**An Effective Machine Learning based Diagnosis System for Multiple Disease Prediction**

Adith Sreeram A S1 and Pappuri Jithendra Sai2

1 VIT-AP University, Inavolu, Andhra Pradesh 522237

2 VIT-AP University, Inavolu, Andhra Pradesh 522237

**Abstract**

The majority of the existing healthcare machine learning models are narrowly focused on a specific disease. In light of this, we have developed a streamlined method for disease prediction using Streamlit and a unified user interface. Major global killers include diabetes, cardiovascular disease, chronic kidney disease, stroke, and cancer. We face danger from these illnesses. However, the lack of check-ups for chronic conditions is responsible for the vast majority of these deaths. In this study, our system can enable users make predictions about the aforementioned diseases in a centralised location. Success in disease prediction is the driving force behind this study, which employs a wide variety of Classification (K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression, Gaussian naive bayes) and Boosting (Gradient Boost, XGBoost, AdaBoost) algorithms. Accuracy of each algorithm is checked and compared to get the best algorithm for prediction. To provide the highest precision in the anticipated outcomes, we use specialised databases for each condition. The Streamlit interface, a platform for developing Web Application based interfaces, is where all the predictive models are placed. The user can input their information and receive a prognosis of the course of a specific condition, giving them the opportunity to take preventative measures before the situation worsens.

**Keywords:** Streamlit, Multiple Disease Prediction, Diabetes, Heart Disease, Chronic Kidney Disease, Stroke, Cancer, Classification Algorithms, Boosting Algorithms, Gradient Boost, XGBoost, AdaBoost.

**1 Introduction**

Data on patients in many different forms (clinical factors, hospital resources, illness diagnostic data, patient records, and medical equipment) are generated as a result of the digital age and technical advancements in the medical field. Information useful for making good decisions must be mined from this mountain of data, which is both large and complex. Medical data mining is a powerful tool for uncovering previously unknown connections within large medical datasets. The widespread use of data mining tools and machine learning techniques for identifying significant patterns, determining correlations and interactions among various variables, and analysing enormous databases has led to significant developments in the healthcare industry. It's a vital resource for doctors to compare and evaluate the latest findings and make educated decisions. In order to maximise the reliability of the projected outcomes, we employ disease-specific datasets.

The purpose of this research is to develop methods for forecasting the occurrence of several different diseases. K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, Gaussian Naive Bayes, Gradient Boost, XGBoost, and AdaBoost are only few of the classification methods employed here for disease detection. All prediction algorithms are tested for accuracy and compared against one another to determine which one performs the best. In short, our aim is to create a web application, the top predicted algorithm for each disease is selected and included. By entering the appropriate input values for that particular disease, the user can quickly forecast the needed disease.

**Diabetes** is a metabolic disease that impairs the body's capacity to generate or use insulin, a hormone that controls blood sugar levels. There are two varieties of diabetes: Type 1, in which the body's immune system attacks the cells in the pancreas responsible for producing insulin, and Type 2, in which insulin resistance or insufficient insulin production lead to the disease. Pimas Indians Diabetes Database and the Kaggle Dataset have been used, both of which include the objective variable Outcome as well as other medical predictor factors. This dataset includes a wide variety of patient-specific characteristics, such as gestational diabetes status, blood sugar, insulin, body mass index, age, and more. There are 768 records in this collection, organised into 9 different categories.

**Heart disease** is an umbrella term for a variety of conditions that affect the heart, including arrhythmias, heart failure, and coronary artery disease. High blood pressure, high cholesterol, smoking, obesity, and insufficient physical activity are all major contributors to this epidemic, making it the top cause of death worldwide. Cleveland, Hungary, Switzerland, and Long Beach V are the locations of the datasets used here. There are 76 attributes in total in this dataset, with the target variable included; however, we are only making use of 14 of them (including Age, Sex, Resting Blood Pressure value, Serum Cholesterol Value, and many more). There are 1025 records in this collection and 14 different types of data.

**CKD, or chronic kidney disease,** is a progressive decline in kidney function that occurs over time as a result of chronic renal disease. The dataset we utilise comes from the UCI Machine Learning Repository, and it has 26 attributes in all, one of which is a target column indicating whether or not a given patient has CKD. Age, BP, Sodium, Potassium, and Blood Glucose Level are just a few of the Attributes. There are 400 records and 26 columns in this data collection.

**Stroke** is a medical illness that occurs when blood supply is cut off or severely decreased to a portion of the brain, killing brain cells. Symptoms of this potentially fatal illness include a loss of the ability to speak or hear speech, paralysis or numbness of the face, arm, or leg, and impaired vision. In this study, we employ a dataset consisting of stroke patients and control subjects who did not experience a stroke. In addition to the target column, it has ten more properties. Age, sex, hypertension, body mass index, coronary heart disease, and many others are among the traits. There are a total of 5109 records and 11 columns in this dataset.

**Cancer** refers to a category of disorders characterised by uncontrolled development of cells, their infiltration into neighbouring tissues, and their destruction of those tissues. The prognosis for successful treatment and survival can be considerably enhanced by early discovery and diagnosis. The dataset we employ comes from Kaggle, and its purpose is to classify tumours as benign or malignant based on their unique characteristics. One of the dataset's 30 variables is "diagnosis," which indicates whether the numbers fall under the "malignant" or "benign" category for cancer. With the exception of the id and the diagnosis, all of the attributes in this dataset describe some aspect of the cancer's appearance. There are 569 records and 32 columns in this data collection.

Each disease's collected data will be used to train a separate model, which will then be tested on a separate dataset. We will next utilise the testing set to make predictions, put our model to the test, and assess how well it performed. Models are ranked by how well they predict outcomes once the data has been analysed. When developing our web app, we will use the pickle module to store the top-performing models for each ailment and load them as needed. Streamlit, a python framework for developing attractive online apps, was used to construct the web app. One advantage of Streamlit is that its use does not necessitate familiarity with other web development frameworks like HTML and CSS. With Streamlit, you can rapidly and easily deploy your machine learning models with minimal effort and code. Streamlit allows us to make applications with multiple pages, so we can separate off each condition onto its own. Including a sidebar will make using the web app much more convenient for the user.

**2 Literature Review**

**MS P Keerthi et al**.[1] presented a single streamlit interface for three disease prediction models: heart disease, diabetes, and pneumonia. The heart disease model utilizes logistic regression, the diabetes model uses SVM classifier, and the pneumonia model employs convolutional neural network (CNN) on chest X-ray images. The models are embedded in a web interface using the Python framework Streamlit, allowing users to input their information and obtain predictions for early intervention.

**F. J. Shaikh and D. S. Rao**[2] conducted survey on the application of ML and DL approaches in cancer progression modeling, focusing on predictions that rely on specific ML techniques, input parameters, and data supervision.

**Rayan Alanazi** [3]proposed a system that uses algorithms like CNN for disease prediction and KNN for accurate prognosis based on symptoms. The paper also includes a comparative analysis of the proposed system with other algorithms such as Naïve Bayes, decision tree, and logistic regression, demonstrating its performance and effectiveness in disease prediction.

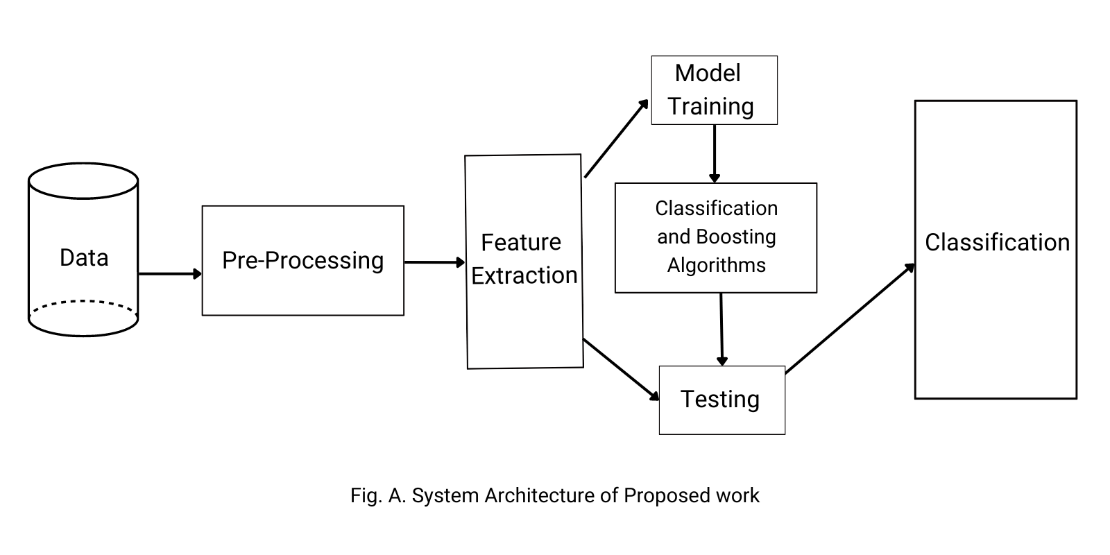
**K. Arumugam et al.**[4] demonstrates that the decision tree model outperforms naive Bayes and support vector machine models, allowing us to optimize it for accurately forecasting heart disease likelihood in diabetic individuals. To enhance accessibility and usability of disease-predicting web app for the general public, **Indukuri Mohit et al.**[5]developed medical test web application. This application leverages machine learning concepts such as logistic regression, SVM, K-Nearest Neighbor’s to provide accurate disease predictions across multiple conditions.

**Nazin Ahmed et al.** [6] aims to develop accurate machine learning models for diabetes diagnosis using clinical data. Several algorithms, including Decision Tree, Naive Bayes, k-nearest neighbor, Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machine, are trained and evaluated using diverse datasets. The data is pre-processed, divided into training and testing sets, and performance metrics are used to assess the models. The best-performing model is then integrated into a web application using Flask, specifically designed for predicting diabetes.

**Gazi Mohammed Ifraz et al.** [7]utilized logistic regression, decision tree classification, and k-nearest neighbor (KNN) algorithms to train three distinct models for prediction. The results showed that logistic regression achieved the highest accuracy among these methods. The focus of the system's design was specifically on predicting kidney disease. This project delves into the utilization of machine learning techniques for the classification of heart diseases. In this study, **Shiva Shanta Mani et al.** [8] proposed an approach that involves comparing the performance of the HRFLM (Heart Disease Risk Factor Learning Model) application with other classification methods, namely K-nearest neighbors, decision tree and random forest. The system is designed to predict heart disease using a specific dataset.

**3 Proposed Work**

Multiple diseases can be predicted with the help of several machine learning techniques, as shown in Figure A. These techniques include K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, Gaussian Naive Bayes, Gradient Boost, XGBoost, and AdaBoost. The best prediction algorithm is selected after rigorously comparing and contrasting the performance of several algorithms. For the sake of convenience, a web application has been developed that can make accurate disease predictions based on the user's input alone.



**Fig. a.** System Architecture of proposed work.

**3.1 Steps followed for our proposed work**

**Step 1:** Collecting datasets from various sources

To build this system we have collected different datasets for different diseases. The datasets are collected from Kaggle and UCI Machine Learning Repository.

**Step 2:** Implementing data pre-processing techniques such as Data Cleaning and Label encoding.

To achieve maximum accuracy and to train our model well we used some data pre-processing techniques like Data Cleaning where we will remove all the missing values and correct all the duplicate value etc. The Label Encoding technique is implemented to convert the categorical data into numerical data such as zero’s and one’s.

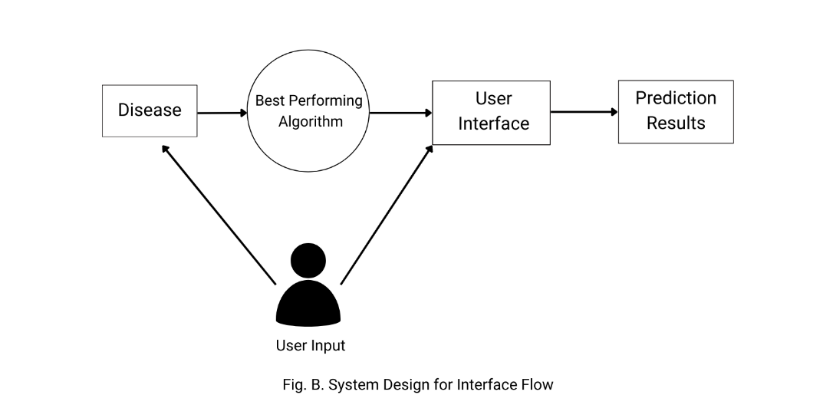
**Step 3:** Using different Classification and Boosting machine learning algorithms, such as K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression, Gaussian naive bayes, Gradient Boost, XGBoost, AdaBoostto create models and train all the models.

Different algorithms are used to create classifier models for different diseases. Once the model is created the dataset is split into two parts namely the training and the testing and the models that we created are trained against its training dataset.

**Step 4:** Finding the best performing model among the trained models.

We need to select the model that is best performing and we will do that by using our model against the testing data and evaluate the models based on the accuracy scores.

**Step 5:** Using Joblib. Joblib is a tool in python which helps us to save and load our models quickly. Here we will save the best performing models for each disease with the help of Joblib separately.



**Fig. b.** System Design for Interface flow

**3.2 Streamlit**

Streamlit is an open-source library that allows us to unique web apps for Machine Learning and Data Science projects fast and efficient. With this framework, you can easily build interactive visualization plots, models, and dashboards without having a worry about the underlying web framework or deployment infrastructure used in the backend. It also provides the users to add widgets which helps the users the interact with the web app and the models that we used. This framework also integrates the popular python and machine learning packages such as NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn and TensorFlow, which enables us to quickly build and deploy our trained models.

*Features of Streamlit:*

**User friendly**: Streamlit provides a simple and intuitive interface for creating interactive data applications with minimal coding.

**Rapid prototyping:** Streamlit is designed for rapid prototyping, enabling data scientists and developers to experiment with different ideas and build fully functional applications.

**Customizable:** Streamlit allows developers to easily customize the user interface, allowing developers to create applications with a unique look and feel.

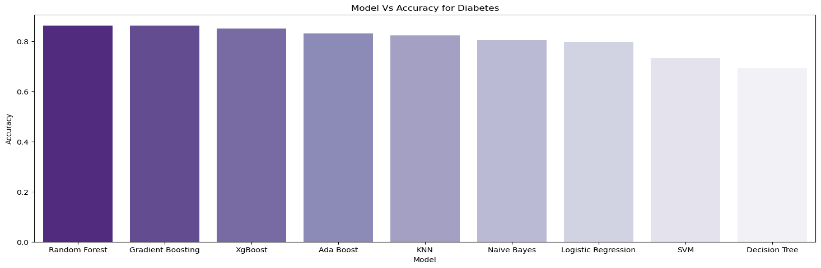
**Real-time collaboration:** Streamlit enables real-time collaboration, allowing multiple users to work on the same project simultaneously.

**Interactive widgets***:* Streamlit offers a wide range of interactive widgets such as sliders, dropdown menus, and checkboxes that allow users to edit and explore data in real-time.

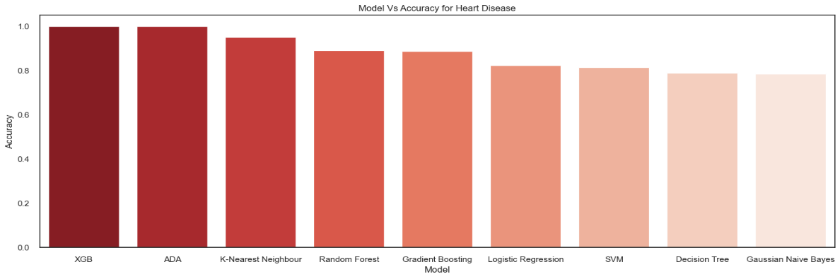
**Data Cache**: The data cache makes computational pipelines easier and faster.

**4 Results and Discussion**

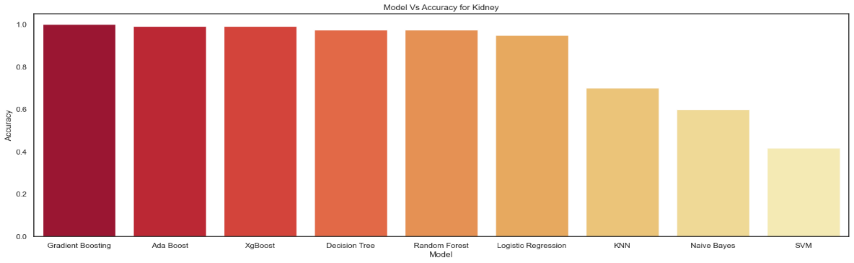
**4.1 Following are the accuracy for each algorithm for various diseases based on the Algorithm used.**

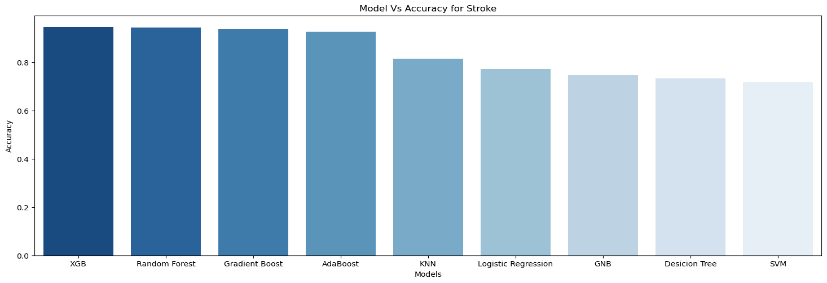
******

**Fig. c.** Model Vs Accuracy for Diabetes

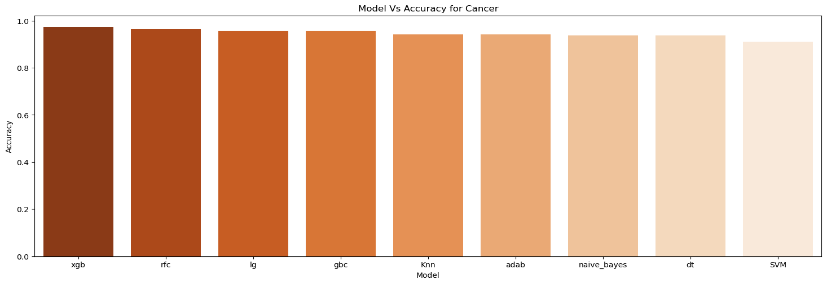
******

**Fig. d.** Model Vs Accuracy for Heart Disease

**Fig. e.** Model Vs Accuracy for Kidney



**Fig. f.** Model Vs Accuracy for Stroke



**Fig. g.** Model Vs Accuracy for Cancer

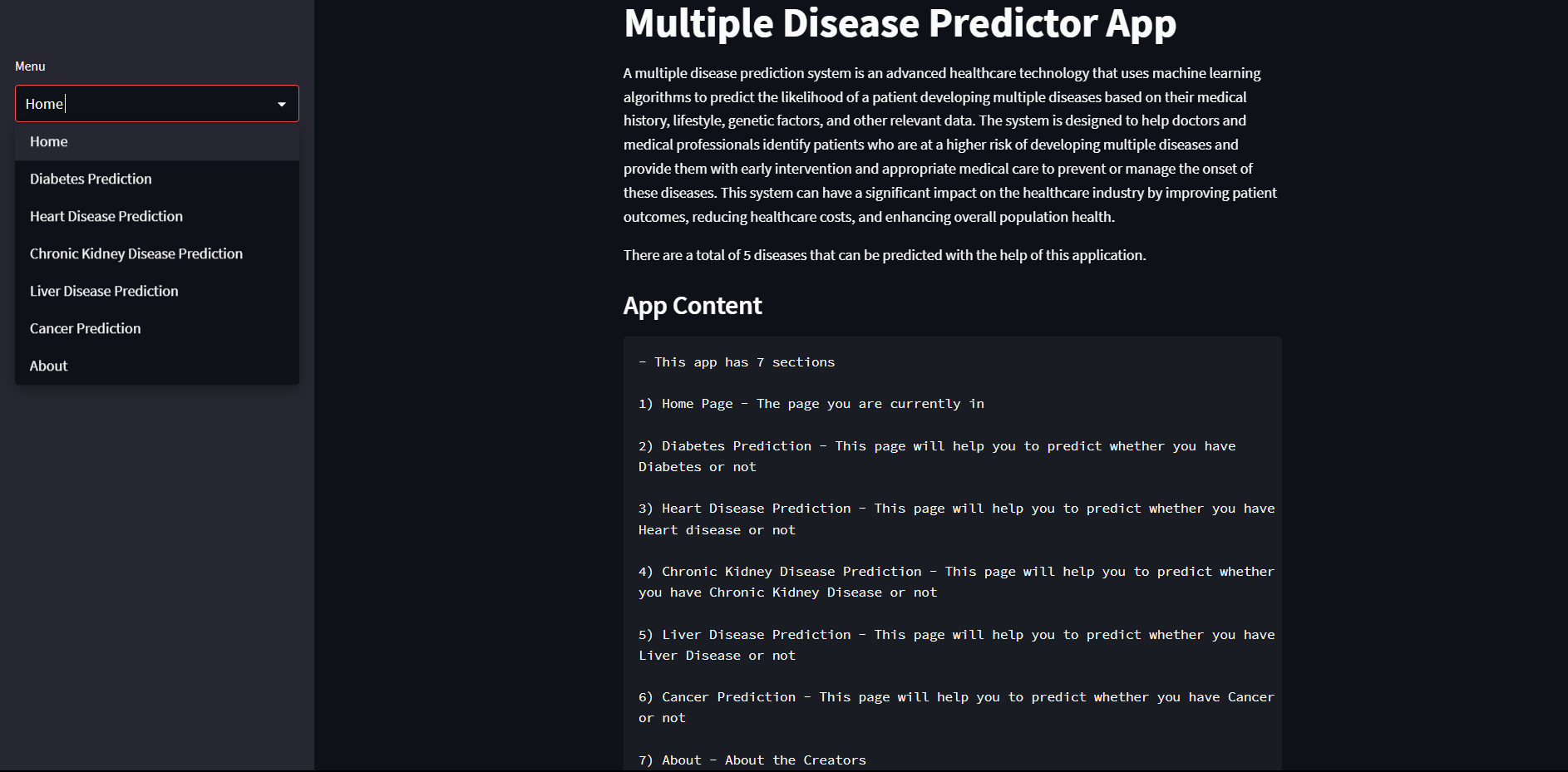
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Disease** | **Machine Learning Algorithms** | | | | | | | | |
|  | Logistic  Regression | Decision  Trees | Random  Forest | KNN | SVM | GNB | Gradient  Boosting | XGBoost | AdaBoost |
| Diabetes | 79.87% | 69.48% | **87.01%** | 82.46% | 73.37% | 80.51% | **87.01%** | 85.06% | 83.11% |
| Heart Disease | 82.44% | 79.02% | 89.26% | 95.12% | 81.46% | 78.53% | 88.78% | **100%** | **100%** |
| Chronic Kidney Disease | 95.83% | 95.83% | 97.50% | 63.33% | 40.83% | 65.00% | 97.50% | 97.50% | **100%** |
| Stroke | 77.20% | 73.38% | 94.53% | 81.50% | 71.91% | 74.85% | 93.83% | **94.71%** | 92.75% |
| Cancer | 95.74% | 93.61% | 96.27% | 94.14% | 90.95% | 93.61% | 95.74% | **97.34** | 94.14% |

**Table. 1.** Accuracy Table

The above Table 1 shows accuracy of each disease against K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression, Gaussian naive bayes, Gradient Boost, XGBoost, AdaBoost. Out of all the algorithms trained we have selected the best performing algorithm which is selected based on the highest accuracy achieved. Now, the best model is saved and loaded into the streamlit interface using Joblib.

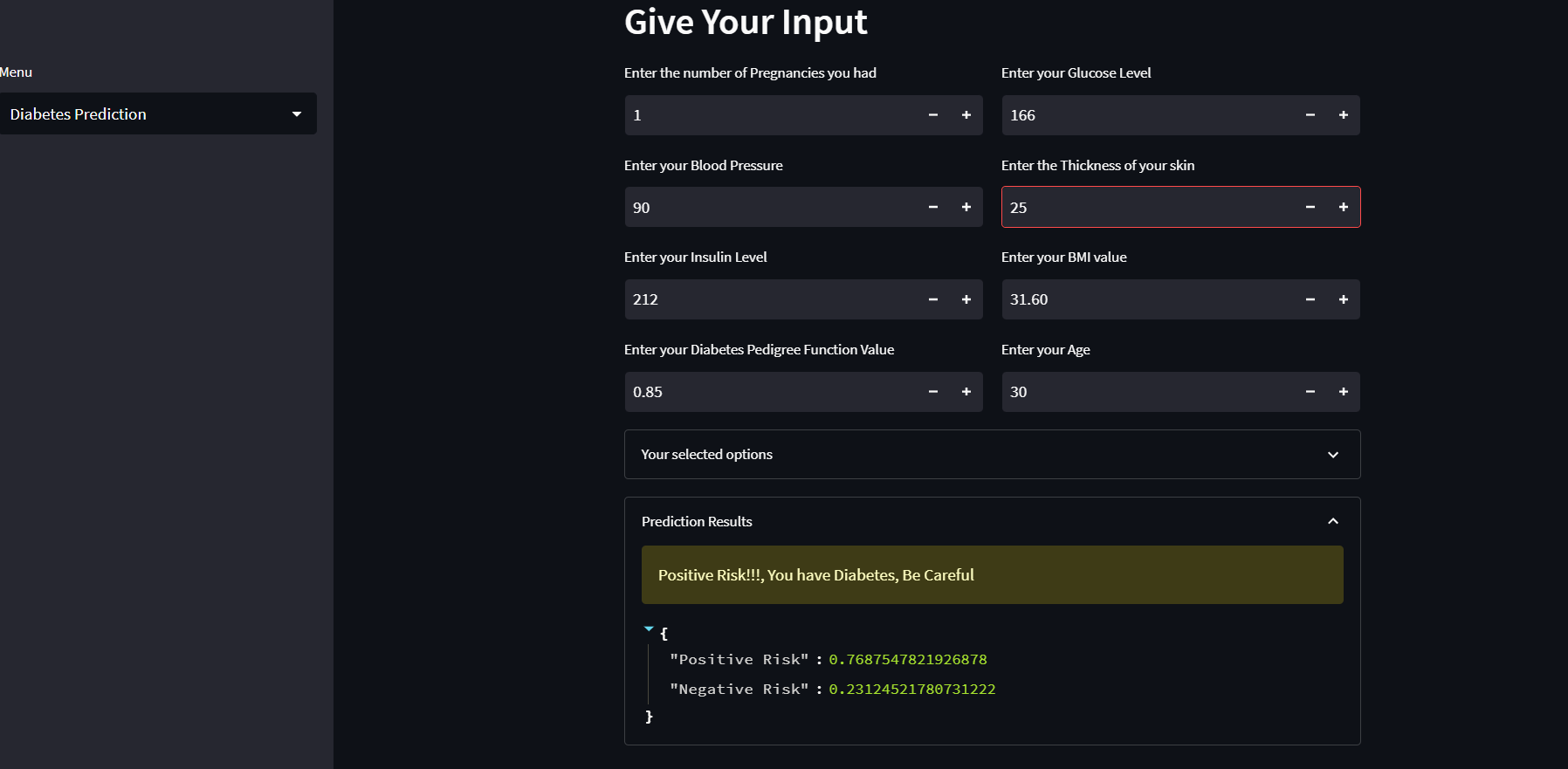
* 1. **Following are the acquired results from the Web application created**

A user interface is developed with Streamlit with which the end user can select a specific disease and enter its respective attributes for prediction.

****

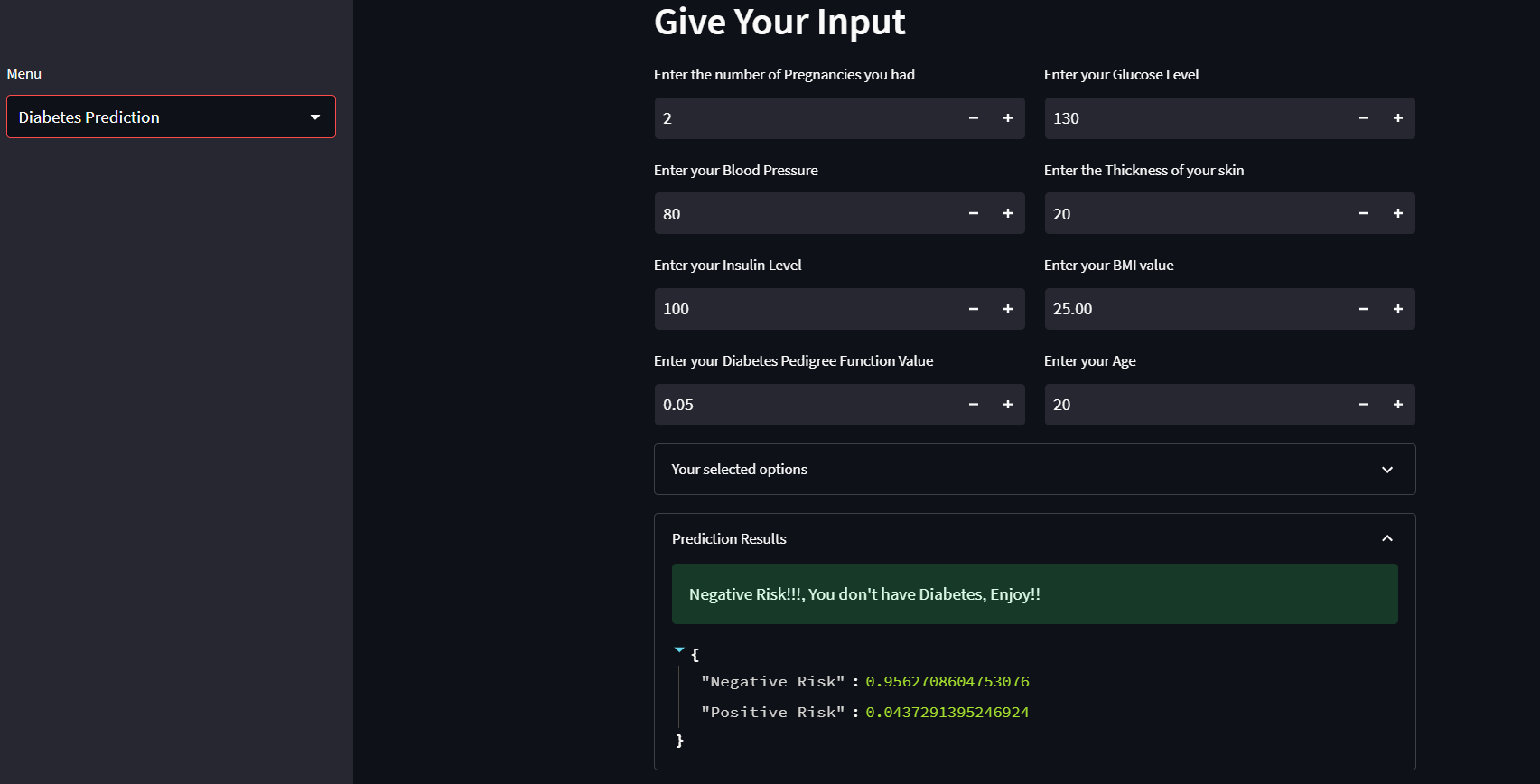
**Fig. h.** Home page Output

With the above picture (Fig. h) we can see the side nav bar contains multiple diseases where we select the disease that we want to predict and enter the values in the input fields.

****

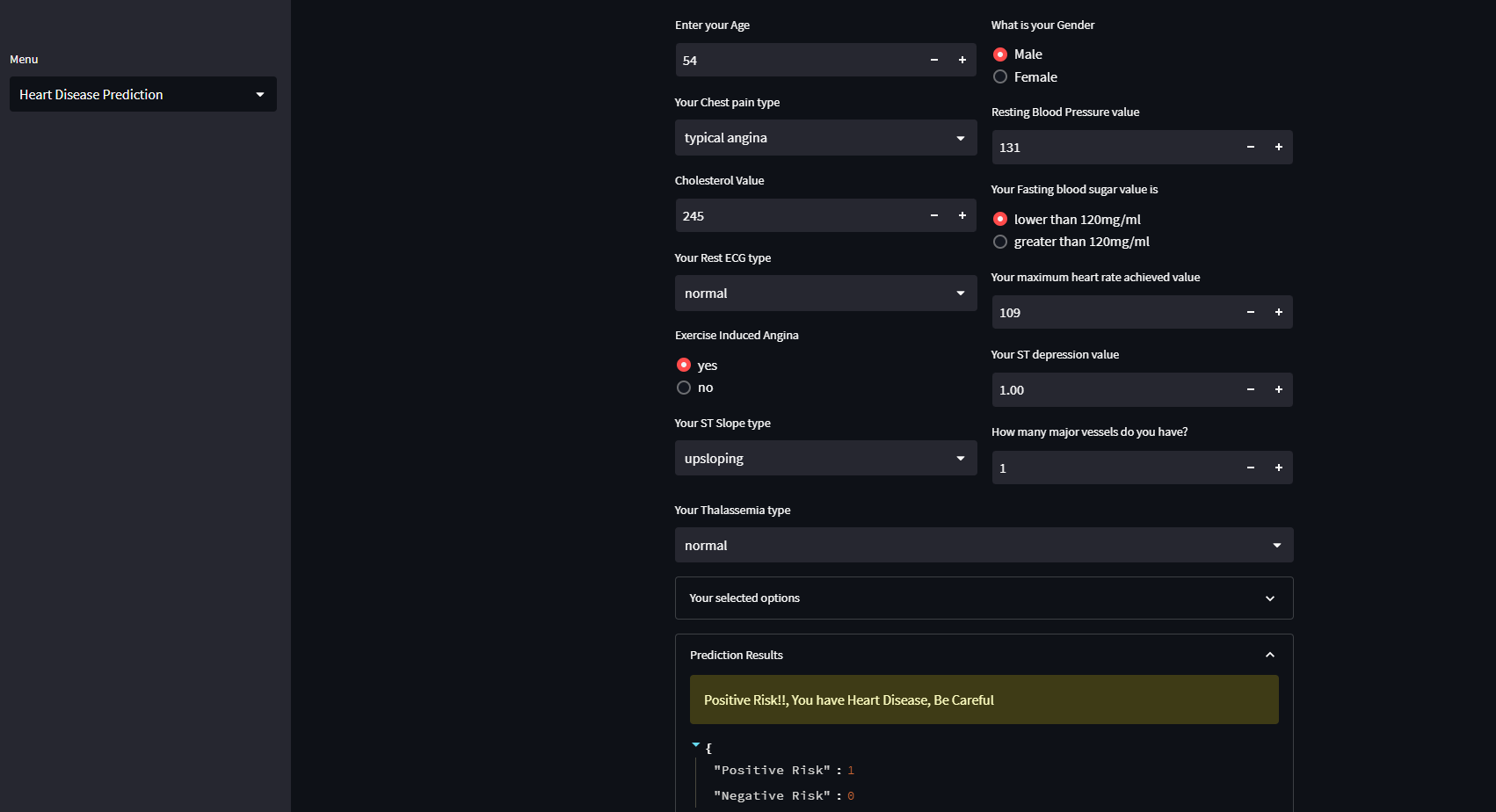
**Fig. i.** Output for diabetes for positive case

Above Fig i, indicates that the person has diabetes



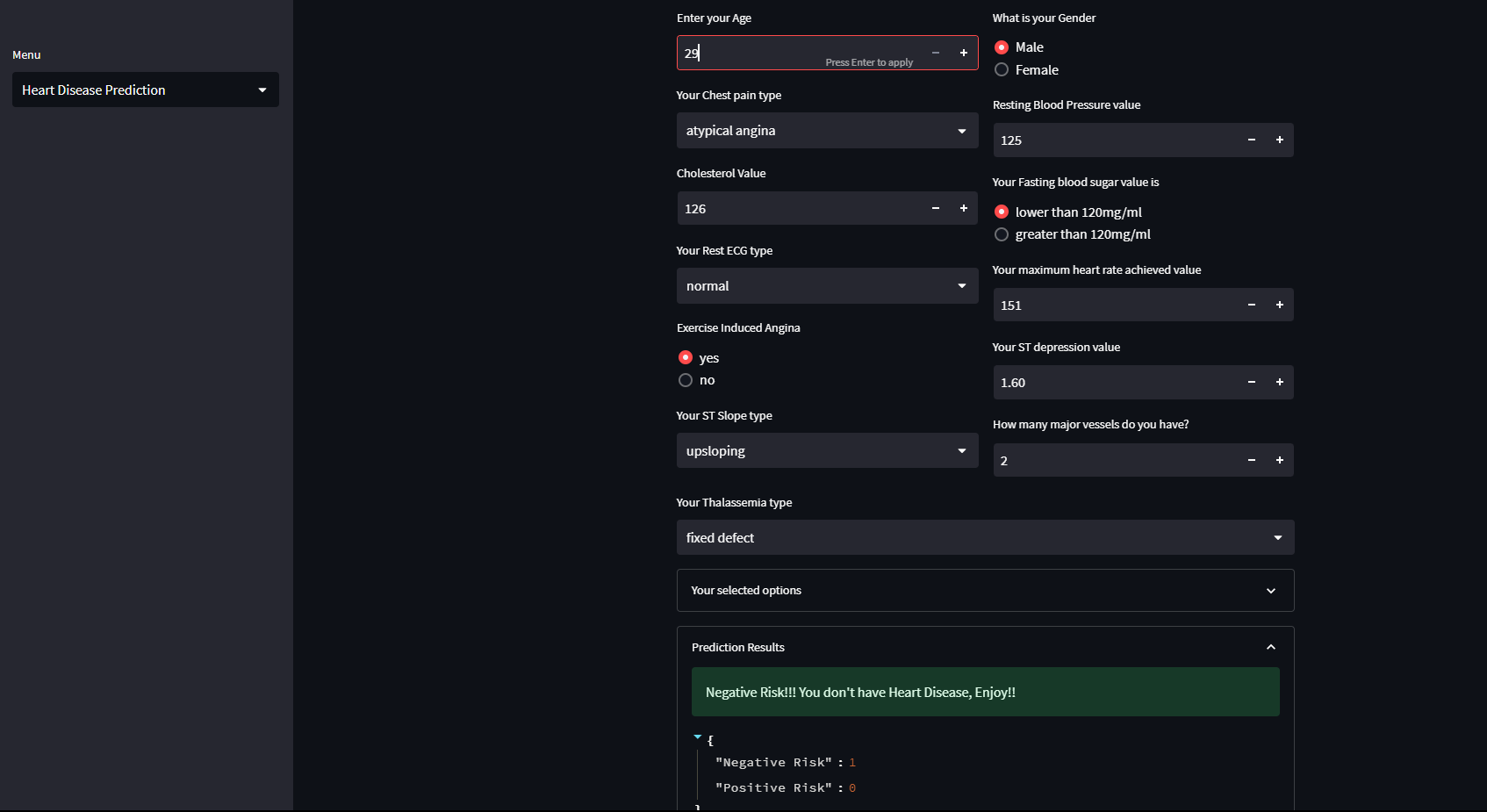
**Fig. j**. Output for diabetes disease for negative case

Above Fig j, indicates that the person doesn’t have diabetes

****

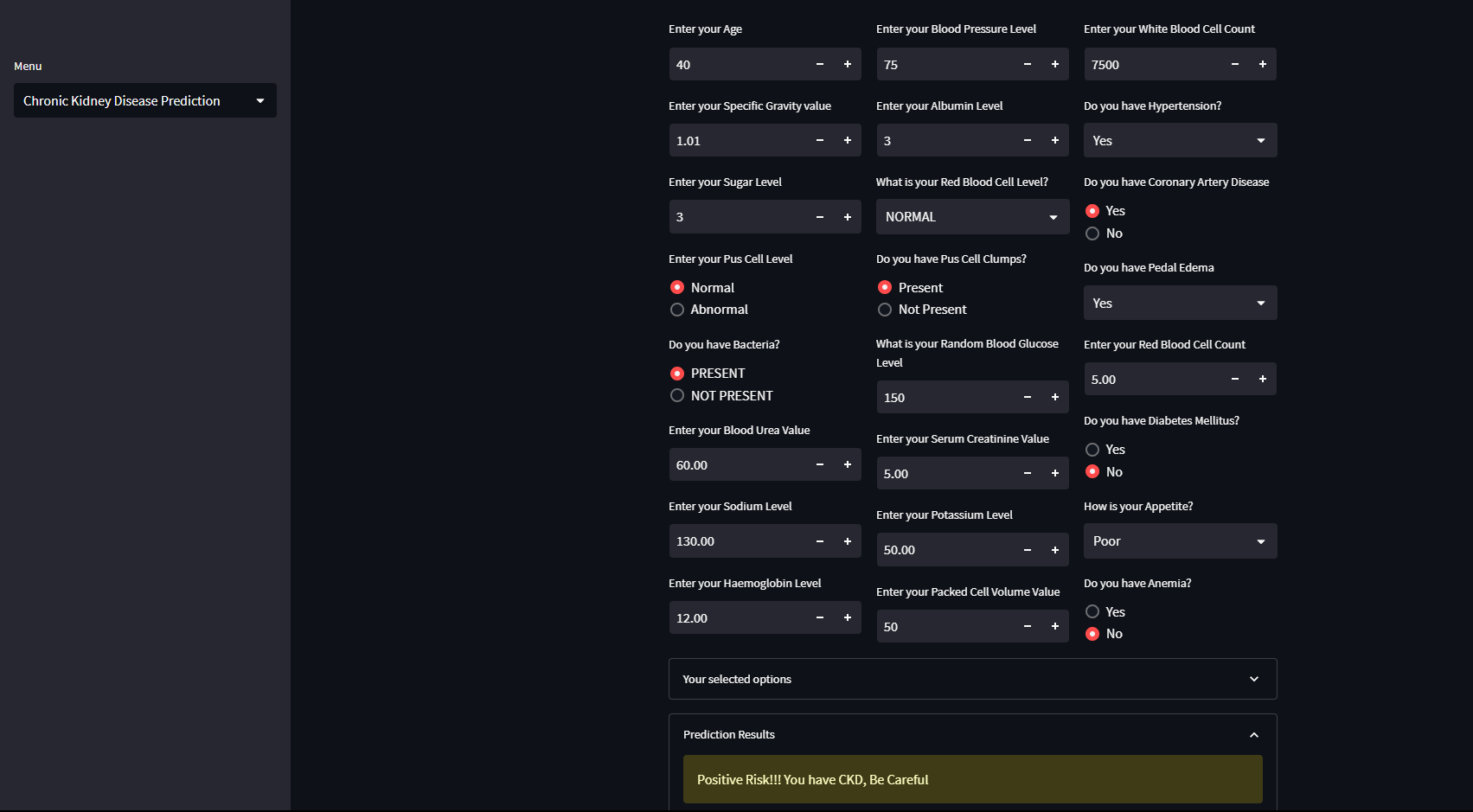
**Fig. k.** Output for heart disease for positive case

Above Fig k, indicates that the person has heart disease.



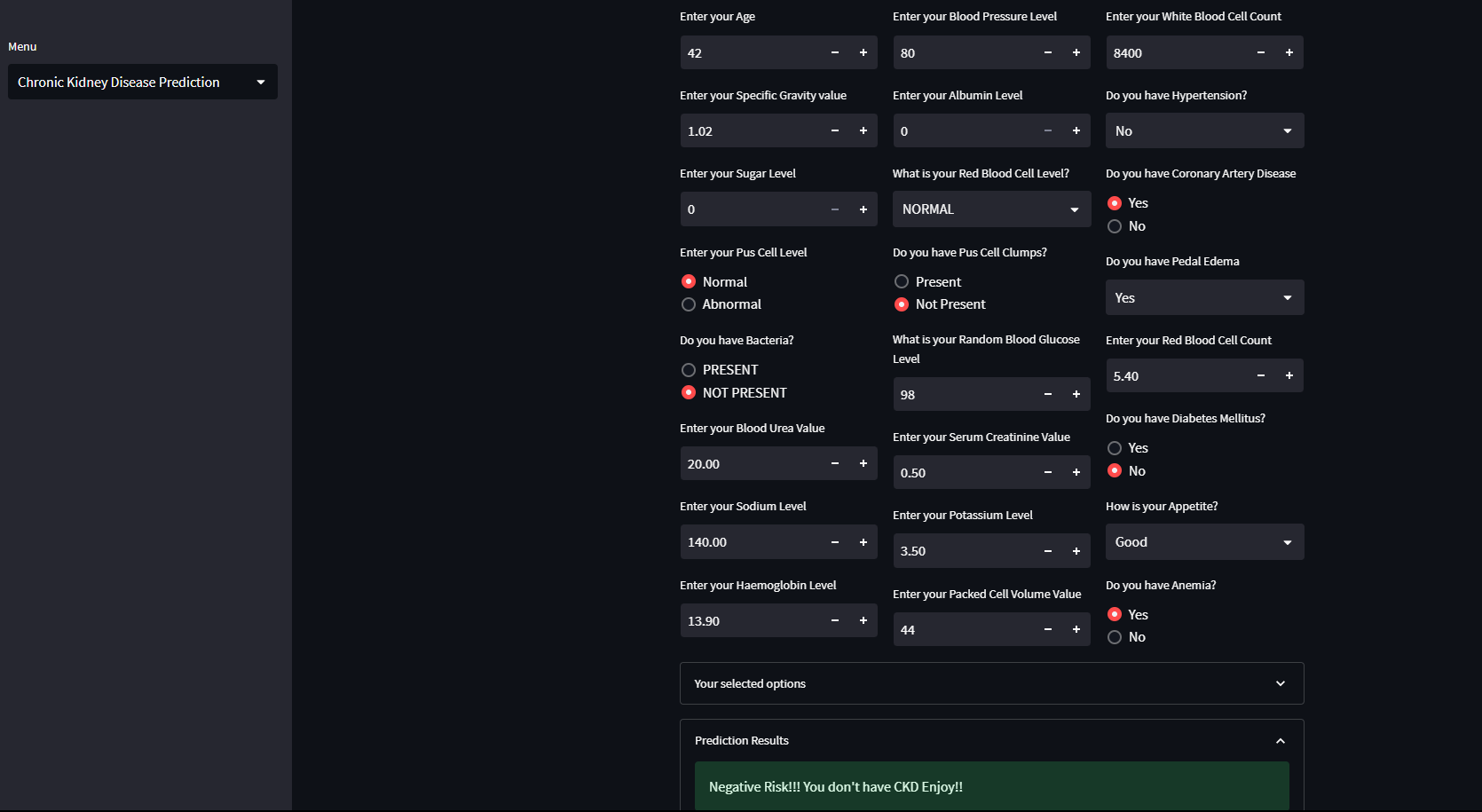
**Fig. l.** Output for heart disease for negative case

Above Fig l, indicates that the person doesn’t have heart disease.

****

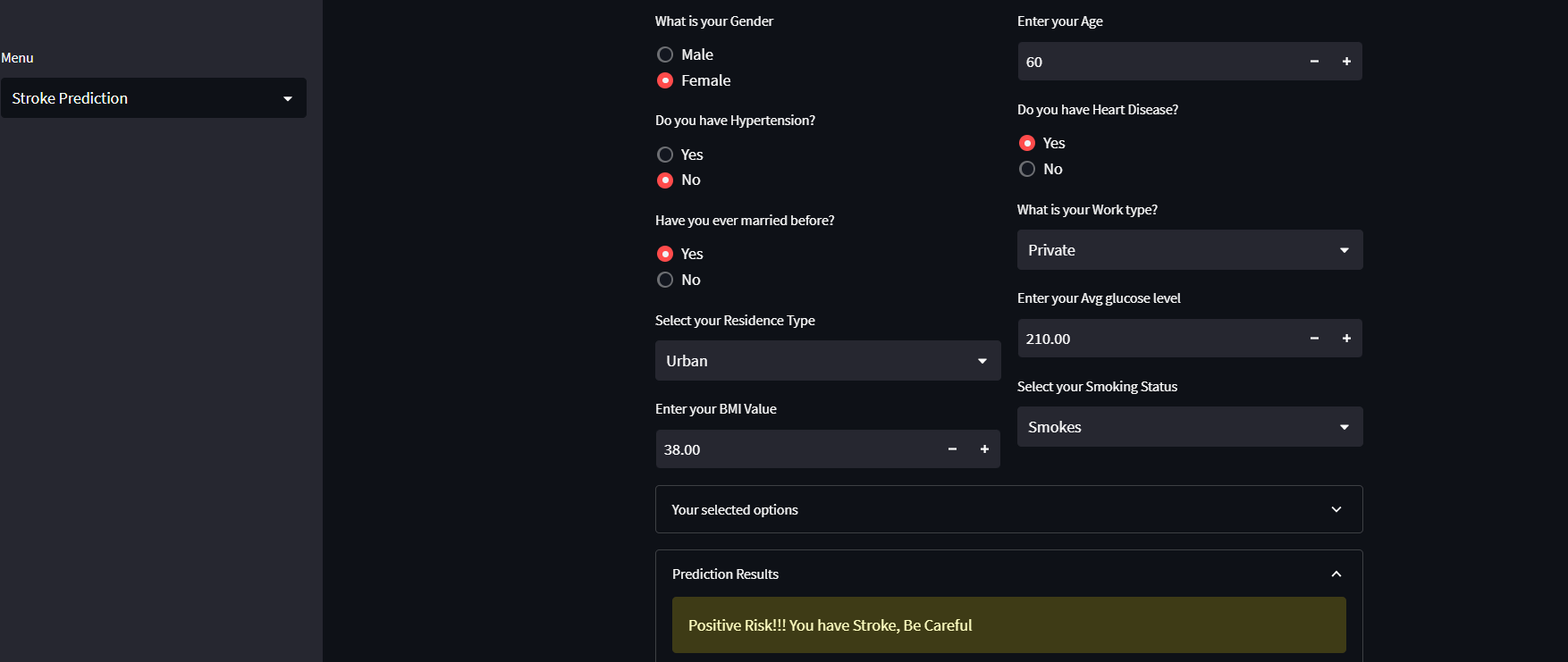
**Fig. m.** Output for chronic kidney disease for positive case

Above Fig m, indicates that the person has chronic kidney disease.



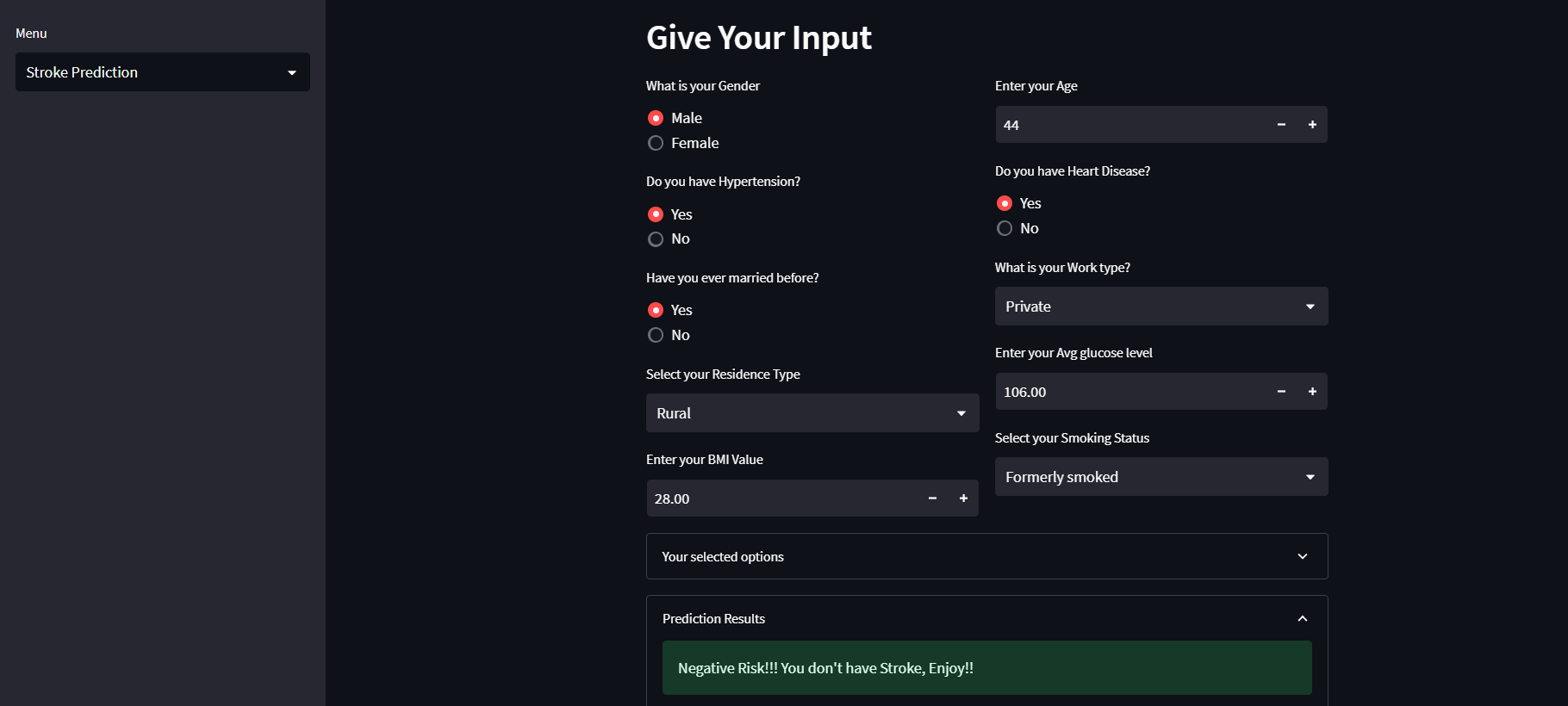
**Fig. n.** Output for chronic kidney disease for negative case

Above Fig n, indicates that the person doesn’t have chronic kidney disease.



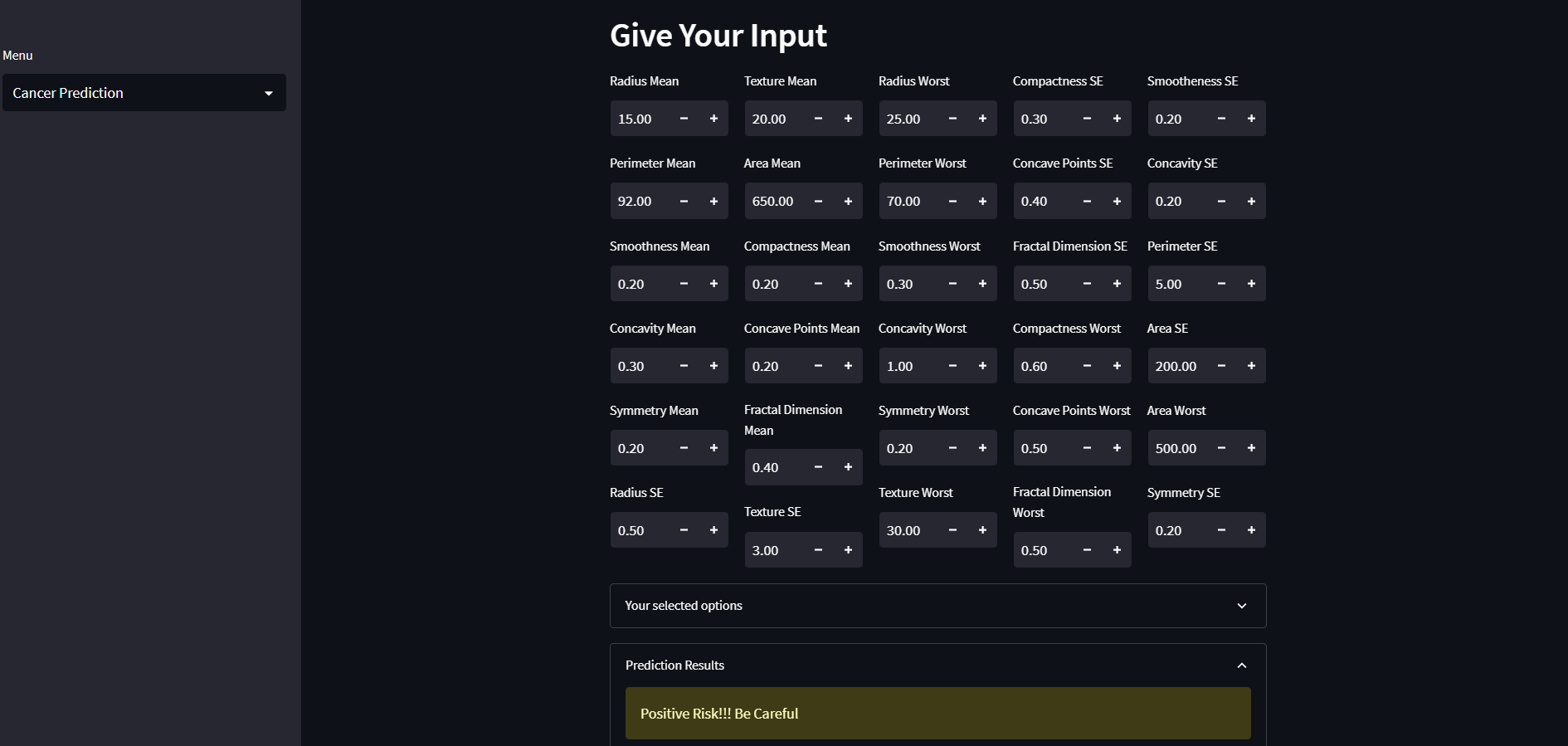
**Fig. o.** Output for Stroke for positive case

Above Fig o, indicates that the person has Stroke.



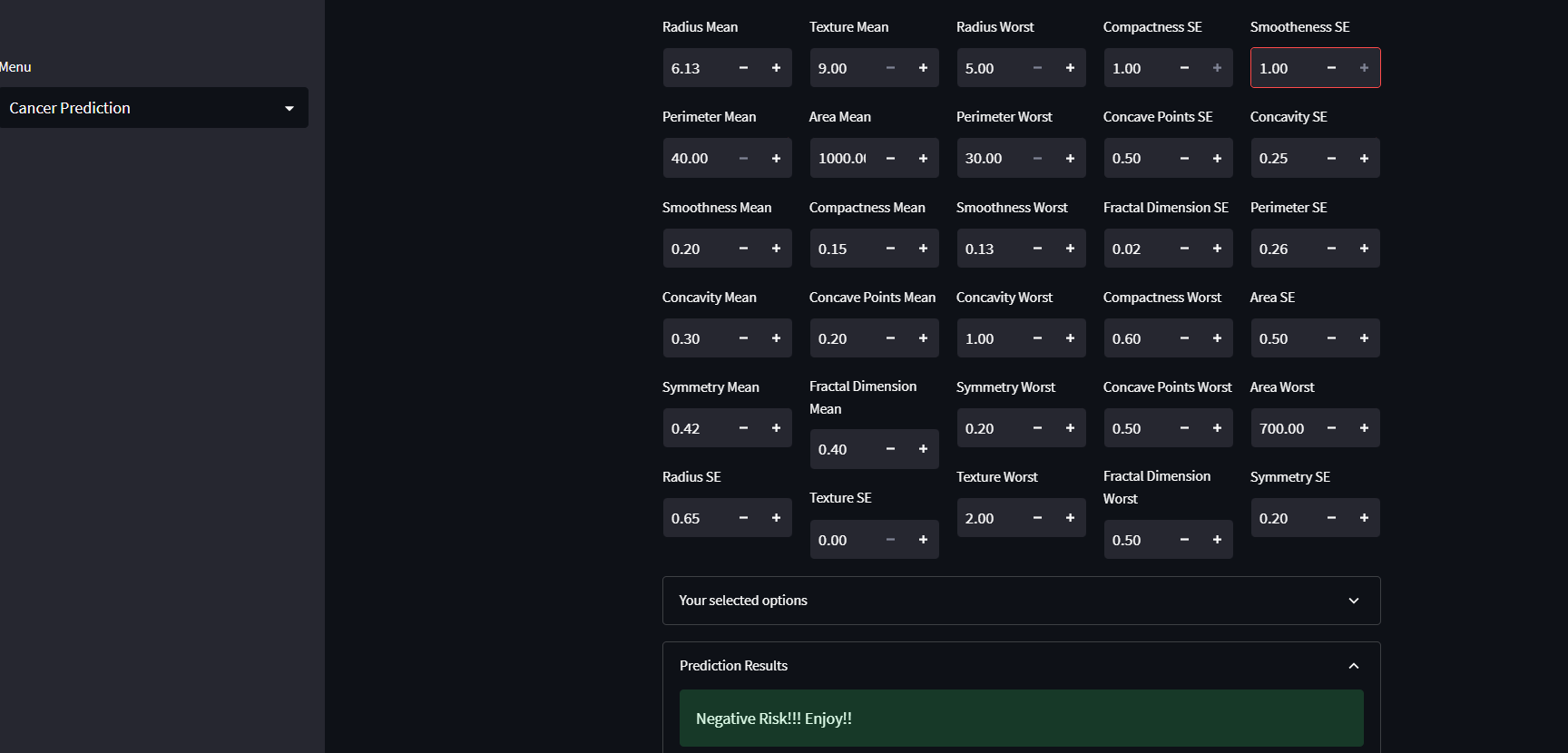
**Fig. p.** Output for Stroke for negative case

Above Fig p, indicates that the person doesn’t have Stroke.



**Fig. q.** Output for cancer, for a positive case

Above Fig q, indicates that the person has Cancer.



**Fig. r.** Output for cancer, for a negative case

Above Fig r, indicates that the person doesn’t have Cancer.

**Conclusion**

Due to modern environmental conditions and lifestyle choices, humans are increasingly vulnerable to a wide range of illnesses. Preventing the worst effects of these diseases requires that they be detected and predicted as early as possible. Manually diagnosing diseases is often a time-consuming and inaccurate process for doctors. In this study, we suggest employing machine learning techniques to detect and predict whether or not a given person has a chronic illness. The suggested system's strength lies in the fact that multiple disease models are integrated into a single user interface, allowing for simultaneous user interaction with each model as needed. The proposed method delivers more accuracy for every disease than pre-existing algorithms, as demonstrated by a comparison of the proposed model's performance to other current works. The suggested approach has the potential to reduce the prevalence of chronic diseases by early diagnosis, and it has the added benefit of lowering the financial burden associated with such an endeavour.

**References**

[1] M. P. Keerthi, G. S. Reddy, V. S. Raghava, and K. B. Reddy, “Streamlit Interface for Multiple Disease Diagnosis,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 2, pp. 1159–1164, 2023, doi: 10.22214/ijraset.2023.49166.

[2] F. J. Shaikh and D. S. Rao, “Prediction of Cancer Disease using Machine learning Approach,” *Mater. Today Proc.*, vol. 50, no. April, pp. 40–47, 2021, doi: 10.1016/j.matpr.2021.03.625.

[3] R. Alanazi, “Identification and Prediction of Chronic Diseases Using Machine Learning Approach,” *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/2826127.

[4] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, “Multiple disease prediction using Machine learning algorithms,” *Mater. Today Proc.*, no. August, 2022, doi: 10.1016/j.matpr.2021.07.361.

[5] I. Mohit, K. S. Kumar, A. U. K. Reddy, and B. S. Kumar, “An Approach to detect multiple diseases using machine learning algorithm,” *J. Phys. Conf. Ser.*, vol. 2089, no. 1, 2021, doi: 10.1088/1742-6596/2089/1/012009.

[6] N. Ahmed *et al.*, “Machine learning based diabetes prediction and development of smart web application,” *Int. J. Cogn. Comput. Eng.*, vol. 2, no. December, pp. 229–241, 2021, doi: 10.1016/j.ijcce.2021.12.001.

[7] G. M. Ifraz, M. H. Rashid, T. Tazin, S. Bourouis, and M. M. Khan, “Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods,” *Comput. Math. Methods Med.*, vol. 2021, 2021, doi: 10.1155/2021/6141470.

[8] B. Shiva Shanta Mani and V. M. Manikandan, “Heart disease prediction using machine learning,” *Handb. Res. Dis. Predict. Through Data Anal. Mach. Learn.*, no. April, pp. 373–381, 2020, doi: 10.4018/978-1-7998-2742-9.ch018.

[9] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012072.

[10] R. A. Godse, S. S. Gunjal, K. A. Jagtap, N. S. Mahamuni, and S. Wankhade, “Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 9, no. 4, pp. 140–144, 2020, doi: 10.17148/IJARCCE.2020.9423.

[11] D. A. Debal and T. M. Sitote, “Chronic kidney disease prediction using machine learning techniques,” *J. Big Data*, vol. 9, no. 1, pp. 4971–4976, 2022, doi: 10.1186/s40537-022-00657-5.

[12] Z. Wang, J. Won Chung, X. Jiang, Y. Cui, M. Wang, and A. Zheng, “Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique,” *Int. J. Eng. Technol.*, vol. 7, no. 4.36, p. 1161, 2018, doi: 10.14419/ijet.v7i4.36.25377.

[13] https://docs.streamlit.io/