# Health Diagnosis Prediction using Machine Learning Model

**[1]Sagar D. Waykar, [2]Mayur S. Khaladkar, [3]Pratik S. Ranaware, [4]Gurudas B. Markad, [5]Prof. Mahesh C. Shinde**

[1,2,3,4]Student, [5]Guide & Professor
Department of Mechanical Engineering,
JSPM's Jayawantrao Sawant College of Engineering, Pune, India

***Abstract*— There are several diseases that are considered threaten to life also chronic and it would help patients if these diseases where dragonised at the earliest stage. Many complications occur if these diseases are not treating and are not identified. The goal of developing a classifier system using Machine Learning (M.L.) algorithm is to help ultimately in solving health related problems by helping doctors to predict and diagnose diseases at earlier stage. Our prediction system is a project that can predict the chances of getting diseases, based on different user test report results. In this paper the risk of contracting diseases such as Diabetes, Heart Disease, Parkinson's & Liver Disease are predicted using M.L. Algorithms. The end result is predicted based on the most accurate M.L. Algorithm. A web page application is designed where the users will select the disease and enter the symptoms or values such as Age, Blood Pressure Value, Glucose level and BMI level. The system then uses the M.L. model which can predict the chances of the selected disease in the patients.**

***Index Terms*— Machine Learning (ML) Model, Algorithms, Disease, Prediction, Symptoms.**

## I. INTRODUCTION

Machine Learning (M.L.) is a very promising approach which helps in early diagnosis of disease and might help the practitioners in decision making for diagnosis. We explore the landscape of recent advances to address the challenges model interpretability in healthcare and also describe how one would go about choosing the right interpretable M.L. algorithm for a given problem in healthcare.

This project extensively covers the definitions, nuances, challenges, and requirements for the design of interpretable and explainable M.L. models and systems in healthcare and how they should be deployed. Prediction by a traditional bio-medical M.L. models typically involve some supervised algorithm which uses guidance data with the label for the prediction of the models. Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier.

The modern approach to healthcare is to prevent the disease with early intervention rather than go for treatment after diagnosis. Traditionally, physicians or doctors use a risk calculator to access the possibility of disease development. The motive of this study is to design models which can prognosticate the likelihood of various diseases in patients with maximum accuracy. We plan to create an end user support and online consultation system. This paper shows us a way of building an application which can help to solve the health-related issues by assisting the physicians and patients to predict and diagnose diseases at an early stage.

## II. LITERATURE SURVEY

The analysis of the related work returns results on various health care data sets, where the analysis and predictions were carried out using various methods and techniques. Several researchers have developed and implemented various prediction models using variants of data mining techniques, M.L. algorithms, or also a combination of these techniques.

Priyanka Sonar, Prof. K. Jaya Malini **[1]**, according to their research diabetes being one of the most dangerous diseases in the world and can cause a wide variety of disorders including blindness etc. In this paper, they have used machine learning techniques to discover diabetes disease as it is easy and flexible. to predict whether the patient has disease or not. Their goal of this analysis was to invent a system that can help the patient to detect the diabetic disease of the patient with accurate results. Here they mainly used 4 major algorithms Decision Tree, Naïve Bayes and SVM and compared their accuracy which is 85%, 77% and 77.3% respectively.

The main focus of the article by Archana Singh, Rakesh Kumar **[2]**, is that the heart plays an important role in living organisms. Therefore, the diagnosis and prediction of heart related diseases must be perfect and correct because what can cause death in heart related cases is very crucial. So, machine learning and artificial intelligence support the prediction of any kind of natural events. Therefore, in this paper they calculate the accuracy of machine learning in predicting cardiac disease using nearest neighbor, decision tree, linear regression, and SVM by using the UCI repository dataset for training and testing. They also compared the algorithm and its accuracy SVM 83%, decision tree 79%, linear regression 78%, k-nearest neighbor 87%.

A. Sivasangari, Baddigam Jaya Krishna Reddy, Annamareddy Kiran, P. Ajitha **[3]** defines that liver diseases are causing a large number of deaths in India and it is also considered as a life-threatening disease in the world. As it is difficult to detect liver disease at an early stage. So, by using an automated program that uses machine learning algorithms, we can accurately detect liver disease.

They used and compared SVM, Decision Tree, and Random Forest algorithms and measured precision, accuracy, and recall metrics for quantitative measurement. The accuracy is 95%, 87%, 92% respectively.

The research paper by Dinesh Kumar G., Santhosh Kumar D **[4]**, contributes to the correlative application and analysis of different machine learning algorithms in R software, providing an out-of-the-box mechanism for the user to use machine learning algorithms in R software to predict cardiovascular diseases.

Mrunmayi Patil, Vivian Brian Lobo, Pranav Puranik, Aditi Pawaskar, Adarsh Pai, Rupesh Mishra **[5]** their study aims to understand the support vector machine and use it to predict lifestyle diseases to which an individual might be susceptible.

Mr. Santhana Krishnan J, Dr. Geetha. S. **[6]**, are applied two supervised data mining algorithms on the data set to predict the chances of having a heart disease of a patient, they were analyzed with the classification model, namely the Naïve Bayes classifier and the decision tree classification. The decision tree model has predicted heart disease patients with an accuracy level of 91% and the Naïve Bayes classifier has predicted heart disease patients with an accuracy level of 87%.

S. Grampurohit and C. Sagarnal **[7]**, considered the medical records of 4920 patients and identified 132 symptoms corresponding to 41 diseases and implemented this data set using Naïve Bayes, Decision Tree and Random Forest algorithms. In this system, Naïve Bayes showed the comparatively highest security of 93.61%.

D. Dahiwade, G. Patle and E. Meshram **[8]** are implemented the convolutional neural network and nearest neighbor algorithms in the disease dataset from the UCI machine repository. The results found through the comparative analysis are that CNN performed better than KNN in precision and time comparison.

## III. MACHINE LEARNING

The study of machine learning revolves around computer algorithms and programs that have the ability to improve their performance as they gain experience.

This field is a subset of artificial intelligence, as it aims to make machines intelligent. Machine learning leverages statistical models to analyze and learn from collected data, enabling it to accomplish tasks such as predictions and classifications.

### 3.1 How Does Machine Learning Work?

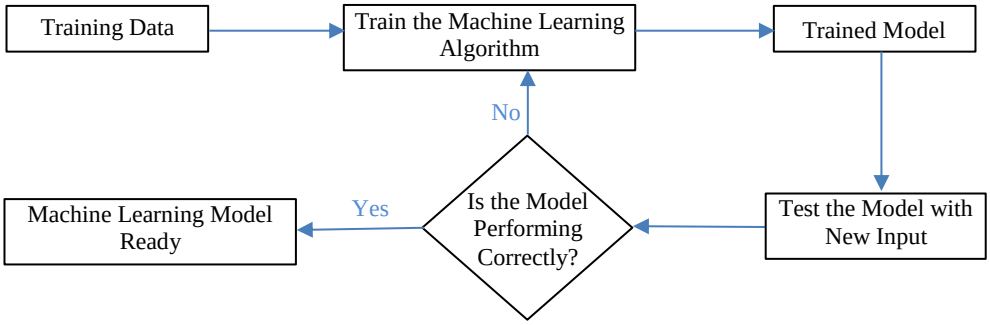


**Figure 01: Machine Learning Model.**

### 3.2 Key Terms Associated with Machine Learning

| Term | Description |
|---|---|
| Features or Attributes | Anything that you can measure and build data for. For example, the typical length of various animals. Feature could be numeric, set of characters, Boolean values, or anything else that describes the data. |
| Training Dataset | (Big data) or the complete set of sample data (training examples) based on which you would train your machine learning model. |
| Testing Dataset | The set of sample data (testing examples) based on which you could test your trained machine learning model for its correctness and determining if further training or adjustment is required to the model. |
| Target variable | The feature or value that you want to predict or identify or the output that the trained machine learning model should produce when an input is fed to it. |

**Table 01: Key Terms Associated with Machine Learning.**

## IV. METHODOLOGY

Processing of prediction system begins with data collection. we visit various websites such as Kaggle, Microsoft Dataset, Awesome Public Datasets Collections, Scikit-learn Datasets and Government Datasets to gather well-verified datasets from authoritative sources and numerous researchers. Then we collect datasets related to Diabetes Disease, Heart Disease, Parkinson's Disease, and Liver Disease and thoroughly study them to understand the information they contain.

For this project, we have chosen to work with the following Machine Learning (ML) algorithms: Support Vector Machine (SVM), Logistic Regression, Random Forest Regressor, and k-Nearest Neighbors (KNN) classification algorithm. After selecting the algorithms, we preprocess the data to prepare it for training the Machine Learning (ML) models.

Next step involves model evaluation. we train and test the datasets using all four algorithms and find calculate the accuracy scores of the machine learning models on both the training data and testing data. This evaluation helps us in developing an accurate prediction system.

Once we have obtained the accuracy scores, we proceed to create the predictive system. We write code that allows users to input their information or symptoms and click on the predict key. The system then provides a binary outcome, indicating whether the person is diabetic or not, has heart disease or not, has Parkinson's disease or not, and has liver disease or not. We follow the same procedure for each disease prediction.

The Final task developing the Machine Learning (ML) models for multiple disease predictions is to deploy the models and create a web application for easy accessibility.

## V. ARCHITECTURE OF DISEASE PREDICTION SYSTEM

### 5.1 Block Diagram of Disease Prediction System

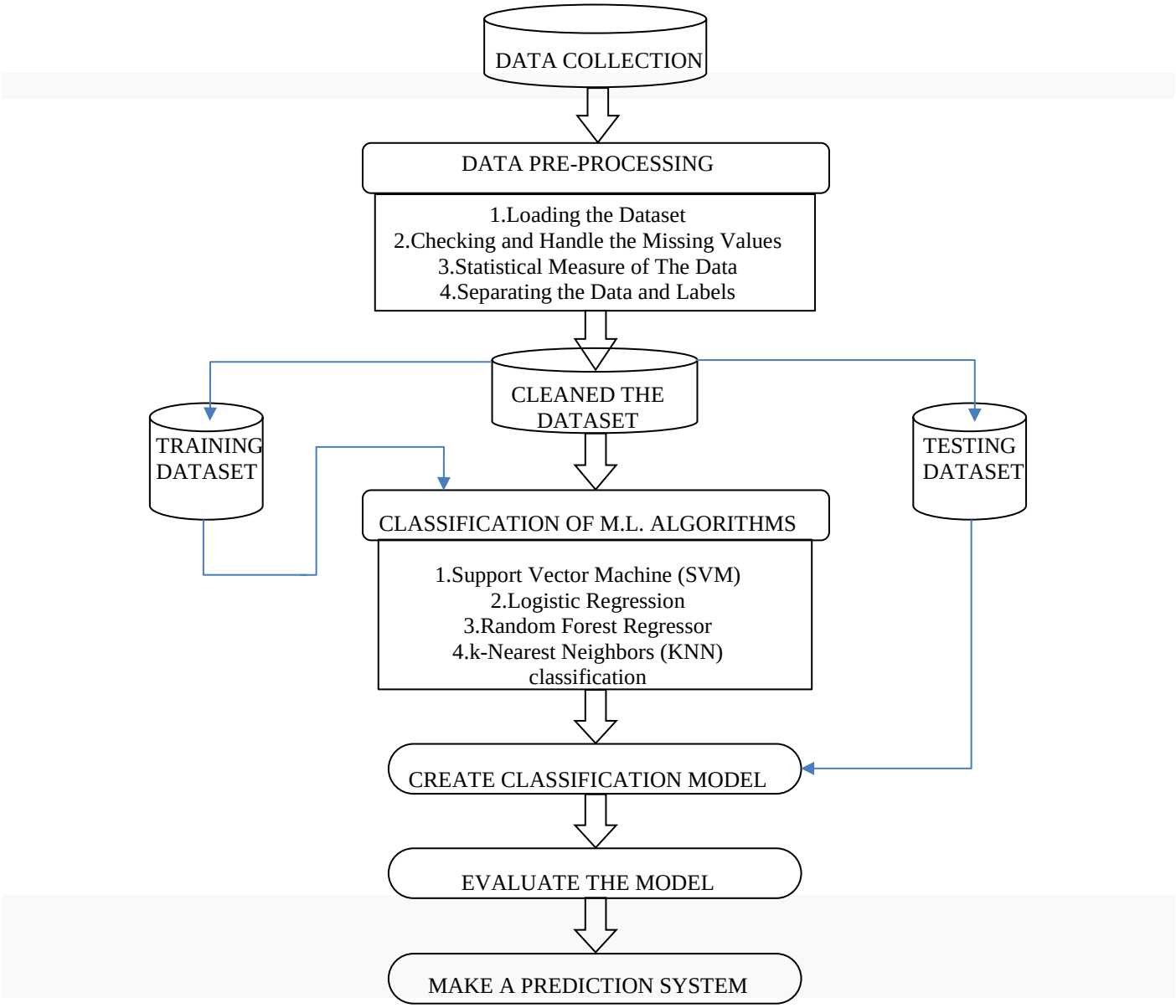


**Figure 02: Block Diagram of Disease Prediction System.**

### 5.2 Front End View of The System

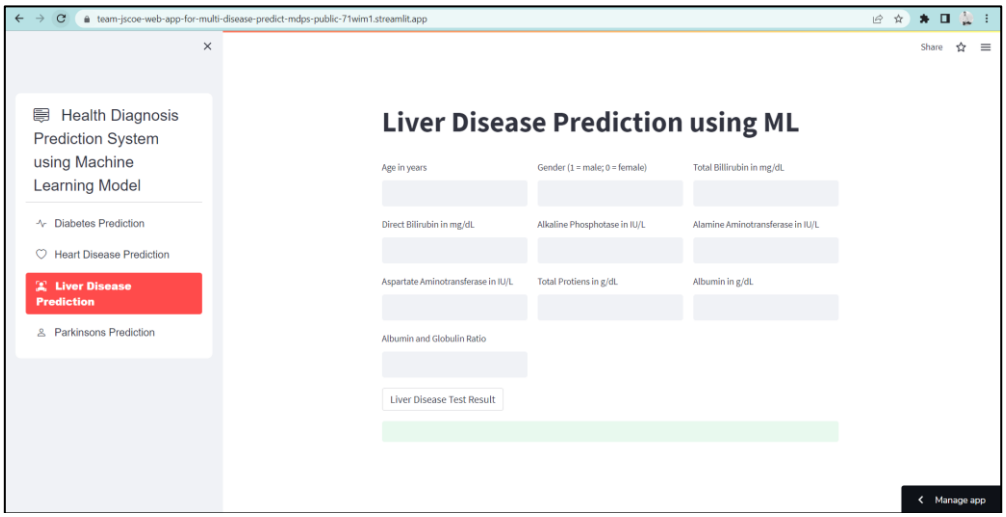


**Figure 03: Front End View of The System.**

## VI. MACHINE LEARNING ALGORITHMS USED

### 6.1 Support Vector Machine (SVM)

Support Vector Machines (SVM) is a classification technique that uses critical boundary datapoints to create a hyperplane that differentiates the data points SVM is a supervised learning method. You plot each datapoint as a point in a n-dimensional space (where n is the number of datapoint attributes you have). Then, you perform classification by finding the hyperplane that adequately differentiates the two classes. The higher the gap between the datapoints (highest boundary datapoints of one class and lowest boundary datapoints of another class), the better.

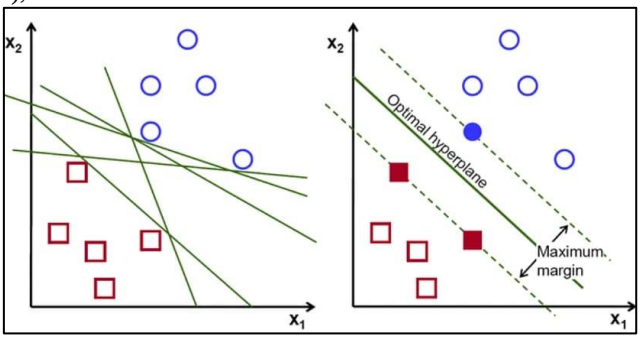


**Figure 04: Support Vector Machine (SVM) Algorithm.**

Hyperplanes are decision boundaries that help in classifying the datapoints. Datapoints falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplanes depends upon the number of data attributes. If the number of input data attributes is 2, then the hyperplane is just a line. If the number of input data attributes is 3, then the hyperplane becomes a two-dimensional plane, and likewise.

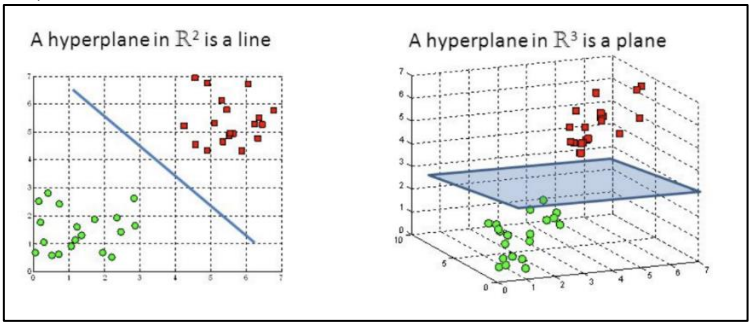


**Figure 05: Support Vector Machine (SVM) Algorithm.**

Support vectors are datapoints that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, you maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help in building the SVM model.

### 6.2 Logistic Regression

Logistic regression is a regression technique used to model and estimate the probability of an event occurring based on the values of the independent variables.

Unlike the linear regression line, the logistic regression line is a curve. This is called Sigmoid curve or 5-curve in short.

All the outcome data points are either concentrated on 0 or 1. The curve depicts various possible probabilities of an event, occurring between 0 (totally uncertain) and (totally certain), on the y-axis.

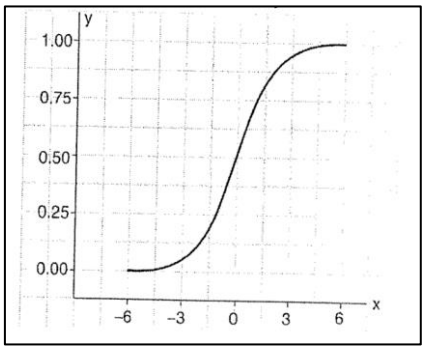


**Figure 06: Logistic Regression.**

The probability prediction formula for logistic regression is as following

$$\hat{P} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

Where,

- $\hat{P}$ (pronounced p-hat) is the predicted probability
- e is the exponential function
- βo and β1 are regression coefficients
- and X, is the value of the independent variable.

Note here that calculating the values of logistic regression coefficients βo and β1 requires statistical software and is impractical to calculate by hand. Also, do not get confused seeing βo + β1 X1 in the logistic regression formula. It has nothing to do with linear coefficient calculation.

### 6.3 K-Nearest Neighbors (KNN) Classification Algorithm

One of the popular variations of K-means clustering algorithm is k-Nearest Neighbors (KNN) classification algorithm. It works on the same principle as K-means clustering algorithm that a data point is likely to resemble its neighbor's and would possibly have the same classification. INN is a supervised learning method.

The way it works is simple and straightforward.

1. Assume that you have already assigned labels to the existing data points in a given data set.
2. Then you are given a new data point to classify
3. You calculate the distance of the new data point with respect to its k neighbors. The number k is chosen based on your data set and requirements but in general higher the better to reduce the noise and avoid incorrect labelling.
4. The new data point is classified based on the classification of the majority of the surrounding neighbor's classification

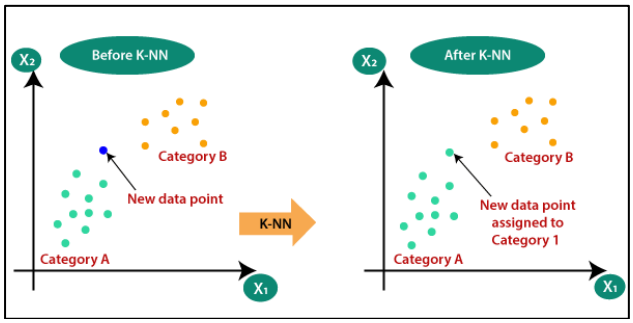


**Figure 07: K-Nearest Neighbors (KNN) Classification.**

### 6.4 Random Forests Regressor

Random Forests is an ensemble technique designed to combine several decision trees to reduce errors and to build a more accurate prediction model.

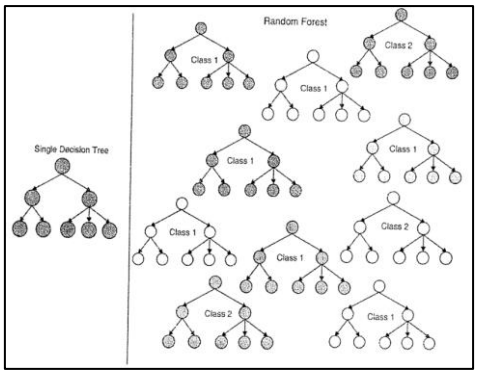


**Figure 08: Random Forest Regressor.**

In random forest, instead of building one decision tree, multiple decision trees are built

1. Each tree in the forest is built using random datapoints drawn from the dataset.

2 The trees in the forest are split such that a fixed subset of Input variables (attributes) is taken each time and the best possible split is performed. For example, if there are 5 attributes in the dataset, you could choose to take 2 attributes at a time and split based on those attributes to create a tree in the forest.

3.Finally, the result from each tree is aggregated to give the outcome. Usually, you count the total number of votes (based on the classification outcome from each tree in the forest) for classification problems and average (based on the outcome value from each tree in the forest) for regression problems.

## VII. RESULTS AND DISCUSSION

### 7.1 The Accuracy Score of used Algorithms

| Disease | Accuracy Score of Algorithms on Train and Test Data in percentage | | | | | | | | Algorithm Which is Used |
|---|---|---|---|---|---|---|---|---|---|
| | Support Vector Machine (SVM) | | Logistic Regression | | Random Forests Regressor | | K-Nearest Neighbors (KNN) Classification | | |
| | Train Data (%) | Test Data (%) | Train Data (%) | Test Data (%) | Train Data (%) | Test Data (%) | Train Data (%) | Test Data (%) | |
| Diabetes | 0.78 | 0.77 | 0.78 | 0.75 | - | - | 0.79 | 0.72 | Support Vector Machine (SVM) |
| Heart | 0.85 | 0.81 | 0.85 | 0.81 | - | - | 0.78 | 0.62 | Logistic regression |
| Parkinson's | 0.88 | 0.87 | 0.87 | 0.82 | - | - | 0.96 | 0.76 | Support Vector Machine (SVM) |
| Liver | 0.71 | 0.71 | 0.71 | 0.71 | 0.87 | 0.87 | 0.78 | 0.78 | Random Forests Regressor |

**Table 02: Accuracy Score of Algorithms.**

### 7.2 Results of Multiple Disease Prediction

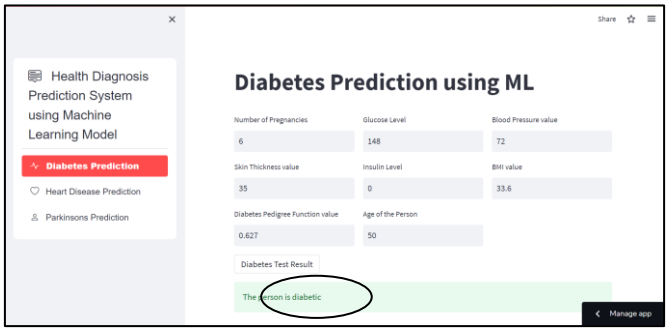


**Figure 09: Results of Diabetes Prediction.**

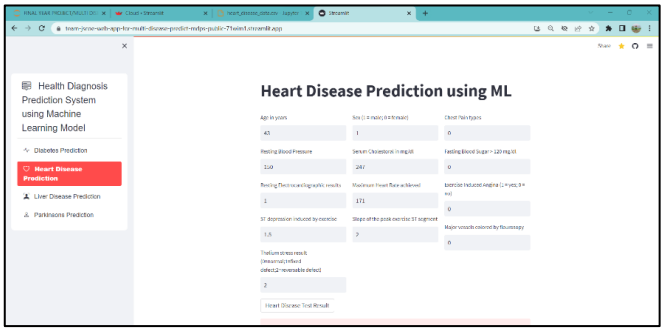


**Figure 10: Results of Heart Disease Prediction.**

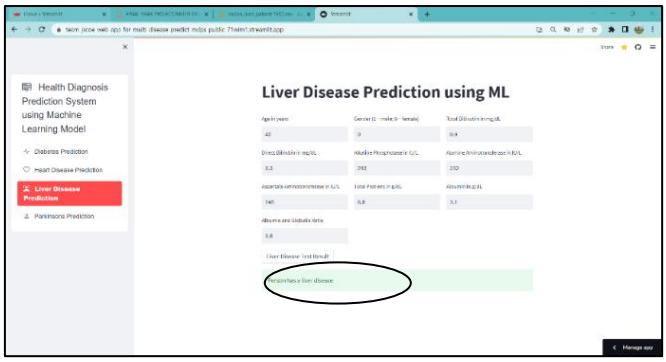


**Figure 11: Results of Liver Disease Prediction.**

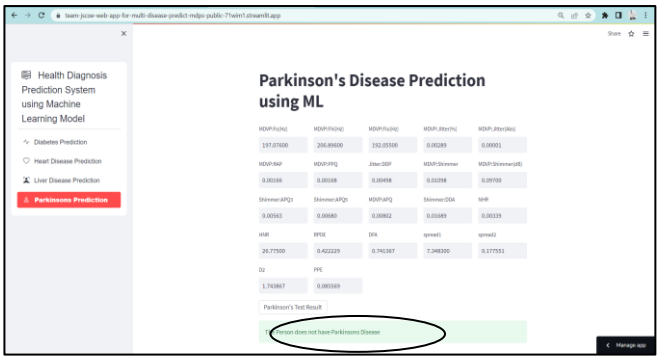


**Figure 12: Results of Parkinson's Disease Prediction.**

## VIII. CONCLUSION

We have proposed a general disease prediction system based on Machine Learning (ML) algorithms. We utilized SVM, Logistic Regression, Random Forest Regressor, and KNN algorithms to classify patient data and predict diseases. Our results and discussion demonstrate accurate disease risk predictions when providing patient records included in the datasets, helping determine the level of risk for the predicted disease.

We calculated the accuracy scores of all four algorithms: Support Vector Machine (SVM), Logistic Regression, Random Forest Regressor and k-Nearest Neighbors (KNN) classification, for each disease dataset. We selected the most accurate and suitable algorithm for each specific disease prediction, resulting in the creation of a machine learning model.

As a result, for the Diabetes and Parkinson's Disease datasets, the SVM algorithm exhibited higher accuracy than the other two algorithms. Logistic Regression yielded high accuracy results for heart disease, while the Random Forest Regressor demonstrated high accuracy results for Liver Disease datasets. Therefore, we can conclude that different algorithms are suitable for different disease predictions in terms of accuracy.

## REFERENCES:

1. Priyanka Sonar, Prof. K. Jaya Malini, "DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC).
2. Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3).
3. A. Sivasangari, Baddigam Jaya Krishna Reddy, Annamareddy Kiran, P. Ajitha, "Diagnosis of Liver Disease using Machine Learning Models" 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).
4. "Prediction of Cardiovascular Disease Using Machine Learning Algorithms" Dinesh Kumar G., Santhosh Kumar D. (2018).
5. "A Proposed Model for Lifestyle Disease Predict Vectorion Using Support Machine" Mrunmayi Patil, Vivian Brian Lobo, Pranav Puranik, Aditi Pawaskar, Adarsh Pai, Rupesh Mishra U.G. Student, Assistant Professor (2018).
6. "Prediction of Heart Disease using Machine Learning Algorithm" Mr. Santhana Krishnan J, Dr. Geetha. S. (2018).
7. S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms,"2020 International Conference for Emerging Technology (INCET),2020, pp. 17, doi: 10.1109/INCET49848.2020.9154130.
8. D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Con- ference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.