# Multiple Disease Prediction System

1. **Prof. Dr Rafath Samrin**, Department of CSE, Deccan College of Engineering and Technology, Hyderabad, India. rafathsamrin@deccancollege.ac.in Ph: +919959003062
2. **Adnan Aleemuddin**, Department of CSE, Deccan College of Engineering and Technology
3. **Syed Sufiyan Ahmed**, Department of CSE, Deccan College of Engineering and Technology
4. **Mohammed Daris**, Department of CSE, Deccan College of Engineering and Technology

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** Machine learning and Artificial Intelligence are playing a huge role in today's world. From self-driving cars to medical fields, we can find them everywhere. The medical industry generates a huge amount of patient data which can be processed in a lot of ways. So, with the help of machine learning, we have created a Prediction System that can detect multiple diseases using the algorithm with the best accuracy scores. Many of the existing systems can predict only one disease and that too with lower accuracy. Lower accuracy can seriously put a patient's health in danger. We have considered five diseases for now that are Heart disease, Liver disease, Breast Cancer, Kidney disease 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, Breast Cancer, Kidney Knn, Random Forest, XGBoost, SVM, Decision tree.

## 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 using only one algorithm to predict the disease. Like one algorithm for diabetes analysis, one for cancer analysis, one for skin disease etc. There is no common system present that can compare the accuracy of multiple algorithms. 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 Flask. In this system, we are going to analyze Diabetes, Heart, Breast, Kidney and Liver disease analysis. Later many more diseases can be included. To implement multiple disease prediction systems, we are going to use machine learning algorithms, and Flask. Python pickling is used to save the behavior of the model. The importance of this system analysis is that while analyzing the diseases all the parameters which cause the disease is 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 health care industry considered using only one algorithm for a disease at a time. For example, one algorithm is used to analyze diabetes. Maximum systems focus on a particular algorithm per disease. This project is also created for public use, not just for medical professionals. When an organization wants to analyze their patient's, healthreports then they have to deploy many models. The approach in the existing system is useful to analyze only particular disease. In multiple diseases prediction system, a user can analyze more than one disease on a single website. The user does not need to traverse different places in order to predict whether he/she has a particular disease or not. In multiple diseases prediction system, the user needs to select the name of the disease, enter its parameters and just click on submit. The corresponding machine learning model will be invoked and it would predict the output and display iton the screen.

### 1.2 Problem system

Many of the existing machine learning models for health care analysis are concentrating on one algorithm per disease per analysis. For example, logistic regression is used for heart disease, SVM is used for liver disease etc. There is no system which can compare the accuracy of multiple algorithms. 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 turnincreases 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 very accurately. So, the user doesn't need to traverse different sites in order to predict the diseases. We are taking five diseases that are Liver, Diabetes, Breast cancer, Kidney, and Heart. To implement multiple disease analyses we are going to use machine learning algorithms and Flask. When the user is accessing this API, the user has to send the parameters of the disease. Flask will invoke the corresponding model and returns the status of the patient.

## 2. LITERATURE REVIEW

1**."Multiple Disease Prediction using Different ML algorithms comparatively (2019)"(IJARCEE)**
**Authors: Karan A. Jagtap, Smita. S, Prof. Suchita Wankhade and Neha S. Mahamuni**

1. "Prediction of Cardiovascular Disease Using Machine Learning Algorithms" (2018). This paper contributes the correlative application and analysis of distinct machine learning algorithms in the R software which gives an immediate mechanism for the user to use the machine learning algorithms in R software for forecasting the cardiovascular diseases.
2. "A Proposed Model for Lifestyle Disease Predict Vectorion Using Support Machine" (2018). This study aims to understand support vector machine and use it to predict lifestyle diseases that an individual might be susceptible to.
3. "Multi Disease Prediction Using Data Mining Techniques" (2017). In this study two different data mining classification techniques was used for the prediction of various diseases and their performance was compared in order to evaluate the best classifier. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications.
4. "Prediction of Heart Disease Using Machine Learning Algorithms" (2018). In this paper, two supervised data mining algorithm was applied on the dataset to predict the possibilities of having heart disease of a patient, were analyzed with classification model namely Naïve Bayes Classifier and Decision tree classification. The Decision tree model has predicted the heart disease patient with an accuracy level of 91% and Naïve Bayes classifier has predicted heart disease patient with an accuracy level of 87%.
5. "Review of Medical Disease symptoms Prediction using Data Mining Technique" (2017). In this paper evaluate the performance of medical disease based on data mining technique. The classifier classified the medical diagnosis of disease data such as cancer, liver problem, heart disease and so on. SVM better classified.

## 3. SYSTEM ANALYSIS

### 3.1Functional Requirement

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

### 3.2 Non-Functional Requirement

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

## 4.DESIGN

### 4.1Architecture Design

-

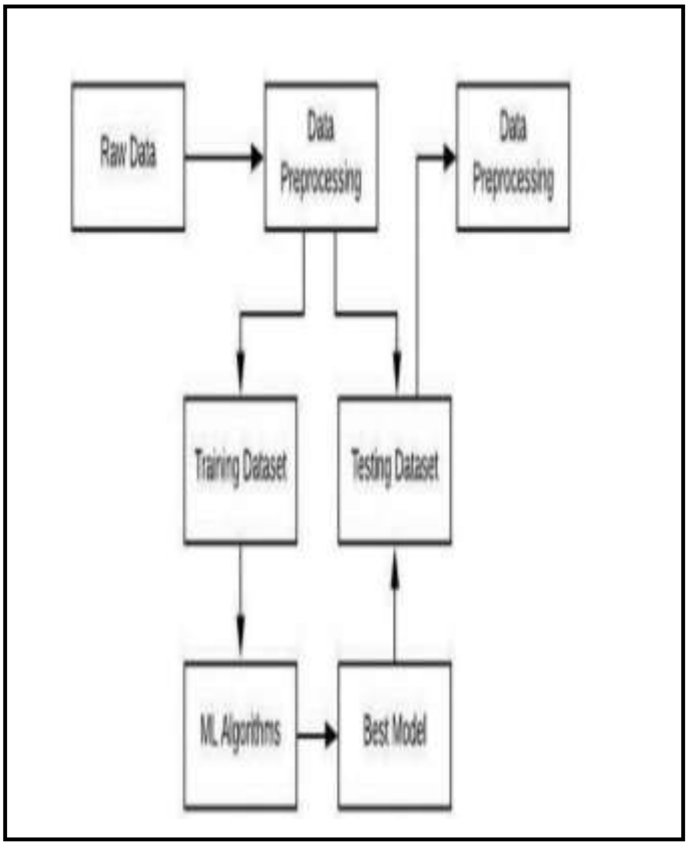
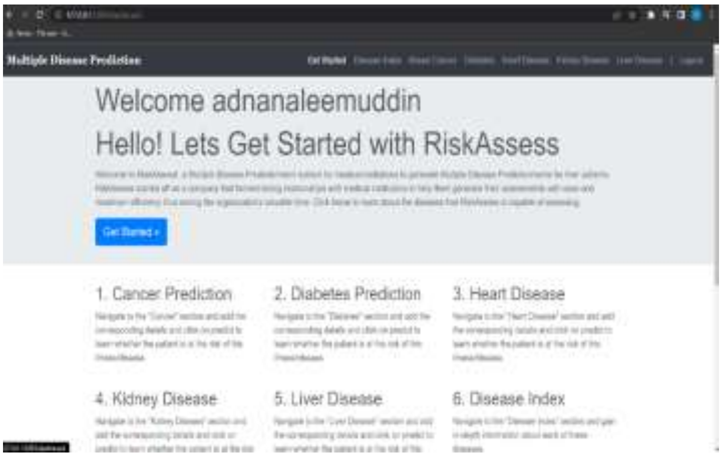


**Figure No4.1: Block Diagram**

In the figure no 4.1 we have experimented on three diseases that is 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 imputed data takes place. After visualization pre-processing of data takes place where we check for outliers, missing values and 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 build a pickle file for all the disease and then integrated the pickle file with the flask framework for the output of the model on the webpage.

### 4.2User Interface Design



Welcome adnanaleemuddin

Hello! Lets Get Started with RiskAssess

1. Cancer Prediction

2. Diabetes Prediction

3. Heart Disease

4. Kidney Disease

5. Liver Disease

6. Disease Index

**Figure No4.2.1: Graphical User Interface**

LOGIN

**Figure No 4.2.2: Login Page**

SIGNUP

Username

Email

Password

**Figure No 4.2.3: Sign up Page**

## 5.IMPLEMENTATION

### 5.1Algorithm

### 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:

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

- 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, found three values for Category A and two values for category B.
- Step-5: Then assign the new point to the category having maximum number of neighbors. For example, Category A has highest number of neighbor so we will assign the new data pointto 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 tree, and second is making prediction 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 choosing the number N for decision trees that you want to build.

**Step-4:** Repeating step 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. XGBoost Algorithm

The working of XGBoost algorithm are as follows:

Step 1: Firstly, creating a single leaf tree.

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

Step 3: Calculating the similarity score using formula:

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

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

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

Step 5: Applying similarity score we calculate Information gain. Information gain help 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:

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

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:

$$\mathit{New\ Residuals} = \mathit{Old\ Residulas} + \rho \sum \mathit{Predicted\ Residuals}$$

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 XGBOOST 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 add a correct value.

**ACCURACY FOR EACH DISEASE:**

**Table No 6.1:Diabetes Disease**

| **ALGORITHM** | **Diabetes** |
|---|---|
| Random Forest | 88% |
| XGBOOST | 89% |

**Table No 6.2:Heart Disease**

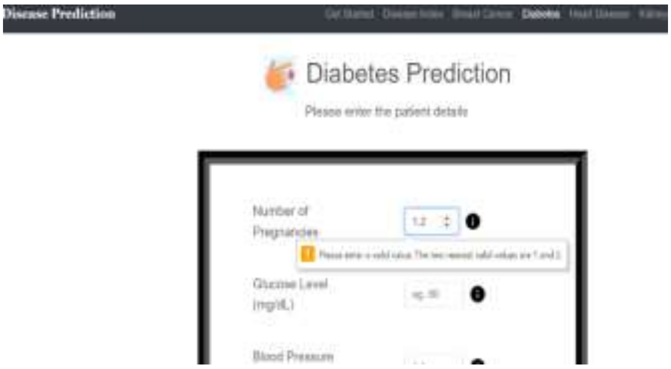
| **ALGORITHM** | **Heart(MEAN ACCURACY)** |
|---|---|
| KNN | 0.971386 |
| SVM | 0.971386 |

**Table No 6.3:Liver Disease**

| **ALGORITHM** | **Liver** |
|---|---|
| Random Forest | 73% |
| XGBOOST | 68% |

**1.Error Message on inputting the incorrect value:**

**Figure No 6.1: Error input**



Diabetes Prediction

Please enter the patient details

Number of Pregnancies: 1.2

! Please enter a valid value. The two nearest valid values are 1 and 2.

Glucose Level (mg/dL): eg. 80

Blood Pressure

## 2.Diabetes Disease :

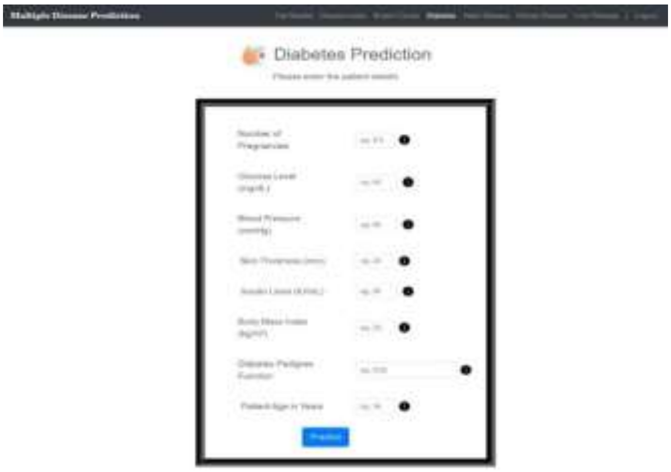
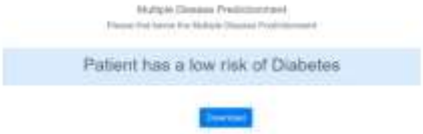


Figure No 6.2:Diabetes Disease Input Data


Multiple Disease Predictionment
Please find below the Multiple Disease Predictionment
Download

Patient has a low risk of Diabetes

Figure No 6.3: Diabetes Disease Output Result

## 3.Heart disease:

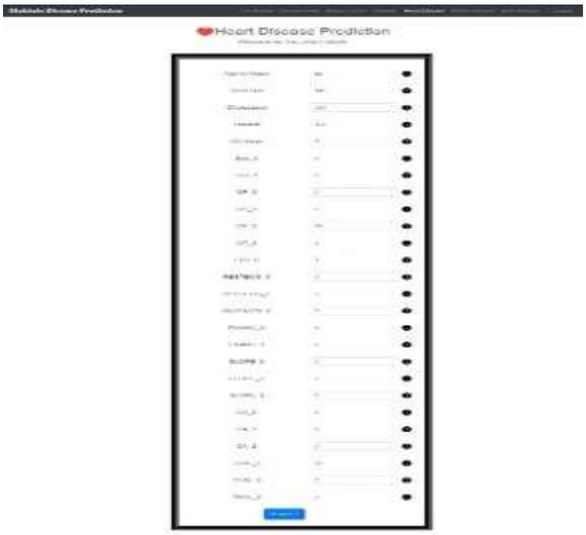
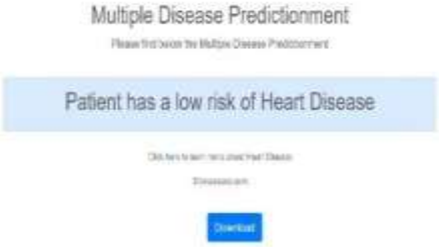


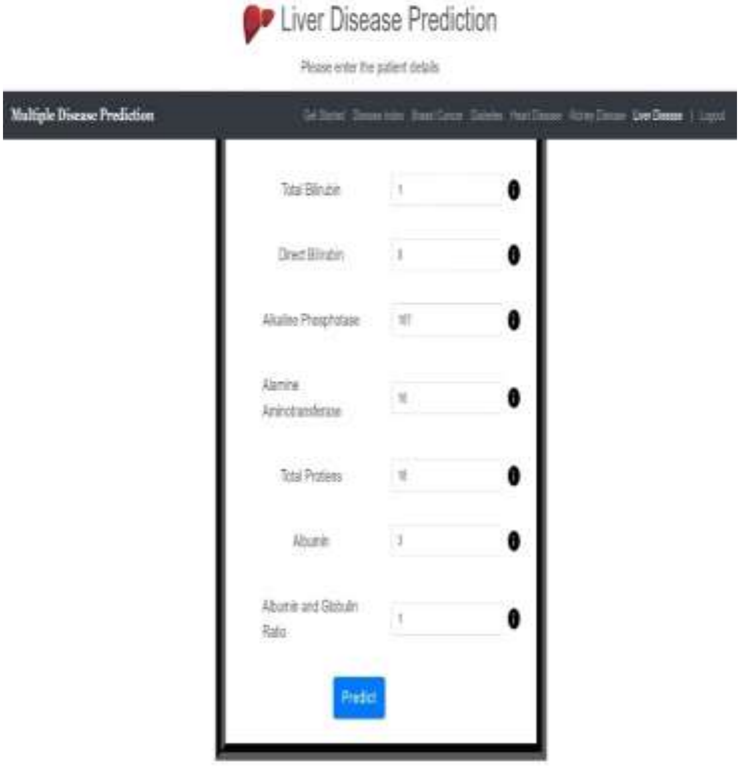
Figure No 6.4: Heart Disease Input Data


Multiple Disease Predictionment
Please find below the Multiple Disease Predictionment
Download

Patient has a low risk of Heart Disease

Figure No 6.5:Heart Disease Output Result

## 4.Liver Disease:



Total Bilirubin 1

Direct Bilirubin 1

Alkaline Phosphotase 187

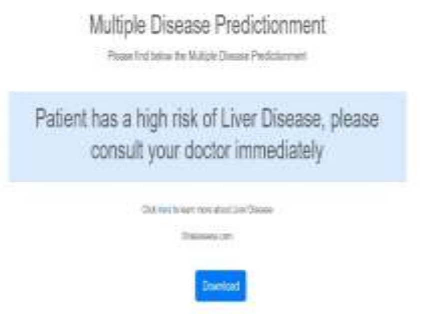
Alamine Aminotransferase 16

Total Protiens 18

Albumin 3

Albumin and Globulin Ratio 1

Figure No 6.6: Liver Disease Input Data



Multiple Disease Predictionment

Please find below the Multiple Disease Predictionment

Patient has a high risk of Liver Disease, please consult your doctor immediately

Click here to learn more about Liver Disease

[illegible]

**Figure No 6.7: Liver Disease Output Result**

## 7.CONCLUSION

The main objective of this project was to create a system that would predict more than one disease and do so with high accuracy. Because of this project the user doesn't need to traverse different websites which saves time as well. Diseases if predicted early can increase your life expectancy as well as save you from financial troubles.

For this purpose, we have used various machine learning algorithms like SVM, KNN etc.

## 8.FUTURE SCOPE

- In the future we can add more diseases in the existing API.
- We can try to improve the accuracy of prediction in order to decrease the mortality rate.
- Try to make the system user friendly.

## ACKNOWLEDGEMENT

We sincerely thank our college **Deccan College of Engineering** for giving us a platform to prepare a project on the topic "Multiple Disease Prediction System" and wouldlike to thank our HOD **Dr. Syed Razi Uddin** for giving us the opportunities and time to conduct and research on the subject. We are sincerely grateful for **Dr. Rafath Samrin** as our guide and for providing help during our research, which would have seemed difficult without their motivation, constant support, and valuable suggestion. Professor Rafath Samrin can be reached at rafathsamrin@deccancollege.ac.in or **+91-9959003062.** Moreover, the complication of this research paper would have been impossible without theco-operation, suggestion, and help of our friends and family.

## 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)