JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Multiple Disease Prediction Webapp

[1]Mohammed Juned Shaikh, [2]Soham Manjrekar, Danish Khan,[4]Muzaffar Khan,[5]Danish Jamadar

[1]Assistant Professor

[2-5]Students, [1-6]Department of Computer Engineering,

[1]Rizvi College of Engineering Mumbai, India

Abstract - Our point is to anticipate the various sorts of illness in a single stage by utilizing the inbuilt python module Streamlit. In this task we are utilizing Naïve Bayes algorithm, random forest, decision tree and svm classifier are utilized for prediction of a particular disease. The calculation which gives more accuracy is used to train the data set before implementation. To implement multiple disease analysis using machine learning algorithms, Streamlit and python pickling is utilized to save the model behavior. In this article we analyze Diabetes analysis, Heart disease and Parkinson's disease by using some of the basic parameters such as Pulse Rate, Cholesterol, Blood Pressure, Heart Rate, etc., and also the risk factors associated with the disease can be found using prediction model with good accuracy and Precision. Further we can include other kinds of chronic diseases, skin diseases and many others. In this work, demonstrating that using only core health parameters many diseases can be predicted. The significance of this analysis is to analyze the maximum diseases to screen the patient's condition and caution the patients ahead of time to diminish mortality proportion. To implement multiple disease analysis used machine learning algorithms, Streamlit. We have considered three diseases for now that are Heart, Liver, and Diabetes and in the future, many more diseases can be added. The user has to enter various parameters of the disease and the system would display the output whether he/she has the disease or not. This project can help a lot of people as one can monitor the persons' condition and take the necessary precautions thus increasing the life expectancy.

Key Words: Diabetes, Heart, Liver, KNN, Random Forest, XG Boost.

1. INTRODUCTION

In this digital world, data is an asset, and enormous data was generated in all the fields. Data in the healthcare industry consists of all the information related to patients. Here a general architecture has been proposed for predicting the disease in the healthcare industry. Many of the existing models are concentrating on one disease per analysis. Like one analysis for diabetes analysis, one for cancer analysis, one for skin diseases like that. There is no common system present that can analyze more than one disease at a time. Thus, we are concentrating on providing immediate and accurate disease predictions to the users about the symptoms they enter along with the disease predicted. So, we are proposing a system which used to predict multiple diseases by using Django. In this system, we are going to analyze Diabetes, Heart, and

malaria disease analysis. Later many more diseases can be included In multiple disease prediction, it is possible to predict more than one disease at a time. So, the user doesn't need to traverse different sites in order to predict the diseases. We are taking three diseases that are Liver, Diabetes, and Heart. As all the three diseases are correlated to each other. To implement multiple disease analyses we are going to use machine learning algorithms and Streamlit. When the user is accessing this API, the user has to send the parameters of the disease along with the disease name. Our Model will invoke the corresponding model and return the status of the patient. Our basic idea is to develop a system which will predict and give the details of the disease predicted along with its severity which as symptoms are given as input by the user. The system will compare the symptoms with the datasets provided in the database. If the symptom matches the datasets, then it should ask other relevant symptoms specifying the name of the symptom. If not, the symptom entered should be notified as the wrong symptom. After this a prompt will come up asking whether you want to still save the symptom in the database. If you click on yes, it will be saved in the database, if not it will go to the recycle bin. The main feature will be the machine learning, in which we will be using algorithms such as Naïve Bayes Algorithm, K-Nearest Algorithm, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine, which will predict accurate disease and also, will find which algorithm gives a faster and efficient result by comparatively-comparing. The importance of this system analysis is that while analyzing the diseases all the parameters which cause the disease are included so it is possible to detect the disease efficiently and more accurately. The final model's behavior will be saved as a python pickle file.

1.1 Description

A lot of analysis over existing systems in the healthcare industry considered only one disease at a time. For example, one system is used to analyze diabetes, another is used to analyze diabetes retinopathy, and another system is used to predict heart disease. Maximum systems focus on a particular disease. When an organization wants to analyze their patient's, health reports then they have to deploy many models. The approach in the existing system is useful to analyze only particular diseases. In multiple disease prediction systems, a user can analyze more than one disease on a single website. The user doesn't need to traverse different places in order to predict whether he/she has a particular disease or not Main objective behind developing a system helps the doctors to cross verify their diagnosed results which gives promising solutions over existing death rates. By using our proposed work try to invent a unique platform and most promising solution for early diagnosis of multiple diseases. Existing work analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease wrong. So, we are giving more accurate solutions by using machine learning and Convolutional neural networks to detect diseases and make predictions.

1.2 Problem System

Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. For example, first is for liver analysis, one for cancer analysis, one for lung diseases like that. If a user wants to predict more than one disease, he/she has to go through different sites. There is no common system where one analysis can perform more than one disease prediction. Some of the models have lower accuracy which can seriously affect patients' health. When an organization wants to analyze their patient's health reports, they have to deploy many models which in turn increases the cost as well as time. Some of the existing systems consider very few parameters which can yield false results.

1.3 Proposed System

In multiple disease prediction, it is possible to predict more than one disease at a time. So, the user doesn't need to traverse different sites in order to predict the diseases. We are taking three diseases that are Liver, Diabetes, and Heart. As all the three diseases are correlated to each other. To implement multiple disease

2. LITERATURE REVIEW

1. According to the paper, diabetes is one of the dangerous diseases in the world , it can cause many varieties of disorders which includes blindness etc. In this paper they have used machine learning techniques to find out diabetes disease as it is easy and flexible to forecast whether the patient has illness or not . Their aim of this analysis was to invent a system that can help the patient to detect the diabetes disease of the patient with accurate results. Here they used mainly 4 main algorithms Decision Tree , Naïve Bayes , and SVM algorithms and compared their accuracy which is 85%,77%, 77.3% respectively . They also used ANN algorithm after the training process to see the reactions of the network which states whether the disease is classified properly or not . Here they compared the precision recall and F1 score support and accuracy of all the models[1] .

2. The main aim of the paper is, as the heart plays an important role in living organisms. So, the diagnosis and prediction of heart related disease should be perfect and correct because it is very crucial which can cause death cases related to heart

.So, Machine learning and Artificial Intelligence supports in predicting any kind of natural events .So in this paper they calculate accuracy of machine learning for predicting heart disease using k-nearest neighbor ,decision tree, linear regression and SVM by using UCI repository dataset for training and testing . They also compared the algorithm and their accuracy SVM 83 %,Decision tree 79%,Linear regression 78%,k-nearest neighbor 87%[2].

3. The system defines that liver diseases are causing a high number of deaths in India and is also considered as a life-threatening disease in the world. As it is difficult to detect liver disease at an early stage .So using automated programs using machine learning algorithms we can detect liver disease accurately .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[3].

3. SYSTEM ANALYSIS

3.1Functional Requirement

- The system allows the patient to predict the disease
- The user adds the input for the particular disease and based on the trained model of the user input the output will be displayed .

3.1 Non-Functional Requirement

- The website will provide a range of the values during the prediction of the disease.
- The website should be reliable and consistent.

4. DESIGN

4.1 Architecture Design

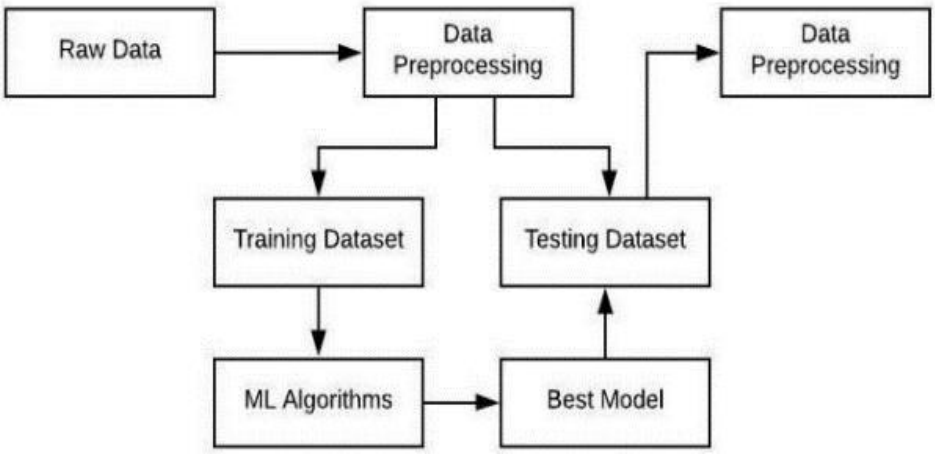

Figure 4.1: Block Diagram

In figure no 4.1 we have experimented on three diseases that are heart diabetes and liver as these are correlated to each other. The first step is to the dataset for heart disease, diabetes disease and liver disease we have imported the UCI dataset, PIMA dataset and Indian liver dataset respectively. Once we have imported the dataset then visualization of each inputted data takes place. After visualization pre-processing of data takes place where we check for outliers, missing values and also scale the dataset then on the updated dataset we split the data into training and testing .Next is on the training dataset we had applied KNN, XGBoost and random forest algorithm and applied knowledge on the classified algorithm using testing dataset. After applying knowledge, we will choose the algorithm with the best accuracy for each of the disease.Then we built a pickle file for all the diseases and then integrated the pickle file with the Django framework for the output of the model on the webpage.

5. IMPLEMENTATION

5.1.1 KNN Algorithm

The working of the K-NN algorithm is as followed:

- Step-1: Start to select the K value for example k=5
- Step-2: Then we will find the Euclidean distance between the points. It is calculated by the as:

$$Euclidean\ Distance = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2}$$

- Step-3: Then we will calculate the Euclidean distance of the nearest neighbor.
- Step-4: Then count the number of the data points in each category .For example, find three values for Category A and two values for category B.
- Step-5: Then assign the new point to the category having the maximum number of neighbors. For example, Category A has the highest number of neighbors so we will assign the new data point to category A.
- Step-6: So finally, our KNN model is ready.

5.1.2. Random Forest Algorithm

Random Forest working is possible in two phases, first is to create the random forest by merging N decision trees, and second is making predictions for each tree created in the first phase.

The working of the random forest is as follows:

Step-1: Firstly, it will select random K data points from the training set.

Step-2: After selecting k data points then building the decision trees associated with the selected data points (Subsets).

Step-3: Then choose the number N for decision trees that you want to build.

Step-4: Repeating steps 1 and 2 .

Step-5: Finding the predictions of each decision tree, and assigning the new data points to the category that wins the majority votes.

5.1.3. XG Boost Algorithm

The working of XG Boost algorithm are as follows:

Step 1: Firstly, create a single leaf tree.

Step 2: Then for the first tree, we must compute the average of the target variable as prediction and then calculate the residuals using the desired loss function and then for subsequent trees the residuals come from prediction that was there in the previous tree.

Step 3: Calculating the similarity score using formula:

$$Similarity\ Score = Gradient \frac{Gradient^2}{Hessian + \lambda}$$

where, Hessian is equal to the number of residuals; Gradient2 = squared sum of residuals; λ is a regularization hyperparameter.

Step 4: Applying a similarity score we select the appropriate node. The higher the similarity score the more homogeneity.

Step 5: Applying similarity scores we calculate Information gain. Information helps to find the difference between old similarity and new similarity and tells how much homogeneity is achieved by splitting the node

at a given point. It is calculated by the formula:

$$Information\ Gain = Left\ Similarity + Right\ Similarity - Similarity\ for\ Roots$$

Step 6: Creating the tree of desired length using the above method pruning and regularization can be done by playing with the regularization hyperparameter.

Step 7: Then we can predict the residual values using the Decision Tree you constructed.

Step 8: The new set of residuals is calculated as:

$$New\ Residuals = Old\ Residulas + \rho \sum Predicted\ Residuals$$

where ρ is the learning rate.

Step 9: Then go back to step 1 and repeat the process for all the trees.

6. RESULT

In the system diabetes disease prediction model used KNN algorithm, heart disease uses the XG Boost algorithm and liver uses the random forest algorithm as these gave the best accuracy accordingly. There when the patient adds the parameter according to the disease it will show whether the patient has a disease or not according to the disease selected. The parameters will show the range of the values needed and if the value is not between the range or is not valid or is empty it will show the warning sign that adds a correct value.

1. User Interface:

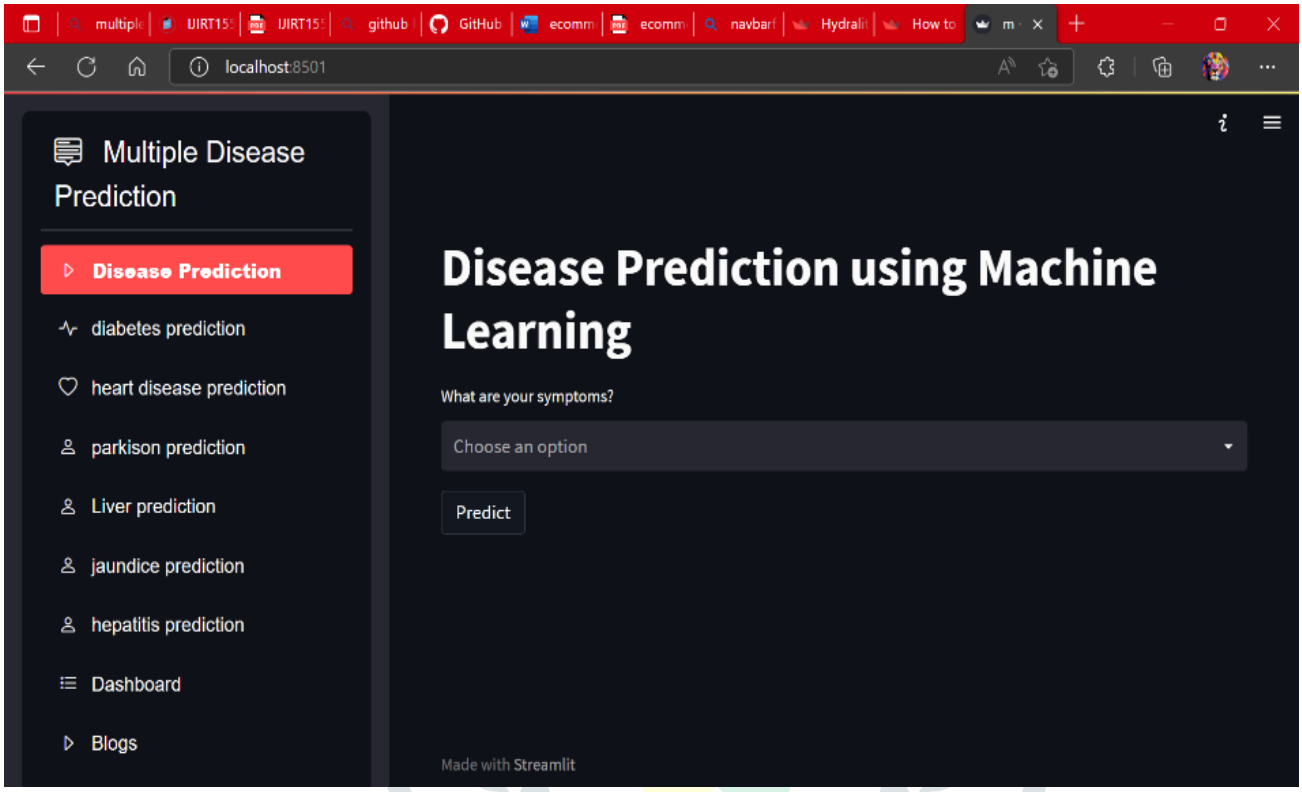

Figure No 6.1: User interface

2. Diabetes Disease:

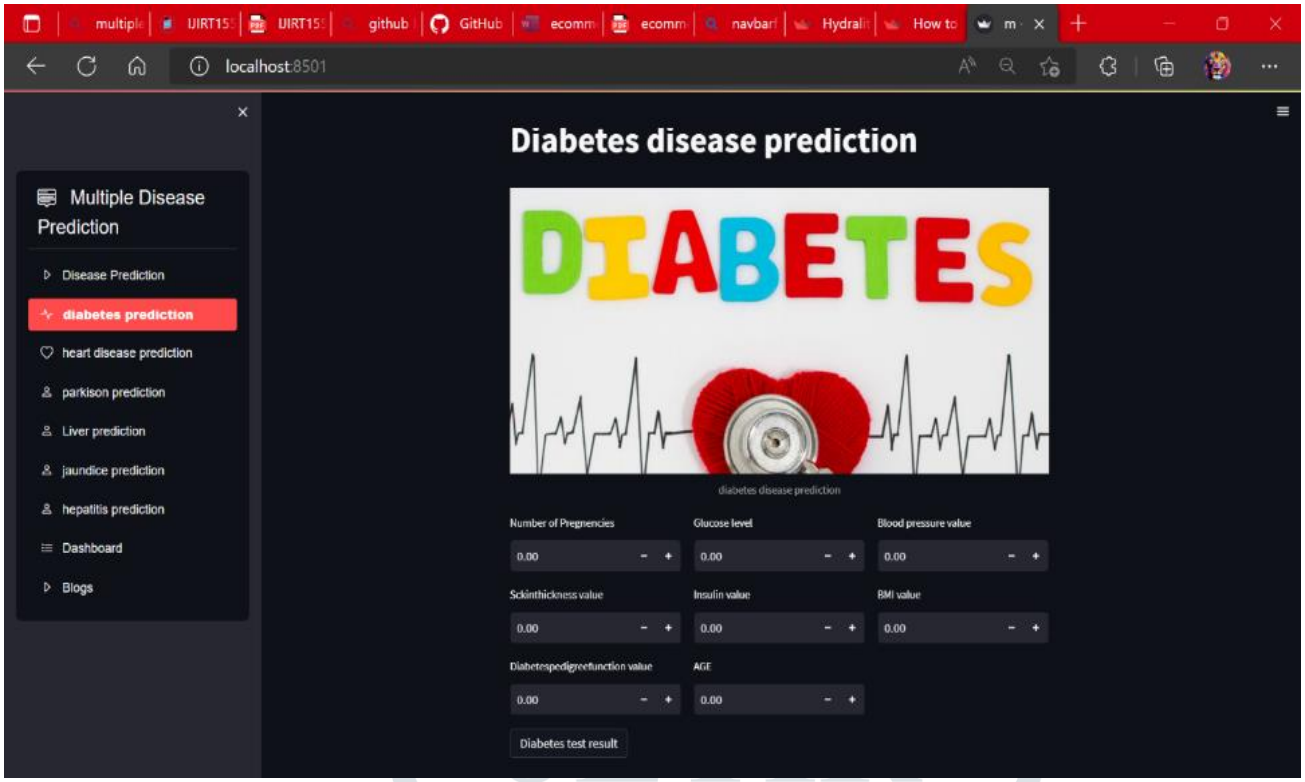

Figure No 6.2: Diabetes Disease Input Data

3. Heart Disease Prediction

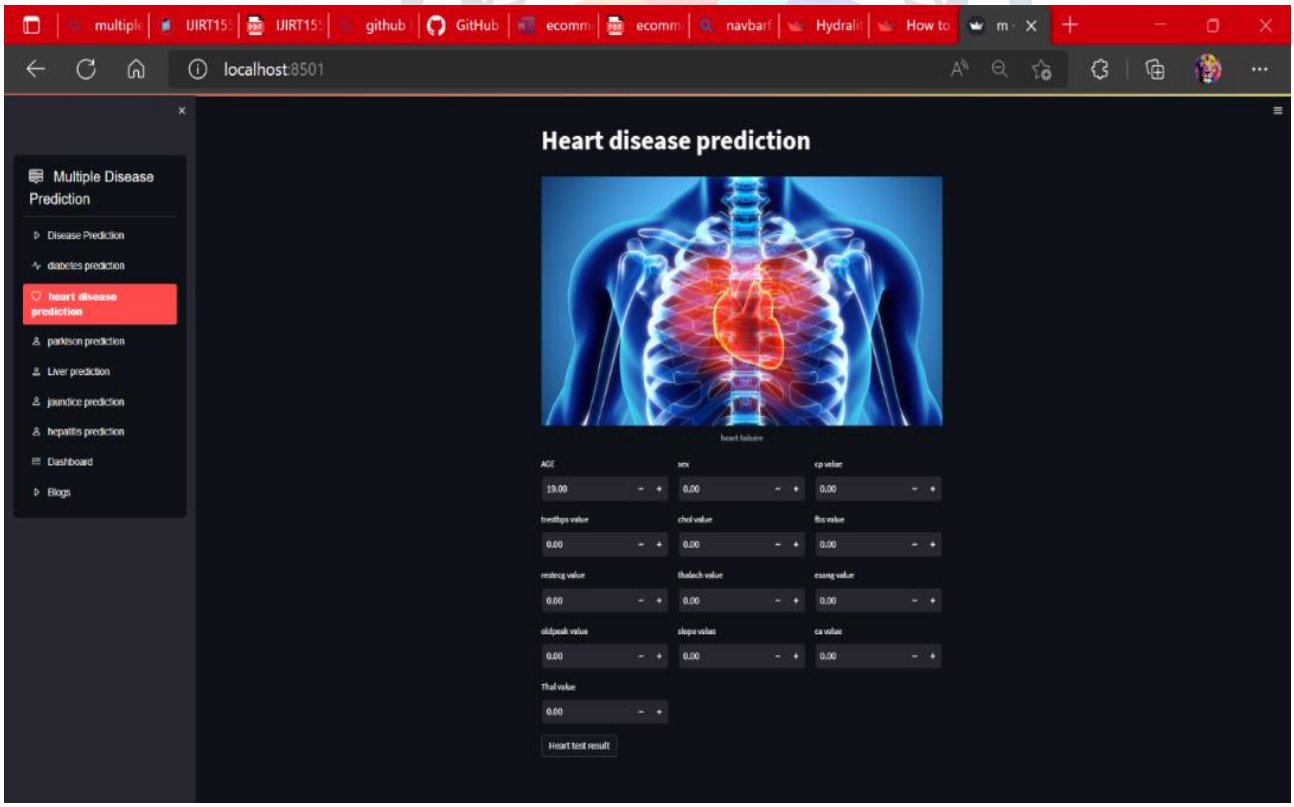

Figure No 6.3: Heart Disease Prediction

4. Parkinson's Prediction

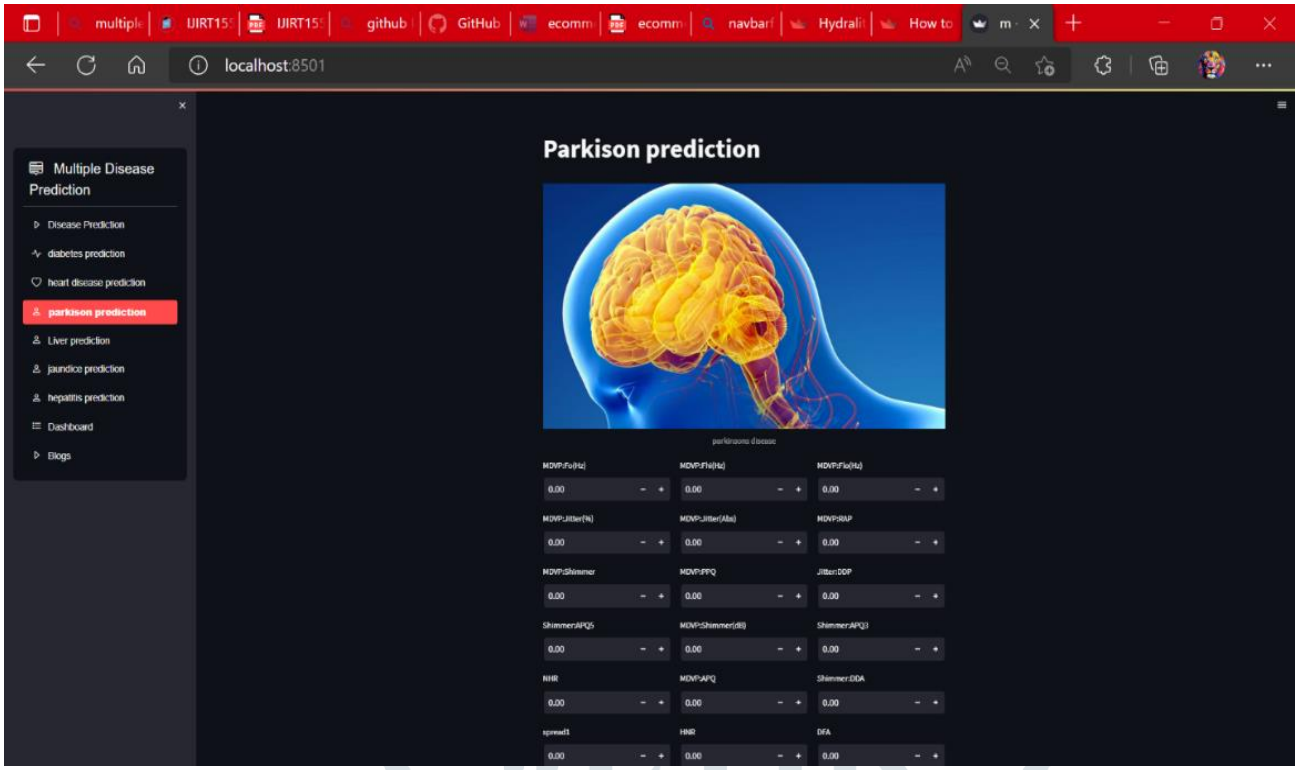

Figure No 6.4: Parkinson's Prediction

5. Liver Disease Prediction

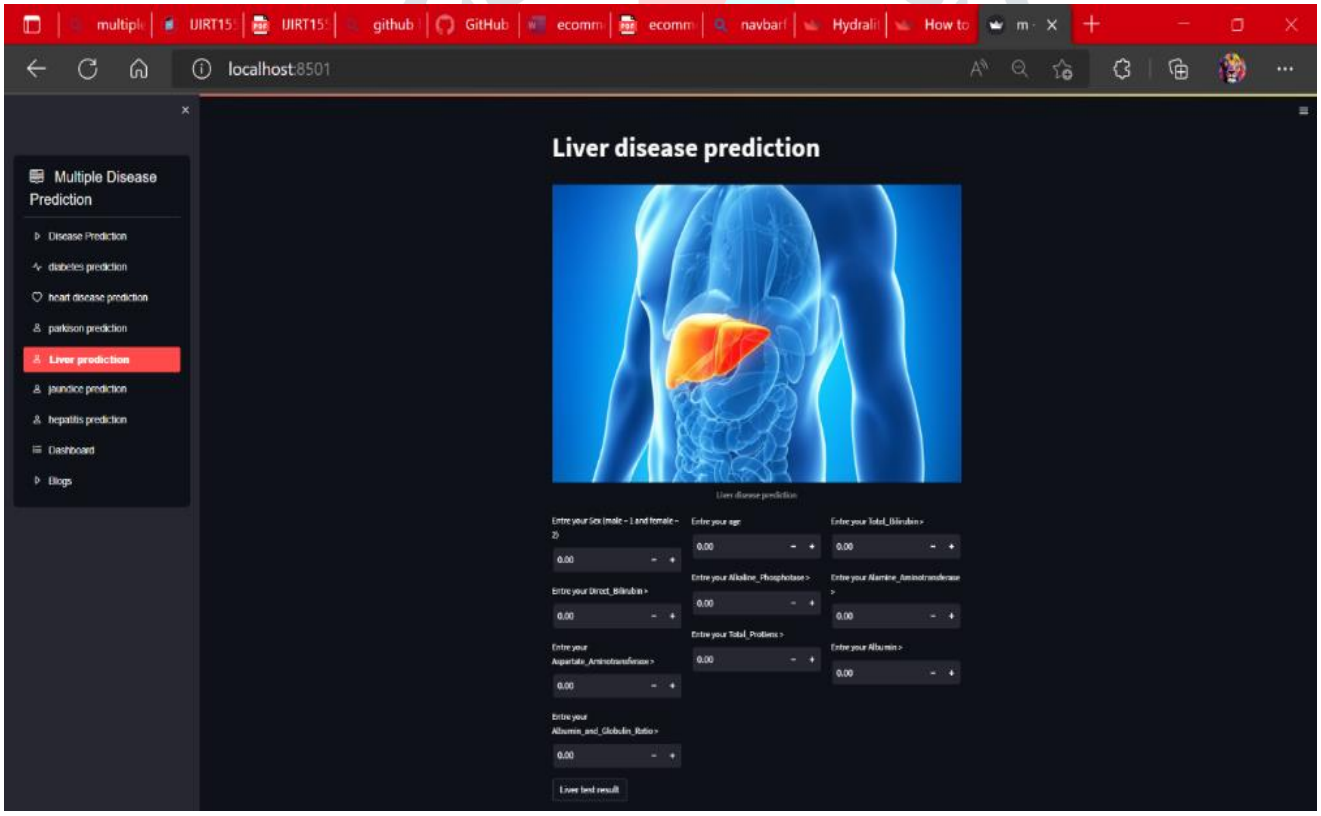

Figure No 6.5: Liver Disease Prediction

6. Jaundice Disease Prediction

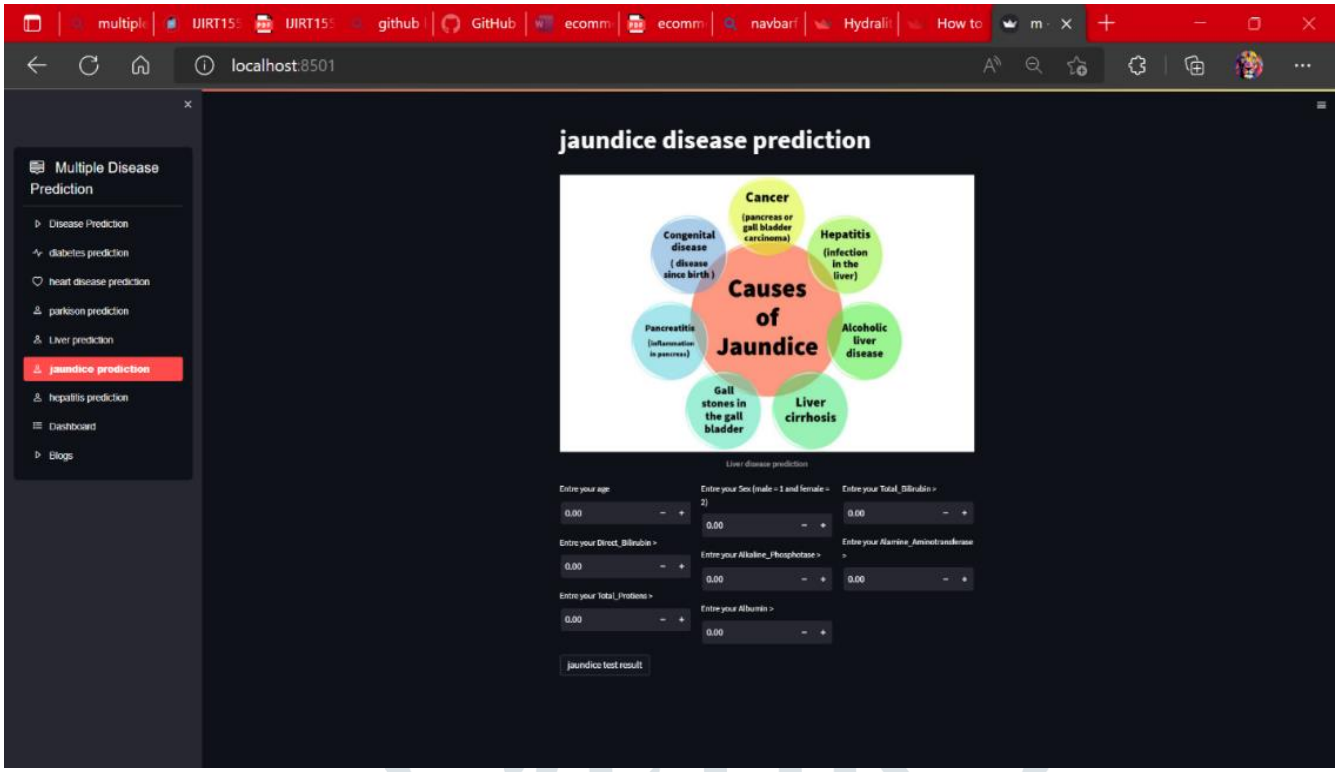

Figure No 6.6: Jaundice Disease Prediction

7. Hepatitis Prediction

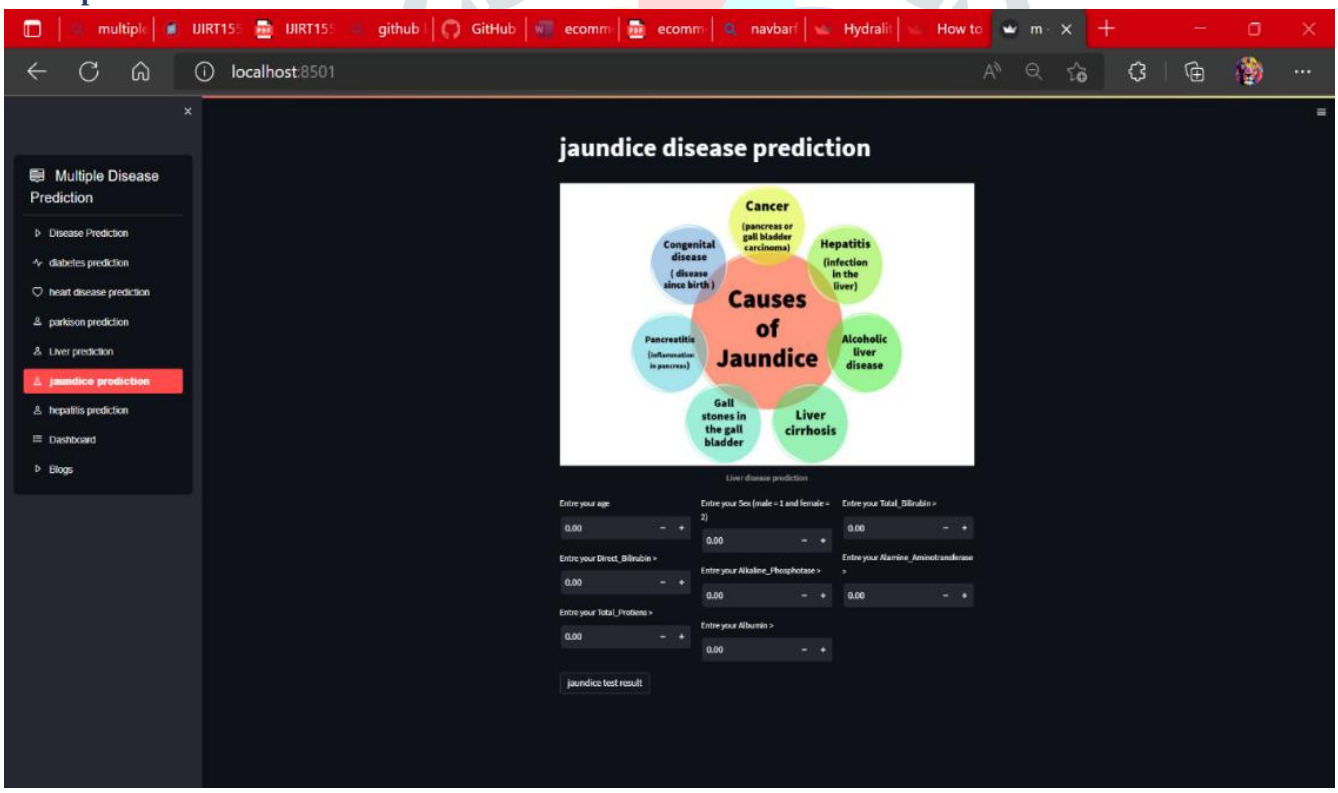

Figure No 6.7: Hepatitis Prediction

7. ACKNOWLEDGEMENT

We sincerely thank our college "**Rizvi College of Engineering"** for giving us a platform to prepare a project on the topic "Multiple Disease Prediction Webapp" and would like to thank our principal **Varsha Shah** for giving us the opportunities and time to conduct and research on the subject. We are sincerely grateful for **Prof. Mohammed Juned** as our guide, for providing help during our research, which would have seemed difficult without their motivation, constant support, and valuable suggestions.

8. 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)