**Multiple Disease Prediction System using Machine Learning Algorithm**

Prachi Kumari1, Jharkhand Rai University, Ranchi

Sahil Singh Munda2, Jharkhand Rai University, Ranchi

Suman Singhdeo3, Jharkhand Rai University, Ranchi

Kumar Amrendra4, Assistant Professor, Jharkhand Rai University, Ranchi

Anuradha Sharma5, Assistant Professor, Jharkhand Rai University, Ranchi

**Abstract**

Multiple Disease Prediction Systems (MDPS) are computer-based tools that use machine learning algorithms to predict the likelihood of patients developing multiple diseases based on various factors such as medical history, clinical symptoms, and laboratory test results. The early identification of such individuals can lead to early interventions and treatments, potentially preventing or delaying the onset of multiple diseases. In this paper, we present a review of MDPS and its potential applications in healthcare. We discuss the various machine learning algorithms that can be used to develop MDPS, as well as the performance metrics that can be used to evaluate its performance. We also highlight the benefits and challenges associated with MDPS implementation in healthcare. Finally, we discuss the future directions of MDPS research and its potential to improve healthcare outcomes by enabling earlier detection and intervention in individuals at risk of developing multiple diseases. We have considered Seven diseases for now that are Heart, Liver, Kidney, Malaria, Pneumonia and Diabetes and in the future, many more diseases can be added. The user must enter various parameters of the disease and the system would display the output whether he/she has the disease or not.

**Introduction**

A Multiple Disease Prediction System (MDPS) is a computer-based tool that uses machine learning algorithms to predict the likelihood of a patient developing multiple diseases. It considers various factors such as medical history, clinical symptoms, and laboratory test results to generate predictions. MDPS can be developed using a wide range of machine learning algorithms, including decision trees, support vector machines, artificial neural networks, and others. These algorithms analyze large datasets of patient information and learn patterns that can be used to predict the likelihood of future health outcomes. The performance of MDPS can be evaluated using various metrics, such as accuracy, sensitivity, and specificity. The use of MDPS has several potential benefits, including improved patient outcomes, reduced healthcare costs, and increased efficiency in healthcare delivery. MDPS can also be used to support research into the relationships between different diseases and their risk factors, potentially leading to new insights into the underlying mechanisms of disease. Overall, MDPS represents a promising approach to improve healthcare outcomes by enabling earlier detection and intervention in individuals at risk of developing multiple diseases.

**Problem Statement**

Multiple disease prediction system is a problem of developing an intelligent system that can accurately predict the likelihood of a patient having one or more diseases based on their symptoms, medical history, and other relevant factors. The goal of such a system is to help healthcare professionals in making accurate diagnoses and providing appropriate treatment options to patients. The development of a multiple disease prediction system involves the use of various techniques and algorithms from machine learning, data mining, and artificial intelligence to process large amounts of medical data and identify patterns that can be used to predict disease outcomes. The system should be able to accurately identify multiple diseases, even those that may have similar symptoms, and provide treatment recommendations based on the patient's condition.

**Proposed system:** In multiple disease prediction System (MDPS), It has capability to predict different types of disease at a time with accuracy like breast cancer, heart, kidney, liver, malaria. So, the user does not need to visit different websites to predict the diseases. for Implementation of multiple disease prediction system (MDPS) we are using various machine learning algorithms and Flask.

**System Analysis:** A system analysis of a multiple disease presentation system is crucial to ensure that the system is effective, efficient, and secure. By evaluating the system's functionality, design, and performance, healthcare professionals can identify potential issues and make necessary improvements to ensure that the system provides accurate and reliable predictions to improve patient outcomes.

Pickle File

Embedded Pickle file in Flask

Input Dataset

**DESIGN**

**Architecture Design**

Malaria

Diabetes

Heart

Input Dataset

Splitting

Visualize the Data

Pre-processing

Liver

**Pneumonia**

Accurate Model

Applied Knowledge

Testing

Classification Algorithms

Training

**Kidney**

Cancer

Embedded Pickle file in Flask

Pickle File

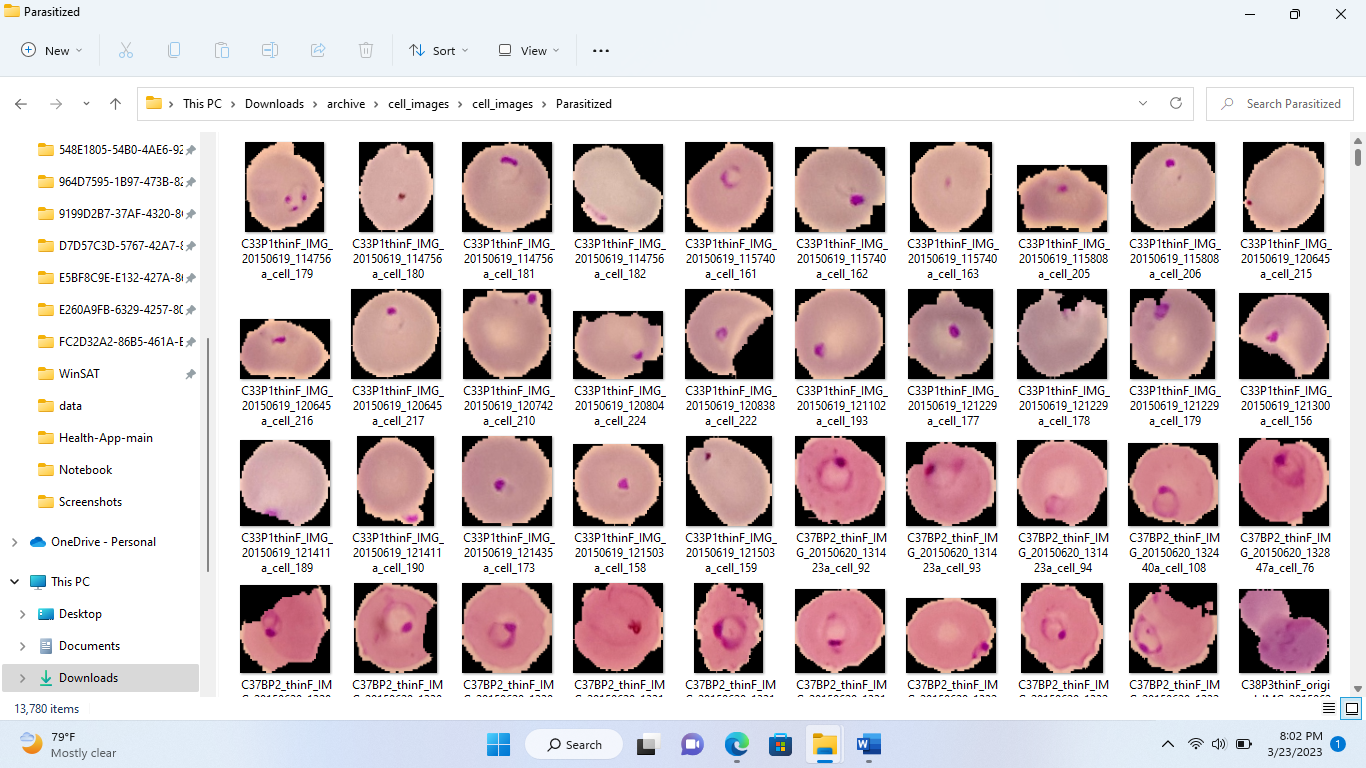
**Figure No: Block Diagram**

In the above architecture we have experimented on seven diseases that is heart, diabetes, Breast Cancer, liver, Kidney, Malaria, and Pneumonia as these are correlated to each other. The first step is to collect the dataset for Breast Cancer, diabetes disease, heart, Kidney, liver disease, malaria, and pneumonia we have imported the Winconsin Breast Cancer Dataset, PIMA (Pima Indian Diabetes Dataset) dataset, Indian liver dataset, Chronic Kidney Disease Dataset, Malaria Cell images Dataset and Chest x-ray image Pneumonia respectively. Once we have imported the dataset then we visualize each inputted data. After visualization, pre-processing of data takes place where we check the outliers, missing values and scale the dataset. then we get updated dataset and then after split the data into two parts training and testing. Next, we had applied Logistic Regression, Random Forest, SVM and random forest algorithm for trained the model and applied knowledge on the classified algorithm using testing dataset. then After, we will choose the algorithm (KNN, SVM, RD, DT) with the best accuracy for each of the disease. Then we build a pickle file for all the disease and then integrated the pickle file with the Flask framework for the output of the model on the webpage.

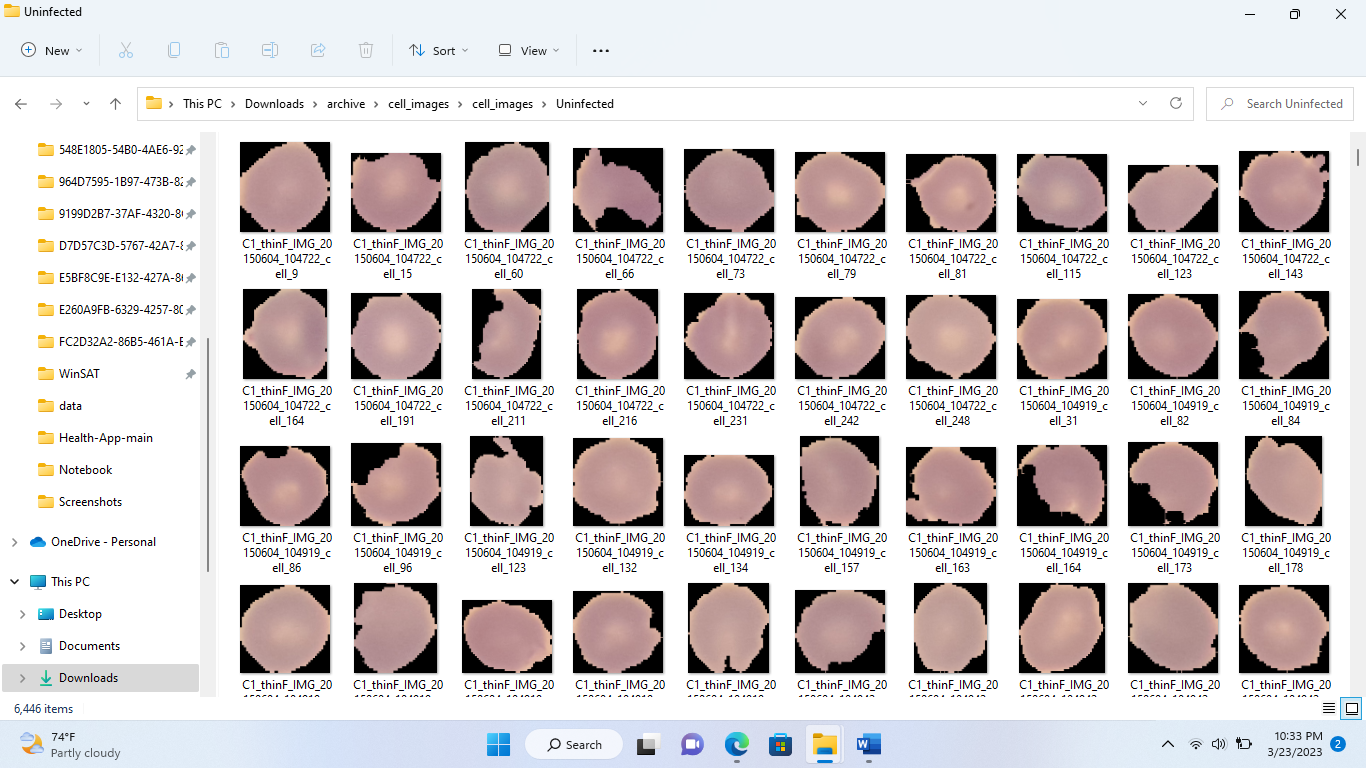
**Dataset Collection and processing:**

Malaria: the dataset collected from NIH National library of medicine that is available on the Kaggle website ([Malaria Cell Images Dataset | Kaggle](https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria)). In this dataset have 27,558 images. It contains of two different classes parasitized (that contains 13,779 images) and uninfected (13,779 images ). The color model of images is RGB.

Sample of parasitized:

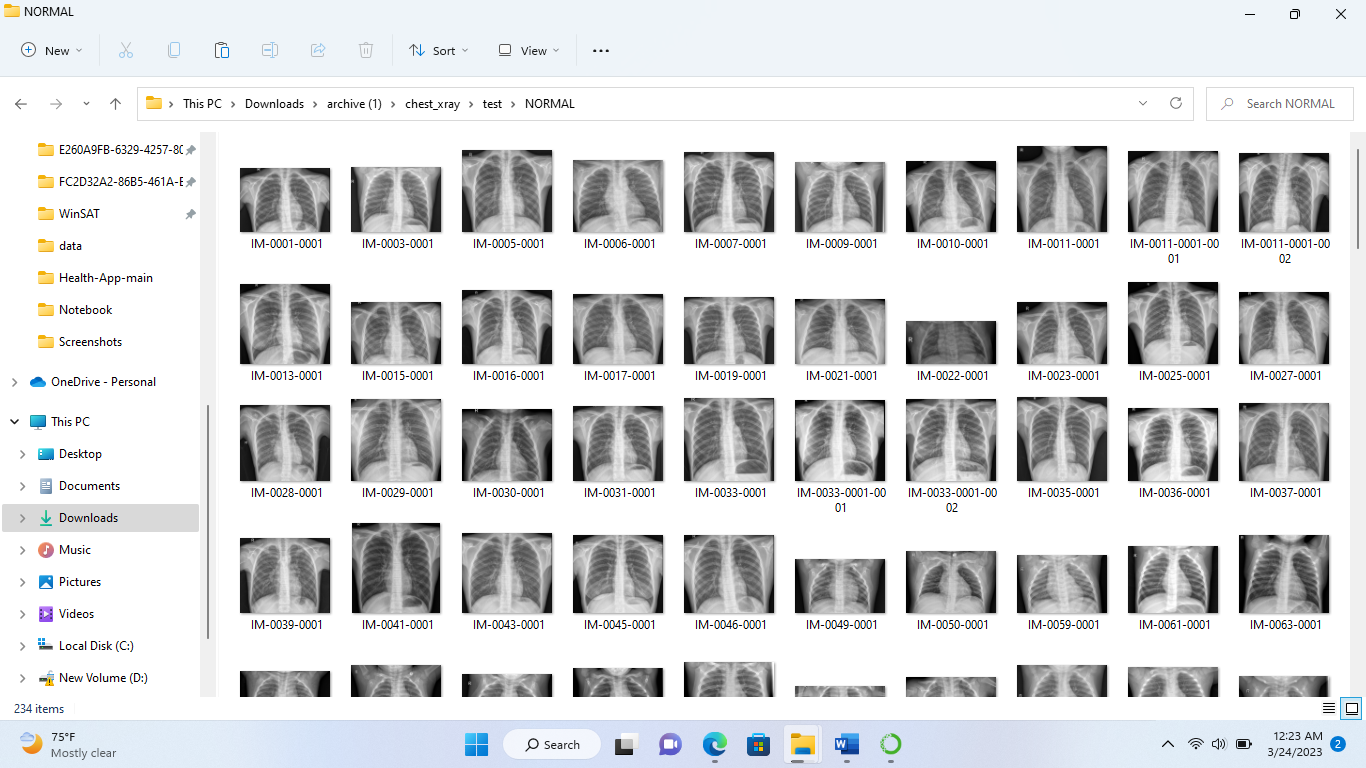


**Sample of uninfected**

****

**Pneumonia:** Pneumonia Dataset has taken from Kaggle website ([Chest X-Ray Images (Pneumonia) | Kaggle](https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia)). There are 5,863 images in jpg format and it contains three types of folder train, test, Val. This folder has 2 categories pneumonia/ normal

**Sample of infected Images**

****

**Methodology:**

**Logistic regression:**

p(y=1|x) = 1 / (1 + exp(-z))

where p(y=1|x) is the probability of the dependent variable (y) being equal to 1 given the independent variables (x), exp is the exponential function, and z is the linear combination of the independent variables and their coefficients:

z = b0 + b1x1 + b2x2 + ... + bn\*xn

where b0 is the intercept and b1...bn are the coefficients for each independent variable (x1...xn). The coefficients are estimated during the model fitting process, where the goal is to find the values that minimize the error between the predicted probabilities and the actual values.

The logistic regression model uses the sigmoid function, which takes the linear combination z as input and maps it to a value between 0 and 1. This allows the model to output a probability that the dependent variable is equal to 1, given the values of the independent variables. If the probability is greater than a chosen threshold, the model predicts that the dependent variable is equal to 1, otherwise it predicts that it is equal to 0

**Random Forest**

Random forest is an ensemble learning algorithm that combines multiple decision trees to make a final prediction. Each decision tree in the random forest is constructed using a random subset of the training data and a random subset of the features.

Here is the formula for random forest:

Prediction = mode(y1, y2, ..., ym)

where yi is the predicted class for the ith decision tree, and m is the total number of trees in the random forest.

**Decision Tree:** a decision tree can be represented by a set of nested if-else statements that split the data into smaller subsets based on the values of the input features.

There are following step used in Decision Tree algorithm

1. Start with the entire dataset.
2. For each feature in the dataset:
3. Calculate the information gain by splitting the data on that feature.

IG (S, A)= Entropy(S)-∑v€A |SV|/ S. Entropy(S)

1. Choose the feature with the highest information gain as the node to split on.
2. Create a branch for each possible value of the chosen feature and recursively repeat steps 1-3 for each branch, using only the subset of the data that corresponds to that branch.
3. Continue splitting the data until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples per leaf node.
4. Assign a class label to each leaf node based on the majority class of the samples that reach that node.

**Support Vector Machine:**

Support Vector Machine (SVM) algorithm used for classification, regression, and outlier detection.

The standard formula of the SVM for binary classification is:

y(x) = w^T x + b

where:

* x is the input vector.
* w is the weight vector.
* b is the bias term.
* y(x) is the predicted output of the SVM for the input vector x.

**CNN And Transfer E- Learning: -** Convolutional Neural Networks (CNN) is a type of deep neural network that is commonly used for image classification, object detection, and other computer vision tasks.

The standard formula of a convolutional layer in a CNN is:

h[l] = σ(W[l] \* x[l-1] + b[l])

where:

* x[l-1] is the input feature map from the previous layer.
* W[l] is the set of learnable weights (filters) for the l-th layer.
* b[l] is the set of learnable biases for the l-th layer.
* h[l] is the output feature map of the l-th layer.
* σ () is the activation function.

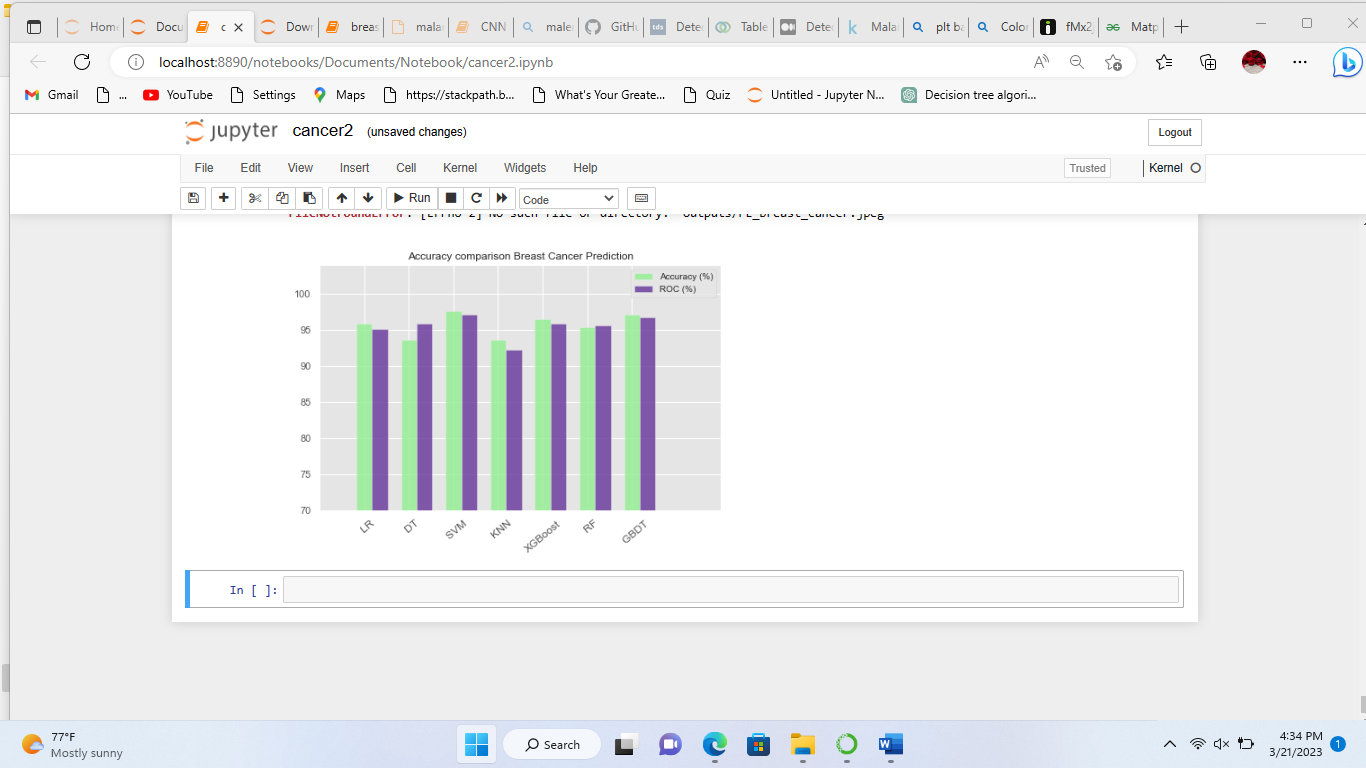
**Transfer Learning:**

Transfer learning is a technique in machine learning that involves using a pre-trained model on a related task as a starting point for a new task. Transfer learning can be used to reduce the amount of data and computation required for training a new model and can improve the performance of the new model.

**Accuracy comparison for each disease:**

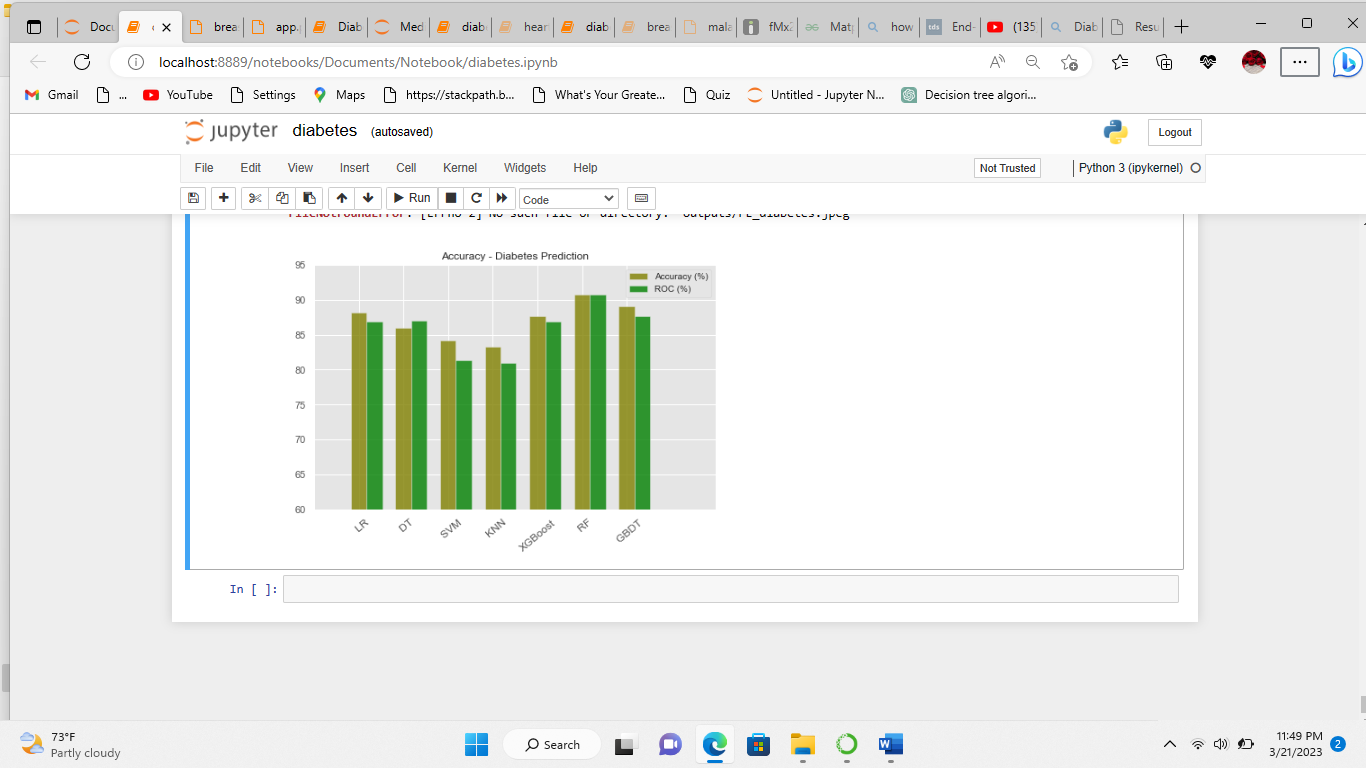
**Breast Cancer:**

For Breast cancer, here we applied following algorithm to compare the accuracy score. After comparison, we get, SVM -97.66, Gradient boosting classifier-97.66, Random Forest Classifier- 96.49, Xgboost- 96.49, KNN-93.57 Decision tree-95.32 logistic-regression-95.51.

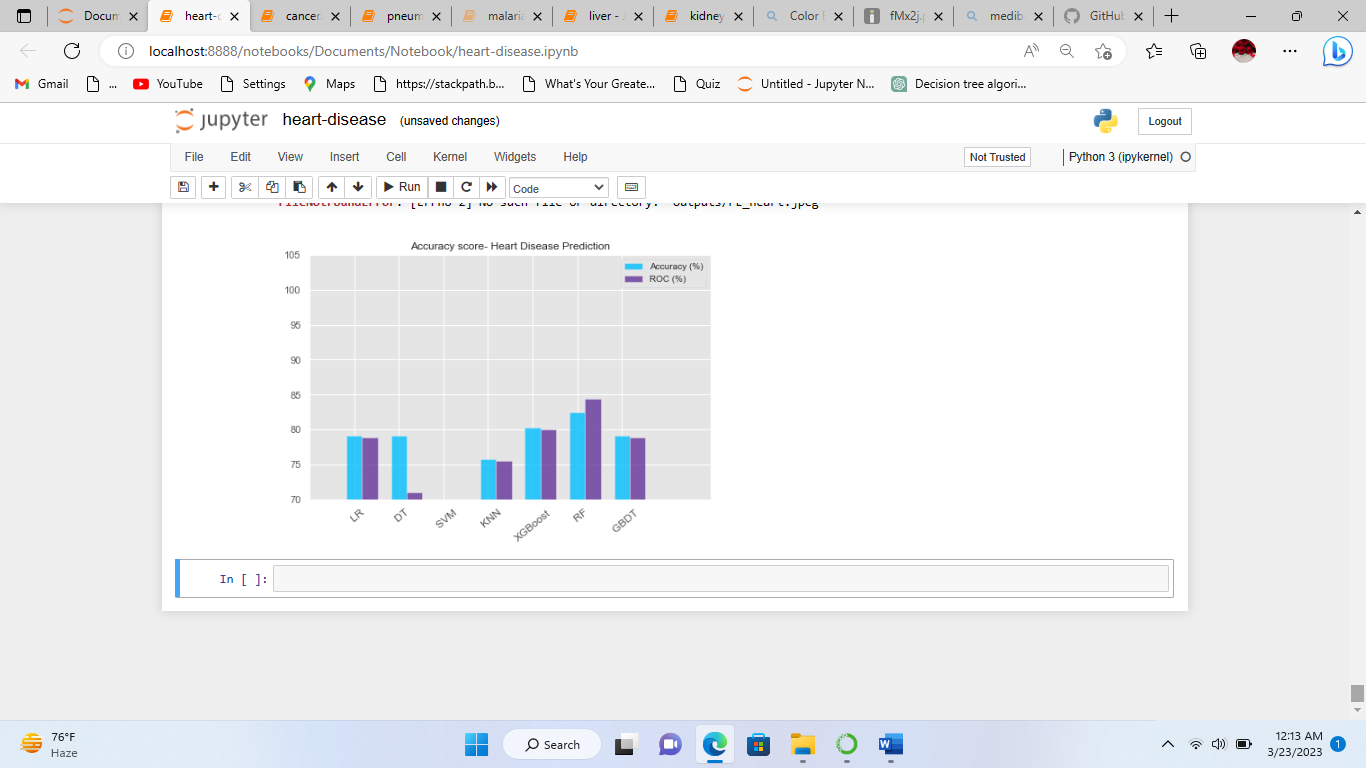


**Diabetes:**

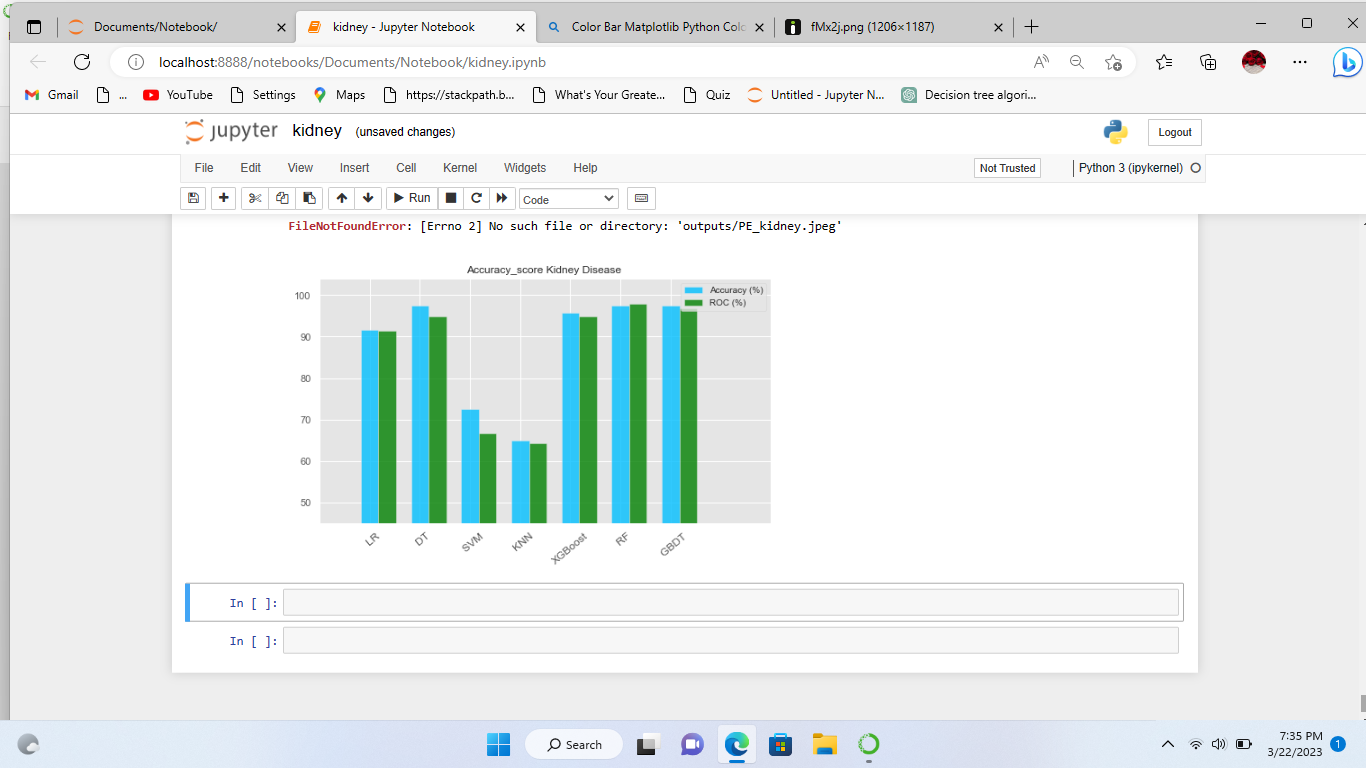
For Diabetes disease, after comparison, we get, Random Forest Classifier- 90.79 with highest accuracy score, SVM -84.21, Gradient boosting classifier-89.04, Xgboost- 87.72, KNN-83.33 Decision tree-85.96 logistic-regression-88.16.

****

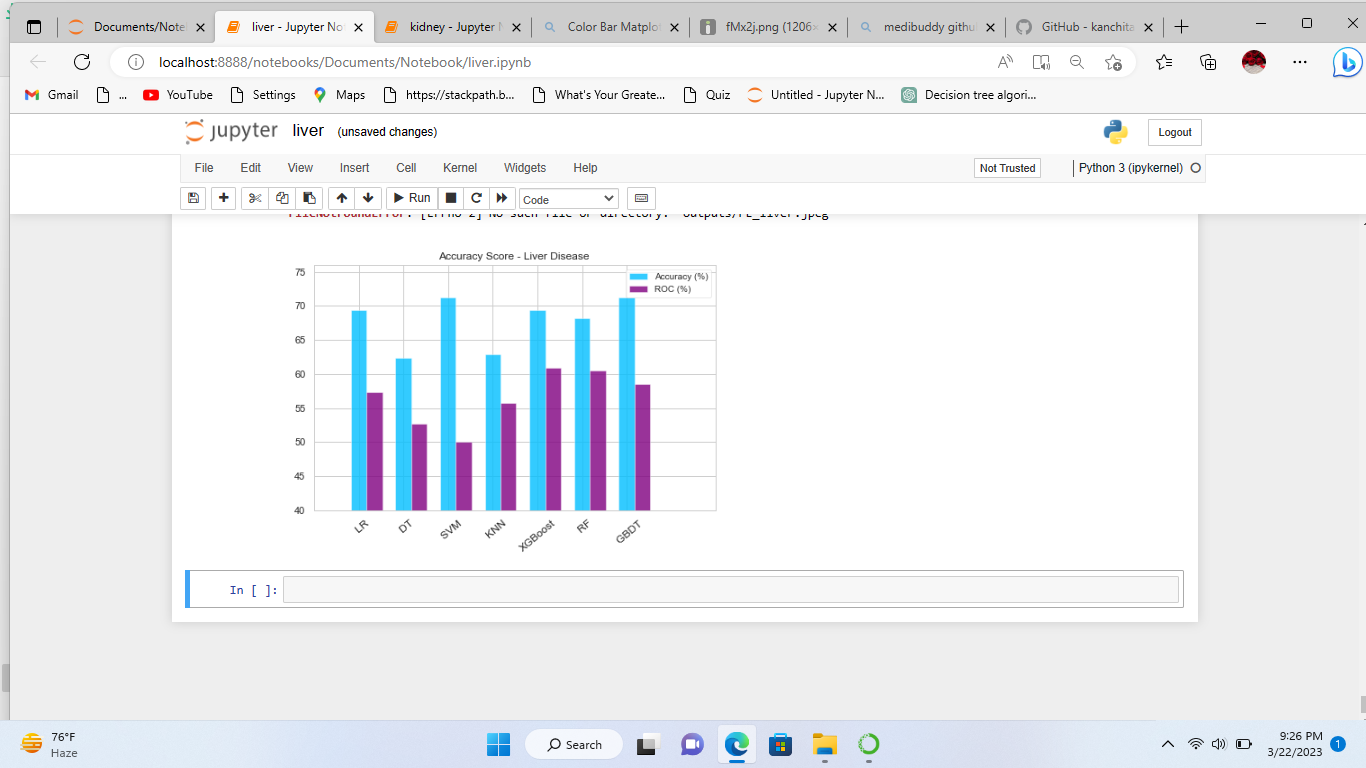
**Heart disease**: After comparison, we get, Random Forest Classifier-0.824176 with highest accuracy score, SVM -0.516484, Gradient boosting classifier- 0.791209, Xgboost- 0.802198, KNN-83.33, Decision tree-0.791209, logistic-regression-0.791209

****

**Kidney:** After comparison, we get, Random Forest Classifier-0.824176 with highest accuracy score, SVM -0.516484, Gradient boosting classifier- 0.791209, Xgboost- 0.802198, KNN-83.33, Decision tree-0.791209, logistic-regression-0.791209

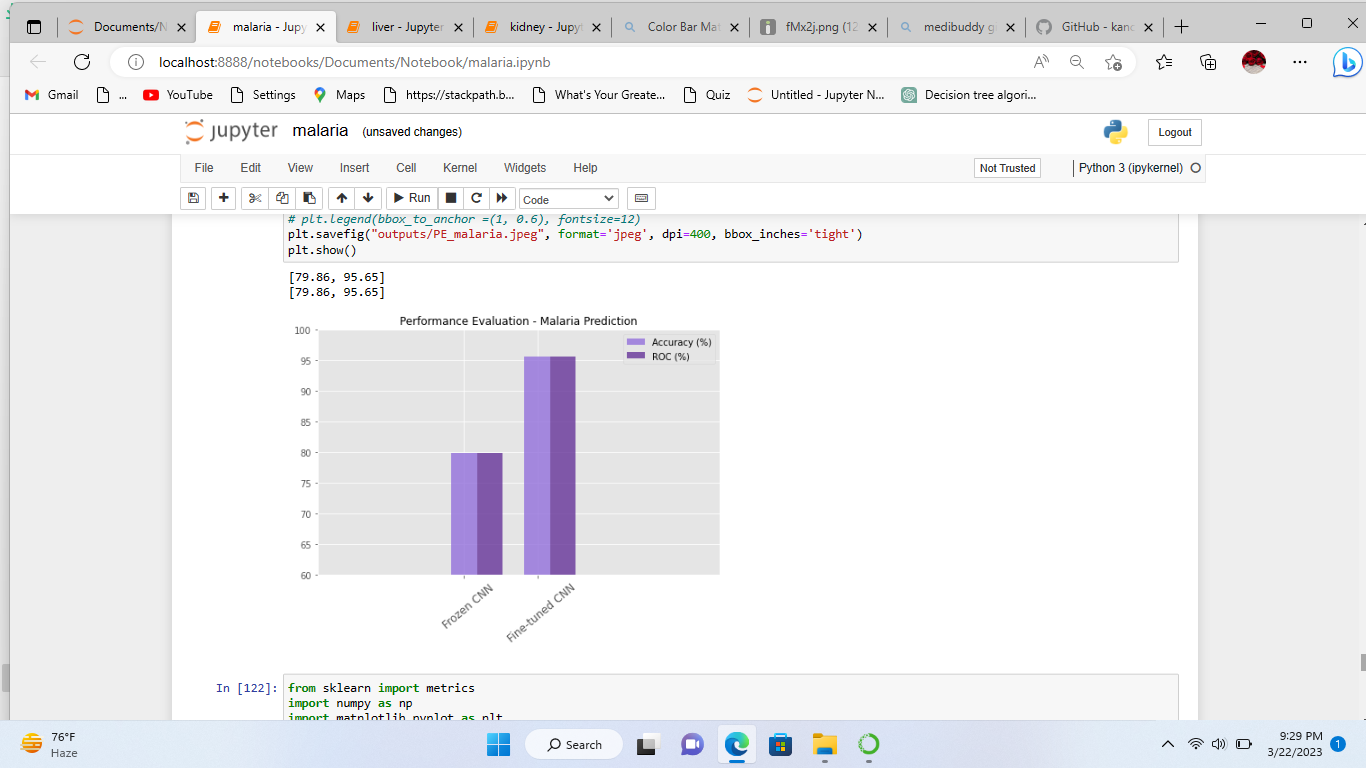
****

**Liver Disease:** After comparison, we get, SVM- 71.18 with high accuracy, Random Forest Classifier-68.24, Gradient boosting classifier- 71.18, Xgboost- 69.41, KNN-62.94, Decision tree-62.35, logistic-regression-69.41

****

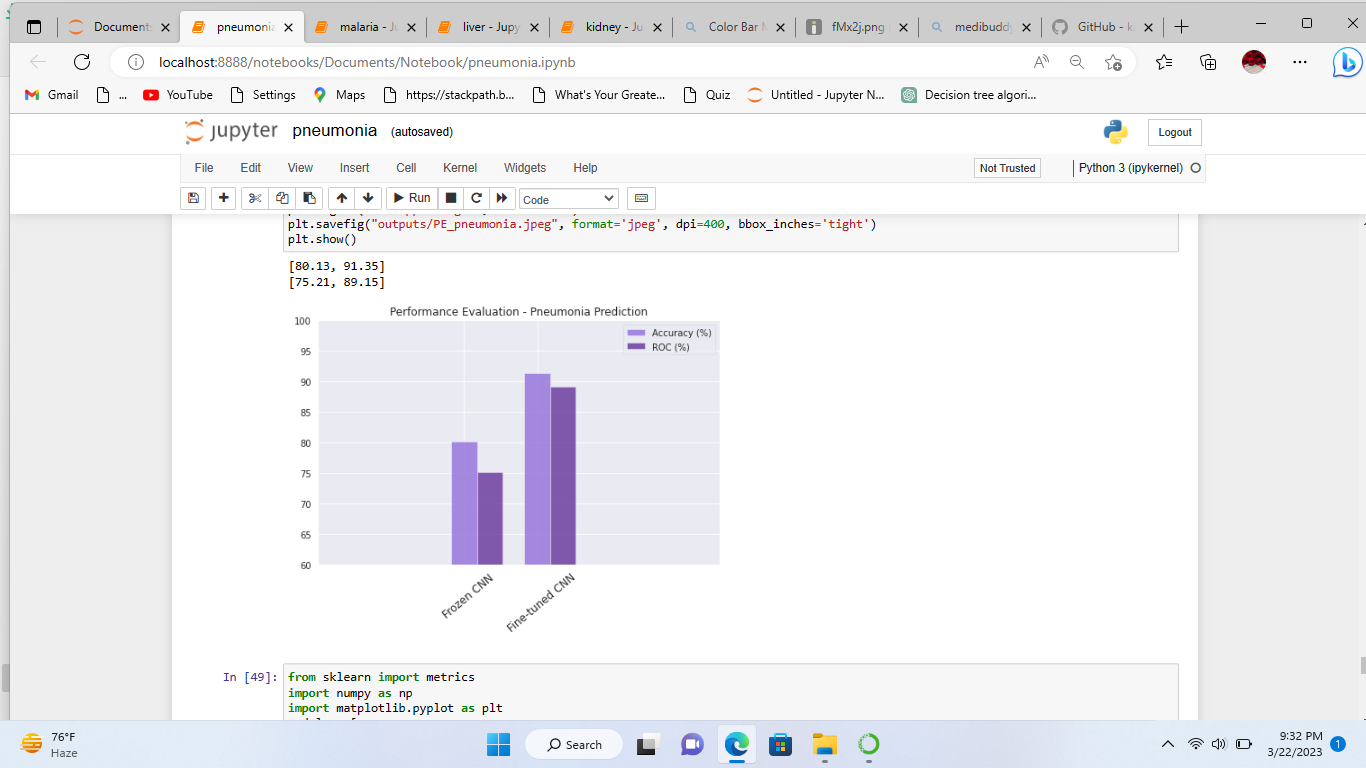
**Malaria:**

Here we use fine- tuned CNN to get the high accuracy of model. The accuracy score of this model is 95.65%.

****

**Pneumonia :**

Here we use fine- tuned CNN to get the high accuracy of model. The accuracy score of this model is 91.35%.

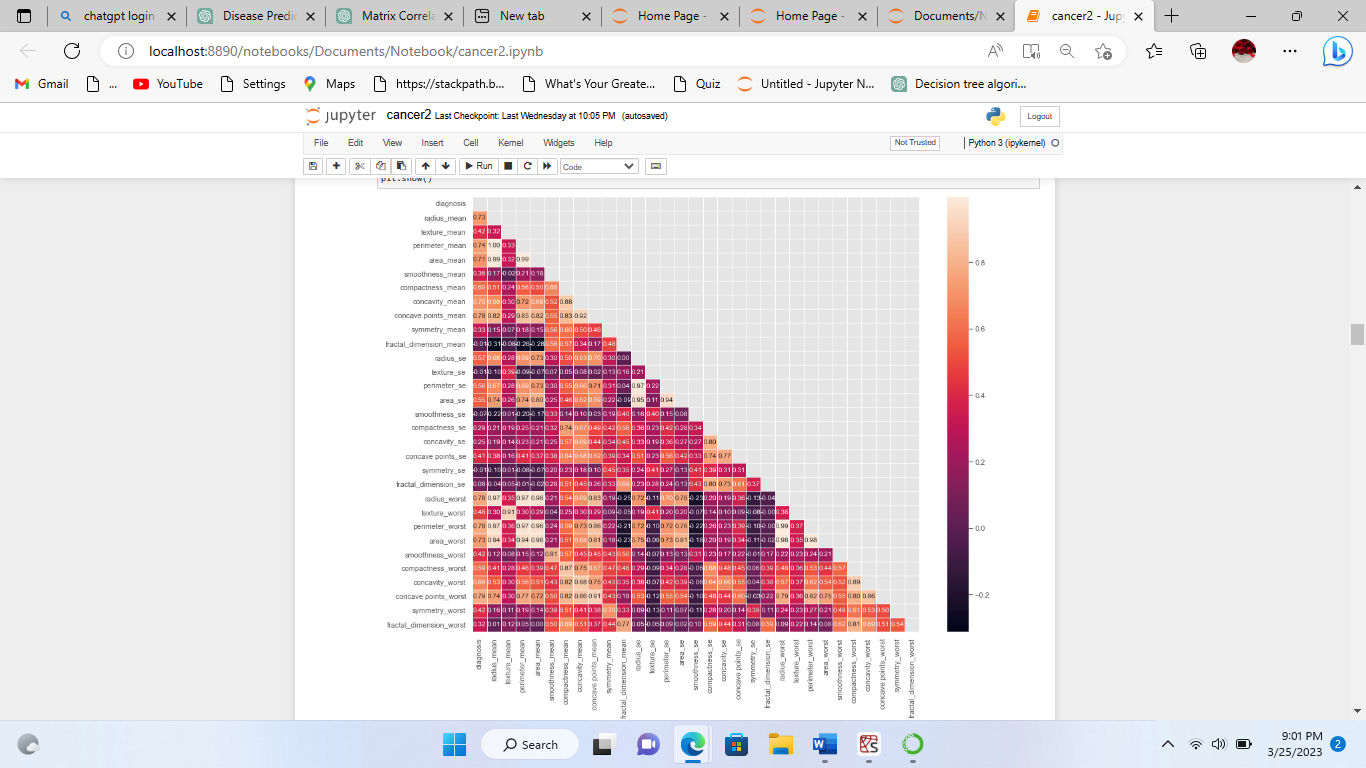
****

**Result**

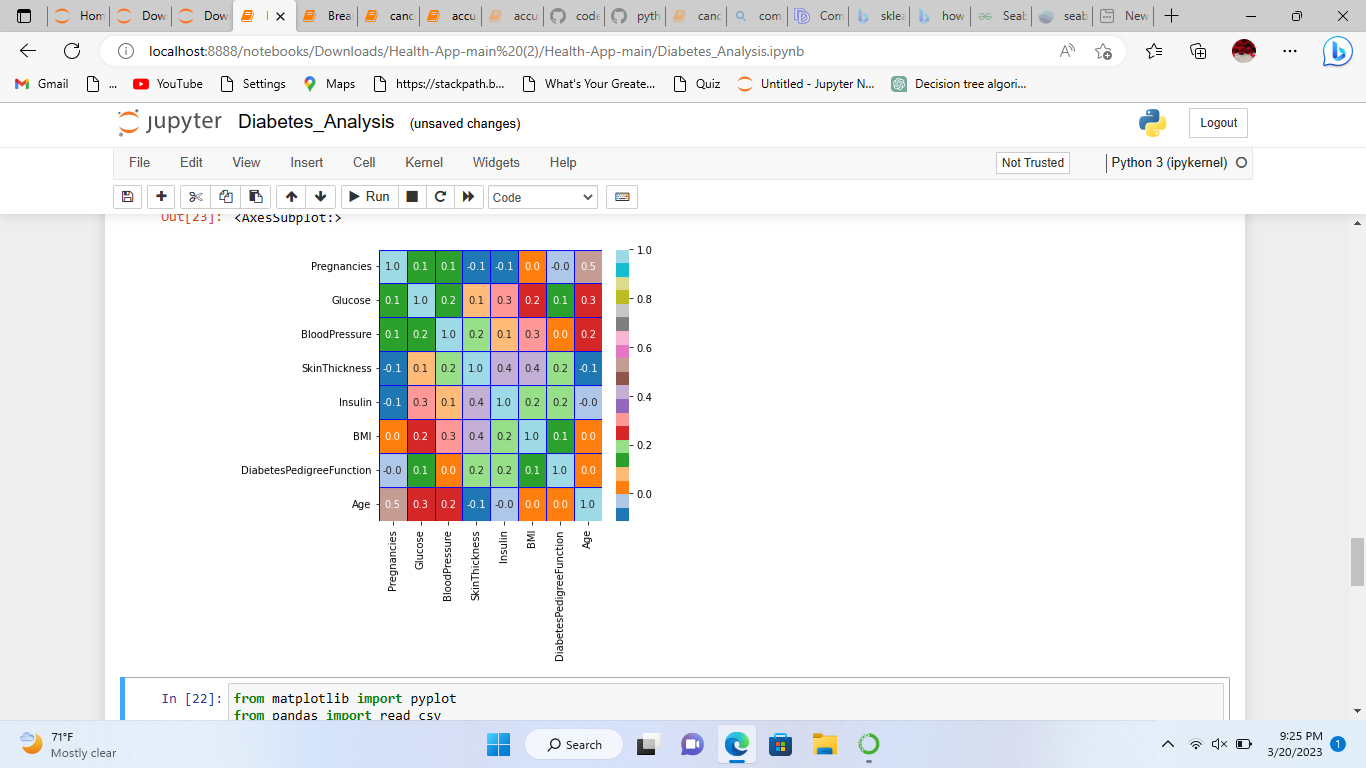
In the Multiple disease prediction system, diabetes disease prediction, heart disease, Kidney, liver, breast Cancer model used SVM, Decision Tree, random forest algorithm and Malaria and Pneumonia uses the CNN and Transfer E-learning for as these gave the best accuracy accordingly. There when the patient adds the parameter (for breast Cancer, diabetes, liver, Kidney, Heart) according to the disease it will show whether the patient has a disease or not. The parameters will show the range of the values needed and if the values are not between the range or is not valid it will show the warning sign that add a correct value.For malaria and Pneumonia, we upload image that it will predict the Output as an infected or uninfected with accuracy of uploaded image of diseases.

Matrix correlation: Matrix correlation metric used in machine learning is the Pearson correlation coefficient, which measures the strength of the linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation.

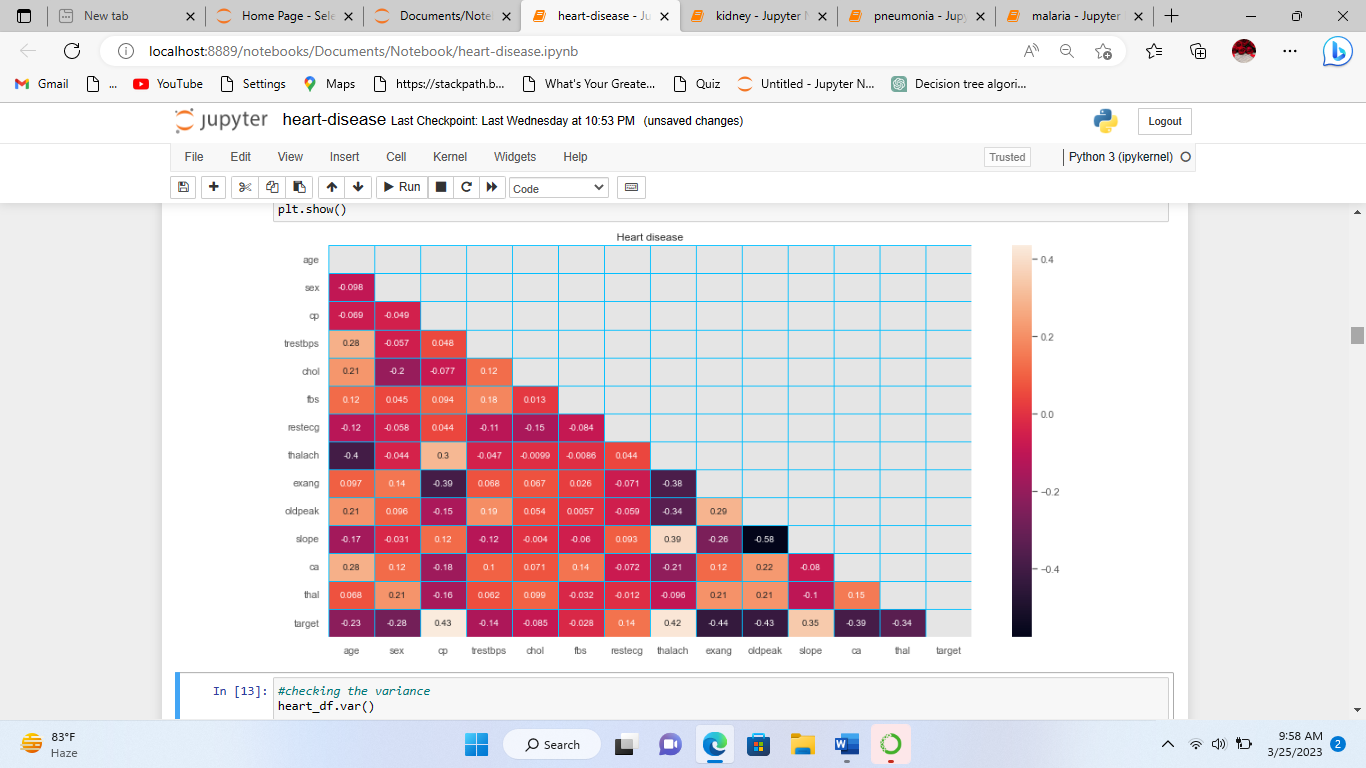
**Matrix correlation of Breast Cancer disease:**

****

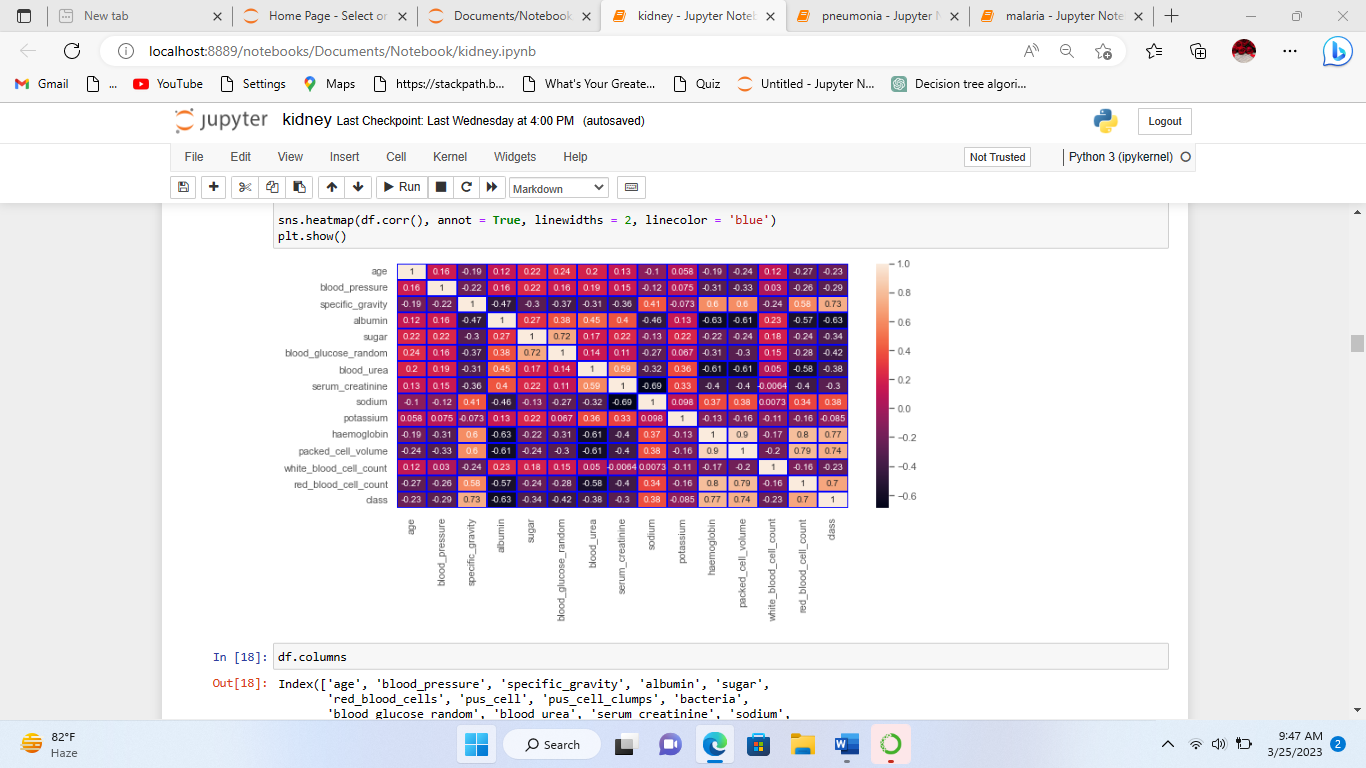
**Matrix correlation of Diabetes disease:**

****

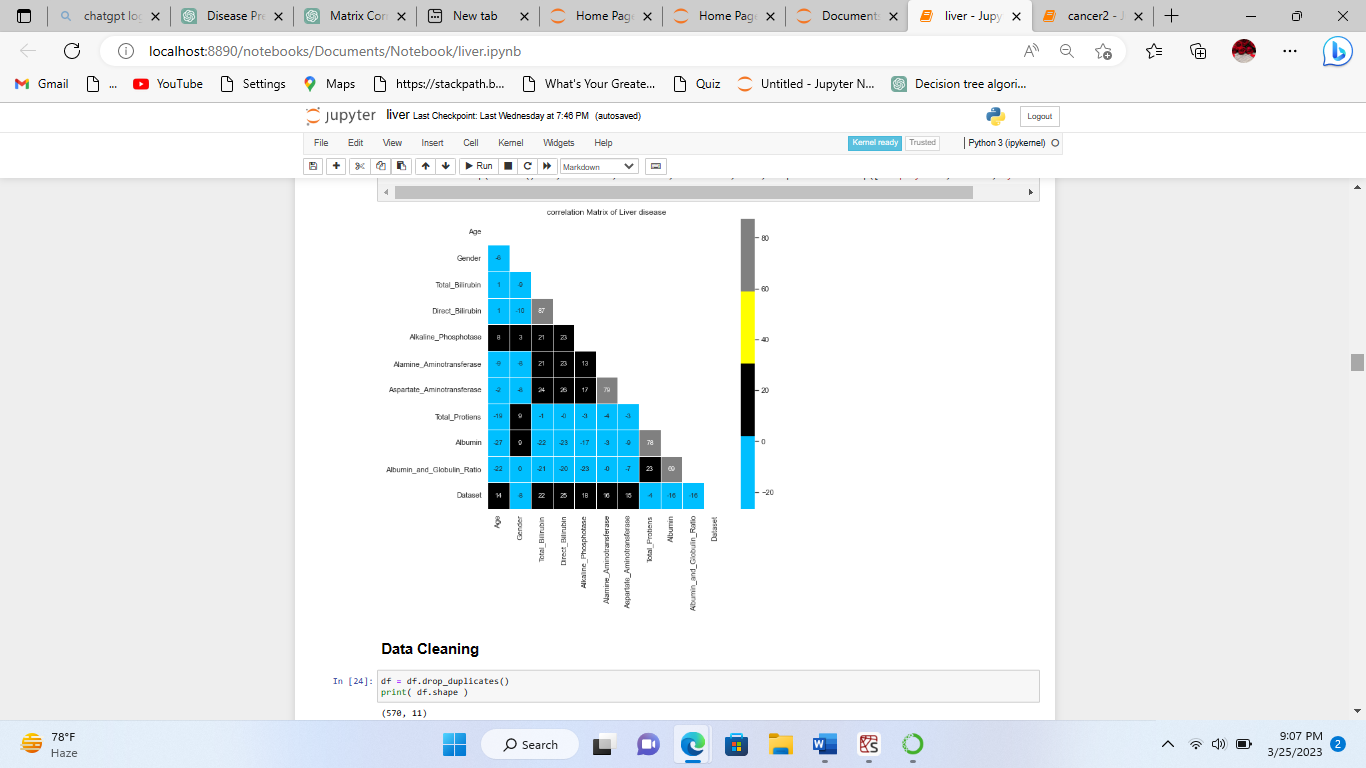
**Matrix correlation of Heart disease:**

****

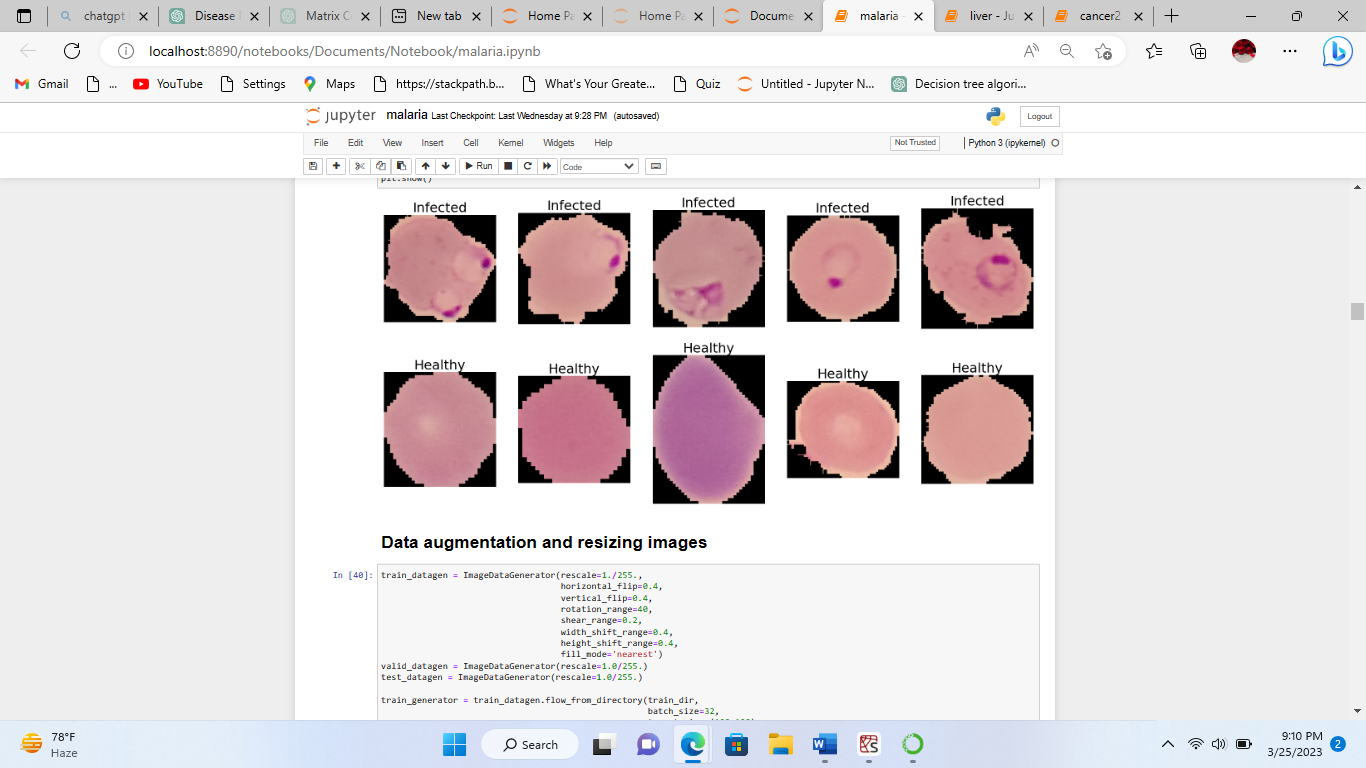
**Matrix correlation of kidney disease:**

****

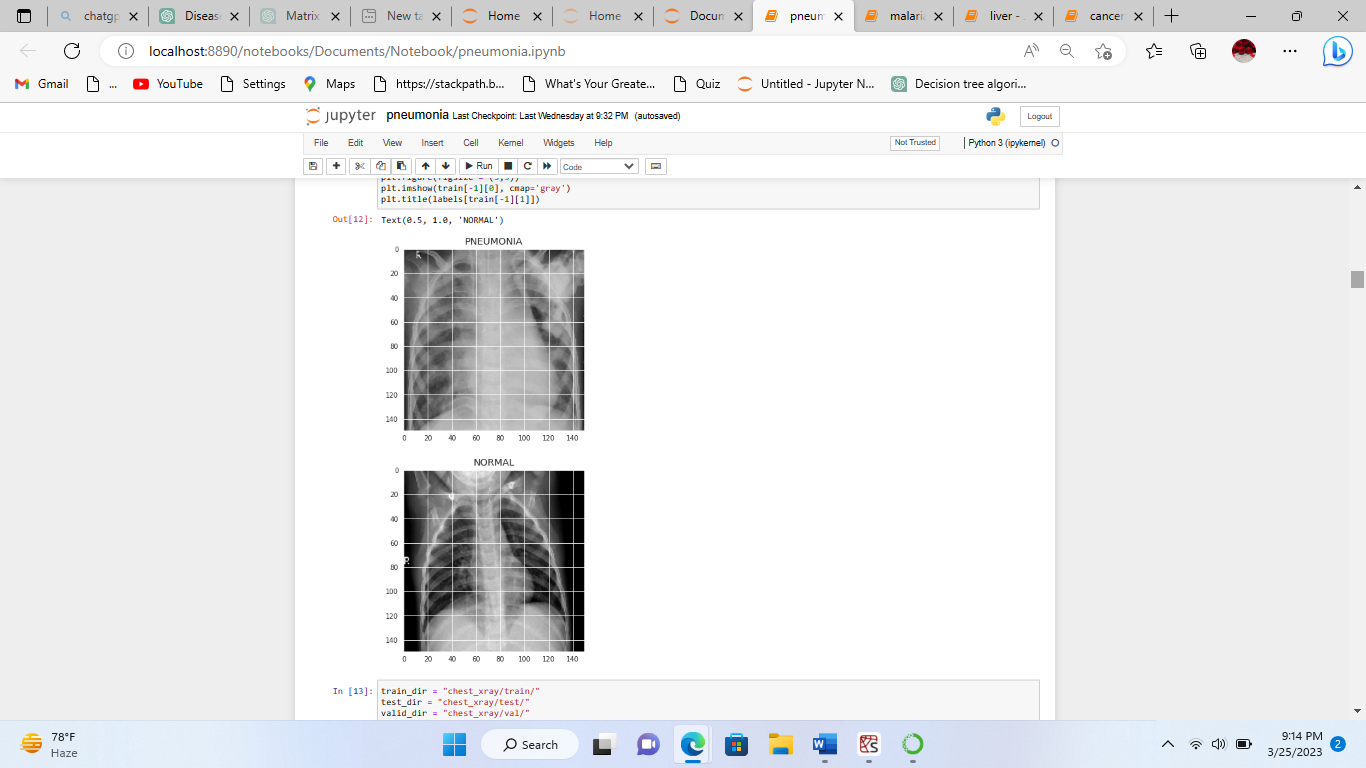
**Matrix correlation of Liver disease:**

****

**Result of Malaria:**

****

**Result of Pneumonia:**

****

**Conclusion:**

A multiple disease prediction system is a valuable tool that can help healthcare professionals and individuals identify potential health issues early and take necessary steps to prevent or treat them. By analyzing a patient's symptoms, medical history, and other relevant data, these systems can provide predictions on potential diseases or conditions that a person may be at risk of developing. Overall, the use of multiple disease prediction systems can lead to improved patient outcomes, reduced healthcare costs, and more efficient use of healthcare resources. However, it is important to note that these systems should not replace the judgment and expertise of healthcare professionals. They should be used as a complementary tool to aid in diagnosis and treatment planning. Further research and development in this area could lead to more accurate and effective disease prediction systems, which could have significant benefits for public health and individual wellbeing.

**References**

* [Likitha KN](https://ieeexplore.ieee.org/author/37088974322); [Nethravathi R](https://ieeexplore.ieee.org/author/37088973415); [Nithyashree K](https://ieeexplore.ieee.org/author/37088974093); [Ritika Kumari](https://ieeexplore.ieee.org/author/37088974952); [Sridhar N](https://ieeexplore.ieee.org/author/37088438850); [Venkateswaran K](https://ieeexplore.ieee.org/author/37088492325), Heart Disease Detection using Machine Learning Technique **Published in:**[2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)](https://ieeexplore.ieee.org/xpl/conhome/9532475/proceeding), [Heart Disease Detection using Machine Learning Technique | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9532705)
* A.Sivasangari, Baddigam Jaya Krishna Reddy,Annamareddy Kiran, P.Ajitha,” Diagnosis of Liver Disease using Machine Learning Models” 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC
* Archana Singh ,Rakesh Kumar, “Heart Disease Prediction Using Machine Learning Algorithms”, 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3).
* Priyanka Sonar, Prof. K. JayaMalini,” DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES”, 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC)
* Watson OJ, Sumner KM, Janko M, Goel V, Winskill P, Slater HC, et al. False-negative malaria rapid diagnostic test results and their impact on community-based malaria surveys in sub-Saharan Africa. BMJ Glob Heal. 2019;4:e001582. <https://doi.org/10.1136/bmjgh-2019-001582>.
* Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. Food & function . 2020;11(5):4561–70. doi: 10.1039/C9FO00976K. - [DOI](https://doi.org/10.1039/c9fo00976k)- [PubMed](https://pubmed.ncbi.nlm.nih.gov/32400770/)
* Chan H.P, Sahiner B, Hadjiyski L, Zhou C and Petrick N 2005 Lung nodule detection and classification U.S. patent application no. 10/504, 197 https://www.who.int/news-room/fact- sheets/detail/pneumonia
* <https://www.kaggle.com/code/junkal/breast-cancer-prediction>
* <https://www.sciencedirect.com/science/article/pii/S2352914820306596>
* Odu Nkiruka, Rajesh Prasad, and Onime Clement (2021), “Prediction of Malaria Incidence Using Climate Variability and Machine Learning”, Informatics in Medicine Unlocked, Vol. 22. 100508.
* <https://www.sciencedirect.com/science/article/pii/S235291482030659>
* <https://github.com>