Grenze International Journal of Engineering and Technology, Jan Issue

A GUI Based Approach to Predict Multiple Diseases using Machine Learning Algorithms

Tiasha Dasgupta[1], Arunava Das[2], Sagnik Chakraborty[3] and Lopamudra Dey[4]
[1-4]Heritage Institute of Technology/ Computer Science Engineering, Kolkata, India
Email: tiasha.dasgupta.cse23@heritageit.edu.in, arunava.das.cse23@heritageit.edu.in, sagnik.chakraborty.cse23@heritageit.edu.in, lopamudra.dey@heritageit.edu

***Abstract*— "Multiple Disease Prediction" is a system based on predictive modeling that will predict the user's disease based on the given symptoms. We have considered four diseases for now, that is Heart disease, Liver disease, Chronic Kidney Disease, and Diabetes. There are existing systems that can predict one disease at a time and perhaps with lower accuracy, which puts a patient's health in danger. In this project, we aim to design a multiple disease prediction system by applying principles of machine learning, where our application incorporates two ways of predicting diseases for two different categories of users. For single users, the user must enter various parameters of the disease, and the system would display the output whether she/he has the disease or not. The project makes it easier for a multitude of users, such as a hospital management system, to anticipate diseases from the results of laboratory blood test reports. Blood analysis is a crucial disease indicator since it comprises several factors that are symptoms of diseases. This project not only emphasizes prediction precision but also opens the possibility of easily introducing new diseases.**

***Index Terms*— Chronic Kidney, Diabetes, Heart disease, Liver, Random Forest Classifier, Catboost**

I. INTRODUCTION

A. Background

Predictive analytics in healthcare is a difficult endeavour, but it can eventually assist practitioners in making timely decisions regarding patients' health and treatment based on massive data. Data in the healthcare industry consists of all the information related to patients. Diseases like diabetes, liver diseases, and heart-related diseases are causing many deaths globally, but most of these deaths are due to the lack of timely check-ups for the diseases. The above problem occurs due to a lack of medical infrastructure and a low ratio of doctors as compared to the population. In support of the statement,the statistical recommendation of the WHO can be stated, which says that the ratio of doctors to patients is 1:1,000. However, India's doctor-to-population ratio is 1:1456, which distinctly indicates the shortage of doctors. The diseases related to the heart, liver, and diabetes can cause a potential threat to mankind if not diagnosed early. Therefore, early recognition and diagnosis of these diseases can save a lot of lives. This work is all about the first step towards predicting such diseases using machine learning classification algorithms

Grenze ID: 01.GIJET.10.1.90

B. Problem Statement

Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. For example, the first is for liver analysis, the second is for diabetes analysis, the third is for heart diseases, and so on. 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 may have low accuracy, which can seriously affect patients' health. When an organization wants to analyze their patients' health reports, they have to deploy many models, which in turn increases the cost as well as time.

The ever-increasing risk factors in the environment and lack of awareness among people are making humans more prone to diseases. It is becoming imperative for everyone to understand the conditions that can lead to this particular health hazard.

C. Objective

To make the work seamless and easy, a multiple disease prediction web application is developed that will make predictions about various diseases using the concepts of machine learning. It is made publicly accessible so that users can get a broad analysis of whether they have a particular disease and take action accordingly. Here, a user can analyze more than one disease on a single website without the hassle of going through different links. He/She just selects the name of the particular disease, enters its parameters which invokes the corresponding machine learning model, and the predicted output is displayed on the screen. Additionally, it can also predict diseases by uploading a patient's health report.

D. Process Workflow

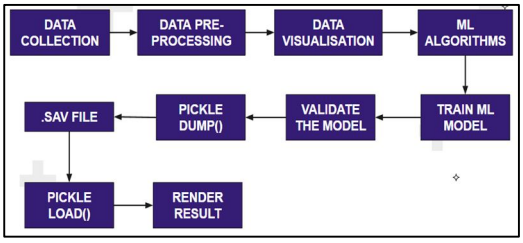


Fig 1. Process Flow

As per the chart given above, the entire project works according to the flow of the flowchart. Data Collection: Details of this section is mentioned in III A section. Data Pre-Processing: Details of this section is mentioned in III B section. ML Algorithms: Details of this section is mentioned in III C section. Validate The Model: Details of this section is mentioned in III D section. Pickle Dump: Details of this section is mentioned in III E section. Creation of .sav File: Details of this section is mentioned in III E section. Pickle Load: Details of this section is mentioned in III E section. Render Result: Details of this section is mentioned in IV section.

II. Literature Survey

The analysis of earlier models that were put forth to forecast the diseases connected to our proposed endeavor is covered in this section. [1] Sonar, P., & JayaMalini, K proposed their work "Diabetes prediction using different machine learning approaches" where they wanted to develop a system that would accurately diagnose a patient's diabetes in order to assist them in managing their condition. Here, they compared the accuracy of four primary algorithms: Decision Tree, Naïve Bayes, and SVM, which were utilized in this study. The accuracy of these algorithms was 85%, 77%, and 77.3%, respectively. [2] The primary goal of the work "Heart disease prediction using machine learning algorithms" proposed by Singh, A., & Kumar, R. is to demonstrate the significance of the heart in living things. The heart-related disease must be accurately diagnosed and predicted because failing so can result in fatalities. They used the UCI repository dataset for training and testing in this research to calculate the accuracy of machine learning for predicting heart disease using k-nearest neighbor, decision tree, linear regression, and SVM. Additionally, they compared the accuracy of the algorithms SVM (83%), Decision Tree (79%), Linear Regression (78%), and K-Nearest Neighbour (87%). [3] In the work "Diagnosis of liver disease using machine learning models" presented by Sivasangari, A., Reddy, B. J. K., Kiran, A., & Ajitha, P., with the use of machine learning algorithms, liver disease can be correctly recognized. For quantitative evaluation, they used and contrasted the SVM, Decision Tree, and Random Forest algorithms, measuring precision, accuracy, and recall metrics. The accuracy is 95%,87%, and 92% respectively. In the model proposed by [13] uses different data mining and the classification algorithms for the prediction of disease. This model is mainly used for the prediction of the Heart

Disease and the algorithms which are used in this model are Decision Tree and the Naïve Bayes algorithm which are used for the prediction of the Disease and various data mining techniques are also used in this model for correcting and balancing the dataset so that the system can work correctly and can predict the correct diseases. In the model proposed by [12] showed important ML approaches to predict the disease but this model which was proposed by [12] works on the K-Nearest Neighbour (KNN)and Convolution Neural Network (CNN) approach of the machine learning algorithm.Both the KNN and CNN approaches are used in this system which is different from the approach which is used in our project. The CNN uses both the structures as well as the unstructured data for the prediction of the disease which makes it more time consuming.The accuracy of the system proposed by [12] comes out to be very high i.e. above 95% for the KNN algorithm and 100% for the CNN algorithm that is very high for a ML model,In such cases the model is said to be overfitting. The SVM method and the Naïve Bayes technique were used to predict kidney disease [11]. The authors attempted to categorize various stages of kidney disease using the suggested ANFIS algorithm. The study's purpose was to design an effective categorization algorithm using several assessment metrics such as accuracy and execution time. While the SVM Algorithm provided higher classification accuracy, the Naïve Bayes fared better since it produced results in less time. The results show that SVM outperforms the Naïve Bayes Approach in predicting renal illness.

III. Proposed Work And Procedure

Existing System - In the past, there have been AI models for medical examinations that are focused on a single illness for each inquiry. One analysis, for example, might be sufficient to investigate diabetes, liver, or a kidney problem of that nature. There isn't a usual framework where one investigation can carry out more than one disease assumption. This process is time-taking and costly. Additionally, if a user has multiple diseases but the system in place can only anticipate one of them, there is a possibility that the death rate may rise as a result of the inability to foresee the other diseases.

Proposed System - It is feasible to anticipate multiple diseases at once using our GUI application. The user won't have to navigate through numerous models for detecting diseases. Time will be cut short, and there is a chance that the death rate could go down because several diseases can be predicted at once. Additionally, our application makes it easier to predict several diseases using a single dataset, such as a blood sample report. It is helpful when there are numerous users, for instance in a hospital management system.

A. Data Preparation

We have made use of four datasets that are all obtained from a publicly accessible website (https://www.kaggle.com/). A sample blood test report is obtained from Google that is published by the National Health and Nutrition Examination Survey (2011 - 2012 Data Documentation, Codebook, and Frequencies).
The four datasets are as follows:

- Heart Disease Dataset
- Liver Disease Dataset
- Diabetes Dataset
- Chronic Kidney Disease Dataset

B. Data Preprocessing

Data pre-processing is carried out to clean the data so as to use it in an efficient format. Most of the data contains redundant, duplicate, or missing values. So, if such data is used to construct a model, it provides incorrect results. Hence, data must be cleaned before being fed to a model in order to get accurate results.

- *Missing/NaN values***:** In the Diabetes dataset, attributes like Glucose, BloodPressure, SkinThickness, Insulin, and BMI consist of a value of 0. But in real life, these attributes cannot have a zero value. So, we replace all the 0 values with NaN using the replace function. Then we replace the Nan values with the median values of the corresponding column. A similar missing value check is done in the case of the other disease predictions.
- *Correlation Matrix:* In order to comprehend the correlation between more than just one pair of variables, a correlation matrix is mapped for each of the four datasets. Values can vary between -1 to 1 where, a value of 1 denotes a completely negative linear correlation between two variables, 0., is the linear correlation between two variables and a value of 1 denotes a perfect positive linear correlation between two variables.
- *Outlier check:* A data item or object that considerably differs from the other objects is called an outlier. We use the boxplot representation to check for outliers in the DataFrame of all three datasets. We implement the InterQuartile Range (IQR) Method to remove the detected outliers in all three datasets.
- *Imbalanced dataset:* In the case of Diabetes and Liver disease prediction, the target feature in both the datasets can be seen to be imbalanced since there is huge parity between the number of 0s and 1s. As a result, the

model will be biased towards the majority class only. In order to handle the imbalanced dataset, we have made use of SMOTE (Synthetic Minority Oversampling Technique).

C. Classification Algorithms

We have made use of the following algorithms to predict results,

- Random Forest Classifier
- LightGBM Classifier
- CatBoost Classifier

D. Cross Validation

Cross validation is a method for checking overfitting. It is done by dividing the dataset into K folds. During the first iteration, first fold is used for testing purpose and rest K-1 folds are employed for training purpose.Similarly, for second iteration, the second fold is used for testing purpose and the remaining K-1 folds are used for training purpose. In this manner, the process is repeated for K iterations.

E. Developing the Client-end Interface

Single User System - The user interface was developed using Streamlit, which is an open-source python-based framework for developing interactive data science dashboards and machine learning models. Machine learning models take a considerable amount of time to train and predict, especially on large datasets with many features. This would have slowed down rendering the data to the front-end since we are dealing with 4 different datasets and models. To work around this problem, we incorporated Pickle to the back-end, which is a Python tool to save ML models thus minimizing lengthy re-training. All we have to do to save the ML model using Pickle is give the model object into the dump() function. As a result, the object will be serialised and transformed into a "byte stream" which is stored as a.sav file. In order to deserialize the model and use it to generate fresh predictions, we later loaded this file using the Pickle load() function.
Multiple User System - We make use of the file_uploader function in Streamlit, which allows us to upload system files into Streamlit. The file then undergoes checking, and column-wise matching takes place to predict the four diseases that we have considered. Additionally, the GUI also facilitates the download of the final result in the form of a CSV file, where each row corresponds to a patient and each column to a disease(can consist of a maximum of four columns corresponding to the four diseases taken into account).

IV. Results And Discussion

The algorithms used in this work are Random Forest Classifier, LightGBM, CatBoost, and the accuracies are calculated using the cross-validation with factor cv as 10. The datasets of diseases are divided into training and test datasets for classification.

TABLE I: COMPARISON TABLE

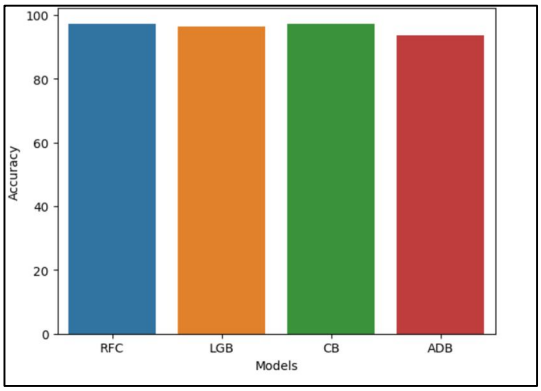
AUTHORS	TECHNIQUE USED	DISEASE	ACCURACY
Purushottam, Kanak Saxena Prof. (Dr.), Richa Sharma	Support Vector Machine	Heart disease	70.59%
K.Ramesh,G.Subbalakshmi and M. Chinna Rao	Naive Bayes	Heart disease	52.33%
V.Rajesh, A.Vamsi Akki Reddy,Naveen Kishore G K.Sumedh, T.Rajesh Sai Reddy.	Random Forest Classifier	Diabetes	74.4%
Deepti Sisodia, Dilip Singh Sisoda	Decision Tree	Diabetes	73.82%
Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, Ankit Chatorikar	Decision Tree	Chronic Kidney Disease	91.75%
Rakshith D B, Mrigank Srivastava, Ashwani Kumar, Gururaj S P	Naive Bayes	Liver Disease	55.56%

The accuracies mentioned in the Table1 have been surpassed by our work. The highest accuracies achieved by our work for Diabetes, Heart, Liver and Chronic Kidney are as follows 95.79%, 94.63%, 92.30%, and 96.25% respectively using LightGBM for diabetes, Random Forest Classifier for Heart disease and CatBoost for Liver and Chronic Kidney Disease.

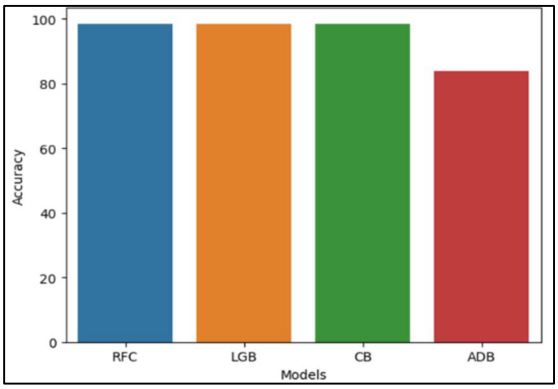
TABLE II: ACCURACIES OF DIFFERENT ALGORITHMS

Disease	RFC	LGBM	CB	Best Accuracy
Diabetes	94.95%	95.79%	95.79%	95.79%
Heart disease	94.63%	91.53%	92.40%	94.63%
Liver disease	92.30%	92.30%	92.30%	92.30%
Chronic Kidney Disease	95.00%	95.00%	96.25%	96.25%

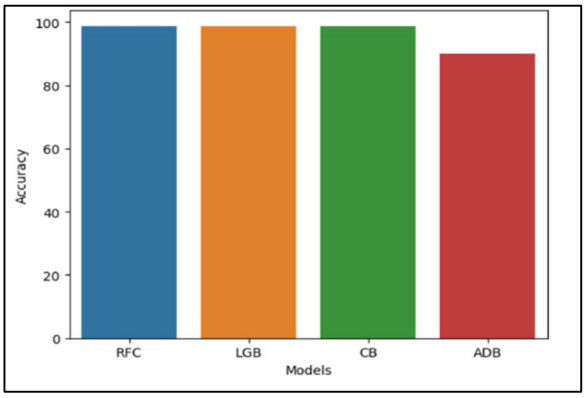
RFC – Random Forest Classifier ; LGBM – Light Gradient Boosting Machine ; CB – CatBoost Classifier



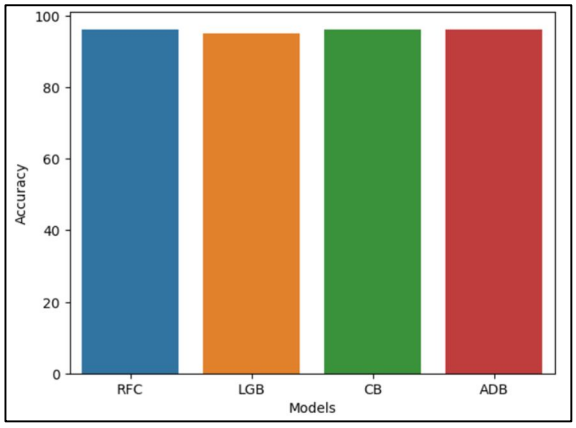
Diabetes



Heart



Liver



Kidney

TABLE III: PRECISION, RECALL, F1 SCORE OF DIFFERENT DISEASES

Disease	Precision = TP/TP+FP	Recall = TP/TP + FN	F1 score = 2*Precision*Recall/ Precision + Recall
Diabetes	0.93	0.96	0.94
Heart	0.92	0.97	0.95
Liver	1	0.97	0.98
Kidney	1	0.94	0.97

TP– True Positive; FP – False Positive; FN – False Negative

V. CONCLUSION

The main objective of this project is 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 our life expectancy as well as save us from financial troubles. Users can be aware about their health conditions and induce lifestyle correction for a healthier future. For this purpose, we have used various machine learning algorithms like Random Forest, XGBoost, and CatBoost to achieve maximum accuracy.
Our entire work is available here: https://github.com/SagnikChak/Multiple-Disease-Detection-System.git

FUTURE PROSPECT

- In future we can add more diseases to the existing GUI.
- We can try to improve the accuracy of prediction in order to decrease the mortality rate.
- Make the system as user-friendly as possible and offer a chatbot for common questions.
- Create a login/signup page to prevent unauthorized access to patient's data.

REFERENCES

[1] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371). IEEE.
[2] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.
[3] Sivasangari, A., Reddy, B. J. K., Kiran, A., & Ajitha, P. (2020, October). Diagnosis of liver disease using machine learning models. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 627-630). IEEE.
[4] Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. Procedia Computer Science, 85, 962-969.
[5] Subbalakshmi, G., Ramesh, K., & Rao, M. C. (2011). Decision support in heart disease prediction system using naive bayes. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), 170-176.
[6] Kishore, G. N., Rajesh, V., Reddy, A. V. A., Sumedh, K., & Reddy, T. R. S. (2020). Prediction of diabetes using machine learning classification algorithms. International Journal of Scientific & Technology Research, 9(1), 1805-1808.
[7] Mohit, I., Kumar, K. S., Reddy, U. A. K., & Kumar, B. S. (2021, November). An Approach to detect multiple diseases using machine learning algorithm. In Journal of Physics: Conference Series (Vol. 2089, No. 1, p. 012009). IOP Publishing.

[8] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.
[9] Tekale, S., Shingavi, P., Wandhekar, S., & Chatorikar, A. (2018). Prediction of chronic kidney disease using machine learning algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 7(10), 92-96.
[10] Rani, A. J. M., Nishanthini, S., Josephine, D. J., Venugopal, H., Nissi, S. G., & Jacintha, V. (2022, October). Liver Disease Prediction using Semi Supervised based Machine Learning Algorithm. In 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC) (pp. 1389-1392). IEEE.
[11] Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE transactions on information technology in biomedicine, 14(4), 1114-1120.
[12] Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.
[13] https://www.researchgate.net/publication/333925535_Prediction_of_Heart_Disease_Using_Machine_Learning_Algorithms
[14] VENKATESH SATHYA HARISYAM, “Liver Disease Prediction”https://www.kaggle.com/code/harisyammnv/liver-disease-prediction/data
[15] DAVID LAPP, “Heart Disease Dataset”https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
[16] MEHMET AKTURK, “Diabetes Dataset”https://www.kaggle.com/datasets/mathchi/diabetes-data-set
[17] MANSOOR IQBAL, “Chronic Kidney Disease Dataset” https://www.kaggle.com/datasets/mansoordaku/ckdisease

GUI SCREENSHOTS

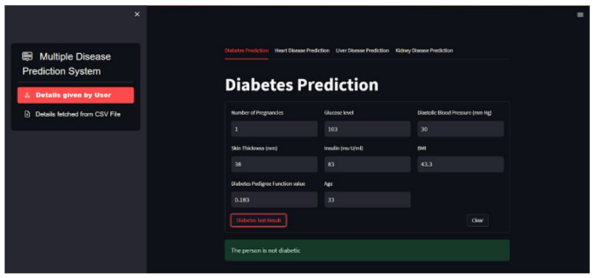


Fig 2. Landing page with default choice set as Diabetes Prediction

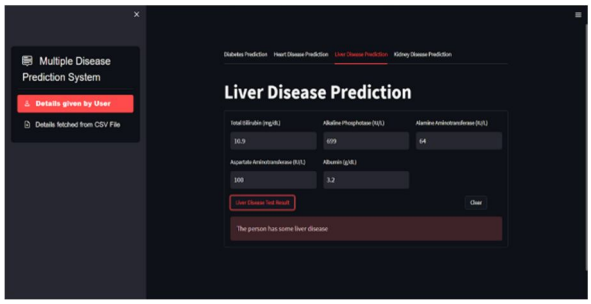


Fig 3. Liver Disease Prediction with a random sample input

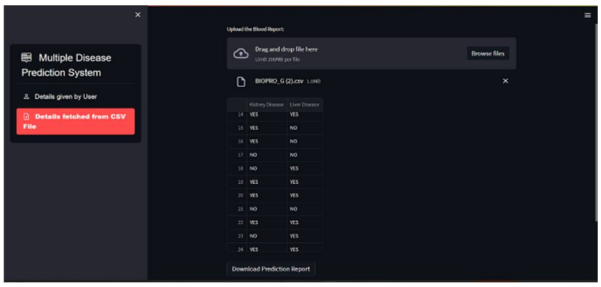


Fig 4. Multiple User System