Disease-Predictor

A Project Report submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Abhinav Bhardwaj *(181500009)*	***Aditya Singh Chauhan*** *(181500046)*
Ankit Parmar *(181500099)*	***Anvit Gupta*** *(181500127)*

Under the Guidance

of

Dr. Mayank Srivastava

Department of Computer Engineering & Applications

Institute of Engineering & Technology

GLA UNIVERSITY
MATHURA
Recognised by UGC Under Section 2(f)

Accredited with A Grade by NAAC

GLA University
Mathura- 281406, INDIA
May, 2022

Department of Computer Engineering and Applications
GLA University, 17 km Stone, NH#2, Mathura-Delhi Road,
P.O. Chaumuhan, Mathura-281406 (U.P.)

Declaration

I hereby declare that the work which is being presented in the ***B.Tech. Major Project “Disease-Predictor”****, in partial fulfilment of the requirements for the award of the* ***Bachelor of Technology*** *in* ***Computer Science and Engineering*** *and submitted to the* ***Department of Computer Engineering and Applications*** *of* ***GLA University, Mathura****, is an authentic record of my own work carried under the supervision of* ***Dr. Mayank Srivastava,*** *Associate Professor***.**

The contents of this project report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree

.

Sign _______________________

Name of Candidate: *Abhinav Bhardwaj*
University Roll.: *181500009*

Sign _______________________

Name of Candidate: *Aditya S. Chauhan*
University Roll No.: *181500046*

Sign _______________________

Name of Candidate*: Ankit Parmar*
University Roll.*: 181500099*

Sign _______________________

Name of Candidate: *Anvit Gupta*
University Roll No.: *181500127*

Department of Computer Engineering and Applications
GLA University, 17 km Stone, NH#2, Mathura-Delhi Road,
P.O. Chaumuhan, Mathura-281406 (U.P.)

Certificate

This is to certify that the above statements made by the candidate are correct to the best of my/our knowledge and belief.

Sign ________________________
Supervisor
Dr. Mayank Srivastava
Designation of Supervisor
Dept. of Computer Eng. & Application

Sign ________________________
Project Co-Ordinator
Dr. Mayank Srivastava
Associate Professor
Dept. of Computer Eng. & App.

Sign ________________________
Program Co-Ordinator
Dr. Rakesh Kumar Galav
Assistant Professor
Dept. of Computer Eng. & App.

Department of Computer Engineering and Applications
GLA University, 17 km Stone, NH#2, Mathura-Delhi Road,
P.O. Chaumuhan, Mathura-281406 (U.P.)

Acknowledgement

It gives us a great sense of pleasure to present the report of the ***B. Tech Major Project*** *undertaken during* ***B.Tech. Final Year****. We owe special debt of gratitude to* ***Dr. Mayank Srivastava****, Associate Professor, Department of Computer Engineering and Applications,* ***GLA University, Mathura*** *for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavour's have seen light of the day. We also take the opportunity to acknowledge the contribution* ***of Dr. Rohit Agrawal****, Head, Department of Computer Engineering and Applications,* ***GLA University, Mathura*** *for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.*

Abstract

Human body is guarded by the immune system, yet it is occasionally insufficient to protect us against infections. Many diseases, which are the primary cause of a large number of fatalities worldwide, are caused by environmental factors and people's lifestyles, and detecting these. Many diseases, which are the primary cause of a large number of fatalities worldwide, are caused by environmental factors and people's lifestyles, and detecting these diseases can be difficult. As a result, predicting illness at an early stage is crucial. However, a doctor's ability to make an accurate diagnosis based on symptoms can be hampered. Doctors' prognoses may be wrong, even if they have much medical experience as they are also human and they can be wrong few times. We cannot expect 100% accuracy from them. The most difficult challenge is making accurate illness predictions. To solve this challenge, data mining is crucial for illness prediction. Each year, the amount of data in medical science grows significantly. The correct analysis of medical data, which has been benefited by early patient care, has increased as the amount of data in the medical and healthcare fields has grown. Data mining discovers hidden pattern information in disease data using disease data.

List Of Figures

CONTENTS

Chapter 1
Introduction

1.1 Overview and Motivation

Currently, medical practitioners prognosticate a patient's ailments using their knowledge’s experience, and understanding gained through time. Doctors' prognoses may be wrong, even if they have much medical experience. Our method will provide a single platform for forecasting a wide range of diseases. The reports will be examined by our system, and conclusions will be drawn from them. The model's updated predictions will be used as a training dataset to improve the system's performance by training it for future inputs. Our technology will serve as a resource for doctors, helping them to be more confident in their decisions and lowering the risk of misdiagnoses.

Because of their lifestyles and the environment in which they live, people nowadays suffer from a variety of ailments. As a result, it's critical to be able to diagnose illness early on. On the other hand, a doctor's ability to make an accurate diagnosis based purely on symptoms is restricted. The most difficult difficulty is predicting sickness accurately. To solve this challenge, data mining is required for disease prediction. Each year, the quantity of data available in medical science grows tremendously.

The necessity for precise medical data analysis has grown as the amount of data in the medical and healthcare professions has grown, which has been supported by early patient care.

The project's major goal is to give medical professionals a web application that will assist them in diagnosing patients. Our online application will make it easier for doctors to diagnose a patient's ailment.

Our platform will create such a model utilising training dataset, and the projected outcome will be included as training data to improve the system's performance over time.

1.2 Objective

- Our system will provide a single platform which can be used to predict different diseases.
- Our system will analyze the reports and will provide the conclusion based on that.
- The new results predicted by the model will also be used as previous dataset to train the model for future input for enhancing the system's performance.
- The new results predicted by the model will also be used as previous dataset to train the model for future input for enhancing the system's performance.
- Our system will serve to doctors as a source which will help doctors to trust their decisions and also reduce the chances of faulty diagnostics.

1.3 Issues and Challenges

In the medical area, using data mining is a very difficult process. Data mining in medical research starts with a theory, and outcomes are altered to meet the hypothesis. This is in contrast to traditional data mining, which begins with datasets and no obvious theory. [28] Traditional data mining is primarily concerned with patterns and trends in datasets, whereas medical data mining is not. Clinical decisions are frequently made based on the doctor's intuition. Due to unintentional prejudice, mistakes, and exorbitant medical costs, the quality of service delivered to patients suffers. Data mining has the ability to create an environment that is rich in knowledge. It has the potential to significantly enhance the quality of clinical decisions. [30] Three supervised machine learning techniques are utilised in the [29] survey. The heart disease dataset was analysed using these methods. For this method, the Classification Accuracy should be compared. This research should be expanded to predict heart disease using fewer variables. The association rule data mining approach was used to predict heart disease in a [29] survey.

The author proposed a method for reducing the number of rules by employing a search restriction. This work should be expanded in the future by employing fuzzy learning models to determine the precision of time in order to reduce the number of rules. In his [31] survey, the author developed a novel approach for categorization that employs a weighted association rule. This research can be furthered in the future by employing the association rule concealment approach in data mining. The author provided a basic group of variables for predicting heart disease in his [30] survey.

This study can be developed and improved in the future to automate heart disease prediction. Real data from health-care organisations and agencies should be obtained in order to compare the best accuracy with all data mining techniques. The author's survey [32] predicts the characteristics of a diabetic patient who develops heart disease. As a consequence of using the Weka tool, the Bayes model was able to accurately categorise 74% of the input examples. Other data mining approaches will be used to expand this study in the future.

Chapter 2
Literature Review

Machine learning has been playing exceptional roles in the area of health over the past few years and continuing to play it in present as well as future. With the growing researches machine learning algorithms has proven amazing results in predictions of various deadly diseases, which in turn made it easy to tackle with the worst situations beforehand. Most of the diseases like breast cancer, brain tumor, chronic kidney disease and Acute Liver Disease which rely on the reports of different tests for the confirmation can be predicted in advance using machine learning algorthms. These advancements in health sector with the help of technologies has provided better treatments to many patients and also saved thousands of lives. And hence, technologies and humans have shaken hands together are creating exceptional outcomes.

2.1 Common Diseases

Dahiwade[9] et al. suggested a machine learning-based approach for diagnosing and predicting common illnesses. The symptoms dataset came from the UCI ML repository, and it included symptoms for a wide range of disorders. To predict numerous illnesses, the system employed CNN and KNN as classification approaches. Furthermore, the proposed solution was complemented with additional information on the tested patient's living habits, which proved to be useful in determining the amount of risk associated with the anticipated disease. In terms of processing time and accuracy, Dahiwade[9] et al. compared the outcomes of the KNN and CNN algorithms. CNN had an accuracy of 84.5 percent and a processing time of 11.1 seconds, respectively. Statistics show that the KNN algorithm performs poorly when compared to the CNN method. In light of this research, Chen[22] et al. [10] found that CNN outperformed traditional supervised algorithms including KNN, NB, and DT. The authors determined that the suggested model was more accurate, which they attribute to the model's capacity to recognise complicated nonlinear interactions in the feature space. Furthermore, CNN finds high-importance characteristics that result in a more precise description of the condition, allowing it to effectively anticipate diseases of great complexity.

This conclusion is well-supported, with actual evidence and statistical reasons to back it up. However, the models offered lacked information, such as Neural Networks characteristics such as network size, architectural type, learning rate, and back propagation technique, among others. Furthermore, the performance analysis is only evaluated in terms of accuracy, which calls into question the veracity of the reported conclusions. Furthermore, the authors failed to account for the bias problem that the tested algorithms encounter. Incorporating additional feature variables, for example, might dramatically improve the performance metrics of underperforming algorithms.

2.2 Kidney Disease

Using The Kidney Function Test (KFT) dataset, Serek et al. planned a comparison investigation of classifier performance for chronic kidney disease (CKD) identification. The classifiers employed in this study are KNN, NB, and RF, and their performance is measured in terms of F-measure, precision, and accuracy. According to the findings, RF outperformed NB in terms of F-measure and accuracy, but NB outperformed RF in terms of precision. Vijaya ani's goal for this project was to use SVM and NB to detect renal disorders. Acute Nephritic Syndrome, Acute Renal Failure, Chronic Glomerulonephritis, and Chronic Kidney Disease were identified using the classifiers. In addition, the study looked at which classification algorithm performed best in terms of accuracy and execution time.

According to the results, SVM outperformed NB in terms of accuracy, making it the best performing algorithm. NB, on the other hand, categorised data in a short amount of time. Several more empirical research focused on identifying CKD; Charlenae et al. and Kotuku et al. determined that the SVM classifier is the best for renal illnesses since it handles semi-structured and unstructured data effectively. As a consequence of this versatility, SVM was able to handle bigger feature spaces, resulting in great accuracy when diagnosing complicated kidney illnesses. Although the data corroborate the conclusion, the previous claim that alternative hyper-parameters were not tested when evaluating the performance of ML algorithms weakens the conclusion. According to Uddin[3], exploring the hyper-parameter space might result in varied accuracy outcomes and higher ML algorithm performance.

2.3 Acute Liver Disease

Liver is the second largest inside organ in the human body. It plays an important role in human body such as for producing protein, clotting blood, as well metabolizing cholesterol, glucose, and iron. Liver also has the function for removing toxins from the body, hence is significant to ensure survival. When a liver fails to operate, many of the body functions cannot be well performed, hence causing significant damage to the body. Liver can be damaged if it is infected with a virus, attacked by its own immune system or injured by chemicals. One fatal liver disease is caused by hepatotropic virus such as the hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus, which can result in chronic liver disease.

According to [27], liver disease is one of the killer diseases in the world. To contain the disease requires an enhanced health analysis through automatic diagnosis of patient record stored in health institutions or organizations. A data mining approach can be used to classify the liver disease into acute or chronic based on the patients’ symptoms. This allows the doctors or medical providers to extract the correct information in order to suggest for effective medical assistance.

2.4 Heart Diseases

Marimuthu etal[16]. used supervised machine learning techniques to forecast cardiac disease. Gender, age, chest pain, gender, goal, and slope were used to organise the data characteristics. The utilised machine learning algorithms were DT, KNN, LR, and NB. According to the research, the LR algorithm had the highest accuracy of 86.89 percent, making it the most successful among the other algorithms. Dwivedi sought to improve the precision of heart disease prediction in 2018 by taking into consideration new characteristics such as resting blood pressure, serum cholesterol in mg/dl, and maximum heart rate reached. The employed dataset came from the UCI ML lab, and it included 120 heart disease positive samples and 150 heart disease negative samples. Artificial Neural Networks (ANN), SVM, KNN, NB, LR, and Classification Tree Dwivedi assessed them all. The findings of Polaraju and Vahid et al., who found that Logistic Regression outperformed other approaches including ANN, SVM, and Adaboost, support this conclusion. The studies excelled in conducting in-depth analyses of machine learning models. For example, for each ML algorithm, several hyper-parameters were evaluated to converge to the greatest feasible accuracy and precision values. Despite this benefit, the learning models are limited in their ability to target illnesses with greater accuracy and precision due to the modest size of the imported datasets.

2.5 Breast Diseases

Shubair attempted to diagnose breast cancer using ML algorithms, particularly RF, Bayesian Networks, and SVM, and Yao[21] came to the conclusion that the RF method performed better than SVM. The researchers used the UCI Repository to access the Wisconsin original breast cancer dataset and used it to compare the learning models in terms of important criteria including accuracy, recall, precision, and ROC graph area. The classifiers were put to the test using the K-fold validation technique, with K set to 10. SVM outperformed other methods in terms of recall, accuracy, and precision, according to the simulation findings. However, the ROC graph indicated that RF had a better likelihood of correctly classifying the tumour. Yao[21], on the other hand, tested a variety of data mining approaches, including RF and

SVM, to find the best algorithm for breast cancer prediction. As a result of the findings, the classification rate, sensitivity, and Random Forest algorithm's accuracy, sensitivity, and specificity were 96.27 percent, 96.78 percent, and 94.57 percent, respectively, whereas SVM's accuracy, sensitivity, and specificity were 95.85 percent, 95.95 percent, and 95.53 percent, respectively.

The former gives more accurate estimations of the amount of data obtained in each feature characteristic. Furthermore, RF is the best method for classifying breast illnesses since it scales effectively for big datasets and reduces the risk of variation and data overfitting. Multiple performance indicators were offered in the research, which helped to solidify the main point. However, including a pre-processing stage to prepare raw data for training has been shown to be detrimental to ML models. According to Yao[21], removing sections of data affects image quality, and hence the ML algorithm's effectiveness is hampered. Parkinson's disease (PARKINSON'S DISEASE) Chen[22] et al. proposed a successful Parkinson's disease (PD) diagnostic method based on Fuzzy k-Nearest Neighbour (FKNN).. The goal of the study was to compare the suggested SVM-based and FKNN-based techniques. For the creation of an optimum FKNN model, the Principal Component Analysis (PCA) was used to combine the most discriminating features. The dataset was collected from the UCI repository and included a variety of biological voice measurements from 31 patients, 24 of whom had Parkinson's disease.

The results of the experiments show that the FKNN strategy outperforms the SVM methodology in terms of sensitivity, accuracy, and specificity. Behroozi[23]'s goal for this study was to develop a novel classification framework for diagnosing Parkinson's disease, which was strengthened by a filter-based feature selection technique that improved 8 | P a g e classification accuracy by 15%. The framework's categorization was determined by using separate classifiers for each subset of the data. dataset to account for the loss of important data KNN, SVM, Discriminant Analysis, and NB were chosen as classifiers. SVM came out on top in every performance indicator, according to the findings. Eskidere[24] also compared the performance of SVM with other classifiers such as Least Square Support Vector (LS-SVM), General Regression Neural Network (GRNN), and Multi-layer Perceptron Neural Network to track the course of PD (MLPNN). The results showed that LS-SVM is the most effective model.

The proper comparison of decoders with their ideal performance metric supports this result. ML algorithms are meant to maximize a variety of performance measures, according to Leveson (e.g., Neural Networks optimizes squared error whereas KNN and SVM optimize accuracy). Furthermore, the writers are very skilled in proposing detailed frameworks. SVM parameters such as the kernel and regularization value, for example, were thoroughly discussed.

2.6 Parkinson’s Disease

Chen[22] et al. [22] reported a successful Parkinson's disease (PD) diagnostic method based on Fuzzy k-Nearest Neighbour (FKNN). The goal of the study was to compare the suggested SVM-based and FKNN-based techniques. For the creation of an optimum FKNN model, the Principal Component Analysis (PCA) was used to combine the most discriminating features. The dataset was collected from the UCI repository and included a variety of biological voice measurements from 31 patients, 24 of whom had Parkinson's disease. The experimental findings have indicated that the FKNN approach advantageously achieves over the SVM methodology in terms of sensitivity, accuracy, and specificity. In line of this study, Behroozi[23] [23] aimed to propose a new classification frame work to diagnose PD, which was enhanced by a filter-based feature selection algorithm that increased the classification accuracy up to 15%. The classification of the framework was characterized by applying independent classifiers for each subset To account for the loss of useful information, use the dataset. KNN, SVM, Discriminant Analysis, and NB were chosen as classifiers. SVM came out on top in every performance indicator, according to the findings. In addition, Eskidere[24] [24] compared the effectiveness of SVM with other classifiers such as Least Square Support Vector (LS-SVM), General Regression Neural Network (GRNN), and Multi-layer Perceptron Neural Network to track the course of PD (MLPNN).

The results showed that LS-SVM is the most effective model. The proper comparison of decoders with their ideal performance metric[25] supports this result. ML algorithms are meant to maximize a variety of performance parameters, according to Lavesson[25] [25]. (e.g., Neural Networks optimizes squared error whereas KNN and SVM optimize accuracy). Furthermore, the writers are very skilled in proposing detailed frameworks. Parameters of SVMs, such as the kernel and the training set, are examples The importance of regularization value was discussed in detail. However, before examining the performances, the ML models were not calibrated. [26] calibration, according to Caruana[26], significantly improves the categorization of a few learning models, particularly NB, SVM, and RF.

2.7 Conclusion

The application of various machine learning algorithms allowed for the early diagnosis of a variety of ailments, including heart, kidney, breast, and brain disorders. SVM, RF, and LR algorithms were the most extensively used predictio algorithms in the literature, with accuracy being the most widely used performance indicator. When it came to forecasting common illnesses, the CNN model proved to be the most accurate. Furthermore, because of its consistency in handling high-dimensional, semi-structured, and unstructured data, the SVM model exhibited improved accuracy in most cases for renal illnesses and PD. Because of its capacity to scale effectively for big datasets and its propensity to prevent overfitting, RF exhibited advantage in the likelihood of correct illness classification for breast cancer prediction. Finally, when it came to forecasting cardiac illnesses, the LR algorithm proved to be the most accurate. To improve illness efficiency in the future, more advanced machine learning algorithms will be required. prediction. Furthermore, learning models should be adjusted more often after the training period to improve performance. Furthermore, to minimize overfitting and improve the accuracy of deployed models, datasets should be enlarged on diverse demo images. Finally, to improve the performance of learning models, more relevant feature selection approaches should be applied.

Chapter 3
Proposed Work

Flow of project starts from collecting dataset from trusted platforms like Kaggle, UCI, Google Dataset Search, etc. After collecting perfect dataset now, it comes step of pre-processing of dataset followed by training and testing the model on different algorithms to compare their performances. Based on the performance we will select our final algorithm.

3.1 Data Source

The dataset used here for predicting disease is taken from Kaggle. Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based datascience environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Gold bloom and Jeremy Howard. Nicholas Gruen was founding chair succeeded by Max Lev chin. Equity was raised in 2011 valuing the company at $25 million. On 8 March 2017, Google announced that they were acquiring Kaggle.

3.2 Architecture Diagram

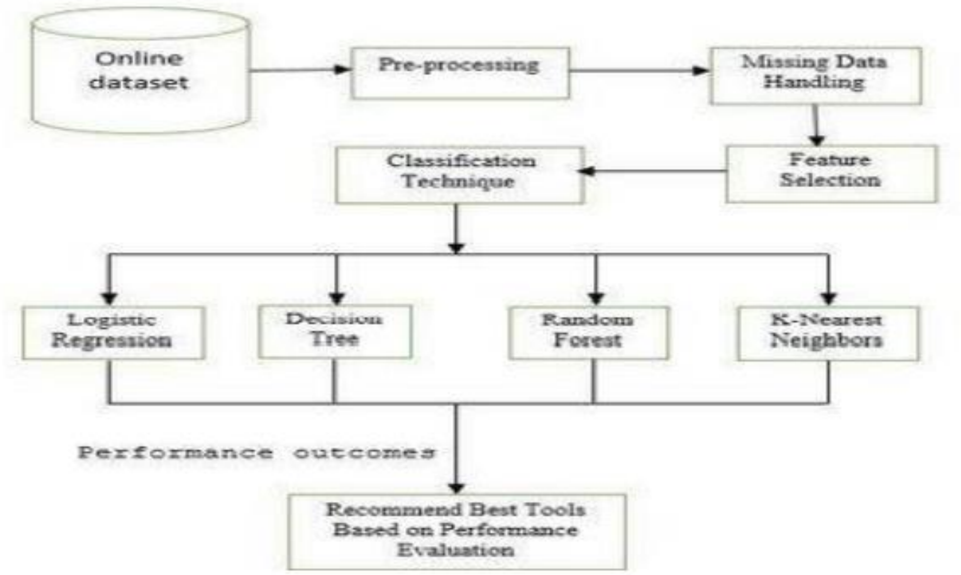


Fig. **3.1**

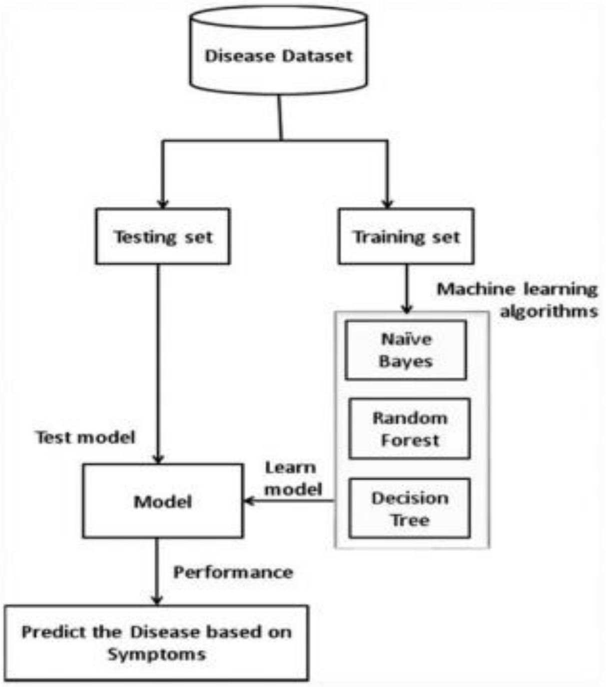


Fig. **3.2**

Architecture Diagrams

3.3 Description of Algorithm

In this section describe four main algorithms are used in this system namely

i) Decision tree Classification Algorithm

ii) Naive Bayes Classifier Algorithm.

iii) Random Forest Classifier Algorithm.

iv) K-Nearest Neighbour Algorithm.

v) Support Vector Machine Algorithm.

3.3.1 Decision Tree Classification Algorithm

The decision tree is a supervised machine learning algorithm. It handles both the categorical data and numerical data. Based on certain conditions it gives a categorical solution such Yes/No, True or false, 1 or 0. For handling medical dataset the Decision tree Classification algorithm is widely used. The result of this model differing from the other models like the knn model, SVM model. The output consists of horizontal and vertical line splits based on the condition depends on the dependent variables. The accuracy level of this algorithm is quite higher than the other algorithms. The reason for the higher accuracy of this algorithm is these model analyses the dataset in the tree shape format. Thus, each and every attribute of the dataset is been analysed. Thus, the accuracy rate of this model is higher. This model analyses the data in the tree-shaped structure. Tree shaped diagram determines the course of actions. The decision tree model analyses the data on the basis of three nodes namely

- **Root node** - this main node, on basis of this node all other perform it function
- **Interior node** - the condition of dependent variables is handled by this node
- **Leaf node** - the final result is carried on a leaf node.

Formula for finding root node (Information Gain)

Information Gain = Class Entropy - Entropy Attributes

To find Class Entropy:

$$(Pi + Ni) = -PP + Nlog2PP + Nlog2NP + N.$$

Here,

=> P, Possibilities of Yes.

=> N, Possibilities of No.

To find Entropy Attributes:

Entropy attribute = $\sum Pi + NiP + N$.

3.3.2 Naive Bayes Classification Algorithm

Naive Bayes classifier is a supervised algorithm which classifies the dataset on the basis of Bayes theorem. The Bayes theorem is a rule or the mathematical concept that is used to get the probability is called Bayes theorem. Bayes theorem requires some independent assumption and it requires independent variables which is the fundamental assumption of Bayes theorem.

Bayes theorem on Mathematical Representation:

$$P(A\backslash B) = P(B\backslash A) * P(A)\ P(B)$$

Here,

P (A) => independent probability of A (prior probability)

P (B) => independent probability of B

P (B\A) => conditional probability of B given A (likelihood)

P (A\B) => conditional probability of A given B (posterior probability).

Naive Bayes is a simple and powerful algorithm for predictive modelling. This model is the most effective and efficient classification algorithm which can handle massive, complicated, non-linear, dependent data. Naïve comprises two parts namely naïve & Bayes where naïve classifier assumes that the presence of the particular feature in a class is unrelated to the presence of any other feature.

3.3.3 Random Forest

Random forest is an ensemble of various decision trees, trained with the bagging methodology. Bagging is used for making the model more stable and accurate by approaching averaging model technique. The random forest classifier is basically a collection of decision tree classifiers where each tree is constructed with a number of random vectors and is able to vote for the most favoured class for prediction. The injection of randomness in the model prevents it from over fitting and provide better result for classification analysis.

3.3.4 K-Nearest Neighbour (KNN)

The K-Nearest Neighbours method is based on the Supervised Learning approach and is one of the most basic Machine Learning algorithms. The K-NN method assumes that the new case/data and existing cases are comparable and places the new case in the category that is most similar to the existing categories. The K-NN method saves all available data and classifies a new data point based on its similarity to the existing data. This implies that fresh data may be quickly sorted into a well-defined category using the K-NN method. The K-NN approach may be used for both regression and classification, however it is more commonly utilized for classification tasks. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data.

It's also known as a lazy learner algorithm since information doesn't learn from the training set right away, instead storing it and retrieving it later.

The k-nearest neighbours algorithm (k-NN) is a non-parametric classification technique created by Evelyn Fix and Joseph Hodges in 1951[1] and later extended by Thomas Cover in statistics. [2] It is employed in the categorization and regression of data. In both circumstances, the input is a data set with the k closest training samples. Depending on whether k-NN is used for classification or regression, the following is the result:

The outcome of k-NN classification is a class membership. An item is categorized by a majority vote of its neighbours, with the object allocated to the most common class among its k closest neighbours (k is a positive integer, typically small). If k = 1, the item is simply assigned to that single nearest neighbour's class. The result of k-NN regression is the object's property value. This number is the average of the k closest neighbours' values. k-NN is a special case of a variable-bandwidth, kernel density "balloon" estimator with a uniform kernel.

The naive version of the algorithm is easy to implement by computing the distances from the test example to all stored examples, but it is computationally intensive for large training sets. Using an approximate nearest neighbour search algorithm makes k-NN computationally tractable even for large data sets. Many nearest neighbour search algorithms have been proposed over the years; these generally seek to reduce the number of distance evaluations actually performed.

k-NN has some strong consistency results. As the amount of data approaches infinity, the twoclass k-NN algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data). Various improvements to the k-NN speed are possible by using proximity graphs.

For multi-class k-NN classification, Cover and Hart (1967) prove an upper bound error rate of

$$R^* \leq R_{knn} \leq R^* \left(2 - \frac{M R^*}{M-1}\right)$$

Here,

=>R* is the Bayes error rate (which is the minimal error rate possible),
=>R_{knn} is the k-NN error rate, and
=>M is the number of classes in the problem

3.3.5 Support Vector Machine

Support Vector Machines, also called Support Vector networks are supervised learning algorithms used for both classification and regression analysis. It classifies the data points plotted in a multidimensional space into categories by parallel lines called the hyperplane. The classification of data points involves the maximization of margin between the hyperplane. There are different kernels available for mapping of linear or no linear data points in a multidimensional space for separation. For our analysis, we have used only the Linear and Radial basis function as kernel.

Chapter 4

Implementation and Result Analysis

Our project's flow began with data collection from reputable sources such as Kaggle. After obtaining the perfect dataset, the next step was to pre-process the dataset if required, then train and test the model on various Machine Learning techniques to compare their performance. After comparing them we had selected our final algorithm based on the comparison results of different algorithm. And at the end we had injected the pickle file of the trained model in the web app designed using ReactJs (for Client side) and Flask (for server side).

4.1 Training and Testing the model for accuracy

Here, the model will be trained using the datasets and tested for finding the accuracy of the model. Optimization will be done to improve the accuracy if needed. In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data. Such algorithms work by making data-driven predictions or decisions, through building a mathematical model from input data. The data used to build the final model usually comes from multiple datasets. But in this project, we have used single dataset for every diseases in different stages of the creation of the model.

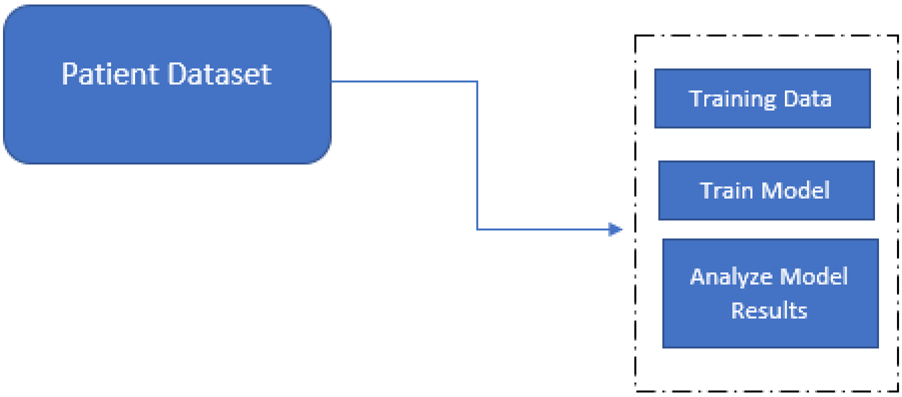


Fig-4.1: Data Flow

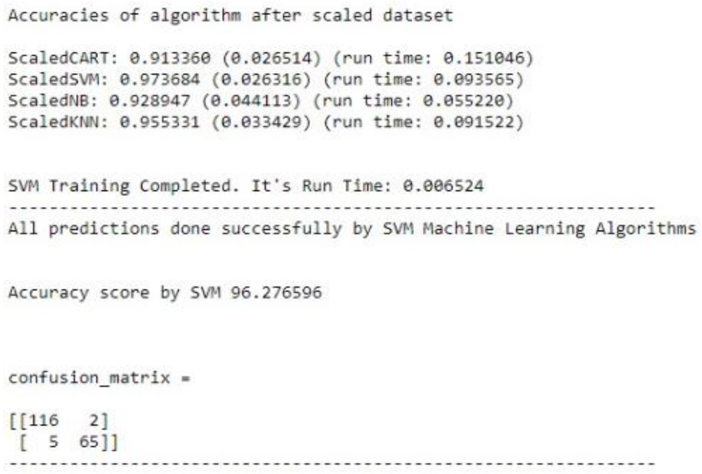
The above fig 2 shows the model is initially fit on a training dataset. The model (e.g. Random Forest Algorithm or SVM or k-Nearest Neighbour Algorithm, depending on dataset) is trained on the training dataset using a supervised learning method. In practice, the training dataset often consist of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), which is commonly denoted as the target (or label).

Before proceeding for the training and testing of the model, we have to split the dataset for training and testing in an appropriate ration so that both the phases have enough data as well as the testing dataset should be unseen / new for the trained model.

The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison of accuracies and standard deviations, the model is selected. The model fitting can include both variable selection and parameter estimation. Successively, the fitted model is used to predict the responses for the observations.

Finally, the test dataset is a dataset used to provide an unbiased evaluation of a final model fit on the training dataset. If the data in the test dataset has never been used in training (for example in cross-validation), the test dataset is also called a holdout dataset and it will be used for testing the trained model (on training dataset).

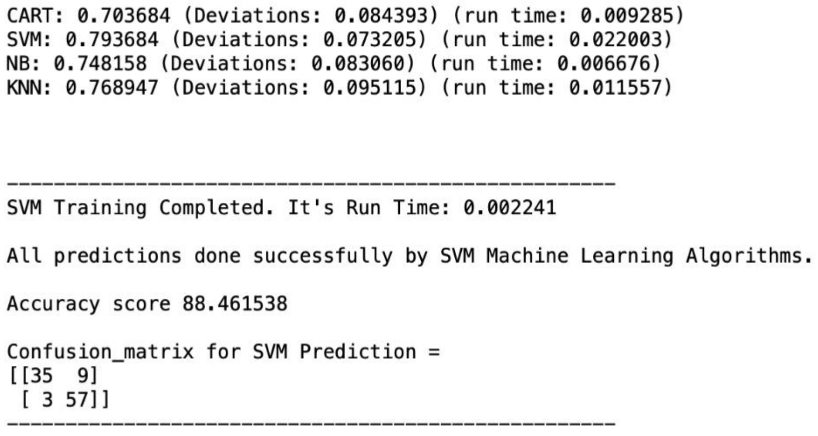
4.2 Result Analysis



```
Accuracies of algorithm after scaled dataset

ScaledCART: 0.913360 (0.026514) (run time: 0.151046)
ScaledSVM: 0.973684 (0.026316) (run time: 0.093565)
ScaledNB: 0.928947 (0.044113) (run time: 0.055220)
ScaledKNN: 0.955331 (0.033429) (run time: 0.091522)


SVM Training Completed. It's Run Time: 0.006524
-------------------------------------------------------------
All predictions done successfully by SVM Machine Learning Algorithms


Accuracy score by SVM 96.276596



confusion_matrix =

[[116   2]
 [  5  65]]
-------------------------------------------------------------
```

Fig. 4.2

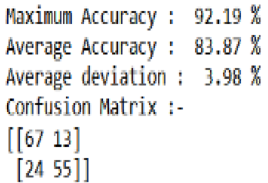
Based on the fig 4.2 of the result of the model for the ***Breast Cancer***, we can see clearly that the ***SVM model*** is giving best accuracy, i.e., ***97.36%*** for the given dataset. So, we have proceeded with the ***SVM Model*** and print the ***Confusion Matrix*** for the test dataset.



```
CART: 0.703684 (Deviations: 0.084393) (run time: 0.009285)
SVM: 0.793684 (Deviations: 0.073205) (run time: 0.022003)
NB: 0.748158 (Deviations: 0.083060) (run time: 0.006676)
KNN: 0.768947 (Deviations: 0.095115) (run time: 0.011557)



----------------------------------------------------
SVM Training Completed. It's Run Time: 0.002241

All predictions done successfully by SVM Machine Learning Algorithms.

Accuracy score 88.461538

Confusion_matrix for SVM Prediction =
[[35  9]
 [ 3 57]]
----------------------------------------------------
```

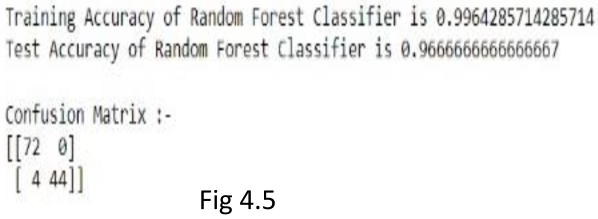
Fig. 4.3

Based on the above fig 4.3 of the result of the model for the ***Heart Attack Disease***, we can see clearly that the ***SVM model*** is giving best accuracy, i.e., ***88.46%*** for the given dataset. So, we have proceeded with the ***SVM Model*** and print the ***Confusion Matrix*** for the test dataset.



```
Maximum Accuracy :  92.19 %
Average Accuracy :  83.87 %
Average deviation :  3.98 %
Confusion Matrix :-
[[67 13]
 [24 55]]
```

Fig 4.4



```
Training Accuracy of Random Forest Classifier is 0.9964285714285714
Test Accuracy of Random Forest Classifier is 0.9666666666666667

Confusion Matrix :-
[[72  0]
 [ 4 44]]
```

Fig 4.5

Similarly, for the datasets of the disease like ***Acute Liver Disease*** and ***Chronic Kidney Disease***, ***Random Forest Algorithm*** is the best suited Model that is giving ***83.87%*** and ***96.67%*** accuracies respectively (refer fig. 4.4 and fig 4.5).

4.3 Cloud based deployment process of the model

Here, the model will be deployed on a cloud server to make the web app accessible across the geographical areas. For the cloud deployment process, I used Heroku Cloud Platform.

Here is the overall representation of the project:

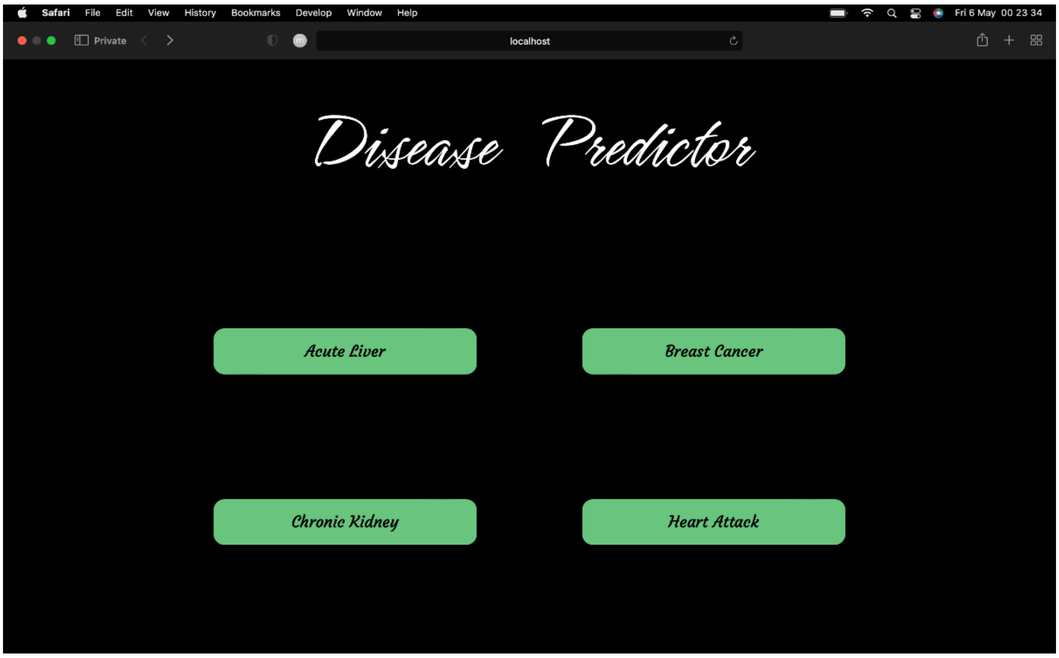


Fig 4.6

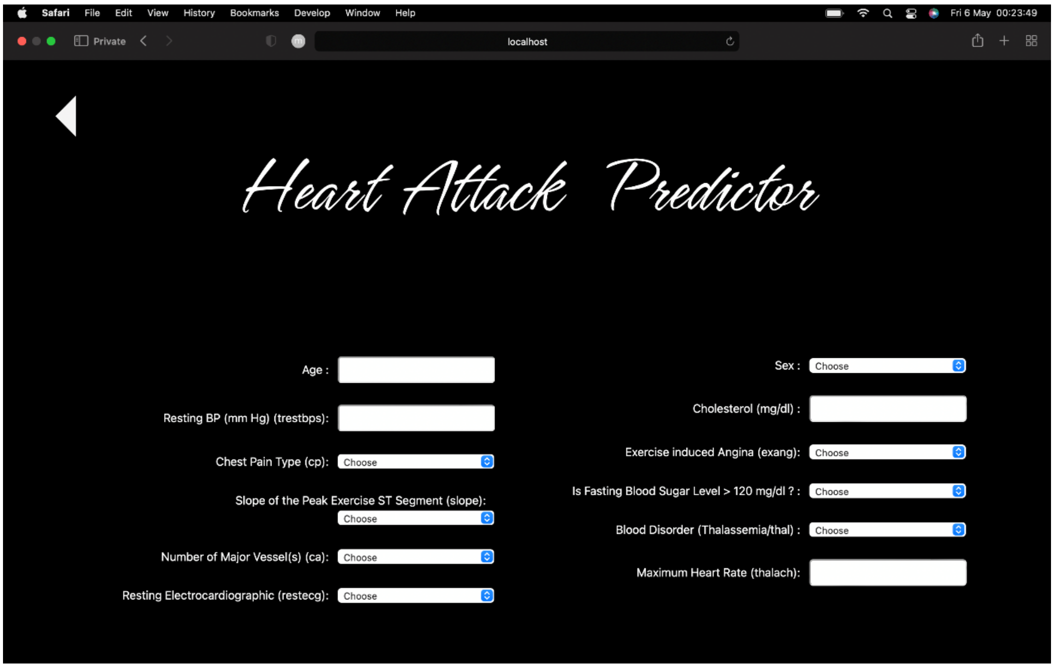
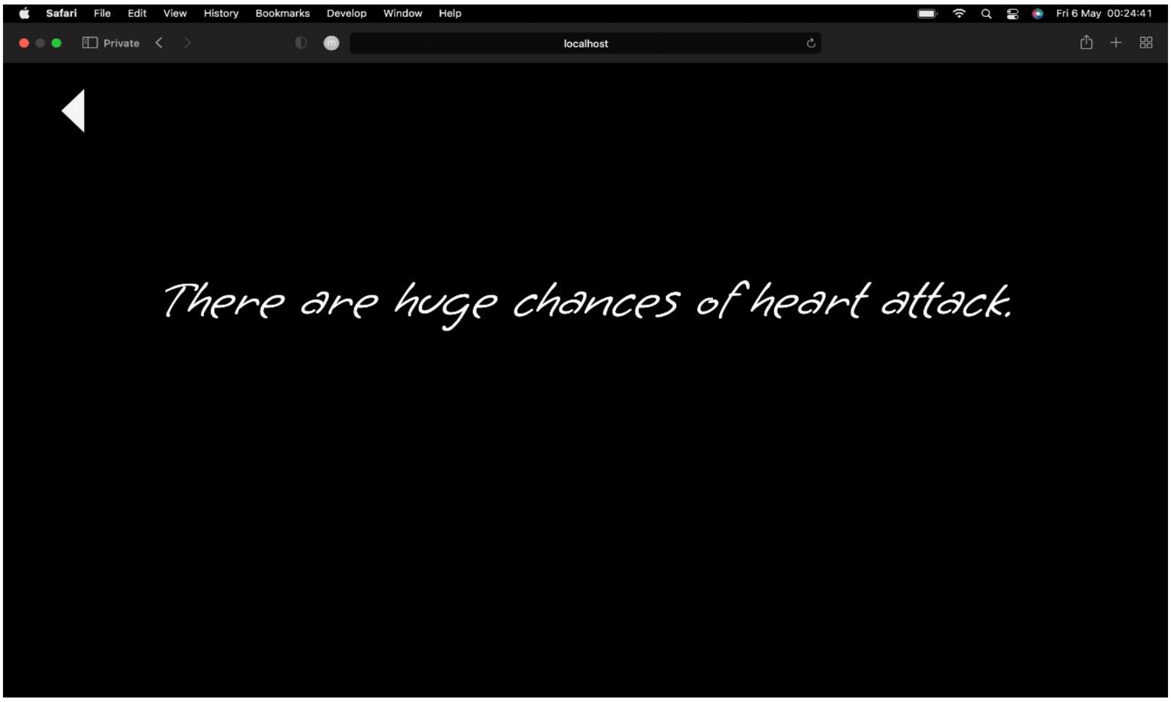

Safari File Edit View History Bookmarks Develop Window Help
Fri 6 May 00:23:49
Private
localhost
Heart Attack Predictor
Age :
Resting BP (mm Hg) (trestbps):
Chest Pain Type (cp): Choose
Slope of the Peak Exercise ST Segment (slope): Choose
Number of Major Vessel(s) (ca): Choose
Resting Electrocardiographic (restecg): Choose
Sex : Choose
Cholesterol (mg/dl) :
Exercise induced Angina (exang): Choose
Is Fasting Blood Sugar Level > 120 mg/dl ? : Choose
Blood Disorder (Thalassemia/thal) : Choose
Maximum Heart Rate (thalach):

Fig 4.7



There are huge chances of heart attack.

Fig 4.8

Chapter 5
Conclusion

Finally, I'd like to state that this project Disease prediction using machine learning is extremely useful in everyone's day-to-day lives, but it's especially important for the healthcare sector, because they're the ones who use these systems on a daily basis to predict the diseases of patients based on their general information and symptoms. Now that the health industry plays such an important role in curing patients' diseases, this is often a useful tool for the health industry to inform the user, and it's also useful for the user if he or she doesn't want to travel to the hospital or other clinics, because simply by entering the symptoms and other relevant information into the form, the user can learn about the disease to which he or she is subjected, and the health industry can benefit as well. Simply ask the user for their symptoms and enter them into the system, and they will give you the precise and, to some degree, accurate diseases in a matter of seconds. If the health sector embraces this idea, doctors' workload will be decreased and they will be able to forecast the patient's sickness more readily. The purpose of disease prediction is to provide predictions for a variety of common diseases that, if left untreated and often ignored, can progress to deadly disease and cause a slew of problems for the patient. This project may be updated in the future by adding additional attributes to the dataset and making it more interactive for the users. It can also be done as a mobile application. We'll make changes to the system by linking it to the hospital's database.

References

[1] A. Gavhane[1], G. Kokkula[1], I. Pandya, and K. Devadkar[1], "Prediction of heart disease using machine learning," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

[2] Y. Hasija[2], N. Garg, and S. Sourav, "Automated detection of dermatological disorders through image-processing and machine learning," in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 1047–1051.

[3] S. Uddin[3], A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019.

[4] R. Katarya[4] and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302–305.

[5] P. S. Kohli[5] and S. Arora, "Application of machine learning in disease prediction," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

[6] M. Patil[6], V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra, "A proposed model for lifestyle disease prediction using support vector machine," in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1–6.

[7] F. Q. Yuan[7], "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2016, pp. 1903–1907. [8] S. Ismaeel

[8], A. Miri, and D. Chourishi, "Using the extreme learning machine (elm) technique for heart disease diagnosis," in 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), 2015, pp. 1–3.

[9] D. Dahiwade [9], G. Patle, and E. Meshram, “Designing disease prediction model using machine learning approach,” Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, no. Iccmc, pp. 1211–1215, 2019.

[10] S. Jadhav[10], R. Kasar, N. Lade, M. Patil[6], and S. Kolte, “Disease Prediction by Machine Learning from Healthcare Communities,” International Journal of Scientific Research in Science and Technology, pp. 29–35, 2019.

[11] R. Saravanan[11]and P. Sujatha, “A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification,” in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 945– 949.

[12] Y. Amirgaliyev[12], S. Shamiluulu, and A. Serek, “Analysis of chronic kidney disease dataset by applying machine learning methods,” in 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1–4.

[13] V. S and D. S, “Data Mining Classification Algorithms for Kidney Disease Prediction,” International Journal on Cybernetics & Informatics, vol. 4, no. 4, pp. 13–25, 2015.

[14] A. Charleonnan[13], T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, “Predictive analytics for chronic kidney disease using machine learning techniques,” 2016 Management and Innovation Technology International Conference, MITiCON 2016, pp. MIT80–MIT83, 2017.

[15] P. Kotturu[15], V. V. Sasank, G. Supriya, C. S. Manoj, and M. V. Maheshwarredy, “Prediction of chronic kidney disease using machine learning techniques,” International Journal of Advanced Science and Technology, vol. 28, no. 16, pp. 1436–1443, 2019.

[16] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, “A Review on Heart Disease Prediction using Machine Learning and Data Analytics

Approach," International Journal of Computer Applications, vol. 181, no. 18, pp. 20–25, 2018.

[17] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Computing and Applications, vol. 29, no. 10, pp. 685–693, 2018.

[18] K. Polaraju, D. Durga Prasad, 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. 2321–9939, 2017. [Online]. Available: www.ijedr.org

[19] S. Pouriyeh[19], S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204– 207.

[20] P. P. Sengar[20], M. J. Gaikwad, and A. S. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, pp. 796–801, 2020.

[21] D. Yao[21], J. Yang, and X. Zhan, "A novel method for disease prediction: Hybrid of random forest and multivariate adaptive regression splines," Journal of Computers (Finland), vol. 8, no. 1, pp. 170–177, 2013.

[22] H. L. Chen[22], C. C. Huang, X. G. Yu, X. Xu, X. Sun, G. Wang, and S. J. Wang, "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor 12 | Page approach," Expert Systems with Applications, vol. 40, no. 1, pp. 263–271, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2012.07.014

[23] M. Behroozi[23] and A. Sami, "A multiple-classifier framework for Parkinson's disease detection based on various vocal tests," International Journal of Telemedicine and Applications, vol. 2016, 2016.

[24] O. Eskidere[24], F. Ertas¸, and C. Hanilc¸i, “A comparison of regression ¨ methods for remote tracking of Parkinson’s disease progression,” Expert Systems with Applications, vol. 39, no. 5, pp. 5523–5528, 2012.

[25] N. Lavesson[25], Evaluation and Analysis of Supervised Learning Algorithms and Classifiers, 2006.

[26] R. Caruana[26] and A. Niculescu-Mizil, “An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics,” Proceedings of the 23rd international conference on Machine Learning, pp. 161–168, 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.3232

[27] N. Lavesson[25], Evaluation and Analysis of Supervised Learning Algorithms and Classifiers, 2006. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1529/3/032002/pdf

[28] Ruben D. Canlas Jr.,”DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES”, August 2009

[29] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction” IJCSE Vol. 3 No. 6 June 2011

[30] M. ANBARASI, E. ANUPRIYA, N.CH.S.N.IYENGAR, “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm”, International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376

[31] Carloz Ordonez, “Association Rule Discovery with Train and Test approach for heart disease prediction”, IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343

[32] G. Parthiban, A. Rajesh, S.K.Srivatsa “Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method”