Proceedings of the 5th International Conference on Smart Systems and Inventive Technology (ICSSIT 2023)
IEEE Xplore Part Number: CFP23P17-ART; ISBN: 978-1-6654-7467-2

Multiple Disease Prediction System using Machine Learning and Streamlit

Laxmi Deepthi Gopisetti
Computer Science & Engineering
VNRVJIET
Hyderabad, India
laxmideepthi_g@vnrvjiet.in

Srinivas Karthik Lambavai Kummera
Computer Science & Engineering
VNRVJIET
Hyderabad, India
srinivaskarthik33@gmail.com

Sai Rohan Pattamsetti
Computer Science & Engineering
VNRVJIET
Hyderabad, India
psairohan21@gmail.com

Sneha Kuna
Computer Science & Engineering
VNRVJIET
Hyderabad, India
kunasneha00@gmail.com

Niharika Parsi
Computer Science & Engineering
VNRVJIET
Hyderabad, India
nihaparsi12@gmail.com

Hari Priya Kodali
Computer Science & Engineering
VNRVJIET
Hyderabad, India
haripriya_k@vnrvjiet.in

2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) | 978-1-6654-7467-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSIT55814.2023.10060903

***Abstract*—There are many existing machine learning models related to health care which mainly focuses on detecting only one disease. Therefore, this study has developed a system to forecast several diseases by using a single user interface. The proposed model can predict multiple diseases such as diabetes, heart disease, chronic kidney disease and cancer. If left untreated, these diseases pose a risk to humanity. As a result, many lives can be saved by early detection and diagnosis of these disorders. This research work attempts to implement various classification algorithms (K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression, Gaussian naive bayes.) to perform disease prediction. The accuracy of each algorithm is validated and compared with each other to find the best one for prediction. Furthermore, multiple datasets (for each disease each dataset) are used to achieve utmost accuracy in the predicted results. The main goal is to create a web application capable of forecasting several diseases by using machine learning, including diabetes, heart disease, chronic kidney disease, and cancer.**

Keywords: Single user interface, Diabetes, Heart disease, Chronic kidney disease, Cancer, K Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, Gaussian naive bayes.

I. INTRODUCTION

This study intends to predict multiple diseases such as diabetes, heart disease, chronic kidney disease and cancer. Various classification algorithms (like KNN, SVM, Random Forest, Logistic Regression and Naive Bayes) are used here to perform disease detection. The accuracy of each algorithm is validated and compared with each other in order to find the best one for prediction. Further, multiple datasets (for each disease each dataset) are used to achieve utmost accuracy in the predicted results. The best predicted algorithm for every disease is chosen and integrated to build a web application. The user can easily predict the required disease by typing respective attribute(input) values of that specific disease.

Here, various datasets are used, wherein each dataset belongs to separate disease.

The Cleveland, Hungary, Switzerland, and Long Beach V databases are used to study the cardiac disease. It has 76 qualities, including the anticipated attribute, however a selection of 14 of them are used for this research. The "target" field alludes to the patient's having heart illness. 0 means there is no disease, while 1 means there is a disease.

The dataset utilised for chronic renal disease had 25 features and was collected over a 2-month period in India (e.g., red blood cell count, white blood cell count, etc). The classification, which can be either "ckd" or "notckd" and refers to chronic kidney disease, is the object of the attack. This collection contains 400 records overall.

For breast cancer, Kaggle dataset which consist of 32 attributes out of which 5 attributes are used. This dataset consists of 570 records.

For Diabetes, Pimas Indians Diabetes Database and Kaggle Dataset which comprised of one target variable, Outcome, and multiple medical predictor factors are used. The patient's BMI, insulin level, age, number of previous pregnancies, and other factors are predictor variables. This dataset consists of 769 records and 9 columns.

On Collecting datasets from various sources, We Performed data pre-processing techniques such as Label encoding and Created models using various machine learning algorithms such as K-NN, Gaussian NB, Decision Trees, Support Vector Machine, Logistic Regression and Random Forest. For each disease various algorithms are used to create classifier models.

For each disease the dataset is split into two parts i.e., training set and testing set and each model is trained against its training dataset. After evaluating each model of a particular disease against testing dataset, the model's accuracy is calculated and the model's accuracy is compared against each other and the best one is selected.

Coming to web application (UI) it mainly includes features like 'sidebar' which is used for navigation and 'forms' to enter input attribute values for a particular disease. The sidebar is created using the option_menu method of streamlit_option_menu module. The input value fields in the forms are simply created using the text_input method of streamlit. We have easily loaded the trained models(classifiers) for prediction into the streamlit editor using the pickle module.

The pickle module has a 'load' method to import the saved classification models.

The application (UI) also includes test result button for each disease prediction page to trigger the prediction function for that particular disease. At last, using the inbuilt prediction function of streamlit, the disease prediction process has been carried out through a web application

II. Related Work

Keniya, Rinkal[1] From an open-source dataset, an excel sheet was created where all the symptoms for the respective diseases were listed. Around 230 diseases with more than 1000 unique symptoms in all were listed. The symptoms, age, and gender of an individual were used as input to various machine learning algorithms.

Godse Rudra A [2] In this system machine learning algorithms such as Naïve Bayes Algorithm, K-Nearest Algorithm, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine are used. The system also suggests the user to consult a doctor based on the report.

Alanazi, Rayan [3] Datasets are divided into training and test data and preprocessing is applied to the collected data. After that, CNN and KNN machine learning algorithms are used to train the training data set. Once the desired target has been reached after several epochs, the generated model is prepared for testing.

Arumugam, K [4] The Cleveland data set is used. After cleaning the data using by preprocessing, Machine learning algorithms like SVM, Nave Bayes, and Decision Tree C4.5 are now fed this data. These classifiers are used to predict heart disease in diabetes individuals.

Mohit, Indukuri [5] This work a web application is developed to detect diseases like breast cancer, diabetes and heart disease using machine learning models such as logistic regression, SVM, K-Nearest Neighbor's.

KM Jyothi Rani [6] Diabetes dataset used in this work contains 2000 cases. Predicting whether the person is diabetic or not is the objective. Different classification algorithms used are KNN, Logistic Regression, Decision Tree, Random Forest, SVM.

Dr. Sunita Varma [7] Goal of this work is to predict diabetes with better accuracy. Various classification and ensemble algorithms like Random Forest, SVM, KNN, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used to predict diabetes. The Pima Indian Diabetes Dataset repository at UCI is where the information was found.

Quan Zou [8] Data from hospital physical examinations in Luzhou, China was used to create the dataset. Diabetes is predicted using machine learning classification techniques including decision trees, random forests (RF), and neural networks. The Luzhou dataset results demonstrate the shortcomings of the blood glucose-free approaches.

Nazid Ahmed [9] First, preprocessing is done on two different datasets. The data is then split into two sets for training and testing. Using a range of machine learning algorithms, prediction models are developed using the training data. The performance is then accessed in relation to several metrics. Finally, flask is used to deploy the best ML model in a web application. Only diabetic disease is predicted by the system's architecture. Accuracy can be raised by using a larger, deeper dataset with more attributes.

Gazi Mohammed Ifraz [10] In this work, three different models were trained for accurate prediction using a variety of physiological variables as well as ML methods as logistic regression (LR), decision tree (DT) classification, and K-nearest neighbor (KNN). Only kidney illness is predicted by the system's design. Accuracy can be raised by using a larger, deeper dataset with more attributes.

S.Revathy [11] Machine learning prediction algorithms can be used intelligently to anticipate the onset of CKD and provide a means of early treatment. This study suggests the best prediction model when attempting to predict CKD utilizing classifiers like Decision Tree, Random Forest, and Support Vector Machine. Only kidney illness is predicted by the system's design. Accuracy can be raised by using a larger, deeper dataset with more attributes.

Zixian Wang [12] Based on a dataset for chronic kidney disease (CKD) from the UCI machine learning data warehouse, this study analyses chronic kidney disease using machine learning techniques. For 400 individuals with chronic kidney disease, the A priori association approach is used to identify CKD. Only kidney illness is predicted by the system's design. Accuracy can be raised by using a larger, deeper dataset with more attributes.

F.J Shaikh [13] The ML & DL techniques utilized in cancer progression modelling are proposed to be reviewed in this research. Many of the predictions discussed are associated with certain ML, input, and data sample supervision. can employ additional cutting-edge machine learning algorithms and extraction techniques to provide more thorough comparison analysis.

Baban Uttamrao [14] The suggested approach compares the outcomes of HRFLM application to other classification methods, such as decision tree and support vector machine. One dataset is used by the developed system, which only predicts heart disease. In some instances, it causes inaccuracies.

Harshit Jindal [15] The K nearest neighbors (KNN), Logistic Regression, and Random Forest Classifiers algorithms utilized in this study can aid practitioners or medical analysts in correctly diagnosing heart disease. The major goal of this project's refinement is to perform as well as possible when predicting the chance of heart disease using a single dataset.

Dhai Eddine Salhi [16] They choose to use Neural Networks (NN), KNN, and SVM to assess the dataset that comprises data on Algerian patients after analyzing these articles' analysis. After examining the earlier findings, they conclude that the neural network algorithm is the ideal one for our study because it consistently produces accurate results. This technique is exclusively used to forecast cardiac disease. They solely used a dataset of Algerian patients.

Jean Sunny [17] On the Wisconsin Diagnostic Breast Cancer dataset, which is taken from a digitised image of an MRI, this research compares six machine learning algorithms: Naive Bayes, Random Forest, Artificial Neural Networks, k-Nearest Neighbour, Support Vector Machine (SVM), and Decision Tree (DT). Using a larger, deeper dataset with more attributes will result in more accuracy than the suggested article.

III. Proposed Work

Fig.1 explains about predicting multiple diseases with the help of various machine learning algorithms such as Naïve Bayes, K-NN, Random Forest, Logistic Regression and SVM to close the gap between patients and doctors so that each can pursue their own objectives. The accuracy of each algorithm is validated and compared with each other in order to find the best one for prediction. Multiple datasets are integrated to achieve utmost accuracy in the predicted results. To make things easier for the end users, a web application has been developed, wherein the user can easily predict the disease, which they want, just by entering the respective attribute(input) values of that specific disease.

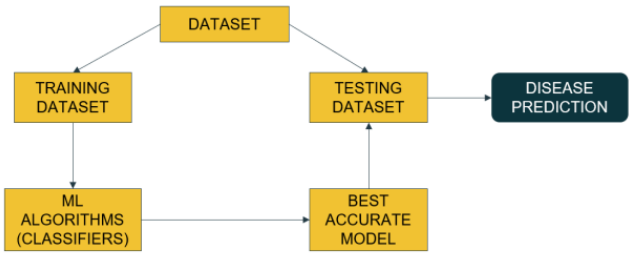

Fig. 1. System Architecture of Proposed work

The Advantages of the Proposed System include the following:

- System can predict multiple diseases using different machine learning techniques.
- Huge amount of data can be analyzed in a minimal time.
- Accuracy of each algorithm is validated and compared with each other in order to find the best one for prediction.

A. Workflow of proposed work

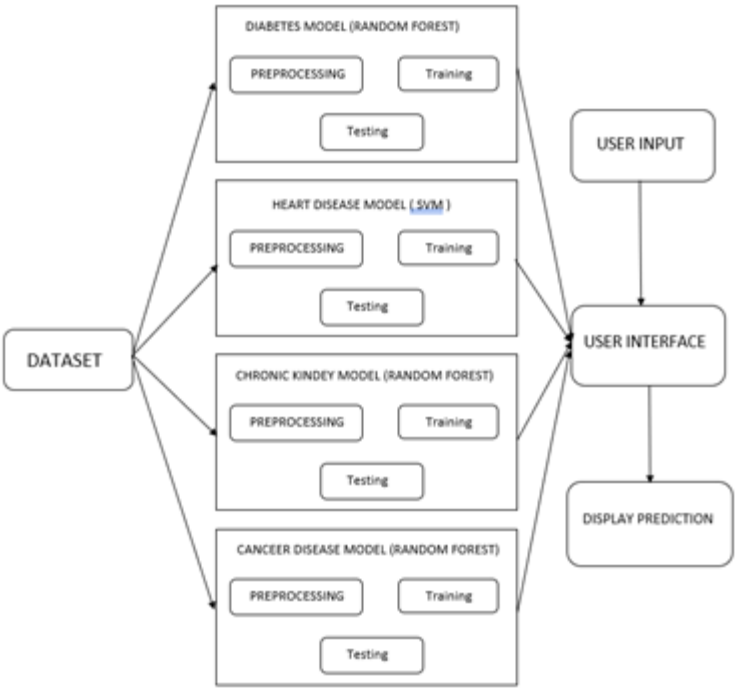

Fig. 2. System Design

Step 1: Collecting datasets from various sources.

For this project datasets for different diseases i.e., heart disease, cancer, diabetes, and chronic kidney is collected from Kaggle dataset

Step 2: Performing data preprocessing techniques such as Label encoding.

Label encoding helped in converting the categorical data such as gender, appetite into numerical data in the form zero's and one's.

Step 3: Creating models using various machine learning algorithms such as K-NN, Gaussian NB, Decision Trees, SVM, Logistic Regression and Random Forest.

For each disease various algorithms are used to create classifier models

Step 4: Training all the models against datasets.

For each disease the dataset is split into two parts i.e., training set and testing set and each model is trained against its training dataset

Step 5: Evaluating the models using metrics such as accuracy score.

After evaluating each model of a particular disease against testing dataset, the model's accuracy is calculated.

Step 6: Pickling the best model.

The model's accuracy is compared against each other and the best one is selected

B. Machine Learning Models

1) Gaussian Naïve Bayes

Gaussian Naive Bayes is a simple method for creating classifiers, which uses models to assign issue instance class labels that are given as vectors of factors values and are selected from a finite set. There isn't a single technique for training these classifiers, but rather a tribe of algorithms founded on a common tenet: given the class variable, all naïve Bayes classifiers infer that a certain feature's value is independent of any other feature's value.

2) K Nearest Neighbor

The K-nearest Neighbors algorithm, is a supervised learning tool that is applicable to both classification and regression tasks. Using the idea of grouping, it places the new one into category that matches available one by measuring similarity between the new data and existing data. Because it is based on the premise that neighboring similar points can be located, it is commonly employed as a categorization strategy.

3) Support Vector Machine

SV-Machine is a procedural methodology which is being used for classification & regression challenges. On the other hand, this algorithm is specifically utilized for classification issues in real-world problems and scenarios. Each data point is represented using this method as a point in n-D space, where n is the total number of features, each of which has a unique feature. To identify the hyper-plane that separates the classes in the supplied data set, the classification technique is then called.

4) Decision Tree

Both classification and regression issues can be solved using decision trees. It is a greedy algorithm, or ID3. Only a question is posed in a decision tree, and subtrees are created based on the answer (yes/no). Recursively, input data is divided up based on chosen properties. The attribute selection measures are used in order to determine the best attribute for the root node and sub-node. Since it uses the same reasoning process that people use when making decisions, it is easy to understand.

5) Random Forest

Random Forest (RF) works by creating various decision trees on different subsets of data and the outputs of those individual decision trees are used together to form a final output. Random forest works on bagging principle, first various subsets are generated from the dataset this process is called bootstrap. Now each model works on these subsets and gives outputs, next random forest combines the outputs of various models and produces output based on voting this step is known as aggregation.

6) Logistic Regression

Logistic-Regression is a method of classification strategy, this model works by finding out probability and performs classification based on that data. Sigmoid Function is used to find out the probability of a row belonging to a class, this probability will be between 0 and 1, it takes a product of theta transpose and parameter vector as input and gives the probability. Threshold is a value based on which classification is decided, a class is given to a row that has probability less than the threshold and other class is given to the row that has more probability than the threshold.

C. Streamlit

Elegant machine learning web apps may be swiftly produced and distributed using an open-source framework called Streamlit. It is a Python library designed primarily with machine learning users in mind. Since the majority of data scientists are familiar with this framework, their work has been made easier because they aren't interested in devoting weeks to learning how to utilise it to build web applications. The good thing about Streamlit is that it allows you to create your own online application right away, even if you don't have any previous experience with web development. Therefore, Streamlit is a great choice if you're interested in data science and want to deploy your models quickly, simply, and with the least amount of code possible.

Features of StreamLit:

- HTML, CSS, and JavaScript are not necessary.
- As opposed to taking days or months to construct a web app, we can create a stunning machine learning or data science software in only a few hours or even minutes.
- It works with most Python libraries, including Pandas, Matplotlib, Seaborn, Plotly, Keras, PyTorch, and SymPy (latex).
- With less code, amazing online applications could be made.
- Data caching makes computation pipelines simpler and faster.

IV. RESULTS

Following are the accuracy tables for each algorithm for various diseases:

A. Gaussian Naïve Bayes

TABLE I. RESULTS OBTAINED USING GAUSSIAN NAIVE BAYES

Disease	Accuracy
Diabetes	79.5
Heart Disease	84.78
Chronic Kidney Disease	95.8
Cancer	90

From the above Table 1 the highest accuracy is obtained for chronic kidney disease i.e., 95.8.

B. *K Nearest Neighbor*

TABLE 2. RESULTS OBTAINED USING KNN

Disease	Accuracy
Diabetes	80.75
Heart Disease	85.3
Chronic Kidney Disease	83.3
Cancer	85.9

From the Table 2 the highest accuracy is obtained for cancer disease i.e., 85.9.

C. *Support Vector Machine*

TABLE 3. RESULTS OBTAINED USING SVM

Disease	Accuracy
Diabetes	76.75
Heart Disease	86.41
Chronic Kidney Disease	95
Cancer	87.7

From the above Table 3 the highest accuracy is obtained for chronic kidney disease i.e., 95.

D. *Decision Tree*

TABLE 4. RESULTS OBTAINED USING DECISION TREE

Disease	Accuracy
Diabetes	98.75
Heart Disease	76.63
Chronic Kidney Disease	87.5
Cancer	88.3

From the above Table 4 the highest accuracy is obtained for diabetes i.e.,98.75.

E. *Random Forest*

TABLE 5. RESULTS OBTAINED USING RANDOM FOREST

Disease	Accuracy
Diabetes	99.5
Heart Disease	85.86
Chronic Kidney Disease	99.16
Cancer	94.1

From the above Table 5 the highest accuracy is obtained for chronic kidney disease i.e., 99.16.

F. *Logistic Regression*

TABLE 6. RESULTS OBTAINED USING LOGISTIC REGRESSION

Disease	Accuracy
Diabetes	77.5
Heart Disease	83.69
Chronic Kidney Disease	95
Cancer	90

From the above Table 6 the highest accuracy is obtained for chronic kidney disease i.e., 95.

A User Interface is developed through which the end user can select a specific disease and enter its respective attributes for prediction.

Fig. 3. Output for diabetes disease for positive case

Above Fig.3 displays the positive case for diabetes disease i.e. (The Person is diabetic)

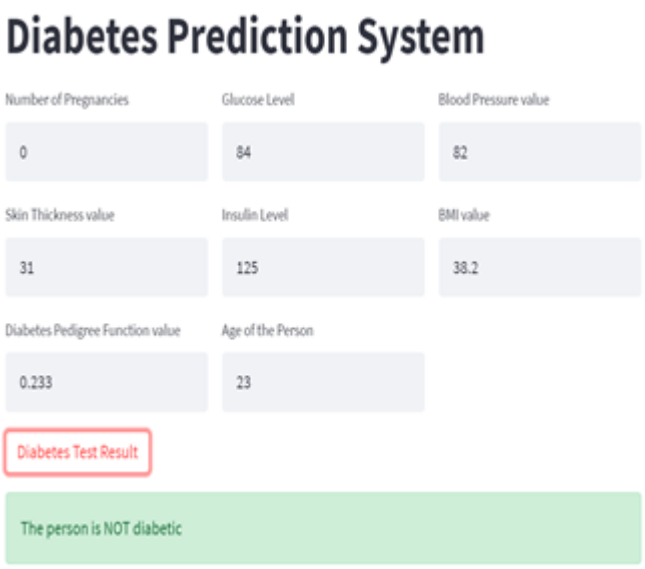

Fig. 4. Output for diabetes disease for negative case

Above Fig.4 displays the negative case for diabetes disease i.e. (The Person is not Diabetic)

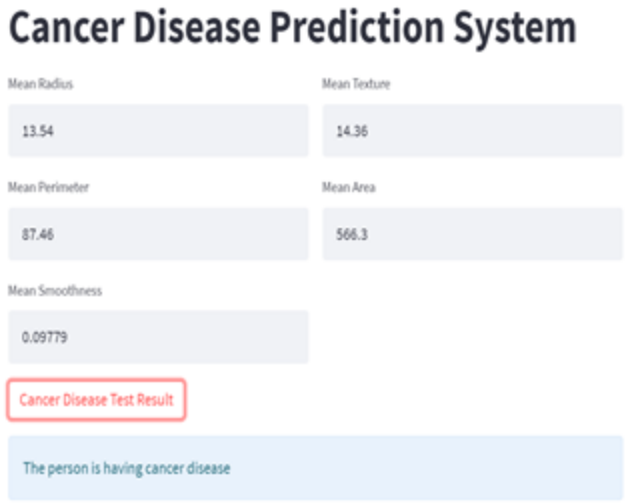

Fig. 5. Output for cancer disease for positive case

Above Fig.5 displays the positive case for Cancer disease i.e. (The Person is having cancer disease)

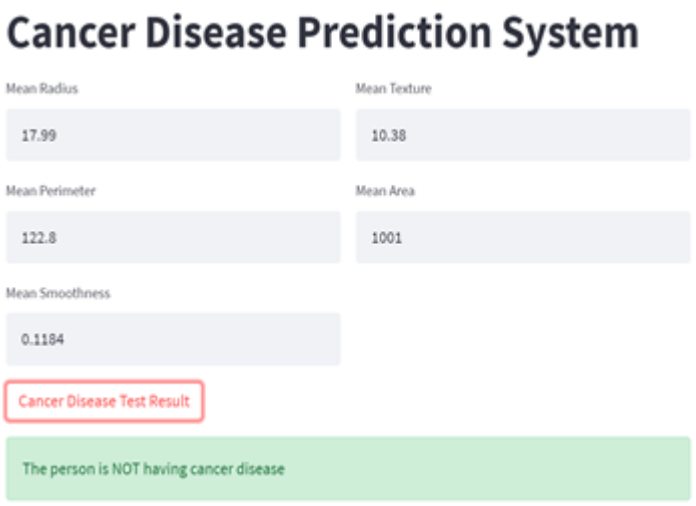

Fig. 6. Output for cancer disease for negative case

Above Fig.6 displays the negative case for Cancer disease i.e. (The Person is not having cancer disease)

Fig. 7. Output for heart disease for positive case

Above Fig.7 displays the positive case for heart disease i.e. (The Person is having heart disease)

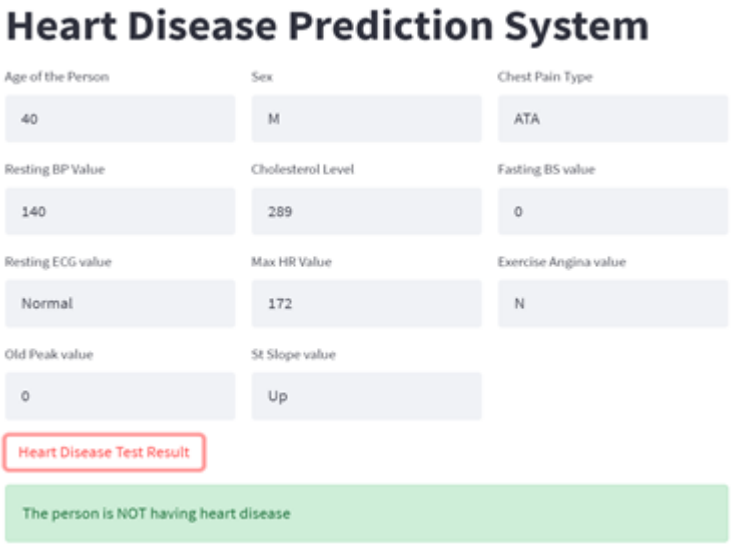

Fig. 8. Output for heart disease for negative case

Above Fig.8 displays the negative case for heart disease i.e. (The Person is not having heart disease)

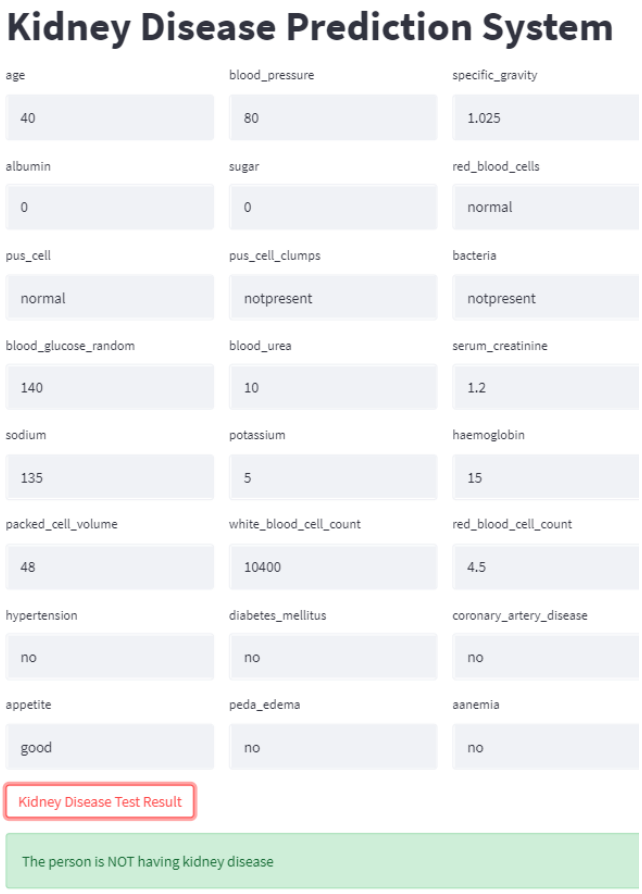

Fig. 9. Output for kidney disease for negative case

Above Fig.9 displays the negative case for kidney disease i.e. (The Person is not having kidney disease)

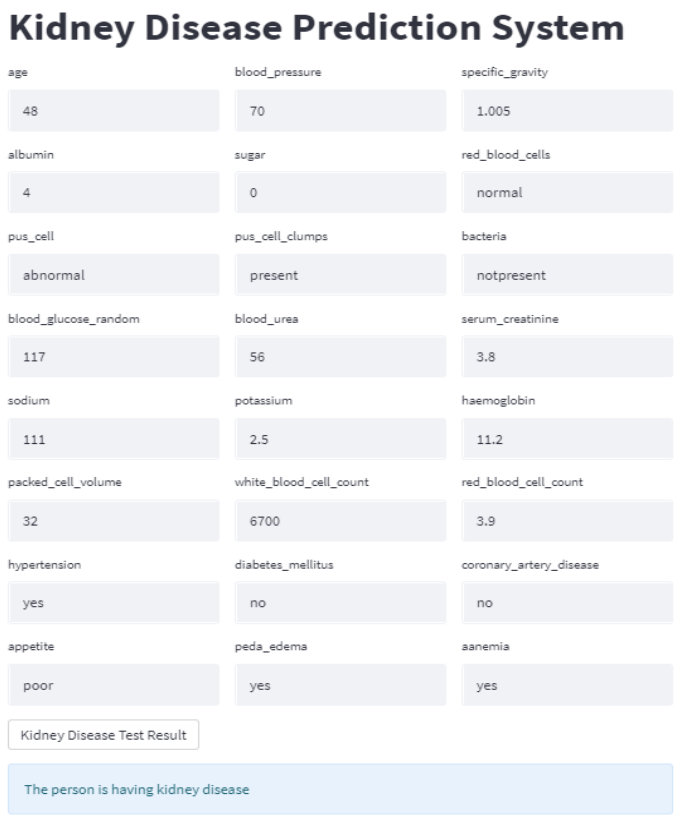

Fig. 10. Output for kidney disease for negative case

Above Fig.10 displays the positive case for kidney disease i.e. (The Person is having kidney disease)

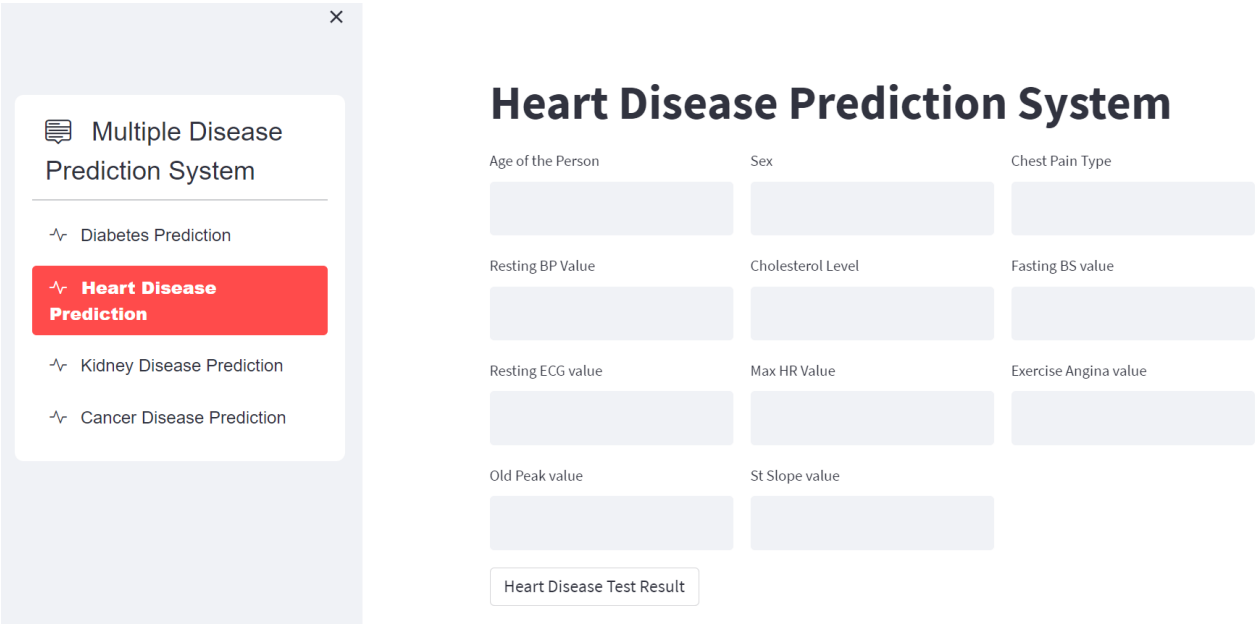

Fig. 11. Single User Interface (UI) of Multiple disease prediction system (Heart Disease Prediction System)

The side nav bar contains multiple diseases where we select the heart disease Prediction and enter the values in the input fields and click on the button Heart Disease Test Result. It will show us whether the patient is having heart disease or not.

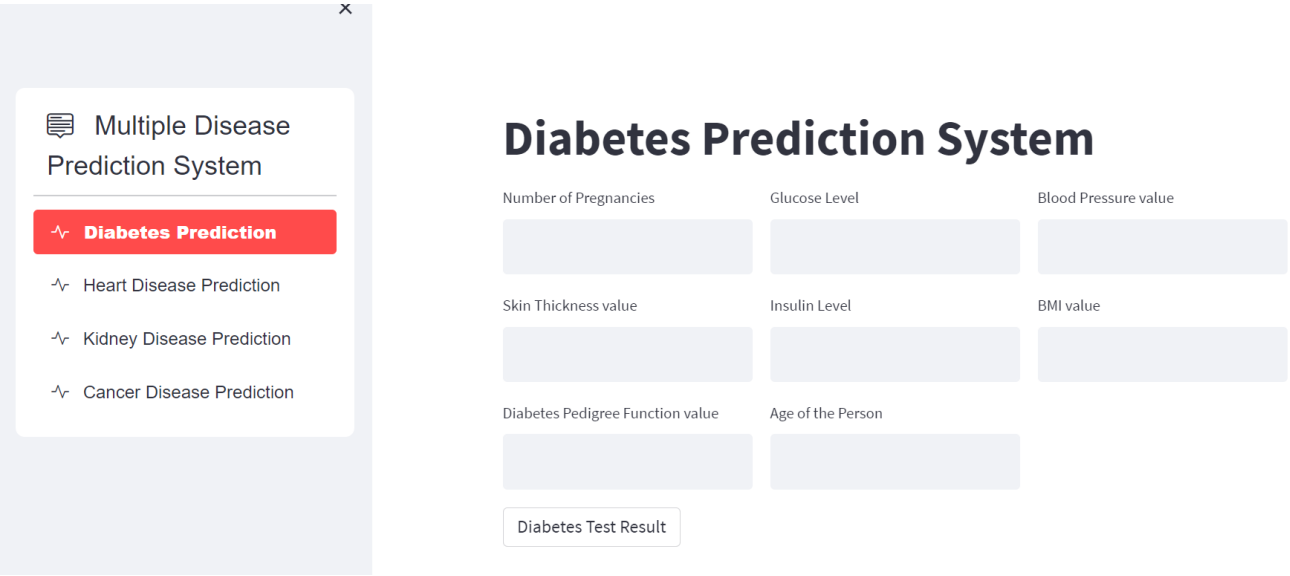

Fig. 12. Single User Interface (UI) of Multiple disease prediction system (Diabetes Disease Prediction System)

The side nav bar contains multiple diseases where we select the diabetes disease Prediction and enter the values in the input fields and click on the button Diabetes Disease Test Result. It will show us whether the patient is having diabetes or not.

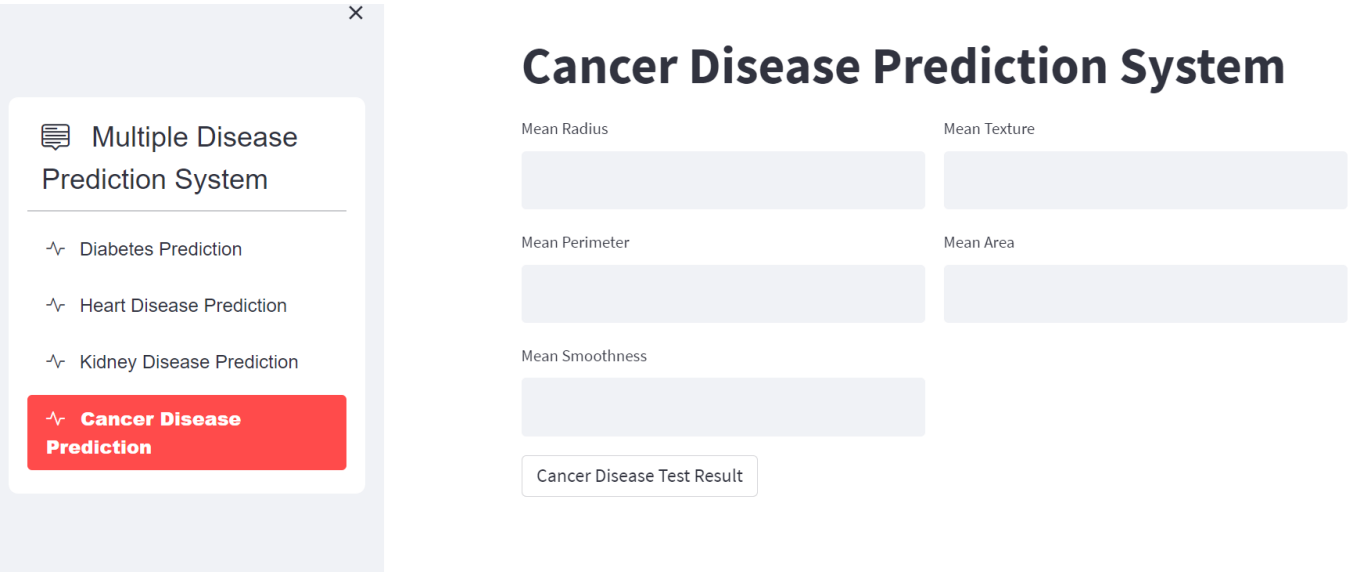

Fig. 13. Single User Interface (UI) of Multiple disease prediction system (Cancer Disease Prediction System)

The side nav bar contains multiple diseases where we select the cancer disease Prediction and enter the values in the input fields and click on the button Cancer Disease Test Result. It will show us whether the patient is having cancer or not.

TABLE 7. ACCURACY SCORE TABLE

Disease /Classifier	KNN	SVM	Random Forest	Decision Tree	Naive Bayes	Logistic Regression
Diabetes Disease	80.75	76.75	**99.5**	98.75	79.5	77.5
Heart Disease	85.3	**86.41**	85.86	76.63	84.78	83.69
Chronic Kidney Disease	83.3	95	**99.16**	87.5	95.8	95
Cancer Disease	85.9	87.7	**94.1**	88.3	90	90

The above Table 7 shows accuracy of each disease against various algorithms. For every disease we have trained with different machine learning models, out of that best accuracy model is taken and loaded into the streamlit editor using the pickle module.

V. CONCLUSION

We observed that for Diabetes Disease we got more accuracy for Random Forest Classifier, for heart disease we got more accuracy for SVM, for chronic kidney disease we got more accuracy for Random Forest Classifier and for Cancer Disease we got the more accuracy for Random Forest.

REFERENCES

[1] Keniya, Rinkal, et al. "Disease prediction from various symptoms using machine learning." Available at SSRN 3661426 (2020).
[2] Godse, Rudra A., et al. "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively." International Journal of Advanced Research in Computer and Communication Engineering 8.12 (2019): 50-52.
[3] Alanazi, Rayan. "Identification and prediction of chronic diseases using ma chine learning approach." Journal of Healthcare Engineering 2022 (2022).
[4] Arumugam, K., et al. "Multiple disease prediction using Machine learning algorithms." Materials Today: Proceedings (2021).
[5] Mohit, Indukuri, et al. "An Approach to detect multiple diseases using machine learning algorithm." Journal of Physics: Conference Series. Vol. 2089. No. 1. IOP Publishing, 2021.
[6] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.
[7] Joshi, Tejas N., and P. P. M. Chawan. "Diabetes prediction using machine learning techniques." Ijera 8.1 (2018): 9-13.
[8] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.
[9] Ahmed, Nazin, et al. "Machine learning based diabetes prediction and development of smart web application." International Journal of Cognitive Computing in Engineering 2 (2021): 229-241.
[10] Ifraz, Gazi Mohammed, et al. "Comparative analysis for prediction of kidney disease using intelligent machine learning methods." Computational and Mathematical Methods in Medicine 2021 (2021).
[11] Revathy, S., et al. "Chronic kidney disease prediction using machine learning models." International Journal of Engineering and Advanced Technology 9.1 (2019): 6364-6367.
[12] Wang, Zixian, et al. "Machine learning-based prediction system for chronic kidney disease using associative classification technique." International Journal of engineering & Technology 7.4.36 (2018): 1161.
[13] Shaikh, F. J., and D. S. Rao. "Prediction of cancer disease using machine learning approach." Materials Today: Proceedings 50 (2022): 40-47.
[14] Rindhe, Baban U., et al. "Heart Disease Prediction Using Machine Learning." *Heart Disease* 5.1 (2021).
[15] Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." *IOP conference series: materials science and engineering*. Vol. 1022. No. 1. IOP Publishing, 2021
[16] Salhi, Dhai Eddine, Abdelkamel Tari, and M. Kechadi. "Using machine learning for heart disease prediction." International Conference on Computing Systems and Applications. Springer, Cham, 2020.
[17] Rane, Nikita, et al. "Breast cancer classification and prediction using machine learning." *International Journal of Engineering Research and Technology* 9.2 (2020): 576-580.