changes in energy consumption generated by residents attitude.
Yichuan Wang(7) advanced a big facts analytics framework for healthcare sectors that identified five massive records analytics competencies such as analytical for patterns, unstructured statistics,analytical,predictive, and traceability.
Gao Zhuetal(8) proposed a new cluster model such as structured output support vector machine to provide classification for object detection. The authors showed a set of clustering algorithms that perform well for large data sets.
Wittekand Daranyi(9) proposed Map Reduce primarily based text mining workflow for powerful use of shared

memory and optimized coalesced reminiscence access.

Javier Andreu-Perezet(10) proposed the new testing hypotheses about disease management from diagnosis to prevention for personalized treatment using characteristics of big data. In this paper, we used Random forest machine learning knowledge of Technique to expect the diverse diseases like Asthama, Stroke, Cholera, Chicken pox and so on., that provide excessive accuracy. Random Forest classifiers hire the chance mechanism that study the data with assumptions for functions that make to offer correct consequences for prediction**.**

Proposed Work

The praposed schemes predicts class of 25 different diseases like ASthama,stroke so on of the test data set.We made our dependent variable into classes Negative-0 and Positive-1.The prapose schemes used statistical classification for the data preprocessing steps.Data sets are collected from the kaggle datasets.Randomforest algorithm was used to build the training model with the data.Table 1 shows the sample disease prediction data set.The attributes used are as follows.

Disease of different types
Fever(1=yes,0=no)
Cough(1=yes,0=no)
Fatigue(1=yes,0=no)
DifficultyBreathing(1=yes,0=no)
Age in Years
Bloodpressure(0=low,1=normal,2=high)
Cholestrrollevel((0=low,1=normal,2=high)
Outcomevariable(Negative=0,Postive=1)

Algorithm 1 (for predictive analytics)

```
1. Imported the data  "Disease_symptom_dataset.csv" using read.csv() method.
2. View the dataset for better understanding of data.
3. Delete the 7th column from the dataset as it is not required for prediction.
4. Convert categorical data to factors and label with numeric.
   4.1. Convert 'Fever', 'Cough', 'Fatigue', 'Difficulty.Breathing', 'Blood.Pressure', 'Cholesterol.Level', and 'Outcome.Variable' columns to factors.
   4.2. convert factor levels to numeric labels.
5. Print summary using the summary() function.
6. Split the dataset into training set and test using sample.split() function with the help of  caTools library.
7. Set the seed for taking test set data randomly.
8. Scale the age column in both the training set and test set using the scale() function.
9. Then import the randomForest library.
10. Create a random forest classifier using the randomForest() function.
   10.1.1. Show the predictor variables (all but the 9th column) and the target variables.
   10.2. Set the number of trees (tree) to 30 .
11. Make a prediction on the test set using the predict() function.
   11.1.1.1. Provide a trained classifier and a test set without target variables.
12. Construct a confusion matrix to obtain the correct number of predictions made by our model.
   12.1.1. Compare the predicted results (y_pred2) with the actual results of the test set.
13. Estimation of Performance Measures:
13.1.1. Exactly: The sum of all the elements on the diagonal divided by the sum of all the elements in the confusion matrix.
13.2. Specifically: True positives divided by all predicted positives.
13.3. Memory (emotion): True positives divided by the sum of actual positives.
13.4. The F1-score: perfect harmonic medium is recalled.
```

Randomforest Classification Algorithm

Random forest classification is a powerful technique for learning clusters that builds multiple decision trees during training and results in the classification of classes of classes or the continuous average prediction for classification functions Each tree in the forest is reared on random training data and available resources, introducing diversity and reducing overcrowding.

The algorithm partitions the feature space repeatedly based on the values of the predictor variables, aiming to maximize the purity of the resulting leaf nodes This process continues until stopping criteria such as maximum tree depth are reached or minimum node size.

Algorithm 2(for tableau visualization)

```
1. Creating data entry:
  1.1. Import a data set containing symptom information into Tableau.
  1.2. Ensure that the list is organized with illness labels, symptoms (e.g., fever, cough), and corresponding values for the presence or absence of symptoms for each illness
2. Create individual cards for each symptom:
  2.1. Create a separate worksheet in Tableau for each symptom (e.g. fever, cough, fatigue, dyspnea, high blood pressure, cholesterol levels).
  2.2. Set the disease color as the focal point on each worksheet.
3. Analysis of attribute classification:
  3.1. For each attribute worksheet:
      3.1.1 Drag the Disease column to the Rows shelf and create rows for each disease.
      3.1.2 and. Drag the characteristics column to the Columns shelf.
      3.1.3 and. Set the collection method to COUNT to count the number of occurrences of each symptom.
      3.1.4 and. Use any devices or colors necessary to distinguish between symptomatic and asymptomatic diseases.
4. Customize the image:
  4.1. Prepare visual attributes (e.g., bar chart, heatmap, treemap) based on disease prevalence and symptom distribution.
  4.2. Include titles, fonts, legends, and a list of tools to make images more clear and defined.
  4.3. Develop color scheme and layout to make the graphic design visually appealing and easy to explain.
5. Steps to repeat for each symptom:
  5.1. Repeat steps 2-4 for each symptom to get separate opinions on fever, cough, fatigue, dyspnea, blood pressure, and cholesterol levels.
6. Dashboard creation (optional):
  6.1. Optionally, put all together and create a dashboard in Tableau.
```

TABLEAU

Tableau is an acclaimed fact visualization and analytics platform that is best used in commercial enterprise intelligence. It prides itself on its flexible connectivity, which helps seamlessly connect data assets with databases, spreadsheets, and mainly cloud-based platforms. The platform's strength lies in its ability to create attractive and interactive visualizations, including bar charts, scatter plots, heat maps, etc., that allow customers to see data from multiple sources Additionally, Tableau empowers customers for integrating dynamic dashboards for multiple visualizations, filters along with interactive features can, which improves statistical analysis and analysis These dashboards can be easily shared with teams or stakeholders , to promote collaboration and aid in data-driven decision-making. Additionally, Tableau's integration with advanced analytics tools and helps cope with large data sets makes it a priority for businesses to derive actionable insights from complex records.

II. Results and Discussions

Accuracy

The proposed scheme provides higher accuracy thanexistingsystem.Theproposedschemeinitializeandsplitthedatasettotraining(80%)andtestingdataset(20%).Theproposedschemeusedthefeatureof Rstudioenvironment and itsin-statistical classification .The proposed system maintains the accuracy above89% while predicting the outcome variable od data . Theproposedscheme usedRandomforest classification algorithm to predict the future status of the health using health symptoms.TheproposedMachinelearningapproachlearns quickly using normalization that reduces the mean absoluteerror. Hence, Randomforest scheme achieves high accuracy andpredictedtheresults.

Tableau Utilization

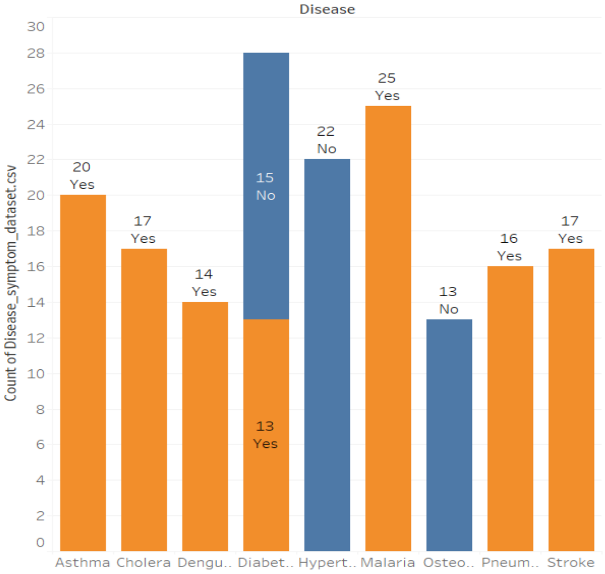


Figure 1

Figure 1 shows the stacked bar chart of Disease vs count of fever (yes or no) and we could see that fever is symptom of certain diseases like Asthama, Cholera, Dengue and Maleria from the graph obtained.

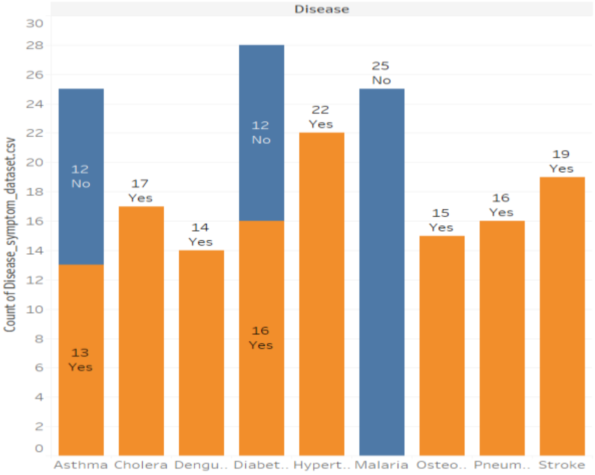


Figure 2

Figure 2 shows the stacked bar chart of Disease vs count of fatigue(yes or no) and we could see that fatigue is symptom of certain diseases like Hypertension, Cholera, Pneumonia and Stroke from the graph obtained.

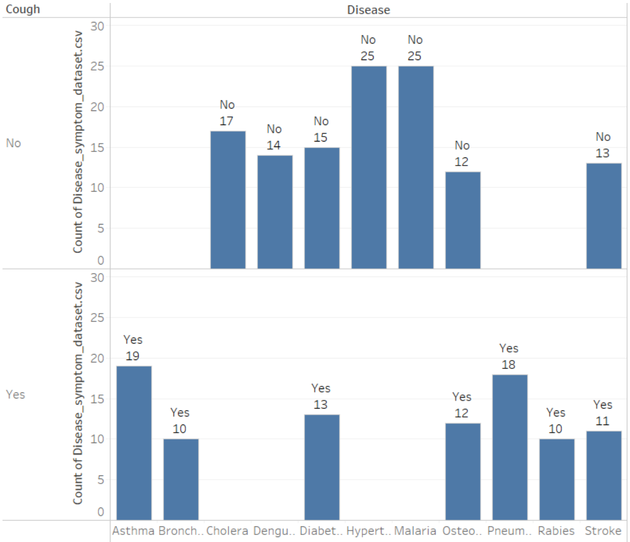


Figure 3

Figure 3 shows the bar graph of Disease vs count of cough(yes or no) and we could see that cough is symptom of certain diseases like Diabetis, Osteoporosis, Pneumonia and Bronchitis from the graph obtained.

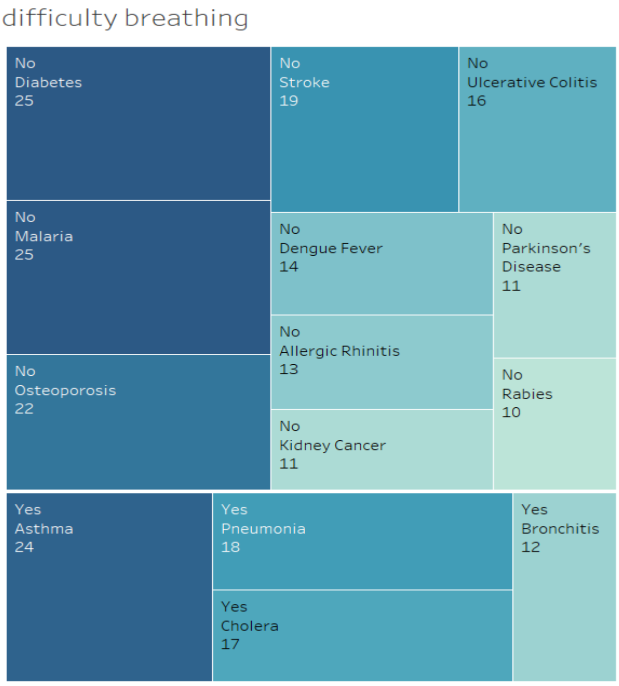


Figure 4

Figure 4 shows the treemap of Disease vs count of DifficultyBreathing(yes or no) and we could see that DifficultyBreathing is symptom of certain diseases like Asthama, Cholera,Pneumonia and Bronchitis from the graph obtained.

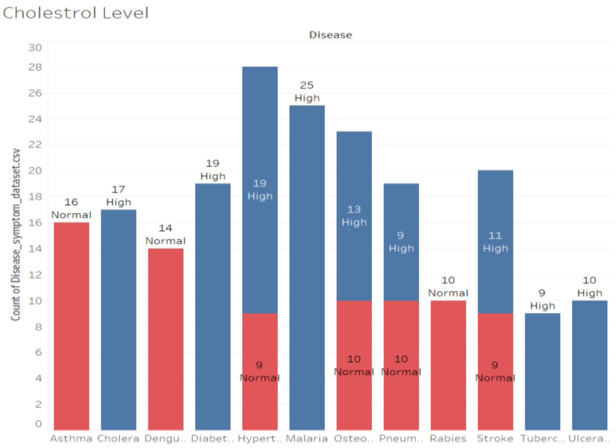


Figure 5

Figure 5 shows the stacked bar graph of Disease vs count of CholestrolLevel(Low or normal or high) and we could see that CholestrolLevel is high for certain diseases like Cholera,Diabetes,Maleria and Tuberclosis and Cholestrollevel is low for certain diseases like Stroke,Allergic Rhinitis and Hypertension from the graph obtained.

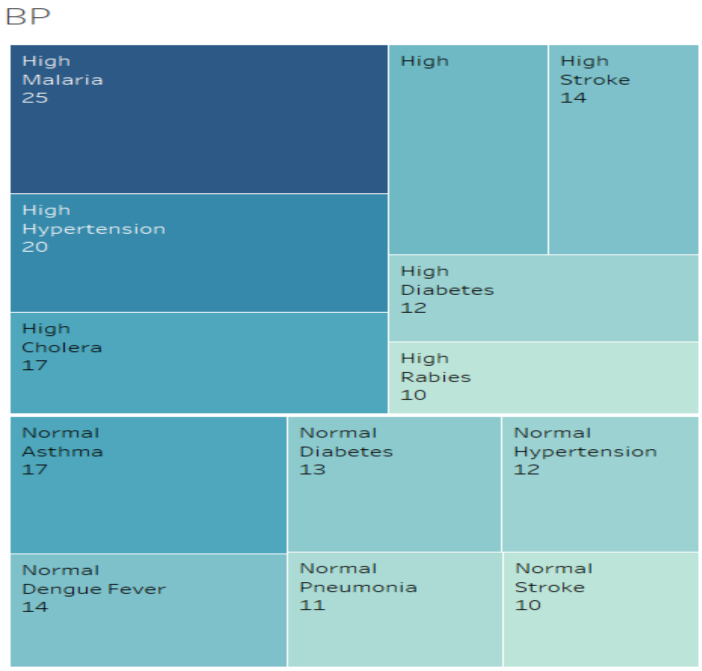


Figure 6

Figure 6 shows the treemap   of Disease vs count of BP(Low or normal or high) and we could see that BP is high for certain diseases like Cholera,Hypertension,Maleria and Stroke and BP is low for certain diseases like Kidney cancer,Liver disease,Diabeties from the graph obtained.

The figures included in our analysis, illustrate the utilization of tableau for analysing the various symptoms of the certain diseases of around 350 paitents.Like if we take Asthama symptoms of asthama from our analysis is fever fatigue and difficulty breathing like that we are able to know all the diseases symptoms.

Classification report and confusion matrix

Table 2 Classification Report for Test Data

	Precision	recall	F1-score	support
0	0.90	0.87	0.88	30
1	0.89	0.92	0.90	39
Avg/total	0.895	0.885	0.89	34.5

Confusion Matrix for Test set

```
[[274]
[435]]
```

III. Conclusion and future work

In this study, we analyzed disease prognosis using a dataset that included symptoms and corresponding disease outcomes. Our analysis includes data preprocessing, model training, and evaluation to predict disease outcomes based on symptom profiles.Initially we preprocessed the data set by encoding categorical variables and divided it into training and test sets. We used the Random Forest algorithm to train the prediction model on the training set and evaluated its performance on the test set. In addition, we used the caret package in R to calculate precision, recall, F1-score, and support metrics, and provided insight into the predictive power of the model.Our results show promising performance in disease prediction, with random forest classification achieving remarkable accuracy and exhibiting a balance of accuracy and recall in disease outcomes. These findings highlight the potential of machine learning in healthcare research, especially in disease prediction tasks based on symptom data. In addition, the visualization tools provided by Tableau play an important role in our analysis. Tableau has simplified the analysis and interpretation of complex health data, enabling us to better visualize disease and its trends. By using interactive dashboards and graphs, we gained deeper insight into the relationship between symptoms and disease outcomes, increasing our understanding of the performance of the predictive model.

While our study provided valuable insights into disease prediction using machine learning, there are several avenues for future research:
Model optimization: Explore methods for model
hyperparameter optimization to further improve prediction accuracy.
Ensemble methods: Assessing the effectiveness of ensemble learning methods in disease prediction tasks.
Real-time prediction: Explore the feasibility of using a predictive model in real-time healthcare systems for timely disease prediction and intervention.

References

[1]. Min Chen, Yixue Hao et.al "Disease Prediction by Machine Learning over big data from Healthcare Communities", IEEE[Access 2017]

[2]. Andrew Alikberov, Stephan Broadly et.al "The Learning Machine", Accessed on: March 26,2020. [Online]. Available: https://www.thelearningmachine.ai.

[3]. IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, " Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Commun. , vol. 55, no. 1, pp. 54–61, Jan. 2017.

[4]. Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[5]. Allen Daniel Sunny1, Sajal Kulshreshtha, Satyam Singh3, Srinabh, Mr. Mohan Ba, Dr. Sarojadevi H " Disease Diagnosis System By Exploring Machine Learning Algorithms", International Journal of Innovations in Engineering and Technology (IJIET) Volume 10 Issue 2 May 2018.

[6]. Shraddha Subhash Shirsath "Disease Prediction Using Machine Learning Over Big Data" International Journal of Innovative Research in Science, Vol. 7, Issue 6, June 2018.

[7]. M. Shouman, T. Turner, and R. Stocker, "Using data mining techniquesin heart disease diagnosis and treatment," pp. 173–177, 2012.;3

[8]. J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, ''Association rule mining to detect factors which contribute to heart disease in males and females,'' Expert Syst. Appl., vol. 40, no. 4, pp. 1086–1093, 2013. doi: 10.1016/j.eswa.2012.08.028.

[9]. S. N. Rao, P. Shenoy M, M. Gopalakrishnan, and A. Kiran B, ''Applicability of the Cleveland clinic scoring system for the risk prediction of acute kidney injury after cardiac surgery in a South Asian cohort,'' Indian Heart J., vol. 70, no. 4, pp. 533–537, 2018. doi: 10.1016/j.ihj.2017.11.022.

[10]. A. S. Abdullah and R. R. Rajalaxmi, ''A data mining model for predicting the coronary heart disease using random forest classifier,'' in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[11]. H. Polat, H. Danaei Mehr, A. Cetin. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, J. Med. Syst. 41(4) 2017 55.

[12]. S. Grampurohit, C. Sagarnal, Disease prediction using machine learning algorithms, 2020 Int. Conf. Emerg. Technol. (INCET) (2020) 1–7, https://doi. Org/10.1109/INCET49848.2020.9154130.

[13]. P. Deepika, S. Sasikala. Enhanced Model for Prediction and Classification of Cardiovascular Disease using Decision Tree with Particle Swarm Optimization, 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1068-1072, doi: 10.1109/ ICECA49313.2020.9297398.

[14]. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018 American Diabetes Association Diabetes Care 2018; 41(Supplement 1): S13–S27. https://doi.org/10.2337/dc18- S002.

[15]. M. Denil, D. Matheson, and N. De Freitas, "Narrowing the Gap: Random Forests In TheDenil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the Gap: Random Forests In Theory and In Practice. Proceedings of The 31st International Conference on Machine Learning, (1998), 665–673. Retrieved from ht," Proc. 31st Int. Conf. Mach. Learn., no. 1998, pp. 665–673, 2014.

[16]. Yao D, Yang J, Zhan X. A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines. J Comput. 2013;8(1):170–7.

[17]. A. M. Mahmood and M. R. Kuppa, "Early Detection of Clinical Parameters in Heart Disease by Improved Decision Tree Algorithm," 2010 Second Vaagdevi Int. Conf. Inf. Technol. Real World Probl., pp. 24–29, 2010.

[18]. Ahmad LG, Eshlaghy A, Poorebrahimi A, Ebrahimi M, Razavi A. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013;4(124):3.

[19]. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.

[20]. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informat. 2006;2:59–77.

[21]. M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, 2017, pp. 300-303.

[22]. J. Aljaaf et al., "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, 2018, pp. 1-9.

[23]. R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 679-684.

[24]. Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," Computer Methods and Programs in Biomedicine, vol. 130, pp. 54-64, 2016.