**Multi-Disease Prediction System using Machine Learning**

**A Project Report**

***Submitted by***

**Vaibhav Raheja C056**

**Viraj Shah C064**

**Mayank Shetty C075**

**Purav Patel C098**

***Under the Guidance of***

# Prof. Manisha Tiwari

# *in partial fulfillment for the award*

# *of the degree of*

## BACHELOR OF TECHNOLOGY (INTEGRATED)

# COMPUTER ENGINEERING

***At***



**MUKESH PATEL SCHOOL OF TECHNOLOGY**

**MANAGEMENT AND ENGINEERING**

**March 2022**

# DECLARATION

We, **Vaibhav Raheja, Viraj Shah, Mayank Shetty, Purav Patel** roll numbers: **C056, C064, C075, C098** respectively B. Tech Integrated (Computer Engineering), X semester understand that plagiarism is defined as anyone or combination of the following:

1. Un-credited verbatim copying of individual sentences, paragraphs or illustration (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.
2. Un-credited improper paraphrasing of pages paragraphs (changing a few words phrases, or rearranging the original sentence order)
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did write what. ( Source:IEEE, The institute, Dec. 2004)
4. We have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of my work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.
5. We affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such a complaint occurs. I understand fully well that the guide of the seminar/ project report may not be in a position to check for the possibility of such incidents of plagiarism in this body of work.

Signature of the Student:

Name: Vaibhav Raheja

Viraj Shah

Mayank Shetty

Purav Patel

Roll No. C056

C064

C075

C098

Place: Mumbai, Maharashtra

Date: 10th March 202

# Certificate

This is to certify that the project entitled “**Multi-Disease Prediction System using Machine Learning**” is the bonafide work carried out by **Vaibhav Raheja (C056), Viraj Shah (C064), Mayank Shetty (C075), Purav Patel (C098)** of B.Tech Integrated (Computer Engineering), MPSTME (NMIMS), Mumbai, during the Xth semester of the academic year 2021-22, in partial fulfillment of the requirements for the award of the Degree of Bachelors of Engineering, Integrated as per the norms prescribed by NMIMS. The project work has been assessed and found to be satisfactory.



Prof. Manisha Tiwari Internal Mentor



Examiner 1 Examiner 2



Director

# Abstract

Chronic diseases are depicted as conditions that require proceeding with medical consideration, limit everyday exercises, or both. Chronic diseases, for example, heart disease, pneumonia, kidney disease and diabetes are the main sources of death and disability on the planet. It is a critical challenge to detect these diseases by regular clinical data analysis. Early prediction of these chronic diseases would help in saving multiple lives. Machine learning models bring effective solutions for accurate predictions and decision making. Machine learning techniques have shown huge development in the medical industry.

The paper attempts to do the predictive analysis of chronic diseases using machine learning. Chronic Kidney Disease will be predicted using Logistic Regression and Random Forest, Diabetes can be predicted using Logistic Regression (LR) and K-Nearest Neighbour (KNN), Heart Disease can be predicted using Random Forest Regression and Decision Tree and finally for Pneumonia can be predicted using a Convolutional Neural Network (CNN) on chest x-rays. An integrated model of the listed algorithms is proposed to predict the listed chronic diseases.

# Table of Contents

| **CHAPTER**  **NO** | **TITLE** | **PAGE**  **NO.** |
| --- | --- | --- |
|  | List of Figures | ii |
|  | List of Tables | iii |
|  | Abbreviations | iv |
| 1 | INTRODUCTION | 1-3 |
| 2 | LITERATURE REVIEW | 4 |
| 3 | ANALYSIS & DESIGN | 5-8 |
| 4 | IMPLEMENTATION & RESULTS | 9-12 |
| 5 | CONCLUSION & FUTURE WORK   * 1. Conclusion   2. Future Work | 13 |

**List of Figures**

| **CHAPTER**  **NO.** | **TITLE** | **PAGE NO** |
| --- | --- | --- |
| 1 | INTRODUCTION | x |
| 3 | ANALYSIS & DESIGN | x |
| 4 | IMPLEMENTATION & RESULT | x |

# List of Tables

| **CHAPTER**  **NO.** | **TITLE** | **PAGE NO.** |
| --- | --- | --- |
| 2 | LITERATURE REVIEW | 4 |

**Abbreviations**

| **Abbreviation** | **Description** |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Chapter 1 Introduction

## Introduction to the Domain Area/Project

Machine Learning is a branch of Artificial Intelligence (AI), where the main objective is to give the computer the ability to learn from a provided set of data. The structure of the data is understood, after which the data is fit into models. These models can be successfully utilized by people for any given application where machine learning is required.

Despite being a field within computer science, it radically differs from the traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers for calculations or problem-solving, whereas Machine Learning algorithms allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range.

This approach in Machine Learning facilitates computers in building models from sample data, in order to automate decision-making processes based on data inputs.

Our project will be harnessing the potential of machine learning, in which a model will be trained in identifying various diseases included in our scope, where the output will be boolean values.

## Problem Statement

During the COVID-19 Pandemic, several hospitals faced a severe shortage of highly- skilled doctors. Despite the hard times, many doctors fulfilled their duty by being the first line of defense against the widespread pandemic.

Most of these doctors often required an expert system to help them aid in diagnosing a patient’s condition. However, it is necessary to realize that these expert systems may be much more useful if they could personalize the diagnosis for each and every patient with the help of an available doctor.

## Motivation and Purpose of the project

There has been an immense load on hospitals and doctors during the COVID-19 Pandemic, and helping the world to our best ability in these tough times is our goal. Our aim is to reduce the pressure on doctors and hospitals by creating an expert machine that detects several major diseases. With a well-defined scope and target to detect the common diseases, we aim to build a helpful prediction machine that will be of great help to medical professionals.

## An Overview of Expert Systems in Medicine

Expert systems in medicine are defined as systems with the ability to capture and store expert knowledge, facts, and reasoning techniques to assist doctors in diagnosing a patient’s condition. These systems attempt to mimic a doctor’s expertise by applying several computational methods to help in decision support and problem solving, by coming up with reasoned conclusions for a patient’s illness or condition.

Our project will incorporate the core elements of an expert system by supporting medical experts with their claims, precursing their diagnosis of a chronic disease in their patients using trained machine learning model

## Project Scope and Objectives

Our model will have a Disease Detection System that will be based on Machine Learning.Our output will show if the patient has that particular disease and will give relative information related to the disease

The Diseases that we are planning on targeting are:

* Pneumonia
* COVID - 19
* Chronic Kidney Disease
* Heart Attack
* Diabetes

# Chapter 2

# Literature Review

### *Review of Literature*

**Papers on machine learning for Chronic Kidney Disease Detection**

| **Sr No.** | **Paper Name** | **Authors** | **Field of Research** | **Algorithm** | **Dataset** | **Paper Summary** | **Reference** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Performance Analysis of Machine Learning  Classifier for Predicting Chronic Kidney Disease | - Rahul Gupta  - Nidhi Koli  - Niharika Mahor  - N. Tejashri | Chronic kidney disease, Prediction, Machine  Learning, Decision Tree, Random Forest and Logistic  Regression. | - Decision Tree  - Random Forest  - Logistic Regression | UCI Chronic Kidney Disease Dataset | - The paper evaluates the performance of Decision Tree, Random Forest, and Logistic Regression on the preprocessed and filtered Chronic Kidney Disease Dataset.  - Accuracy = Decision Tree (98.48%), Random Forest (94.16%), Logistic Regression (99.24%).  - The authors conclude the paper by stating that Logistic Regression provides highest accuracy and recall, while Decision Tree provides highest precision. | International Conference for Emerging Technology (INCET) |
| 2 | Implementation of Machine Learning Algorithms to  Detect the Prognosis Rate of Kidney Disease | - F. M. Javed Mehedi Shamrat  - Pronab Ghosh  - Mahbubul Hasan Sadek  - Md. Aslam Kazi  - Shahana Shultana | Logistic Regression, Decision Tree, Random  Forest, K-Nearest Neighbors, Accuracy. | - Decision Tree  - Random Forest  - Logistic Regression  - K-Nearest Neighbors (KNN) | UCI Chronic Kidney Disease Dataset | - The paper compares the accurate prediction rates of Chronic Kidney Disease using Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN) over the presented dataset.  - Accuracy = Decision Tree (97.91%), Random Forest (100%), Logistic Regression (100%), KNN (95.82%).  - Random Forest takes the most time to predict, and has the best rating in Receiver Operating Characteristic (ROC) curve. | International Conference for Innovation in Technology (INOCON) |
| 3 | A Machine Learning Methodology for  Diagnosing Chronic Kidney Disease | - Jiongming Qin  - Lin Chen  - Yuhua Liu  - Chuanjun Liu  - Changhao Feng  - Bin Chen | Chronic kidney disease, machine learning, KNN imputation, integrated model. | - Random Forest  - Logistic Regression  - K-Nearest Neighbors (KNN) Imputation  - Integrated Model (Random Forest + Logistic Regression) | UCI Chronic Kidney Disease Dataset | - The paper proposes an integrated model, which combines Logistic Regression and Random Forest, to accurately diagnose Chronic Kidney Disease.  - Accuracy = Logistic Regression (98.95%), Random Forest (99.75%), Integrated Model (99.83%).  - The authors conclude by stating that the integrated model has the highest accuracy and can be further perfected by the increase of size and quality of the data, as it currently cannot diagnose the severity of CKD because only two categories of data samples exist in the dataset (ckd and no ckd). | Institute of Electrical and Electronics Engineers (IEEE) |

**Papers on machine learning for Diabetes Detection**

| **Sr No.** | **Paper Name** | **Author** | **Field of Research** | **Algorithm** | **Dataset** | **Paper Summary** | **Reference** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Diabetes Disease Prediction Using Data Mining | -Deeraj Shetty,  -Kishor Rit,  -Sohail Shaikh,  -Nikita Patil | Diabetes, Prediction,  Naïve Bayes,  KNN. | - Naive Bayes Algorithm  - K-Nearest Neighbours | Pima Indians Diabetes Database | -The Paper recommends using a larger dataset for better prediction  -The larger dataset has proved to  increase the accuracy of the algorithm, therefore  working on some more attributes to better diagnose diabetes. | International Conference on Innovations in Information, Embedded and Communication Systems |
| 2 | Diabetes Prediction using Machine Learning Algorithms | -Aishwarya Mujumdara  -Dr. Vaidehi Vb | Diabetes Mellitus,  Big Data Analytics,  Healthcare  Machine Learning, | - Logistic Regression  - LDA  - Random Forest  - Extra Trees Classifier | Self created Dataset based on Pima Indians Diabetes Database | -In this study, various machine learning algorithms are applied on the dataset and the classification has been done  using various algorithms of which Logistic Regression gives the highest accuracy of 96%.  - Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%. | International Conference On Recent Trends In Advanced Computing |
| 3 | Implementation of a Web Application to Predict  Diabetes Disease: An Approach Using Machine  Learning Algorithm | -Samrat Kumar Dey  -Ashraf Hossain  -Md. Mahbubur Rahman | Diabetes,  SVM,  ANN,  Naïve Bayes,  Min Max  Scaling | - ANN  - Support Vector Machine  - Naive Bayes Algorithm  - KNN | Pima Indians Diabetes Database | - From different  machine learning algorithms Artificial Neural Network  (ANN) provide us highest accuracy with Min Max Scaling  Method on Indian Pima Dataset. | International Conference of Computer and Information Technology |
| 4 | A comparison of machine learning algorithms for diabetes prediction | -Jobeda Jamal Khanam, -Simon Y. Foo | Machine learning,  Data Mining,  K-fold Cross Validation, | - Decision Tree  - K-Nearest Neighbors  - Random Forest  - Naive Bayes  - A/B  - Linear Regression  - Support Vector Machine | Pima Indians Diabetes Database | - The paper evaluates the performance of various algorithms and measures to get a high accuracy model for Diabetes prediction. All models give an average accuracy of 70%.  -LR and SVM have 77-78% accuracy in both K-fold and train-test split  -Among all the proposed models, the NN with two hidden layers is  considered the most efficient and promising for analyzing diabetes | The Korean Institute of Communications and Information Sciences (KICS). |

**Papers on machine learning for Heart Disease Detection**

| **Sr No.** | **Paper Name** | **Author** | **Field of Research** | **Algorithm** | **Dataset** | **Paper Summary** | **Reference** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Implementation of Machine Learning Model to Predict Heart Failure Disease | - Fahd Saleh Alotaibi | Machine learning model, heart failure diagnosis, KNN method. | - Decision Tree  - Naïve Bayes  - Random Forest  - Support Vector Machine  - Logistic Regression | UCI Heart Disease Dataset | - The paper aims to improve the heart failure (HF) prediction using the UCI heart disease dataset which is available on the internet. Multiple machine learning approaches were used to predict HF.  - Accuracy = Decision Tree (93.19%), Random Forest (89.14%), Logistic Regression (87.36%), SVM (92.30%) & Naïve Bayes (87.27%).  - In comparison this study showed significant improvement and higher accuracy than previous work. | International Journal of Advanced Computer Science and Applications (IJACSA) |
| 2 | Heart Disease Prediction using Hybrid machine  Learning Model | - Dr. M. Kavitha  - G. Gnaneswar1  - R. Dinesh  - Y. Rohith Sai  - R. Sai Suraj | Cleveland Heart Disease Database, Hybrid algorithm, Machine learning | - Decision Tree  - Random Forest  - Hybrid (Decision Tree + Random Forest) | Cleveland Heart Disease Dataset | - The interface is designed to get the user's input parameter to predict heart disease, for which the authors used a hybrid model of Decision Tree and Random Forest.  - Accuracy = Decision Tree (79%), Random Forest (81%) & Hybrid Model (88.7%). | Institute of Electrical and Electronics Engineers (IEEE) |
| 3 | Heart Disease Diagnosis using Extreme Learning Based  Neural Networks | - Muhammad Fathurachman  - Umi Kalsum  - Noviyanti Safitri  - Chandra Prasetyo Utomo | Heart Disease, Extreme Learning  Machine, Medical Diagnosis. | - Artificial Neural Networks  - Support Vector Machine  - Decision Tree  - Extreme Learning Machine | - Cleveland Heart Disease Dataset  - Hungarian Institute of Cardiology Dataset  - University Hospital Zurich Dataset  - Medical Centre Long Beach Dataset  (All available on UCI Machine Learning  Repository) | - The authors divided the experiments into five parts. For each part, we used a different training and testing dataset which had been divided using K-Fold Cross Validation. Then we set the hidden node of our ELM model for predicting heart disease  - Accuracy = Decision Tree (75%), ELM (80%), SVM (68%), BP ANN (77%)  - From the experimental results and analysis, the authors concluded that the performance of the ELM algorithm tends to be better when compared with SVM, DT and BP ANN. With an average of 83% accuracy, 88% sensitivity and 82% specificity, the application of ELM algorithm can be an alternative solution to help clinicians predict heart disease. | International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA) |

**Papers on machine learning for Pneumonia:**

| **Sr No.** | **Paper Name** | **Author** | **Field of Research** | **Algorithm** | **Dataset** | **Page Summary** | **Reference** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Pneumonia Detection Using CNN based Feature Extraction | -Dimpy Varshni  -Rahul Nijhawan  -Kartik Thakral  -Ankush Mittal  -Lucky Agarwal | DensetNet, Deep Convolutional Neural  Networks, SVM, Transfer Learning, Random Forest, Naive  Bayes, K-nearest neighbors, Feature extraction. | - CNN  - SVM | -The dataset used is ChestX-ray14 released by Wang et al. (2017) also publicly available on the Kaggle | - The authors have used a customised model which is a combination of CNN based feature extraction and supervised classifier algorithm. - The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. The C parameter trades off correct classification of training examples against maximisation of the decision function’s margin. The Area Under the ROC curve (AUC) is an aggregated metric that evaluates how well a logistic regression model classifies positive and negative outcomes at all possible cutoffs. It can range from 0.5 to 1, and the larger it is the better. - Results = SVM (rbf kernel), C = 3.5, gamma = 2e-05 & AUC = 0.7904 | Institute of Electrical and Electronics Engineers (IEEE) |
| 2 | Pneumonia Detection using CNN with Implementation in Python | -Muhammad Ardi | Computer science undergraduate student of Universitas Gadjah Mada | - CNN | -Chest X-Ray Images (Pneumonia) | - This paper successfully detected pneumonia using CNN  - The model is able to predict pneumonia caused by bacteria pretty well since 232 out of 242 samples are classified correctly  -Which gives a accuracy of 95.86% | Pneumonia Detection using CNN with Implementation in Python |
| 3 | Detection of Pneumonia using ML & DL in Python | -A Sharma,  -M Negi,  -A Goyal,  -R Jain,  -P Nagrath | Neural network, confusion matrix, keras, recall, hyper parameters | - CNN | -Chest X-Ray Images (Pneumonia) | -Detection of diseases with the assistance of computers from various Machine and Deep learning techniques are very beneficial in such places where there is shortage of people who are skilled in techniques like radiology Validation Accuracy 0.8410 and Validation Loss 0.8395  -Accuracy 0.9806 and Loss 0.0742 | IOP Conference Series: Materials Science and Engineering |

Machine learning algorithms have not only proven to be an essential tool in the field of mathematics and engineering, but also in the field of medicine as well.

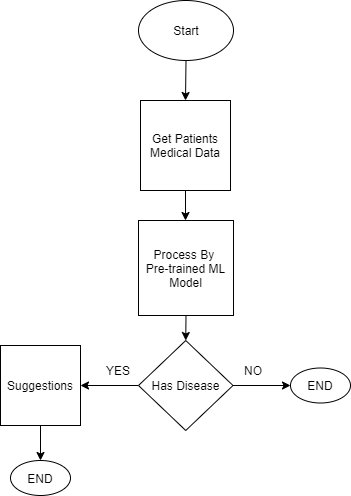
The authors in paper [1] survey various machine learning algorithms and evaluate their accuracy on the Pima Indian Diabetes Database. Among the various models, Logistic Regression and Support Vector Machine provided an accuracy of 78% in both K-fold and test-train split. All models give an average accuracy of 70%. In paper [3], the authors have used a larger dataset based on the same parameters as the PIMA Indian Diabetes dataset which gave the accuracy of 96% for LR.The authors in paper [5] design an interface to get the user's input parameters and predict whether the user is healthy or has a probability of getting a heart disease. With Decision tree the authors achieved an accuracy of 79% and with Random Forest they the accuracy achieved was 81%. Along with these two algorithms the authors experimented with a Hybrid model which was a combination of Decision Tree and Random Forest. The Hybrid model was the most efficient and had the highest accuracy between the three algorithms with 88.7%.

In paper [7], the authors establish different machine learning algorithms to diagnose chronic kidney disease using the UCI Chronic Kidney Disease dataset using their proposed model. The performance of several high-accuracy algorithms was analysed, after which only two were shortlisted for an integrated model, namely Logistic Regression and Random Forest. The accuracy of both these algorithms were 98.95% and 99.75% respectively, while the accuracy of the integrated model was 99.83%, which was higher than the individual accuracies of the algorithms used in the integrated model.

The authors of paper [11] utilised a pneumonia dataset from Kaggle, which consisted of chest X-ray images of several patients. The pre-processing of the dataset consisted of resizing the images and creating a function which stores the pre-processed images. The images of the chest X-ray were converted to an array of numbers, after which the authors used Convolutional Neural Network (CNN) to train the model to detect pneumonia, which ultimately provided an accuracy of 98.06%.

# Chapter 3 Analysis & Design

### *Proposed System*

******

*Fig 3.1: Flow Chart for disease detection model*

* + 1. ***Get Patients Medical Data***

This state deals with the collection of input parameters for the pre-trained model. A specific set of biological parameters are collected and fed to the pre-trained machine learning model, depending on the selected chronic disease.

* + 1. ***Process By Pre-Trained Machine Learning Model***

This stage contains a pre-trained model for each chronic disease considered in our project. Appropriate features are selected for each chronic disease, which helps in further classification of a patient having or not having the specified chronic disease.

* + 1. ***Suggestions***

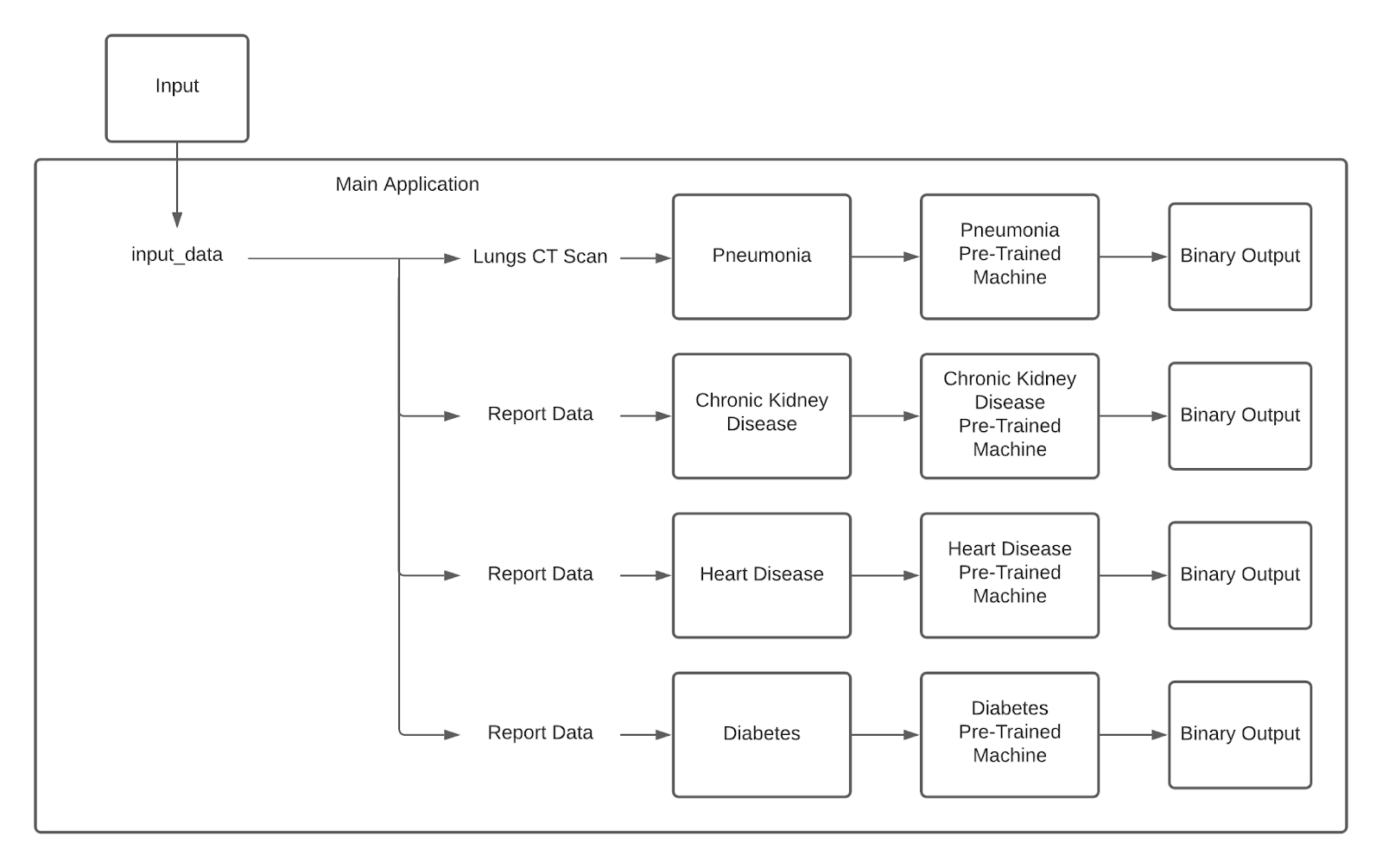
This stage is the output of the pre-trained model for the disease. In the event of a positive result, appropriate suggestions are provided to the medical expert, as there is a high probability of the patient having a chronic disease. If the result is negative, the patient has a low probability of the patient having a chronic disease.

* + 1. ***Experimental setup***

We have used a different dataset to conduct supervised learning on them to find the best of the classifier for each of the diseases. The attributes of the dataset consist of the medical features that are required to diagnose the chronic disease. We have worked on feature reduction to select the best features for the classification. For Chronic Kidney Disease we have used the UCI Chronic Kidney Disease Dataset following the guidelines provided by UCI. For Diabetes we have used the PIMA Indian diabetes. The UCI heart disease dataset is used for detection of heart disease. The researchers of Qatar University have compiled the COVID-QU-Ex dataset, which consists of 33,920 chest X-ray (CXR) images including: 11,956 COVID-19, 11,263 Non-COVID infections (Viral or Bacterial Pneumonia) and 10,701 Normal images of chest X-rays.

Once the models for detection are trained, we have used Anvil to provide the frontend for the prediction system. Anvil is a platform that can be used for full stack development and deployment of machine learning models with python based back end and web based front end.

### *Architecture Diagram*

******

*Fig 3.2: Architecture Diagram for Expert System*

***Disease Detection System***

The user will first decide which disease machine they want to use and upon selecting the disease, the data is input into Disease Detection Systems' Pre-Trained Machine model. The machine will then give a binary output, stating whether the patient has the selected disease, or they are healthy.

The Pre-Trained Machines will be trained on a pre-processed dataset. The dataset will go through multiple processing steps before it will be fed to the machine learning models to train them. The data will be checked for any noisy and missing data. K Nearest Neighbor (KNN) will be used for imputing missing values, outlier detection methods will be used to estimate any noise in the data and rapid miner will be used to remove noise in the dataset. The dataset will also be checked for discrepancies, data transformation, discretization and binning techniques will be used.

## *3.3 UML Diagram*

Chart

Description automatically generated with medium confidence

*Fig 3.3: Sequence Diagram for CVD Detection*

A Sequence Diagram depicting the sequential events of cardiovascular disease (CVD) detection system has been illustrated in Figure 3.3.

The patient’s medical data consisting of 13 parameters collected from various tests are entered into the user interface by the medical expert, where the model predicts the probability of the patient having a CVD. The model will rely on values from the data set to predict the probability of a CVD, based on the set of data entered by the user.

***3.4 Analytical Model***

As we have proposed a system for predicting multiple chronic diseases we have experimented with Diabetes, Chronic Kidney Disease (CKD), Heart Diseases, COVID-19, and Pneumonia. The datasets that are procured are used in accordance with the recommendations and guidance provided by the organisations.

The main objective of our system is to reduce the diagnostic time taken and in turn help the doctor administer treatment at the earliest. To measure the correctness of the models that are used for the prediction we would be computing the Confusion matrix of models, A Confusion matrix consist of true positives, true negatives, false positives, and false negatives,

True Positives are the patients having the disease that are rightly diagnosed by the system. True negatives are the patients that are not having the disease that is rightly diagnosed by the system. False Positives are the patients not having the disease that are diagnosed by the system to have the disease. False negatives are the patients that have the disease but are wrongly classified as not having the disease. Our attempt is to increase the True positives and True negatives while we are reducing False positives and False negatives to a minimum.

From the values procured from the confusion matrix we can find the Accuracy, Misclassification rate, Recall and Precision of the model, helping us judge and compare the various machine learning models.

1. *Accuracy*

The classifier's accuracy is expressed as the number of samples that were correctly corrected out of a total number of samples, and it is represented as follows:

1. *Misclassification Rate*

It is defined as the classifier's error rate that is the case where classification is not at par with reality.

1. *Recall*

It is defined as the classifier’s predicting the positive values to actual positive values.

1. *Precision*

It is defined as the classifier's rate of predicting the positive values to the total positive values.

# Chapter 4 Implements and Results

## Machine learning algorithm for Chronic Kidney Disease Detection

**Logistic Regression**: Logistic Regression is a statistical machine learning technique used for binary classification of categorical dependent variables, by using a set containing independent variables. Assuming the probability of a particular class, Logistic Regression creates a regression model that distinguishes between several samples using the sigmoid function, where one or more variables are used to determine the expected outcome in probabilistic values which lie between 0 and 1. Logistic Regression has previously been implemented to detect chronic kidney disease in patients, and from referenced papers it was found to have a detection accuracy of 98.95% [7], 99.24% [8], and 100% [9] respectively on the UCI dataset, making it a high-accuracy machine learning model for implementation of the expert system.

**Random Forest**: Random Forest is a supervised learning classification technique and is often considered as an ensemble machine learning method, as it is used for classification, regression, and probability. Random Forest can find missing values from many datasets, and it can provide a more accurate value by creating a forest of decision trees during the learning phase, where the number of trees indicate the robustness of the forest as well as the accuracy of the algorithm. Random sampling is used to train the characteristics for sampling nodes, by combining several decisions to form a single decision. The algorithm consolidates multiple forests on different subsets of a dataset and averages the results to enhance the performance of the dataset’s detection accuracy. The accuracy of the machine can be further improved by combining several classifiers.

Step 1: Select 'm' random features from a total of 'n' features, such that m << n.

Step 2: Use the best split to calculate node 's' from the selected 'm' features.

Step 3: Use best split to further split node ‘s’ into daughter nodes.

Step 4: Repeat until the number ‘1’ is reached.

Step 5: Repeating the preceding processes ‘n’ number of times, until a forest of 'n' trees is built.

Random Forest has been chosen as it comparatively takes less time to train and has also been used to detect chronic kidney disease in patients from our referenced papers, where it was found to have a detection accuracy of 94.16% [8], 99.75% [7], and 100% [9] respectively on the UCI dataset, making it a suitable high-accuracy machine learning model for the implementation of the expert system. It can be combined with a Logistic Regression classifier to form an Integrated Model, which offers an average detection accuracy of 99.83%, which is better than the average detection accuracies of both the algorithms combined [7].

## Machine learning algorithm for Diabetes Detection

**K-Nearest Neighbour**: KNN algorithm arranges another information guide in view of comparability toward accessible information. This infers that new information can be handily arranged into a classification of values utilising the KNN algorithm. KNN received an accuracy of 79% [1] and 79% [3] in detection of diabetes on PIMA dataset.

Step 1: Select a value from the dataset.

Step 2: Determine the value of K, such that it is the value of the nearest data point.

Step 3: For each of the K in test data do:

* Figure the distance between every K and each column of the preparation information.
* Arrange the distances in ascending order.
* Select top n columns from the arranged cluster
* Allot a class to the test point in view of the most frequent class of the row.

**Logistic Regression**: Logistic Regression predicts the result of a categorically reliant variable, where the result should be a categorical or discrete value. The output will have probabilities of either 0 or 1. Logistic regression is a highly efficient method for linear classification as well as binary classification problems like the problem at hand. Logistic Regression receives an accuracy of 78% [1] in detection of diabetes on PIMA dataset. Hence, it can be used for a high-accuracy machine learning model to detect diabetes in patients. Pre-processing the data using K-means and clustering can also increase the accuracy to 96% [3], making it a suitable model for implementation of the expert system.

## Machine learning algorithm for Heart Disease Detection

**Decision Tree**: A classification model which is like a tree. It builds a structure of nodes and branches from the evidence collected during the learning phase of the model for each attribute. The connection of the nodes and branches is determined by the number of entities in the dataset. Each attribute has a set number of values used by the forwarding process. Decision for each transaction is reached by following the rules described on each node and branch. At last, the class label will be assigned to the records according to the decision node. This procedure has multiple iterations and will repeat till a class category is assigned to each transaction. In all, attributes are converted to nodes and branches and one of them is selected at the decision.

Step 1: Import dependencies like the libraries and dataset

Step 2: Split the dataset into a Train set and a Test set for learning and validation.

Step 3: Use the train set to train a Decision Tree model

Step 4: Use the Test set to predict results using the trained model.

Step 5: Compare the Actual Values with Predicted Values

Step 6: Visualise the Results of the Decision Tree Regression model.

We are going to use the decision tree for Heart disease detection, as the outputs of the referenced papers [6] had good accuracy, while being easy to read and interpret without requiring any statistical knowledge. It takes less effort to prepare the data, and once the variables have been created, the data cleaning process is minimal. The Decision Tree algorithm achieved an accuracy of 79% [5].

**Random Forest Regression**: Aggregate of multiple decisions is taken in Random Forest Regression. A complex problem is solved by combining multiple classifiers and that also helps improve the machine's accuracy. Amidst the learning phase a forest is generated which consists of multiple trees. This classifier consolidates multiple forests for different subsets of a dataset and averages out the results to increase the accuracy of the machine's detection capabilities.

Step 1: Import dependencies like the libraries and dataset

Step 2: Split the dataset into the Train set and Test set for learning and validation.

Step 3: Use the train set to train the Random Forest Regression model.

Step 4: Use the Test set to predict results using the trained model.

Step 5: Compare the Actual Values with Predicted Values

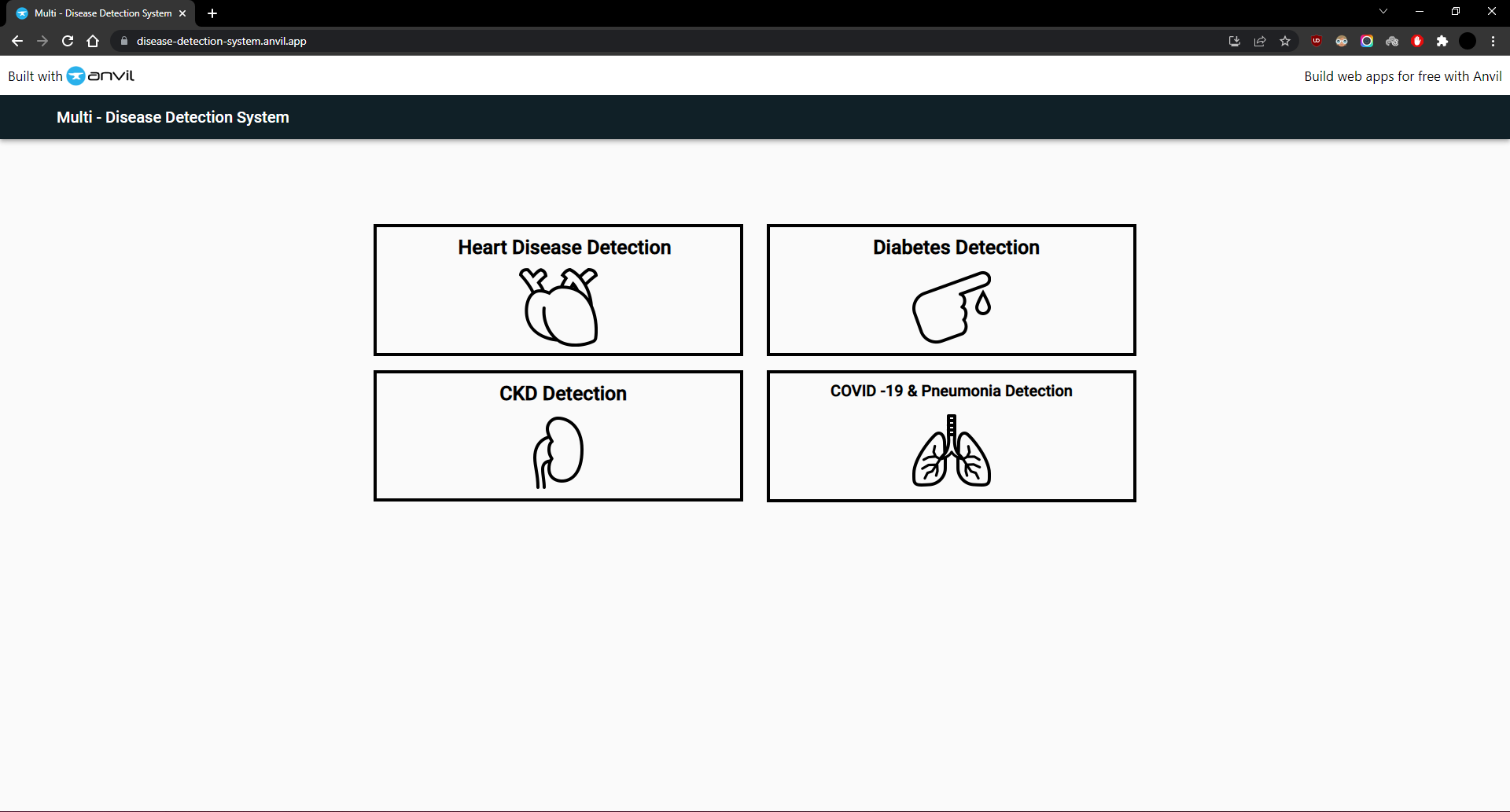
Step 6: Visualise the Random Forest Results

Random Forest Regression was selected from the referenced papers [6], as it comparatively takes less time to train while providing high accuracy, which increases the efficiency of the model. Random Forest Regression achieved an accuracy of 81% [5].

## Machine learning algorithm for Pneumonia and COVID-19 Disease Detection

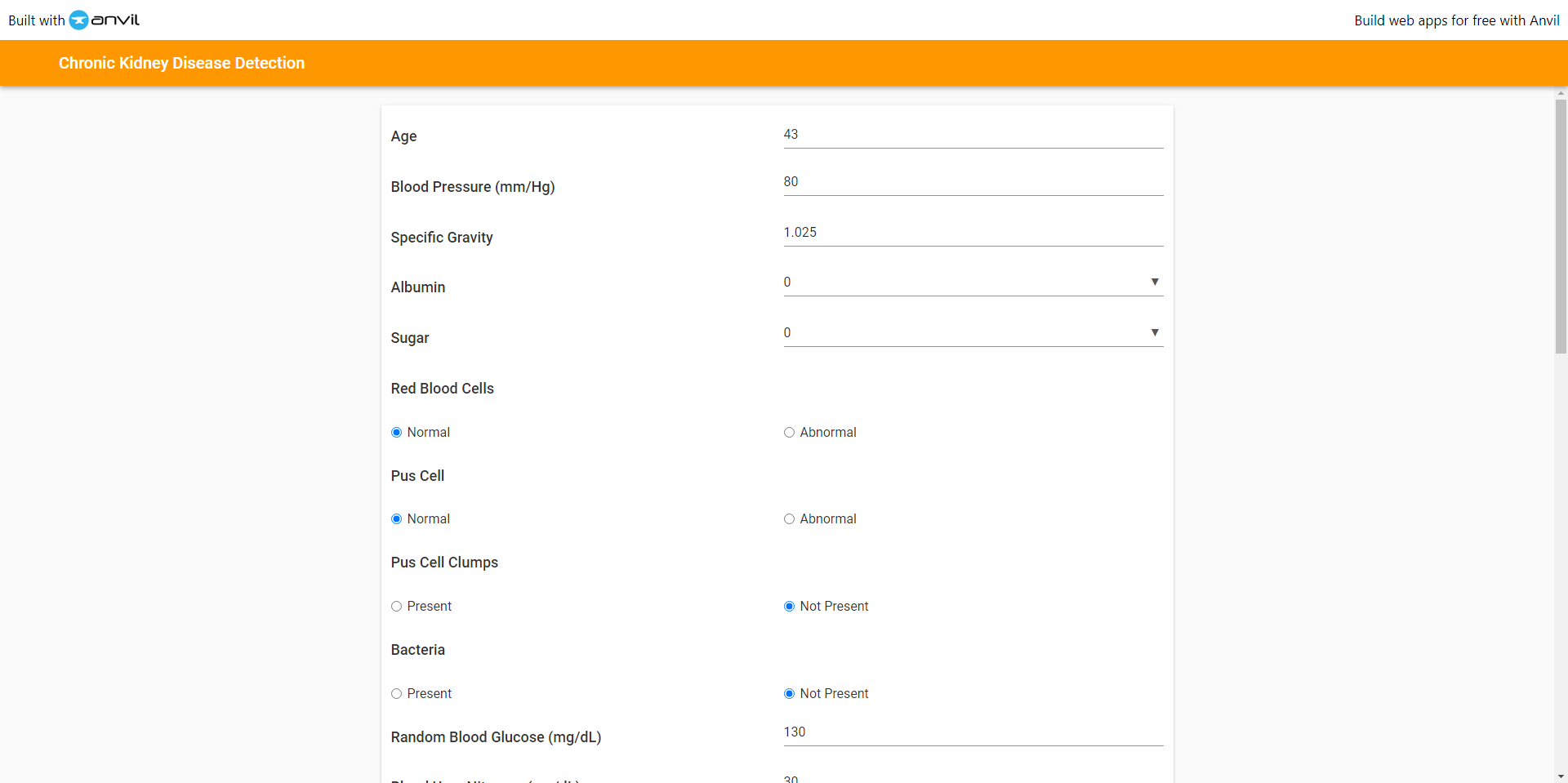
Convolutional neural network: The system figures out how to perform feature extraction in a convolutional neural network. The central idea of CNN is to utilize convolution of images and filters to create invariant features which are pushed to the following layer. The qualities of the following layer are joined with various filters to create more invariant and abstract features. This work is done till the last component/yield (say face of X) is accomplished. Convolutional neural networks comprises a few structure components, for example, pooling layers, convolution layers, and completely associated layers, which are intended to gain proficiency with the spatial ordered progressions of elements by using a background application.

## UI Implementation of Disease Detection

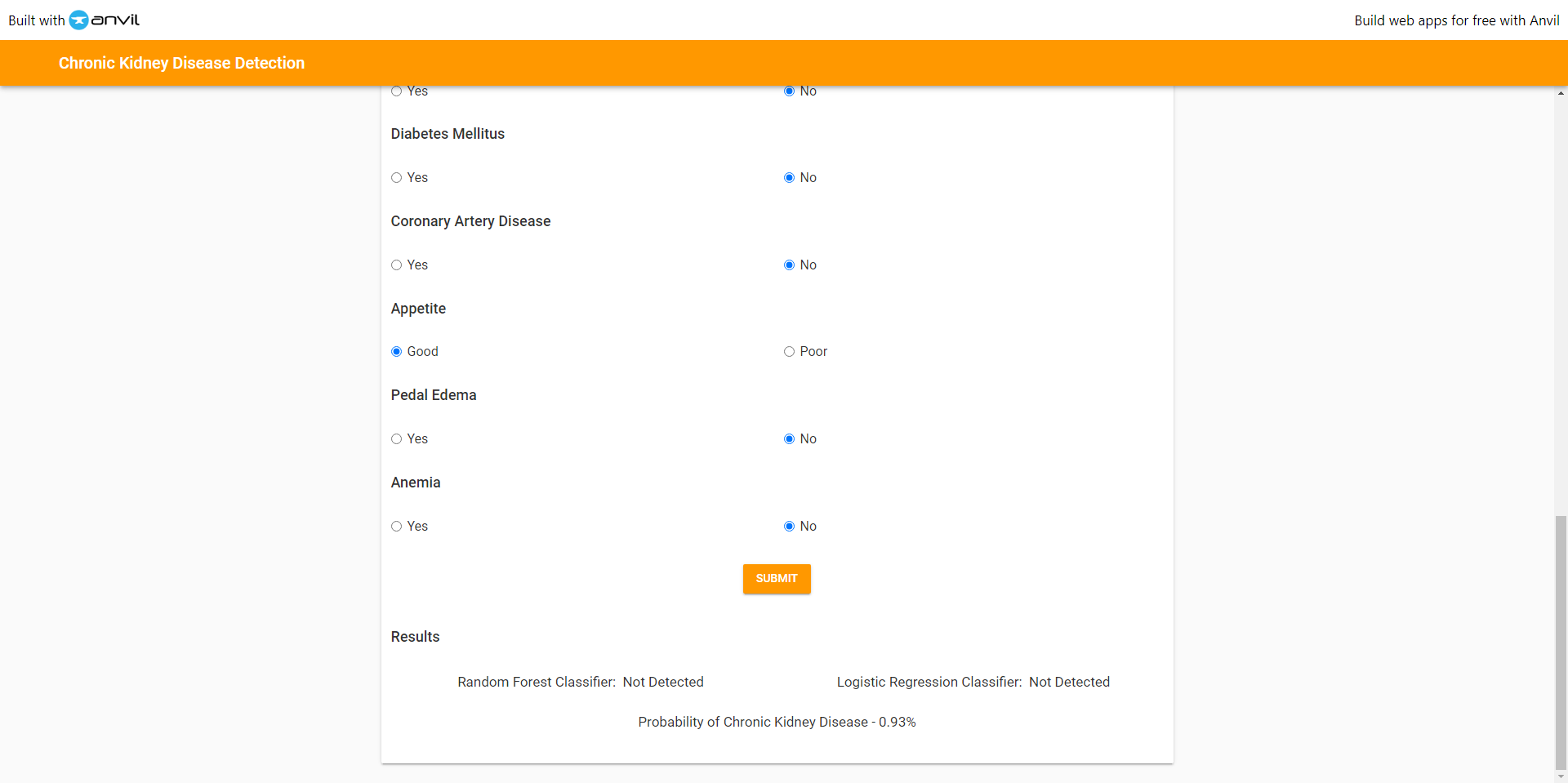


*Fig 4.5: UI for Disease Detection System*

* + 1. ***UI Implementation for Chronic Kidney Disease Detection***

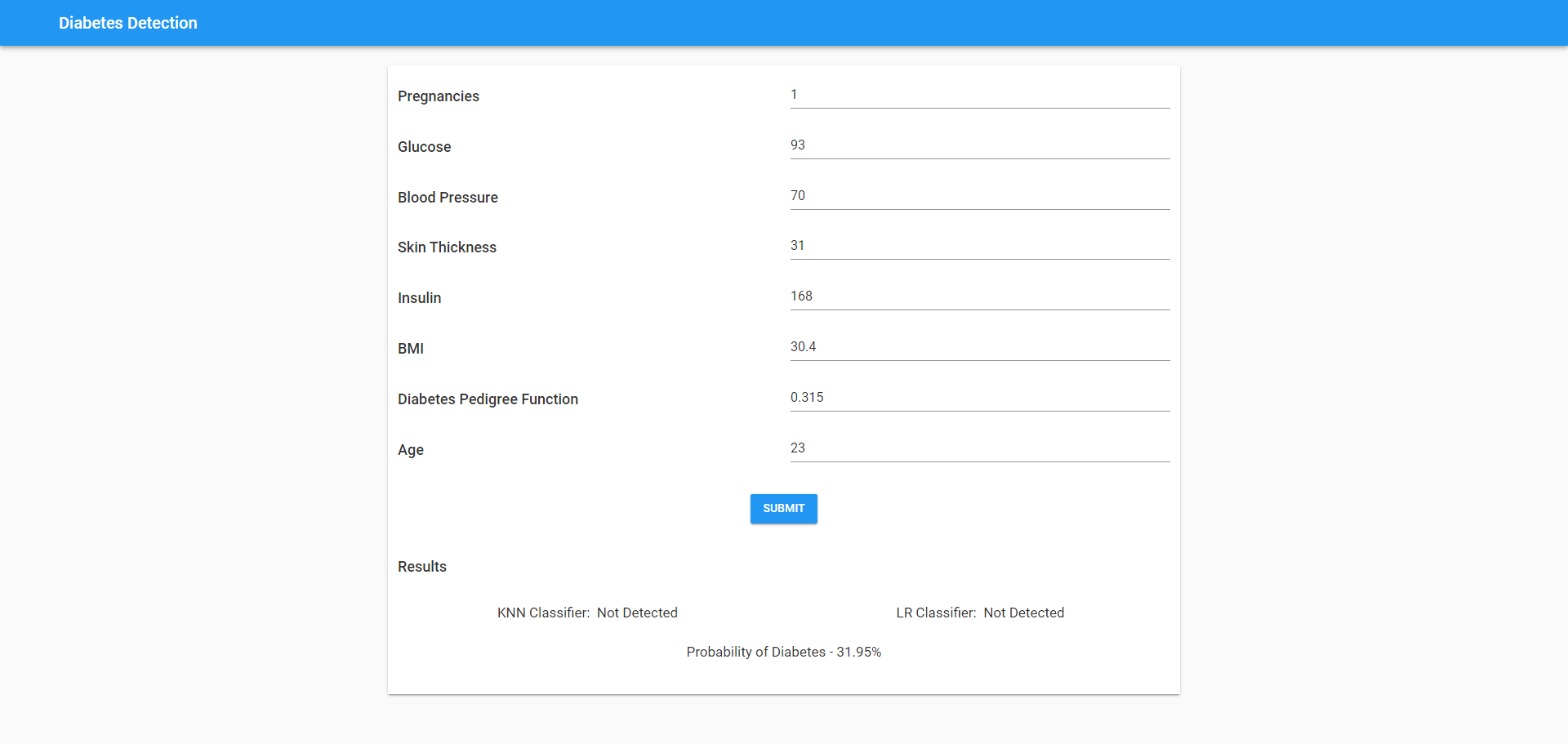


*Fig 4.5.1.1: UI for Chronic Kidney Disease Detection System*



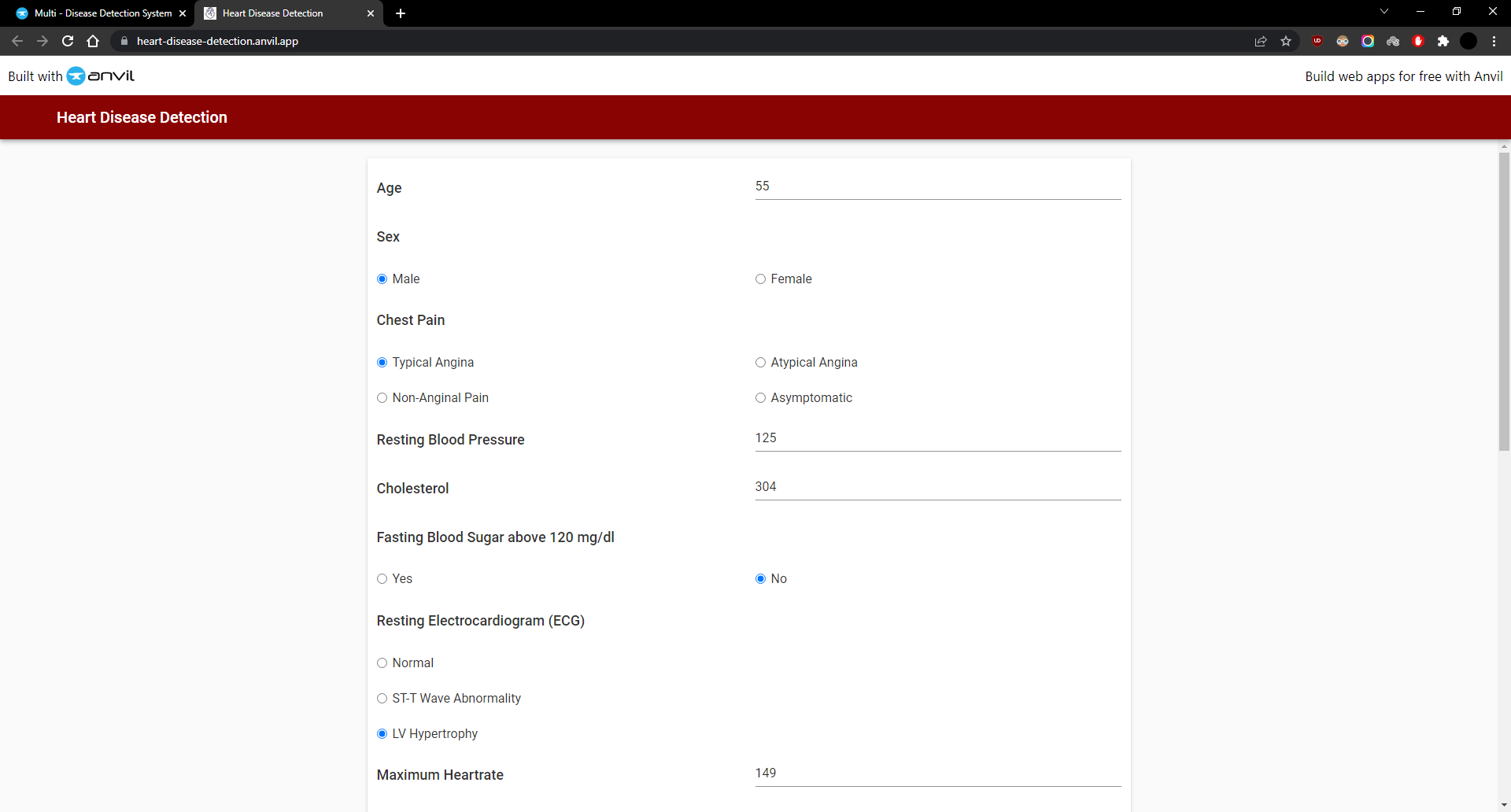
*Fig 4.5.1.2: UI for Chronic Kidney Disease Detection System*

* + 1. ***UI Implementation for Diabetes Detection***

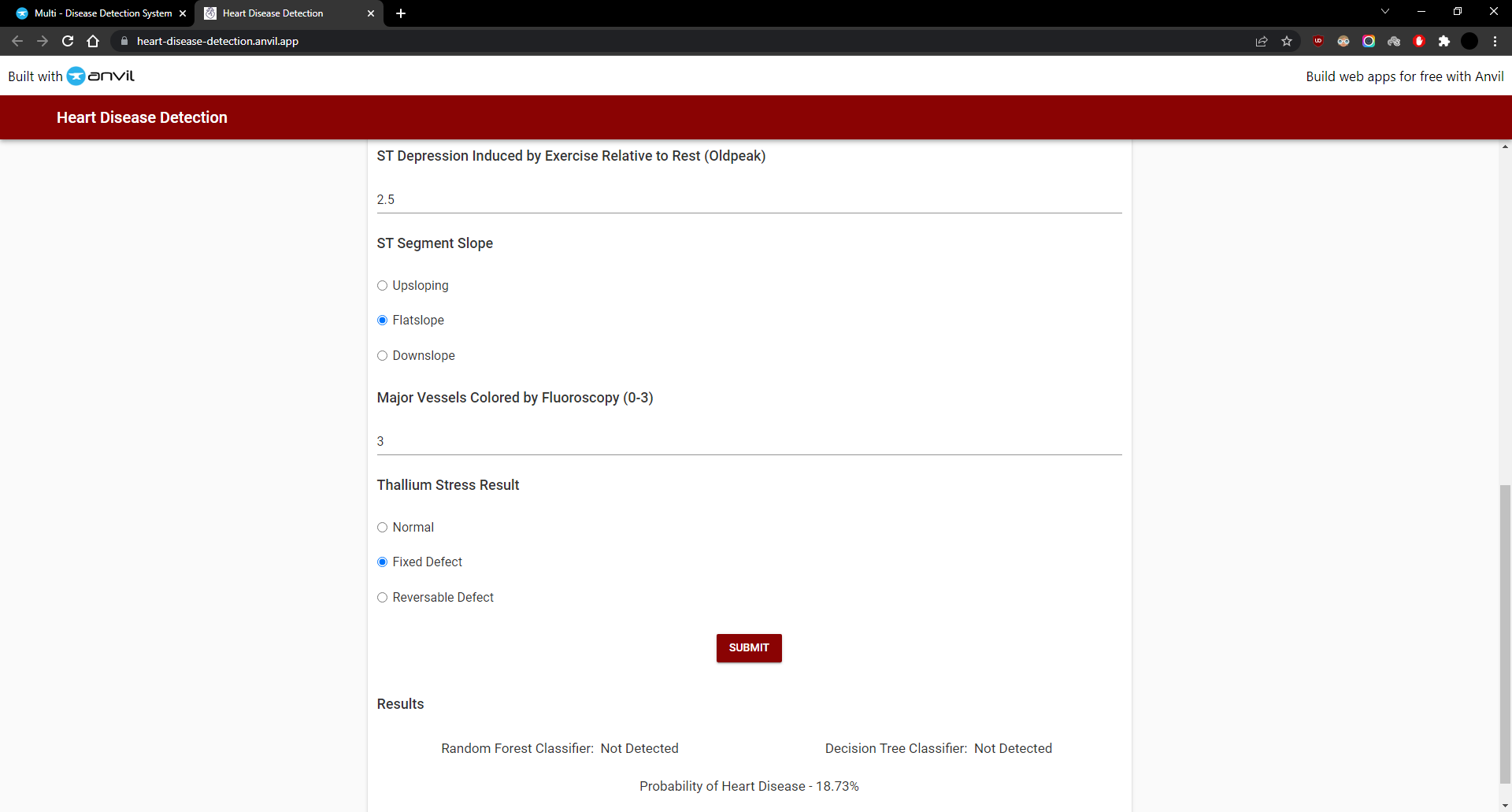


*Fig 4.5.2: UI for Diabetes Detection System*

* + 1. ***UI Implementation for Heart Disease Detection***

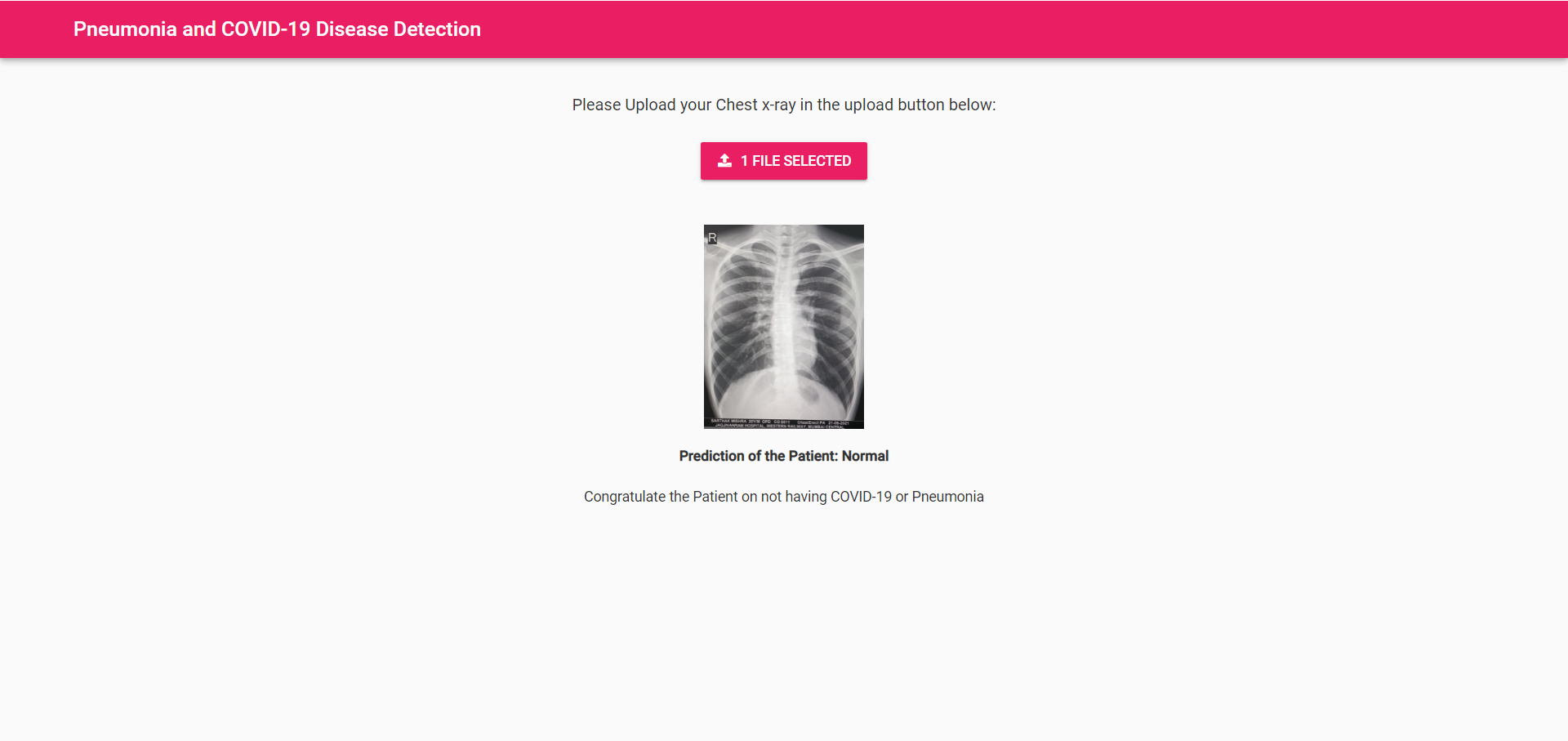


*Fig 4.5.3.1: UI for Heart Detection System*

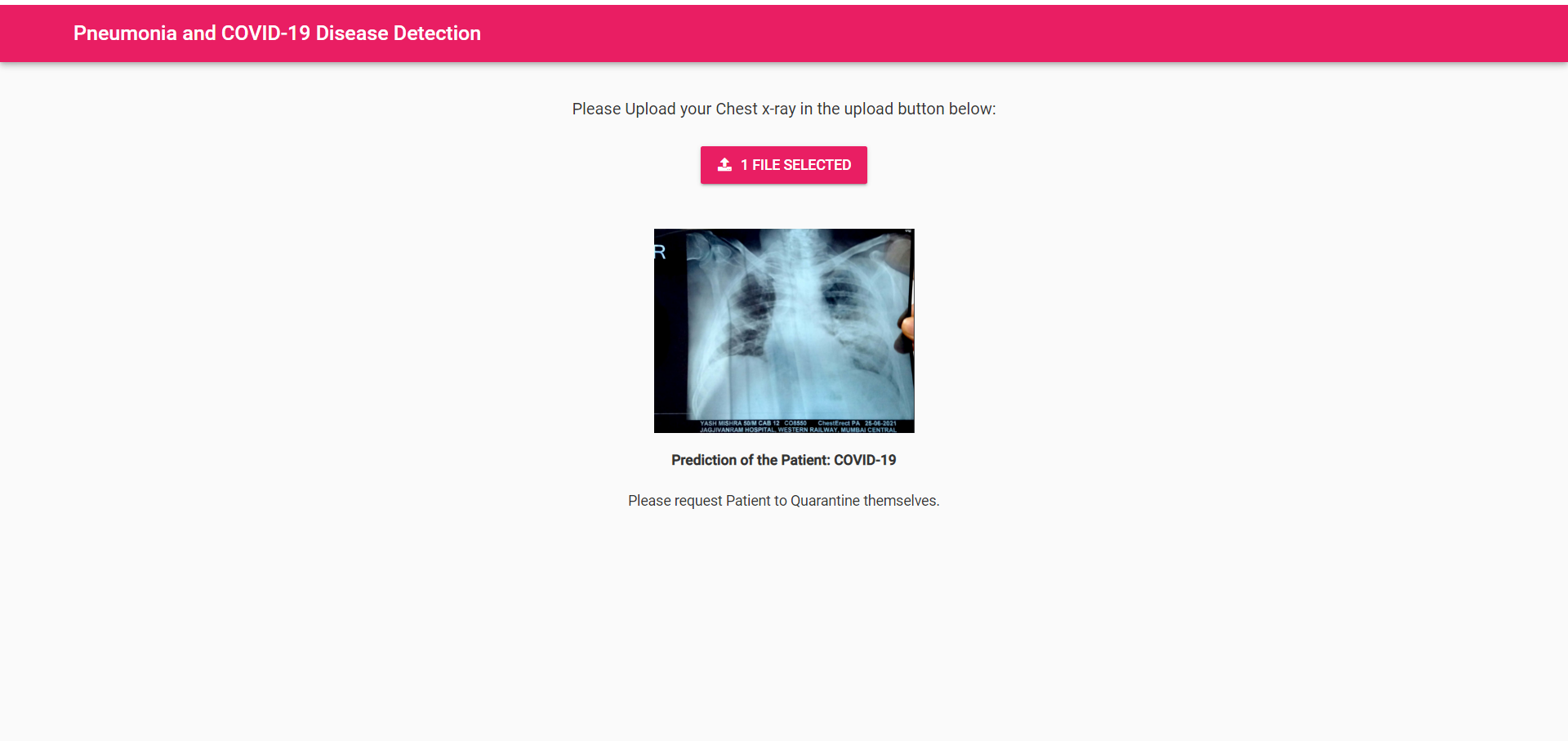


*Fig 4.5.3.2: UI showing output for Heart Detection System*

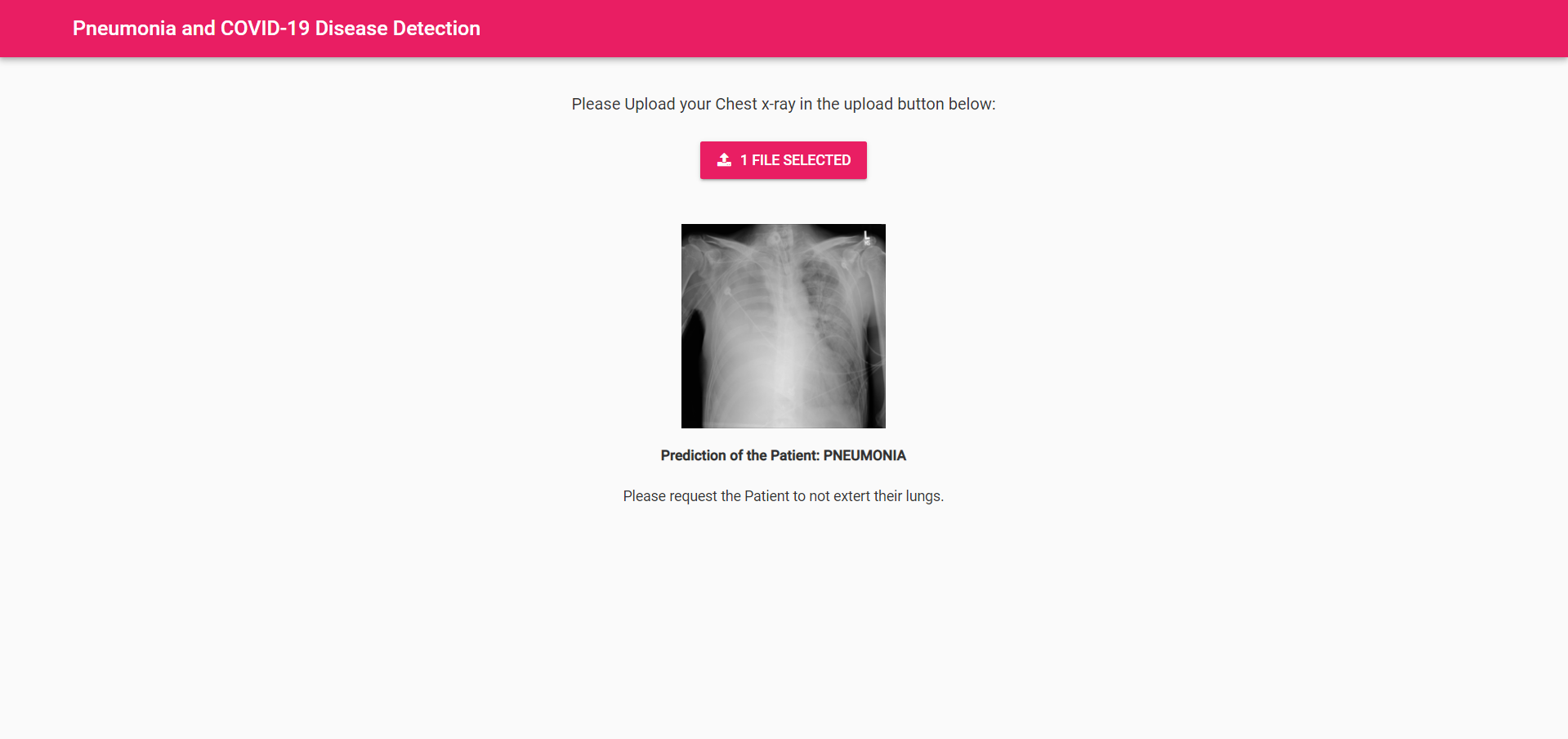
* + 1. ***UI Implementation for Pneumonia and COVID-19 Disease Detection***



*Fig 4.5.4.1: UI showing Normal for Pneumonia and COVID-19 Disease Detection*



*Fig 4.5.4.2: UI showing COVID-19 detected for Pneumonia and COVID-19 Disease Detection*



*Fig 4.5.4.3: UI showing Pneumonia detected for Pneumonia and COVID-19 Disease Detection*

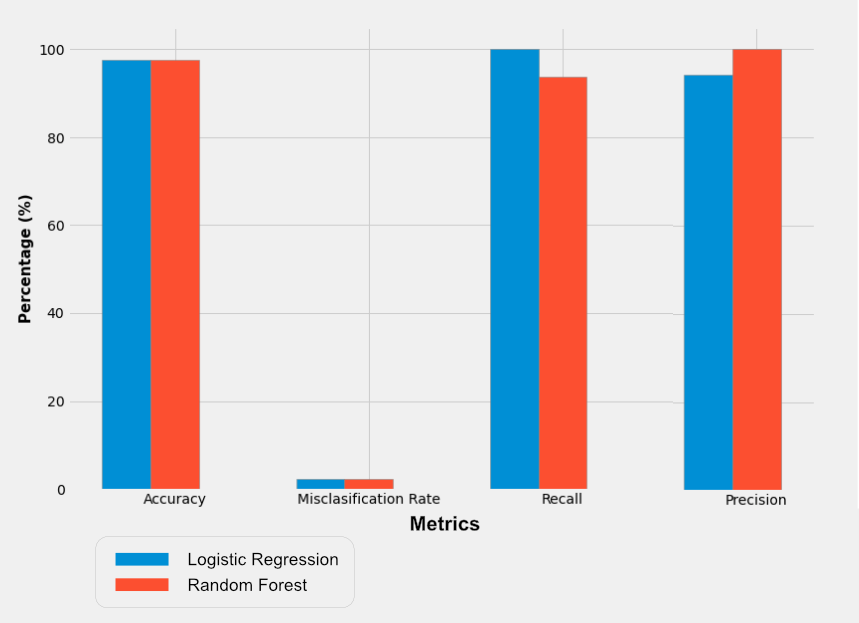
## Results of Chronic Kidney Disease Detection

It was found that the implemented models for the detection of CKD had an accuracy of 97.5% in both, Random Forest Classifier, as well as Logistic Regression. The resulting accuracies were a statistical average of all the accuracies of our referenced papers. Hyperparameter tuning was performed along the lines of the referenced papers to improve the performance of the models, while also maintaining a higher level of accuracy for the detection of CKD.

Chart, bar chart

Description automatically generated

*Fig 4.6.1. Correlation between Chronic Kidney Disease and Numeric features*



*Fig 4.6.2. Metrics for models to predict Chronic Kidney Disease*

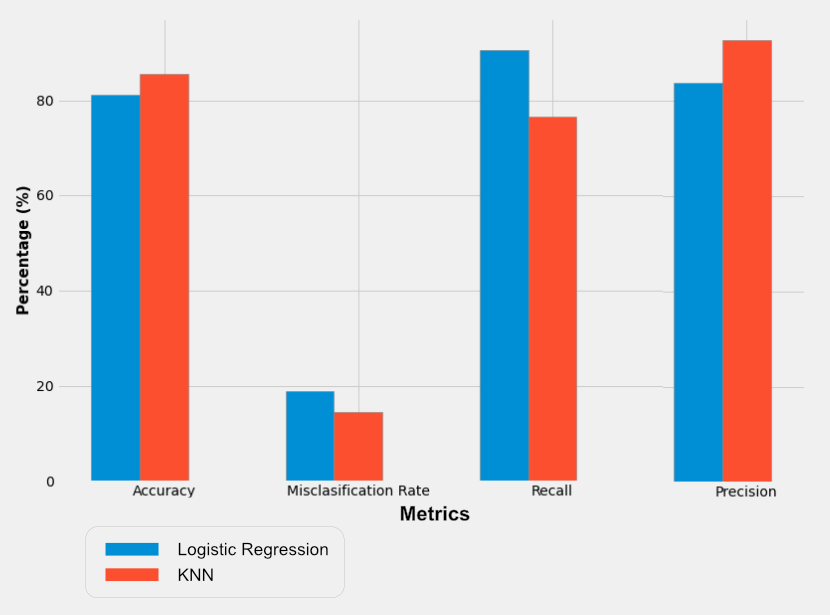
## Results of Diabetes Detection

The detection of Diabetes was performed using Logistic Regression, as well as KNN, where the implemented models had an accuracy of 81% and 85% respectively. During data pre-processing, it was observed that the number of samples with outcome 0 (Diabetes) were significantly higher than those samples with outcome 1 (No diabetes). Up sampling was performed to balance the dataset with respect to the target variable, while averting the possibility of misclassification of any other attribute. This resulted in getting a higher accuracy than the referenced papers which implemented KNN for their detection model. Hyperparameter tuning was performed to increase the accuracy of detecting Diabetes using Logistic Regression, while a standardscaler function was utilised to standardise the data across the dataset, boosting the average accuracy of the model by 3% as compared to the referenced papers which implemented LR for their detection model.

Chart, bar chart

Description automatically generated

*Fig 4.7.1. Correlation between Diabetes and Numeric features*



*Fig 4.7.2. Metrics for models to predict Diabetes*

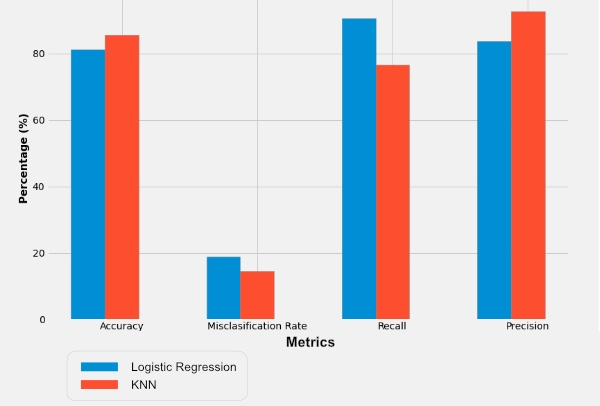
## Results of Heart Disease Detection

Decision Tree Classifier & Random Forest Classifier were implemented for the detection of heart disease, we achieved an accuracy of 78.57% and 84.29% respectively. Accuracies achieved by us are above the average accuracies from our referenced papers. Further we performed hyperparameter tuning on the models to raise the accuracy. Post tuning the accuracy of both the models was raised by 2.8%.

Chart, bar chart

Description automatically generated

*Fig 4.8.1: Correlation between Heart Disease and Numeric features*



*Fig 4.8.2: User Interface showing output for Heart Detection System*

## Results of Pneumonia and COVID-19 Disease Detection

We have implemented CNN with transfer learning which has given us an accuracy of 94%. Transfer learning is utilized when the machine stores the information acquired while taking care of one issue and applying it to an alternate yet related problem. In our case the machine has applied its learnings from Pneumonia to Covid 19 detection. Initially we used only CNN which gave us an accuracy of 88% which was boosted by applying transfer learning. We have implemented a heatmap on the output image to show where the layers convolute as shown in Figure 4.9.



*Fig 4.9: Heatmap of output image to showing layers convolution*

# Chapter 5 Conclusion & Future Work

## Conclusion

Expert systems in medication are characterised as systems with the capacity to catch and store expert information, realities, and thinking strategies to help doctors in diagnosing a patient's condition. These systems endeavour to mirror a specialist's expertise by applying a few computational strategies to help in decision making and critical thinking, by concocting contemplated determinations for a patient's disease or condition. Our project incorporates the core elements of an expert system by supporting medical experts with their claims and helping them in early diagnosis of a chronic disease in their patients using trained machine learning models giving high accuracy of disease detection and prediction.

## Future Work

We plan to take this project ahead by implementing a model to predict other chronic diseases, therefore expanding the reach of the project.

We would further like to improve the validation of each of these models by collecting live data and collaborating with medical professionals and hospitals.

A future step would be to take the project open source and have community development for increasing the project's reach and give back to the community.

# References

[1] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432-439.

[2] Shetty, D., Rit, K., Shaikh, S., & Patil, N. (2017, March). Diabetes disease prediction using data mining. In *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)* (pp. 1-5). IEEE.

[3] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.

[4] Dey, S. K., Hossain, A., & Rahman, M. M. (2018, December). Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In *2018 21st international conference of computer and information technology (ICCIT)* (pp. 1-5). IEEE.

[5] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)* (pp. 1329-1333). IEEE.

[6] Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, *10*(6), 261-268.

[7] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access*, *8*, 20991-21002.

[8] Gupta, R., Koli, N., Mahor, N., & Tejashri, N. (2020, June). Performance analysis of machine learning classifier for predicting chronic kidney disease. In *2020 International Conference for Emerging Technology (INCET)* (pp. 1-4). IEEE.

[9] Shamrat, F. J. M., Ghosh, P., Sadek, M. H., Kazi, M. A., & Shultana, S. (2020, November). Implementation of machine learning algorithms to detect the prognosis rate of kidney disease. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-7). IEEE.

[10] Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., & Mittal, A. (2019, February). Pneumonia detection using CNN based feature extraction. In *2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)* (pp. 1-7). IEEE.

[11] Sharma, A., Negi, M., Goyal, A., Jain, R., & Nagrath, P. (2021). Detection of Pneumonia using ML & DL in Python. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012066). IOP Publishing.

[12] Tahir, A. M., Chowdhury, M. E., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., & Hamid, T. (2021). COVID-19 infection localization and severity grading from chest X-ray images. *Computers in biology and medicine*, *139*, 105002.

[13] Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., & Chowdhury, M. E. (2021). Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in biology and medicine*, *132*, 104319.

[14] Degerli, A., Ahishali, M., Yamac, M., Kiranyaz, S., Chowdhury, M. E., Hameed, K., & Gabbouj, M. (2021). COVID-19 infection map generation and detection from chest X-ray images. *Health information science and systems*, *9*(1), 1-16.

[15] Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., & Islam, M. T. (2020). Can AI help in screening viral and COVID-19 pneumonia?. *IEEE Access*, *8*, 132665-132676.