Diabetes & Heart Disease Prediction Using Machine Learning

BhaveshDhande[1], Kartik Bamble[2], Sahil Chavan[3], Tabassum Maktum[4]

[1,2,3,4]Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Navi Mumbai, India

Abstract One of the root causes of mortality in today's world is the culmination of several heart disease and diabetes illnesses. In clinical data analysis, predicting multiple diseases is a significant challenge. The machine learning approach has proved to be functional in assisting in the decision-making and governing of large amounts of data generated by the healthcare field. The various experiments scratch the surface of machine learning to predict different diseases. The papers present a novel method for identifying significant features using machine learning techniques, which improves the diagnosis of multi-purpose disease prediction. The different features and many well-known classification methods are used to implement the prediction model to predict the heart disease and diabetes. The proposed method utilizes ensemble approach for achieving a higher degree of accuracy rates for by using classification algorithms and feature selection methods. The proposed method implements voting classifier that has sigmoid SVC, AdaBoost, and Decision tree algorithms. The paper also implements the traditional classifiers and presents the comparison of different models in terms of accuracy. The web application is also developed for users to avail its services very easily and make it convenient for their use, particularly in the prediction of heart and diabetes collectively.

Keywords: Machine Learning, classification, feature selection, prediction, heart disease, diabetes

1. Introduction

The answer to calculate risk through diseases via model-based prediction is very difficult. Due to this, the examination of several medical datasets and their forecasting using soft computing is very handy and a cheaper way for professionals in the healthcare industry. These techniques include exploratory analysis and constructive models that support the professional in statistical decision-making, which is a massive requirement of the medical industry. The high blood pressure, high cholesterol, diabetes, peculiar pulse rate, and various other risk factors make it harder to detect illnesses [1]. The severity of heart disease in humans was determined using various data mining and neural network methodologies. The decision trees, genetic algorithm, Naive Bayes, and K-nearest neighbor algorithm are all used to classify the severity of the disease. Because the features of their problems are complex, the diseases must be treated with caution. Failure to do so can reduce the effectiveness of organs or lead to early death. The various metabolic diseases are identified using a medical science approach and raw data mining. The data mining techniques with classification plays a vital role in predicting heart disease and data research [2]. The random forest is an ensemble machine learning algorithm. It is perhaps the most popular and widely used machine learning algorithm with excellent performance across a wide range of classification and regression predictive modeling problems. Also, random forest approach has a brute way of tuning parameters that helps in easier feature selection [3]. The process of aggregating the votes for class labels from individual models and predicting the class with the most votes is called a hard voting ensemble. Whereas, in a soft voting ensemble, the predicted probabilities for class labels are added up, and the class label with the highest sum probability is predicted. The voting ensemble method assumes that all the models have equal contributions in predictions, which is a drawback because some models give better results in some scenarios and poor in others. The Adaptive Boosting (AdaBoost) algorithm is a boosting technique in machine learning that is employed as an ensemble method. The weights are re-allocated to each instance, with higher weights applied to improperly identified instances. This is termed as adaptive boosting. In supervised learning, boosting is used to reduce bias and variation. It is based on the concept of sequential learning [4]. The linear SVM algorithm is used for linearly separable data, which implies that if a dataset can be categorized into two classes using only a single straight line, it is called linearly separable data, and the classifier employed is called linear SVM. Under the supervised learning approach, one of the most prominent machine learning algorithms is logistic regression. It is a method for predicting a categorically dependent variable from a set of independent factors and to describe data and explain the connection between one dependent binary variable and one or more independent variables. As a result, the result must be a discrete or categorical value such as Yes or No, 0 or 1, true or false, and so on. But, instead of giving exact values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1. In order to

predict and diagnose the recurrence of cardiovascular disease, ensemble learning incorporates the model approaches of five classifiers, including support vector machine, artificial neural network, Nave Bayesian, regression analysis, and random forest [7]. Dataset Cleveland's cardiovascular data records were retrieved from the UCI repository, and for Diabetes, the PIMA dataset has been used. The findings of the experiments showed that an ensemble model is a superior strategy in terms of diagnostic performance predictability and accuracy.

In this paper an ensemble approach is proposed to detect the heart disease and diabetes. The different algorithms combined include ADA-boost, decision tree and sigmoid SVC. The classifiers are combined by varying their weights. The major contribution of the paper is as follows:

- Provide a new approach to concealed patterns in the medical data.
- To predict the chance of heart disease with the highest accuracy of prediction.
- To predict the chance of diabetics with the highest accuracy of prediction.
- Error rate compression for the results found to make it relatively exact in accuracy.

The rest of the paper is organized as follows:
The section 2 gives the survey of existing systems for predicting heart diseases and diabetes. The section 3 demonstrates the proposed system along with the ensemble approach. The results of the proposed system are presented in section 4. Finally, the conclusion and future work is expounded in Section 5.

2. Literature Survey

The most general causes of mortality on the planet have been heart and diabetic problems. Furthermore, today, the prediction of the same or even hinting at a minute probability of it is a problem that needs a solution. In the medical field, machine learning has paved its purpose by helping make choices and predict by training over large amounts of data existing in the form of datasets.

The study in [1] represent that diabetes mellitus and hypertension were moderately associated while cardiovascular diseases are strongly associated with severity and mortality for COVID-19. The paper helps to gain relation between diabetics and heart diseases and create a link or gain experience to handle data for both diseases at the same time as it gives an idea of immunity prediction of COVID through the data of diabetics and heart. The quantitative estimate of severity outcomes and or deaths in COVID-19 patients was performed with Comprehensive Meta-Analysis Software (CMA) version 3.0. In paper [2] the K-means clustering algorithm is used for predicting heart diseases and analysis is carried out using visualization tool tableau. The Cleveland heart disease raw dataset with 76 features of 303 patients was pre-processed with exploratory data analysis which narrowed down the dataset to 209 records and 7 important features. The study includes 4 types of chest pain with age, maximum heart rate, and chest pain type which are considered as vital features in prediction. In paper [3] HRFLM method is proposed that stands for union of Linear Method (LM) and Random Forest (RF), which boosts efficiency by improving selection. The study involves pre-processing of Cleveland UCI repository with use of R rattle GUI (Feature Selection and Classification modelling) which provides an easy-to-use visual graphics, working environment for the user of the dataset, and building the predictive analytics. The several approaches are presented in [4] for predicting heart diabetes. The methodology with logistic regression provides 96% precision. This was the first paper that observed the study of more than one dataset and competes between algorithms, with pipeline affected to 98.8% fidelity using Adaptive Boost classifier. In paper [5] the difficulties in the diabetic analysis were convened in relation to the COVID-19 rate. The conclusion derived from the study is that different categories of diabetes have a unique effect on the percentage of mortality rate. The paper [6] Bhavesh Dhandeproposes an approach to predict diabetes mellitus by applying machine learning techniques. The paper concludes that minimum redundancy maximum relevant approach is better than principal component analysis. It cements random forests as a better algorithm than others. The two datasets, Luzhou and Pima were utilized, with 80.84% and 77.21% accuracies fetched respectively. The ensemble approach with various classification algorithms such as KNN, Adaptive boost (AB) Gradient boost (XB), decision tree and random forest is proposed in [7]. Based on the analysis of different algorithms, it can be concluded that the proposed system on this research edged are under cover (AUC) promisingly. The perfect couple for prediction turned out to be an ensemble of (AB+XB) classifiers. In paper [8], for predictive data mining for medical diagnosis various techniques such as KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, etc. are used. An overview of all prediction models is studied in this research paper. According to a performance study of data mining algorithms Naïve Bayes, Decision Tree, KNN provide the highest accuracy rate. The Weka 3.6.0 tool was used for conducting the research. In paper [9], the methodology is for finding out the best algorithm to extract best features from the medical dataset. Whenever, data collection is followed by data pre-processing, data mining, and pattern evaluation, the suitable and highest accuracy is achieved. The data extraction was performed with the WEKA software tool then compared using predictive accuracy, ROC curve, ROC value. The approach for prediction of heart disease by applying various

algorithms like ANN, random forest, SVM is presented in [10]. By using 3-fold cross-validation along with SVM algorithm maximum accuracy achieved was 83.17%. The application of decision tree algorithm with 37 splits and 6 leaf nodes led to an accuracy of 79.12%, and when used with 5-fold cross-validation technique accuracy achieved was 79.54%. By using random forest algorithm, the accuracy achieved was 85.81%, which is maximum as compared to all other algorithms. The method presented in [11] utilizes XGBoost, AdaBoost, gradient boosting, extra trees, light gradient boosting Lightgbm, SGDC, Nu SVM algorithms for prediction of cardiovascular effects. The data pre-processing was performed on UCI repository and Framingham dataset. The data pre-processing using the Multiple Imputation Chain Equation model for filling inexistent values proved to be efficient way of data pre-processing and the accuracy of 95.83% was obtained by using the stacking algorithm. In [12] research was carried out using three methods KNN, Neural Networks and SVM on real dataset of Algerian people. Neural Network algorithm gave highest accuracy of 93%. [13] paper predicted diabetes based on different human body attributes. Study showed that Body Mass Index (BMI) and growing age are major factors in the development of risk for diabetes.

The limitations of existing system are as follows: The prediction of possibilities is not accurate for disease aggregated inputs and hence thereby cannot handle enormous datasets for patient records effectively. As the machine learning approach is based on predicting outcomes using existing data, we cannot be sure of whether it will apply to the current specimen on which it would be experimented, because evolution is a constant. To add up, poor results on very small datasets causes frequent overfitting occurrences. To conclude, these previous studies focus on the particular impacts of specific machine learning techniques and not on the optimization of these techniques using optimized methods.

3. Proposed Methodology

The paper presents a methodology to predict heart diseases and diabetes by applying machine learning techniques and perform classification by using ensemble classifiers over the datasets involved. This will provide an easy glance at the data involved to build analysis. Thus, the proposed methodology will then materialize into a model evaluation phase based on performance, only after bits like feature selection and data pre-processing. By considering the features extracted, the proposed method will frame an idea of the presence of either disease. Furthermore, all the basic models will be analyzed by ensemble techniques for the prediction. Innovative models like the Random Forest and Adaptive Booster will be added, which will further be evaluated using voting classifiers with weights and without weights.

3.1. System Architecture

The process of making new observations or categorizations from the given available data with the help of supervised learning approach is known as classification. The training dataset is hatched within several classifiers and passed through them.
These 1-n classifications are then used through a combined powerful algorithm. This algorithm then further analyzes the data by considering the ensemble considered model. This ensemble model has a build of various boosters/classifiers/outliers to improve results. Then, after building a model with the available test dataset, the prediction of any random data set can be performed. The system architecture of the proposed method is shown in Figure 1. The working of the proposed method is as follows:

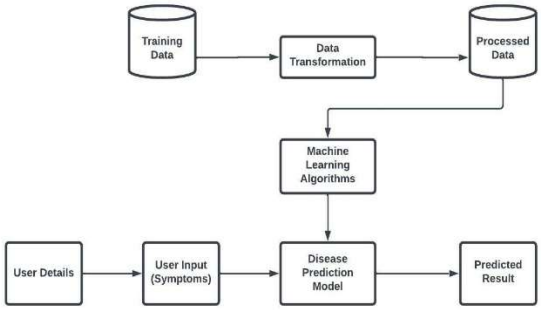

Figure 1: System Architecture

Training Data is undergoing the transformation to generate useful information i.e., processed data by various data analysis and preprocessing using various basic methods. After obtaining processed data, the algorithms can be implemented on the data involved. User then interacts by giving input which is symptoms to the application as a feed to predict results. Then, the algorithms work upon comparisons through various models in the disease prediction model to fetch the required Predicted Result.

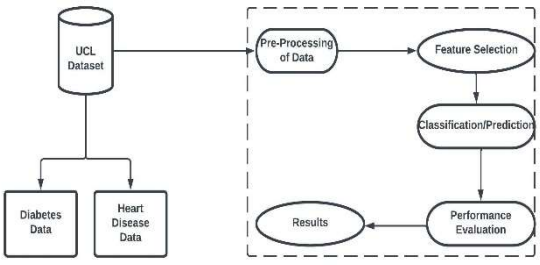

Figure 2: Dataset Handling

The paper considers the two datasets i.e., Heart and Diabetics Datasets. The advantage here is that since both have been taken from the UCL repository and belong to the same result, we can club their results.

The data from these datasets is preprocessed as per the need of exploratory data analysis and the right information is extracted. The process of feature selection is then applied, where the most important features and the impactful ones are thoroughly selected by using required classifiers and outliers.
Further, the different classification models are built using random forest, Ada boost, KNN, logistic regression and decision tree approach to predict the disease. Further ahead, these results are then performance evaluated to check their accuracies and precision, to obtain the final declarative results with their legitimacy decided. After calculating the accuracy of all the above-mentioned models, we selected the top-performing models which are sigmoid SVC, AdaBoost, and Decision tree, and combined them into a single voting classifier.For diabetes prediction after calculating all model accuracies we created a voting classifier with the following weights follows KNN - 2, Decision Tree – 1, Logistic Regression – 2, and Random Forest – 2.

4. Implementation Details& Results

The proposed method is implemented by the use of the environment provided by Google Colab and key Python libraries. The various ways in which data has been analyzed are listed as follows: Plots of Attributes for Detailed Analysis, Voting Classifier: Pair, Scatter Plots, KDE/Target Visualization Plot, Correlation Matrix, Box/Pair Plotting, Plotting Comparison of Models Before & After Standardization. The various models in which data has been analyzed are listed as follows: Ensemble Classifiers, XGBoost, Decision Tree, Logistic Regression, KNN, Random Forest, AdaBoost, Sigmoid SVC, Polynomial SVC, RBF SVC & Linear SVC.

Before beginning EDA, we double-checked the dataset's dimension and variable data types, as well as looked for null values. Diabetes affects about a third of the persons in the dataset, and this division of the 'Outcome' will help our algorithms forecast more accurately for both classes (1 and 0). Co-relational Matrix is plotted to compute relation between the features used for prediction. Co-relation between the features is directly proportional to the accuracy of prediction. The co-relation between the features is improved by scaling all the feature in a specific range of values.

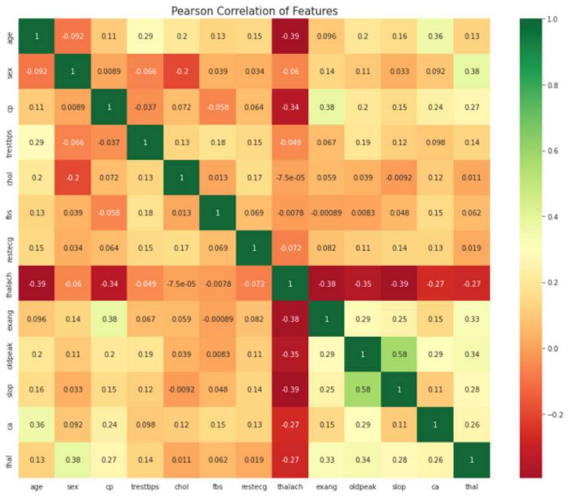

Figure 3: Heart Correlation Matrix

Out of the raw data of 303 patients, 6 had null values in either ca or thal; for this being so few instances of missing data, those data points were just dropped, making a total of 297 data points. The original labels, ranging from 0, no heart disease, to 4, the most advanced stage of heart disease, was redesigned to range from 0, no heart disease, and the original values of 1, 2, 3, and 4 were squished into a single category, 1, "presence of heart disease." In these figures, we are extracting data from the dataset using dat.info so that the dataset can be studied in an analytical manner.

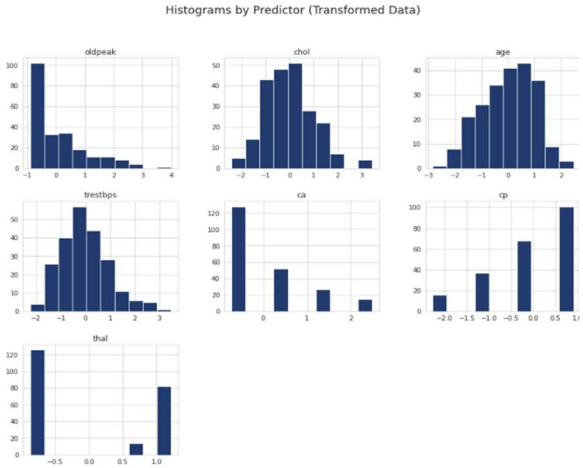

Figure 4: Heart Histogram Attribute Analysis

After selecting the top 7 features for prediction using the random forest classifier. From correlation matrix (Figure 3) and the histogram (Figure 4), we can observe that the features are not correlated to each other. Hence, we scaled all the values between -2 to 2. After data preprocessing the accuracies and recall of most of the algorithm has increased significantly.

Table 1: Model Accuracies for Heart (Before Parameter Tuning)

Algorithm	Accuracy	Recall
Random Forest	84.00%	0.8571
Logistic Regression	85.33%	0.8857
XGBoost	80.00%	0.8000
Decision Tree	74.67%	0.7714
AdaBoost	76.00%	0.8285
KNN	84.00%	0.8285
SVC	81.33%	0.8571

Table 2: Model Accuracies for Heart (After Parameter Tuning)

Algorithm	Accuracy	Recall
Random Forest	84.00%	0.8571
Logistic Regression	85.33%	0.8857
Decision Tree	85.33%	0.8571
AdaBoost	84.00%	0.8285
KNN	84.00%	0.8285
Linear SVC	85.33%	0.8571
RBF SVC	82.67%	0.8857
Sigmoid SVC	88.00%	0.8571
Polynomial SVC	81.33%	0.8571

Comparing Table 1 and Table 2, most models are seeing a lot of improvement all models are up on recall score; the KNN model's accuracy increased from 59% to 84%; the SVC's accuracy increased from 55% to 81% with the exception of XGBoost, which saw a 4% decrease.We have selected the three top-performing models: the sigmoid kernel from the SVCs, AdaBoost from the boosters, and the decision tree over the random forest, for a total of 3 individual models. We combined them into a single voting classifier below. The accuracy which we achieved was 88.57 %. This ensemble method does as well as two of the individual models, the decision tree and AdaBoost, but not as well the sigmoid SVC

To demonstrate the link between different variables, we produced a correlation matrix shown as Figure 5. This graph enables us to see which characteristics have a high correlation with others and hence eliminate them from the model; however, none of the variables had a high correlation with any of the others.The dataset has 8 features which are Pregnancies, Glucose, blood pressure, skin thickness, Insulin, BMI, Diabetes Pedigree Function, Age.

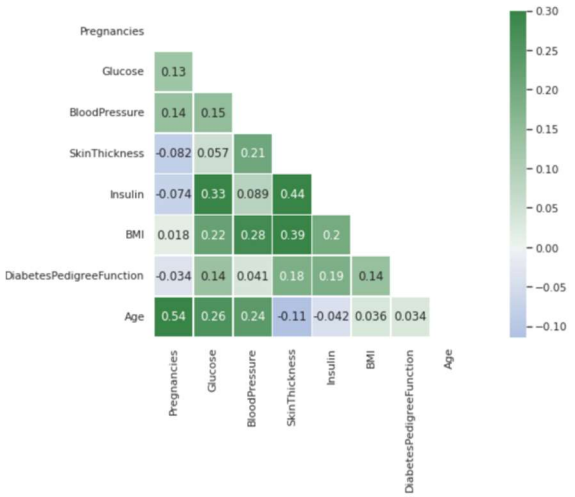

Figure 5: Diabetes Co-relational Matrix

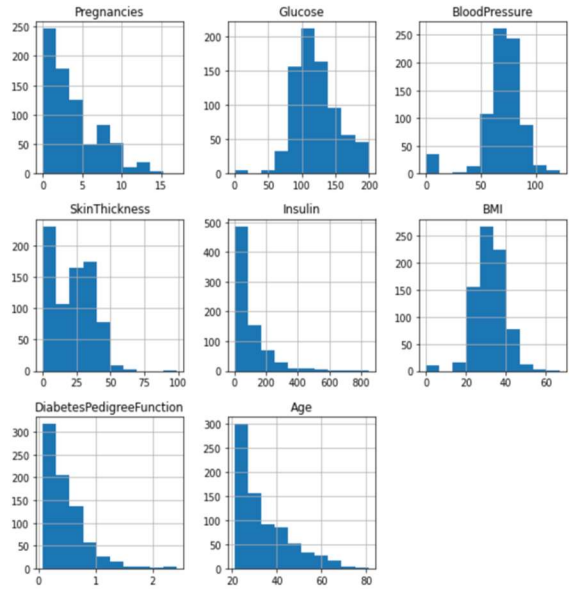

Figure 6: Diabetes Histogram Attribute Analysis

We have developed prediction models using the following classifiers Random Forest, Adaptive Boosting, KNN, Logistic Regression and Decision Tree.

Table 3: Model Accuracies for Diabetic Prediction

Model	Accuracy
Random Forest	72.73%
KNN	77.93%
Logistic Regression	77.92%
Decision Tree	74.46%
AdaBoost	72.73%
VC (without weights)	80.52%
VC (with weights)	80.95%

In Table 3, the two ensemble models used are Adaptive Booster (Boost) & Random Forest (Mean)

- as the foundation. By the combination of different classifiers using Voting Classifier with and without weights, precision can be enhanced. The precision boosts to 80.95% and 80.52% for voting classifiers with or without weights respectively. The same weights were applied on other foundations with the hefty weights delegated for the effective models. The weights used in voting classifier are as follows KNN - 2, Decision Tree – 1, Logistic Regression – 2 and Random Forest – 2.

5. Conclusion

The heart disease and diabetes can be synonymously aggregated to drive a patient's conclusions. In this paper sufficient exploratory analysis and pre-analysis of normalized models has been carried out to understand the need of these predictions using ensemble technique. The system promises to handle and correlate both events of heart and diabetes to drive to quicker prediction using machine learning concepts .For heart disease, it can be concluded that the Voting Classifier of Decision Tree, Sigmoid SVC, and Adaboost has the highest accuracy of 88.57 % and for diabetes, the voting classifier has an accuracy of 80.95 %. The proposed methodology can be extended since it has a scope to conclude immunity of a patient from COVID through the study conducted. The development of a robust model with the help of automated feature selection to work on possibility of COVID through the analysis of both diseases can be carried out in future.

References

[1]Bianca de Almeida-Pititto, Patrícia M. Dualib, Lenita Zajdenverg, JoanaRodrigues Dantas,Filipe Dias de Souza , Melanie Rodacki and Marcello Casaccia Bertoluci on behalf of Brazilian Diabetes Society Study Group (SBD), “Severity And Mortality Of COVID 19”, In Patients With Diabetes, Hypertension And CardiovascularDisease: A Meta-analysis, Diabetology & Metabolic Syndrome Research, (2020)

[2] R. Indrakumari, T. Poongodi, Soumya Ranjan Jena, “Heart Disease Prediction using Exploratory Data Analysis, International Conference” on Smart Sustainable Intelligent Computingand Applications under (ICITETM 2020)

[3] Senthilkumar Mohan 1, Chandrasegar Thirumalai1, And Gautam Srivastava 2,3, (Member,IEEE), “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”,(IEEE ACCESS)

[4] Aishwarya Majumdar, Dr. Vaidehi , “Diabetes Prediction using Machine Learning Algorithms, International Conference” On Recent Trends In Advanced Computing (2019, ICRTAC2019)

[5] Emma Barron, Chirag Bakhai, Partha Kar, Andy Weaver, Dominique Bradley, Hassan Ismail, Peter Knighton, Naomi Holman, Kamlesh Khunti, Naveed Sattar, Nicholas J Wareham,Bob Young, Jonathan Valabhji, Associations of type 1 and type 2 diabetes with COVID-19 related mortality in England: a whole-population study, (Lancet Diabetes Endocrinol 2020)

[6] Quan Zou, Kaiyang Qu , Yamei Luo, Dehui Yin, Ying Ju,Hua Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” Bioinformatics and Computational Biology, (Frontier Genetics Journal, 2018)

[7] Md. Kamrul Hasan , Md. Ashraful Alam , Dola Das, Eklas Hossain, (Senior Member, IEEE), And Mahmudul Hasan, “Diabetes Prediction Using Ensembling of Different MachineLearning Classifiers”, (IEEE ACCESS 2020)

[8] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, Predictive Daa Mining for Medical Diagnosis”, International Journal of Computer Applications, (Volume 17, #8,2011)

[9] Himanshu Sharma, MA Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey” ,International Journal on Recent and Innovation Trends in Computing and Communication (2016)

[10] Baban U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade, “Heart Disease Prediction Using Machine Learning”, IJARSCT, (2021)

[11] Nabaouia Louridi, Samira Douzi, Bouabid El Ouahidi, “Machine Learning-Based Identifcation Of Patients With A Cardiovascular Defect”, Journal of Big Data (2021)

[12] Dhai Eddine Salhi, Abdelkamel Tari, M-Tahar Kechadi, “Using Machine learning for heart disease prediction”, researchgate.net.

[13] Minakshi R. Rajput ,Sushant S. Khedgikar, “Diabetes prediction and analysis using medical attributes: A machine learning approach”, Journal of Xian University of Architecture and Technology.