EasyChair Preprint

№ 5248

Personalized Heart Monitoring and Reporting System

Megha Rathi, Kushagra Nigam, Deepshi Sharma, Kirti Godani and Saransh Khandelwal

March 30, 2021

Personalized Heart Monitoring and Reporting System

Megha Rathi
Department of Computer Science
Jaypee Institute of Information Technology
Noida,India
megha.rathi@jiit.ac.in

Kushagra Nigam
Department of Computer Science
Jaypee Institute of Information Technology
Noida,India
megha.rathi@jiit.ac.in

Deepshi Sharma
Department of Computer Science
Jaypee Institute of Information Technology
Noida,India
megha.rathi@jiit.ac.in

Kirti Godani
Department of Computer Science
Jaypee Institute of Information Technology
Noida,India
megha.rathi@jiit.ac.in

Saransh Khandelwal
Department of Computer Science
Jaypee Institute of Information Technology
Noida,India
megha.rathi@jiit.ac.in

***Abstract*— Treatment cost for heart ailments is very high and there are people in India who can't bear the cost of general treatment for their heart ailments. Indeed, even there are ones who live in remote places and can't look for a decent specialist. All the customary advancements are presently being updated with the time of web and innovation. Innovation make things look less demanding and advantageous to utilize. Along these lines we intend to apply the Machine Learning algorithms and distinctive classifiers to anticipate the probability of heart diseases and altogether analyze them and give exercises and medications to counteract it. Heart disease is taken into consideration as one of the important reasons of dying across the globe. Most of the cases include lack of money for routine check-up, lack of transportation facilities, although most of the problems can be cured with the help of medications and exercises. With the motive to provide names of medications and appoint exercises for persons to stay healthy and eventually reduce the death toll caused by heart disease, we have analyzed and compared data mining techniques of Random Forest (RF), K-Nearest Neighbor (KNN), Gradient Boosting Machine (GBM), Generalized Linear Model (GLM). The performance of each of these algorithms were measured and compared with respect to factors like accuracy, confusion matrix, ROC curve, and AUC value. The best algorithm is then used to predict the probability of heart disease on the parameters provided by the user to generate a report with medications to prevent coronary illness and exercises to stay healthy and fit.**

Keywords*— Gradient Boosting Machine, Generalized Linear Model, Random Forest, K-Nearest Neighbor, Machine Learning, Heart Disease.

I. INTRODUCTION

Data mining is defined as the practice of looking at substantial previous databases so as to generate new information [1]. With the innovation in technology, the medical industry gathers a huge amount of data and stores it in their databases. This data can be further analyzed to produce new information. Hence data mining techniques has a huge scope in the field of medical science for diagnosis of diseases. Different research works has already been published for cancer, diabetes, heart disease with the help of data mining classifiers for prediction and analysis [2]. In this paper we focus on predicting the heart disease with best accuracy and henceforth appoint medications and exercises to prevent the coronary illness. In the starting time of any heart disease you never feel any symptoms, so you may not find yourself in danger of the condition until a standard check-up uncovers you have hypertension or high cholesterol. Getting the likelihood of illness from early ages can assist us with prevention of major cardiovascular diseases.

Nowadays people are more prone to taking stress and stress is a gateway for most of the risk factors which includes rising blood pressure and cholesterol levels, physical inactivity, smoking, alcohol, poor diet [3]. Every heart disease has its side effects and miseries. A healthy lifestyle and medications can have a gigantic effect in improving well-being of an individual. Heart Disease is the largest killer of population in many countries. As an example, coronary heart disorder reasons 4 out of 10 deaths within the USA which is more than all kinds of cancers put together [4]. This sort of high mortality rate itself makes it critical to have a measure of person's risk for heart disease. Diagnosis of Heart Disease isn't so straight forward. Diagnosis can be made on foundation of physician's experience and expertise which could consequently incur high prices of clinical treatment to patients. Our proposed application provides an automated heart disease prognosis which will be efficient and beneficial to patients. The application will display the percentage of person's risk for coronary heart disease and what medicinal drugs and physical activities an individuals should go about to get healthy and avoid any major disease. All the measures taken to predict the best suitable machine learning technique depends upon 14 most prominent attributes taken to introspect the demands. Clinical elements provided by the application depends upon clinical research data which have been founded by the sources available online. Data mining has revolutionized the healthcare domain, proving to be the future of evaluating and deciphering symptoms, reactions of the sicknesses like heart disease. All the responses are extracted from the various data sets and assigned to the rows so as to get the application for predicting heart disease. We have split the data into training and test set by evaluating the Mahalanobis distance which is Euclidean distance between the point and class mean divided by the covariance matrix for the class. This implies that the data points from large covariance attracts data points from a larger area than those with small co-variances. All the training has been done using the data set provided by University of California, Irvine. Dataset for various heart disease medications have been created with the help of online material through websites [5]. The rest of the paper progresses as follows. Literature Survey in section II, Section

III contains Methodology, then System Architecture is described in Section IV. The paper contains the algorithm used to fetch medications from created datasets in Section V and Results and Analysis in Section VI. It finally closes with confusion in Section VII and future advancements in Section VIII.

II. LITERATURE SURVEY

Data mining frameworks for heart disease prediction is presented in the study [6]. The makers found that neural framework with disconnected model planning are sensible for disease prediction in starting period. The proposed algorithm in their paper assigns the weight using Hyperlink-Induced Topic Search (HITS) model and makes a smaller number of rules which improves the request precision.

The pertinence of Data Mining in the Medical field was recognized, and steps were created to apply material techniques in the Disease Prediction. In the proposed research work [7] ANN, KNN, CART, C4.5 were implemented to predict the risk percentage of heart disease.

A heart disease prediction system using various data mining techniques and analysis of the results obtained for all implemented techniques has been recorded [8]. Authors evaluated the popular and efficient heart disease prediction methods from the literature survey and finally selected the most effective algorithm Naive Bayes and Genetic Algorithm for their performance analysis on the heart disease prediction. To develop a distinctive heart disease prediction model, several machine learning algorithms were applied in yet another novel work [9].Prediction models were trained using J48, Naive Bayes and Multi-layer Neural Network inside WEKA software which is a machine learning tool based on 10-fold cross validation. J48 outflanked Naive Bayes classifier and Neural Networks performance.

Heart disease dataset is collected from the UCI Repository and various algorithms applied on the dataset and it has been found that decision tree outperformed in predicting heart infection [10].The problem of heart disease has been resolved in the research [11] using several data mining algorithms to predict heart illness in his research paper. From the experimented results, it was found that SVM classifier with Genetic Algorithm gives better accuracy of prediction when compared with Naive Bayesian, C 5.0, KNN, Fuzzy Algorithm, Neural Network, and Decision Tree. The review paper of heart disease prediction for medicinal service framework by using a few data mining strategies is presented [12]. They have also proposed the model by utilizing 14 out of 76 attributes from the dataset, picked from UCI repository. In another significant contribution artificial neural network is implemented for the prediction of heart disease in individuals [13]. Combination of back-propagation algorithm and feed forward algorithm makes this neural network-based model. Features selection plays significant role in improving classifier's performance. Proposed work presented how most relevant features can be selected out of the set of attributes [14]. The feature selection and classification algorithms were applied on Hepatitis disease data and accuracy is recorded with improved outcome.

CART (Classification and Regression Tree), ID3(Iterative Dichotomizer), DT(Decision Tables) used to build model and 10- fold Cross Validation is used to evaluate classifiers performance in the domain of heart disease detection [15] and hence predicted the risk of heart disease. In other work [16] algorithms like Decision Tree, Fuzzy Logic, Support Vector Machine, Naive Bayes and KNN have been studied and used in the implementation to predict the risk of heart disease and develop a clinical system.

The proposal of a model stating that heart disease can be predicted in a much better way using multiple regression was presented in the paper [17]. They have used 13 attributes in total and total dataset was split into training and testing data ratio being 7:3. An idea of predicting heart disease using machine learning and data mining techniques was suggested by identifying hidden patterns using data mining algorithms [18] which concluded that J48 algorithms accuracy is better than LMT. The recommendation system using ANN, Naive Bayes, Decision Tree, KNN, Logistic Regression has been built to examine the heart prediction disease. According to the authors of [19], Logistic Regression is most suitable for prediction. In yet another novel work [20] following machine learning techniques is implemented: Naive Bayes Classifiers, Decision Tree, K- Nearest Neighbors and Support Vector Machine in order to find out best classifier for heart disease prediction. Various tools and software like Orange, WEKA, MATLAB and KNIME were explored and used.

III. METHODOLOGY

The datasets of Cleveland, Hungary, Switzerland and Virginia are taken from University of California Irvine [21] repository to get the statistical analysis of heart problems. These 4 datasets were merged into one to get a larger dataset. Data preprocessing plays a crucial role in providing meaning to the dataset which included replacing missing values and removing less significant attributes from the main dataset. The 14 fundamental attributes selected in the final dataset are mentioned with their column names in the dataset: Age(age), Sex(sex), Chest Pain(cp), Resting Blood Pressure(trestbps), Serum Cholesterol(chol), Fasting Blood Sugar(fbs), Resting electrocardiographic results(restecg), maximum heart rate achieved(thalach), angina induced due to exercise(exang), ST depression induced due to exercise(oldpeak), the slope of ST segment during peak exercise(slope), No. of major vessels affect(ca) and the presence of thalassaemia(thal).

Later final dataset was split into training and testing set in the ratio of 7:3 respectively. The performance, of each algorithm applied to predict the test data, is measured upon the factors like accuracy, precision, ROC curve and recall.

Parallelly, a new dataset of medications which consist of uses of the medication, side effects of the medication, precautions to take care with the medication has been created with the help of information gathered through the web. The datasets created for medications include BP, Cholesterol, Sugar, and chest pain. These include drugs in the increasing order of effectiveness so that an effective drug is appointed if the probability of heart disease is high. Binary search algorithm has been applied to provide the most suitable drug. Dataset for the exercises with respect to different attributes of heart disease has been created to provide the most appropriate exercise according to the condition

IV. SYSTEM ARCHITECTURE

The proposed research work has application-based machine learning architecture which trains the dataset

according to the best algorithm and delivers the report for the user. The user/patient interacts with the app and enter his/her required details to get the output of the probability of his heart disease and various medications and exercises are displayed which should be followed to prevent the next state of heart disease.

Following are the steps:-

1) Enter the details of user.
2) Train the dataset of heart disease prediction.
3) Predict probability for user using the trained dataset.
4) Get the appropriate medications and exercises.
5) Display in a form of report.

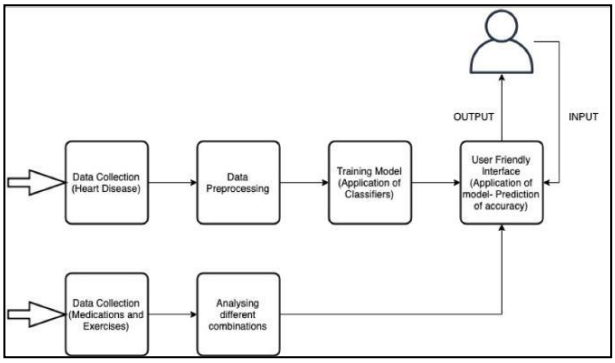

Fig. 1. Architecture of proposed model.

Following pre-processing is done to remove anomalies from the dataset:

Data Cleaning: The data collected by us had so many NA values in each column which makes the dataset unfit for training. Hence these needs to be fixed before the dataset can be used. These NA values were replaced by mean of the column, and this way all the NA values were removed from the dataset.

Factorization: The columns of the dataset were converted to desired type. Example age is a numeric column and hence the algorithm does not get confuse to take anything greater than zero as 1 and lesser than or equal to zero as 0.

The information about the various attributes from the dataset is given in TABLE 1.

TABLE I
DATASET INFORMATION.

Attribute	Description
age	Determining the age of an individual in years.
Sex	Binary attribute for gender choice 0. Female 1. Male
cp	It determines the chest pain suffered by an individual 1. typical angina 2. unstable angina 3. non-anginal pain 4. asymptomatic
Tresbps	An individual's resting blood pressure on the scale given in mmHg.
chol	Determines the Blood Cholesterol (mgdl) level of the body.
fbs	Binary attribute for fbs >120 mgdl 0. No 1. Yes
restecg	Resting ECG measurement of an individual 1. Normal 2. ST elevation or depression of >0.05 mV 3. Left ventricular hypertrophy
thalach	Maximum heart rate achieved.
exang	Binary attribute if angina is induced while exercise. 0. No 1. Yes
oldpeak	Slope of the ST segment induced by exercise relative to rest.
slope	The slope of the peak exercise ST segment 1. Upslope 2. Flat 3. Downslope
ca	Number of heart vessels affected (range from 0-3)
thal	Thalassemia 3. Normal 6. Fixed defect 7. Reversible defect
num	Binary attribute for heart disease presence. >0 : Present = 0 : Absent

A. *Techniques Used*

1) Random Forest (RF): This is one of the ensemble learning methods used for regression, prediction and classification. It builds prediction model which is in the form of group of decision trees created from subset selected randomly from training set. The model is trained using Bagging method. This method is comparatively easy to use as require a smaller number of parameters to be passed to create the prediction model.

2) Generalized Linear Model (GLM): The generalized linear model is used for analyzing the relationship between variable Y with a set of numerous variables so that

$Y = b0 + b1X1 + b2X2 + ... + bkXk.$

3) K-Nearest Neighbors (KNN): In K-NNclassification, the points are classified whether it belongs to one class or the other on the basis of distance calculated. Distance measure can be Euclidean distance or Manhattan distance. K signifies the number points to be considered while classifying the point and majority choice is taken. Its accuracy is often degraded by the presence of noisy or irrelevant features.

4) Gradient Boosting Machine (GBM): Gradient Boosting Algorithm is one of the ensemble methods used for prediction and regression. Prediction model in this is built in the form of group of small regression trees. weak learners are given more weights so that error is minimized, and accurate prediction is made. method is little bit complex to use as requires many parameters to be passed to create the model.

B. *Analysis of factors affcting Heart Disease*

Various factors affect the probability of people of having heart disease. The prediction made by the system depends upon different set of values, it is been observed that the resting ECG results have major role in predicting the probability of heart disease. Males have higher chances of occurrence of heart disease than females. Children having abnormally high levels of be at risk of heart disease. People with age in between 50 and 65 years have higher risk of heart disease. Among all 4 kinds of chest pains (Atypical angina consequence of exercise done for long duration,

asymptomatic angina shows higher probability of heart disease. People with higher value of ST Depression induced by exercise are at risk of heart disease. The serum cholesterol levels above 140 have abnormally high prediction for heart disease. The variation in the body should be noted and provided with the suitable medical advice. The health monitoring system plays a major role in determining the probability of sufferings and thereby helping them to assign drugs and exercises.

C. Data Visualization

We have visualized the training data and plotted the curves for better understanding of the attributes related to heart disease.

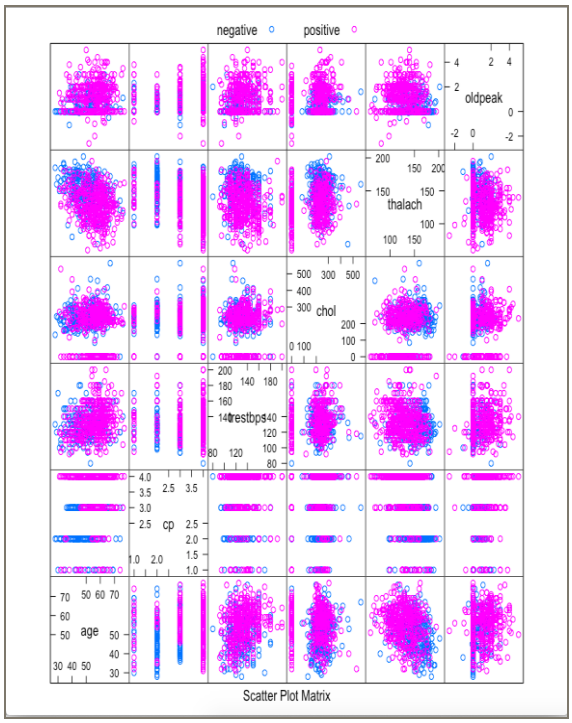

Fig. 2. Pairwise visualization of age, cp, trestbps, chol, thalach, oldpeak

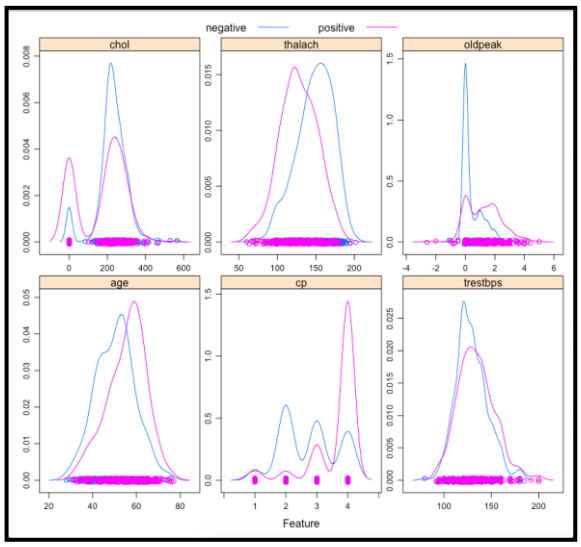

Fig. 3. Variation of data with respect to positive and negative presence of heart disease.

a) Sactter Plot: In this scatter plot(Figure 2), every attribute is plotted against every other attribute(pairwise graphs are displayed) and the dots represent the class being positive(pink) or neg-ative(blue). Here, for instance, we can see pink dots are clustered together which represents that those points in the plot are categorized as positive.

Density Plot: The figure shown below (Figure3) is basically a density plot which represents how the number of records is related to attribute values in continuous interval of time. The data is visualized as Probability Density Function.

V. RECOMMENDATION SYSTEM

The first and most crucial step for developing a recommendation system for heart disease is collecting data by gathering information available from internet. The data has been collected on the basis of considering all the measures affecting heart problems like serum cholesterol, exercised induced angina pain, resting blood pressure, age and many more factors to get a system recommendation for suggesting drugs, side effects, precautions, overdose and exercises in order to aware the pros and cons of drugs prescribed by the application. All the precautionary measure and impact of overdose has been recommended by the application on the basis of all the information gathered.

1. The data has been accumulated from the which is produced manually in order to extract data to enhance our learning after predicting the probability of heart diseases.

2. The mainly consists of 5 attributes -Disease, generic name, common name, precautions, side effects and overdose.

3. The medications are relegated by each arrangement of prescriptions considering all the 14 factors recorded in the UCI.

4. The recommendation has been provided to the patient to enhance their knowledge of drugs, precautions and side effects before taking.

5. Provided exercise proposals would assist them with improving their invulnerability to manage heart issues and would lead them to the path of healing and gaining strength by staying healthy.

As discussed in the section of Methodology, binary search algorithm is used to provide the appropriate drug. The datasets of blood pressure, sugar, and cholesterol medications are created in the increasing order of effectiveness.

START

1. Upper Bound (number_of_rows, predicted_result)
2. Left = 0, Right = number_of_rows
3. While Right - Left > 1 :
    1. Row_being_searched = (Right + Left)/2
    2. If Row_being_searched is lesser than predicted result
        a) Left = Row_being_searched
    3. Else
4. Right = Row_being_searchedReturn Right

END

VI. RESULTS AND ANALYSIS

The study has successfully compared four machine learning algorithms and plotted the ROC curve for each (combined into one for the comparison). It gives the confusion matrix for each algorithm. It was observed that Gradient Boosting Machine (GBM) achieves the best accuracy of up to 90%. Moreover, the reporting of the medications and exercises on the basis of heart disease prediction and different attributes will help lower the cases of coronary heart disease, cardiac arrest, stroke, arrhythmia, other heart diseases in the persons. The objective to provide recovery of the patients on the go has been fulfilled.

a) ROC Curve: This curve gives the performance measurement of all the four algorithms used. So it compares the best algorithm out of all.

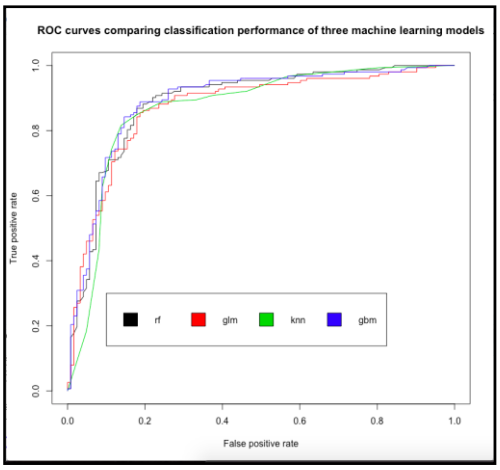

Fig. 4. ROC Curve for the algorithms used in the study

b) Confusion Matrix: The confusion matrix helps us with the visualization of the performance of various algorithms. The key attributes of a confusion matrix are as followed:

- True positive: The number of values predicted to be true and the set of values are actually true are TP values.
- False positive: The number of values predicted to be true while they are actually false is FP value.
- False negative: The numbers of values predicted to be false while they are actually true are FN value.
- True negative: The number of values predicted to be false and the Set of values are actually false are TN values

TABLE II
Confusion Matrix for Random Forest.

	Actual True	**Actual False**
Predicted True	136	16
Predicted False	19	136

TABLE III
Confusion Matrix for GLM.

	Actual True	**Actual False**
Predicted True	132	20
Predicted False	15	132

TABLE IV
Confusion Matrix for KNN.

	Actual True	**Actual False**
Predicted True	132	20
Predicted False	23	132

TABLE V
Confusion Matrix for GBM.

	Actual True	Actual False
Predicted True	139	20
Predicted False	15	132

VII. CONCLUSION

The best suitable machine learning algorithm turned out to be Gradient Boosting Algorithm. We have executed four AI models - Support Vector Machine, K-Nearest Neighbors, Gradient boosting Model and Generalized Linear Model. The interpretation of heart disease depends upon the 12 attributes which were taken from UCI to get the probable predictions. Further on the basis of the probability, exercises and medicines are suggested. All the measures, side effects and overdose issues are specified along with the drug in order to come up with sustainable solution of the dangerous disease like heart. Getting the world free from coronary ailments requires innovation like AI which has huge potential of making the world a healthy place to live in.

VIII. FUTURE ADVANCEMENT

With a furthermore efficiency and enhancement of data set the accuracy of the model could be improved in order to get the accurate personalized health monitoring system. Patients previous health record should be considered to track their health recovering status. The medications, meals and practice routine ought to be noted in order to follow the recouping status of heart which would drop down the infection prompting prosperous and healthy life.

REFERENCES

[1] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[2] M. A. Khaleel, S. K. Pradham, G. Dash et al., “A survey of data mining techniques on medical data for finding locally frequent diseases,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 8, 2013.

[3] “Stress and heart health,” https://www.heart.org/en/healthy-living/healthy-lifestyle/stress-management/stress-and-heart-health, 2003.

[4] “Heart disease facts,” https://www.cdc.gov/heartdisease/facts.htm, 2005.

[5] “Heart disease medications,” https://www.webmd.com/heart-disease/heart-disease-medications, 1999.

[6] R. Thanigaivel and D. K. R. Kumar, “Review on heart disease prediction system using data mining techniques,” Asian Journal of Computer Science and Technology (AJCST) Vol, vol. 3, pp. 68–74, 2015.

[7] A. Kaur and J. Arora, “Heart disease prediction using data mining techniques: A survey.” International Journal of Advanced Research in Computer Science, vol. 9, no. 2, 2018.

[8] N. Singh, P. Firozpur, and S. Jindal, “Heart disease prediction system using hybrid technique of data mining algorithms,” International Journal of Advance Research, Ideas and Innovations in Technology, vol. 4, no. 2, pp. 982–987, 2018.

[9] A. kumar Dwivedi, “Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation,” 2016.

[10] J. Rohilla and P. Gulia, “Analysis of data mining techniques for diagnosing heart disease”,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 7, 2015.

[11] K. Manimekalai, “A proficient heart disease prediction method using dif-ferent data mining tools,” International Journal of Engineering Science, vol. 2676, 2016.

[12] R. Chadha and S. Mayank, “Prediction of heart disease using data mining techniques,” CSI transactions on ICT, vol. 4, no. 2-4, pp. 193–198, 2016.

[13] K. U. Rani, “Analysis of heart diseases dataset using neural network approach,” arXiv preprint arXiv:1110.2626, 2011.

[14] P. Nancy, V. Sudha, and R. Akiladevi, “Analysis of feature selection and classification algorithms on hepatitis data,” Int J Adv Res Comput Eng Technol, vol. 6, no. 1, pp. 2278–1323, 2017.

[15] V. Chaurasia and S. Pal, “Early prediction of heart diseases using data mining techniques,” Caribbean Journal of Science and Technology, vol. 1, pp. 208–217, 2013.

[16] Q. D. Buchlak, N. Esmaili, J.-C. Leveque, F. Farrokhi, C. Bennett,

[17] Piccardi, and R. K. Sethi, “Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review,” Neurosurgical review, pp. 1–19, 2019.

[18] K. Polaraju, D. Durgaprasad, and M. Tech Scholar, “Prediction of heart disease using multiple linear regression model,” International Journal of Engineering Development and Research, vol. 5, no. 4, pp. 1419–1425, 2017.

[19] J. Patel, D. TejalUpadhyay, and S. Patel, “Heart disease prediction using machine learning and data mining technique,” Heart Disease, vol. 7, no. 1, pp. 129–137, 2015.

[20] M. Marimuthu, M. Abinaya, K. Hariesh, K. Madhankumar, and

[21] Pavithra, “A review on heart disease prediction using machine learning and data analytics approach,” International Journal of Computer Applications, vol. 975, p. 8887.

[22] B. Umadevi and M. Snehapriya, “A survey on prediction of heart disease using data mining techniques,” international Journal of Science and Research (IJSR) ISSN (Online), vol. 6, no. 4, 2017.

[23] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[24] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.