**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) refers to a computer-based tool that harnesses machine learning algorithms for forecasting the probability of a patient contracting various illnesses. It takes into account a range of elements such as medical history, clinical indications, and lab test outcomes to formulate predictions. Diverse machine learning algorithms, such as decision trees, support vector machines, artificial neural networks, and others, can be employed to create MDPS. These algorithms analyze extensive sets of patient data and acquire patterns that facilitate prognosticating the likelihood of future health conditions. The effectiveness of MDPS can be assessed through metrics like precision, sensitivity, and specificity. The utilization of MDPS offers numerous potential advantages, encompassing enhanced patient results, diminished medical expenses, and heightened healthcare operational efficiency. Additionally, MDPS can aid in investigating the interconnections between distinct diseases and their associated risk elements, potentially yielding fresh insights into the fundamental mechanisms of illnesses. On the whole, MDPS represents a promising strategy for enhancing healthcare outcomes by enabling early detection and timely intervention for individuals susceptible to multiple diseases.

**Problem Statement**

The concept of a multiple disease prediction system revolves around creating an intelligent solution capable of effectively anticipating the likelihood of a patient having one or more medical conditions. This anticipation is established through an evaluation of symptoms, medical background, and pertinent aspects. The main purpose of this system is to aid healthcare professionals in arriving at precise diagnoses and suggesting suitable treatment courses for patients. The creation of such a system encompasses the utilization of diverse methodologies and algorithms hailing from fields like machine learning, data mining, and artificial intelligence. These techniques are employed to handle substantial volumes of medical data and to discern patterns that facilitate the anticipation of disease outcomes. The system's competence should extend to accurately distinguishing among various diseases, even those that may exhibit comparable symptoms, and delivering treatment recommendations that correspond to the patient's specific condition.

**Proposed system:** Within the realm of a Multiple Disease Prediction System (MDPS), it possesses the ability to concurrently forecast diverse ailments with precision. These encompass conditions such as breast cancer, heart issues, kidney problems, liver ailments, and even malaria. This eliminates the necessity for users to navigate through distinct websites for disease predictions. To bring the Multiple Disease Prediction System (MDPS) to fruition, a mixture of machine learning algorithms and Flask are employed for implementation. This amalgamation enables the efficient realization of the predictive system.

**System Analysis:** Conducting a thorough assessment of a multiple disease presentation system holds paramount importance in guaranteeing its efficacy, productivity, and security. Through a meticulous examination of the system's operational capabilities, design, and overall performance, healthcare experts can pinpoint potential concerns and implement essential enhancements. This ensures that the system furnishes precise and dependable predictions, thereby enhancing 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**

The architecture outlined above encompasses experimentation with seven interconnected diseases: heart conditions, diabetes, Breast Cancer, liver ailments, Kidney issues, Malaria, and Pneumonia. The initial phase involves the acquisition of datasets for each respective disease, drawing from sources such as the Winconsin Breast Cancer Dataset, PIMA Indian Diabetes Dataset, Indian liver dataset, Chronic Kidney Disease Dataset, Malaria Cell images Dataset, and Chest x-ray images for Pneumonia.

Once the datasets are obtained, they undergo visualization to gain insights into the input data. Following visualization, the data undergoes preprocessing, encompassing outlier detection, handling missing values, and dataset scaling. This process yields a refined dataset, subsequently divided into two segments: training and testing data.

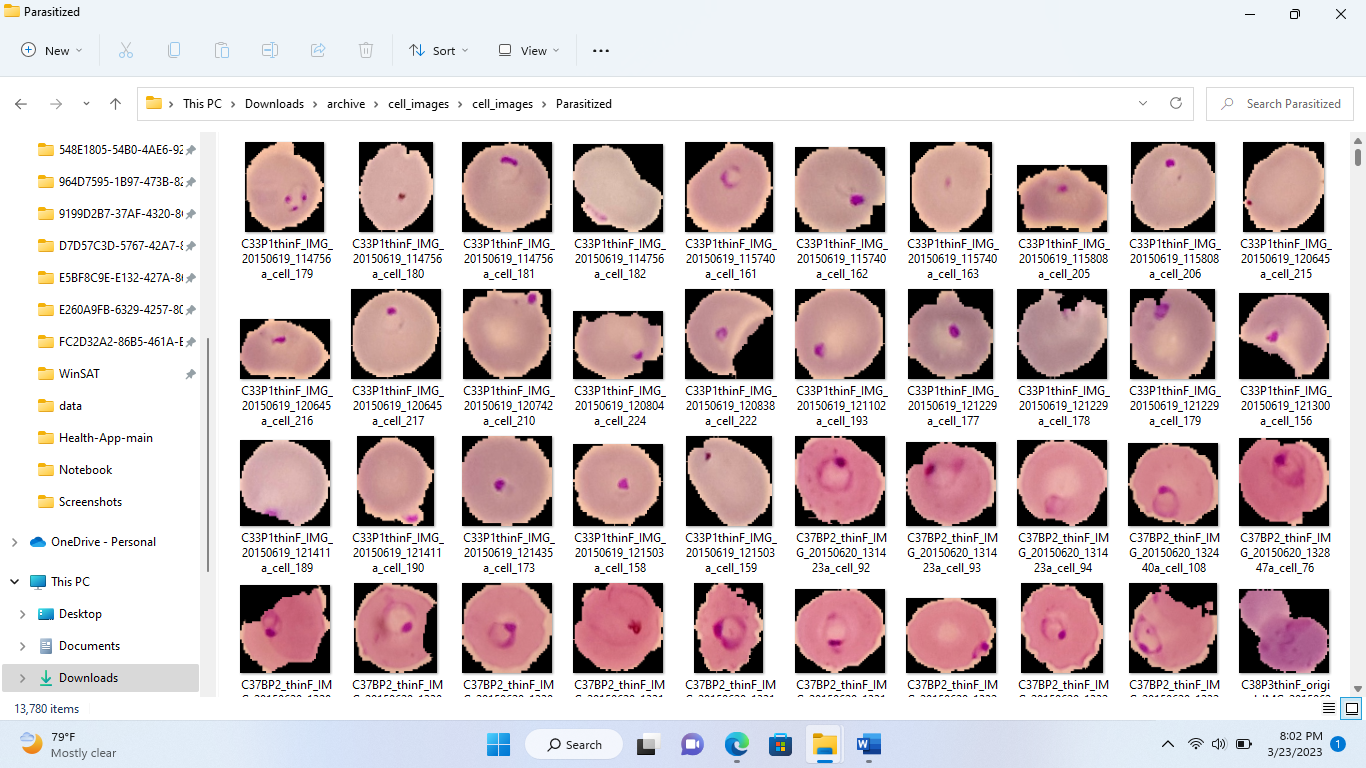
Moving forward, the architecture employs various algorithms including Logistic Regression, Random Forest, and SVM to train the models. These models are then tested using the designated testing dataset. Afterward, the most accurate algorithm (KNN, SVM, RD, DT) is selected for each disease.

The final steps involve creating pickle files for all diseases and integrating these files with the Flask framework. This integration facilitates model output display on a webpage. This entire process is undertaken to ensure the accuracy, reliability, and accessibility of disease predictions through the developed architecture.

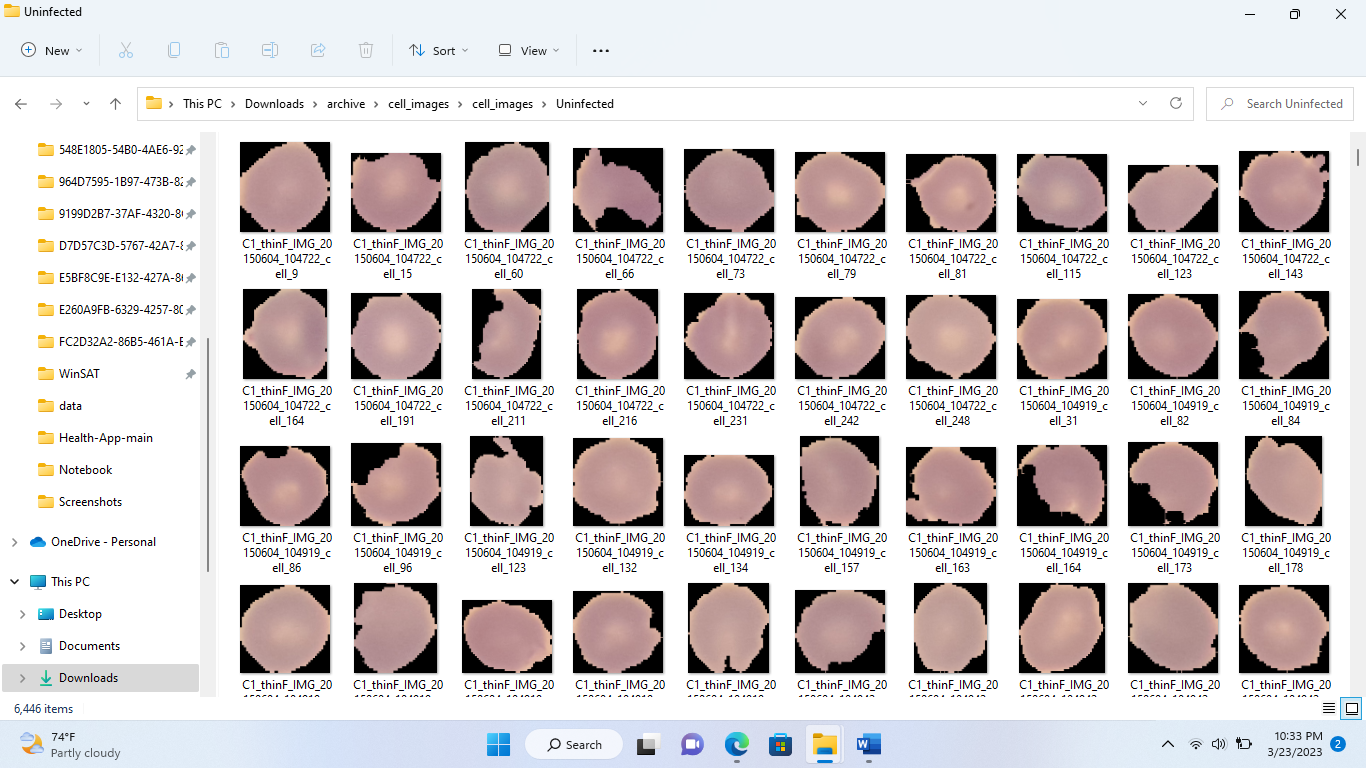
**Dataset Collection and processing:**

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:

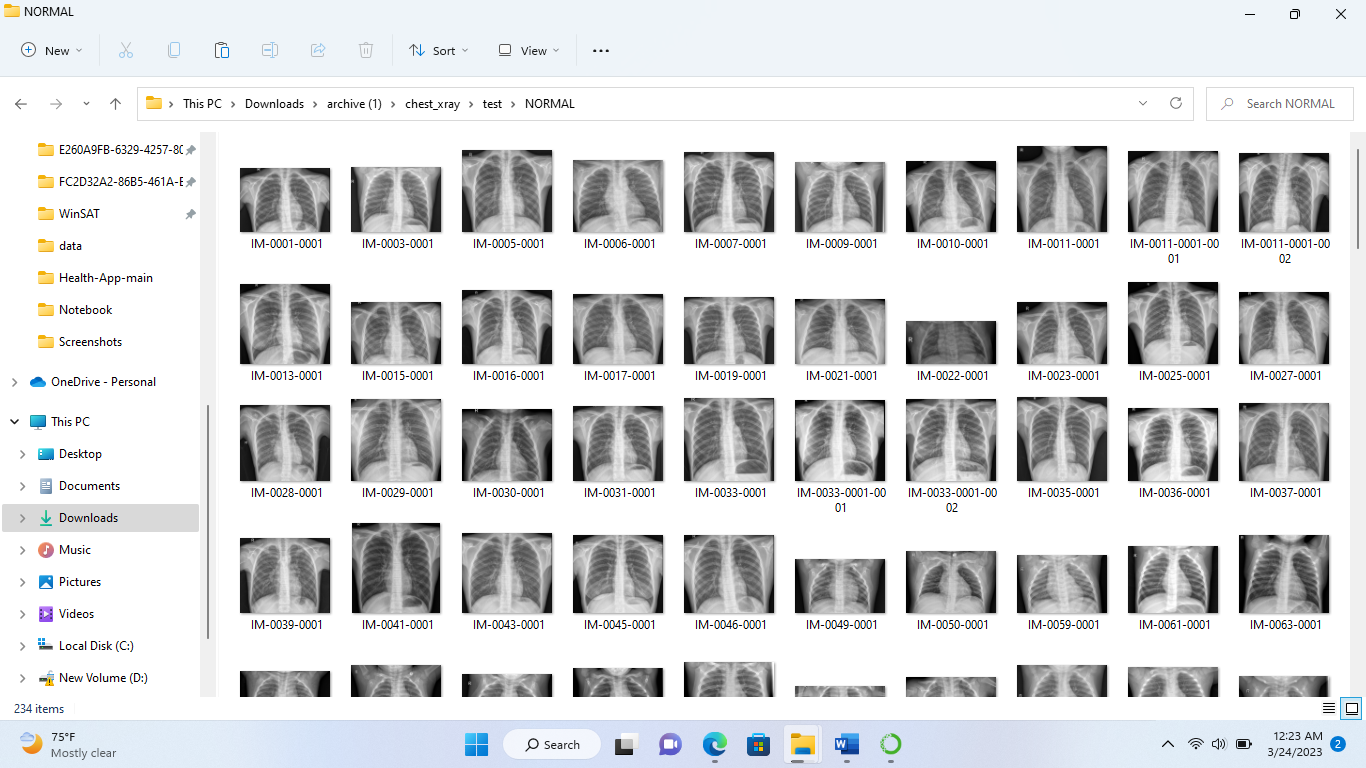


**Sample of uninfected**

****

**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 employs the sigmoid function, which accepts the linear combination denoted as z as input and transforms it into a value confined within the range of 0 to 1. As a result, the model is capable of generating a probability indicating the likelihood that the dependent variable equals 1, based on the independent variable values. Should this probability exceed a specified threshold, the model predicts that the dependent variable is 1; otherwise, it predicts the value as 0.

**Random Forest**

The random forest algorithm constitutes an ensemble learning technique that amalgamates numerous decision trees to produce a conclusive prediction. Within the random forest framework, each individual decision tree is crafted employing a randomized subset of the training data and a random subset of the available 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.

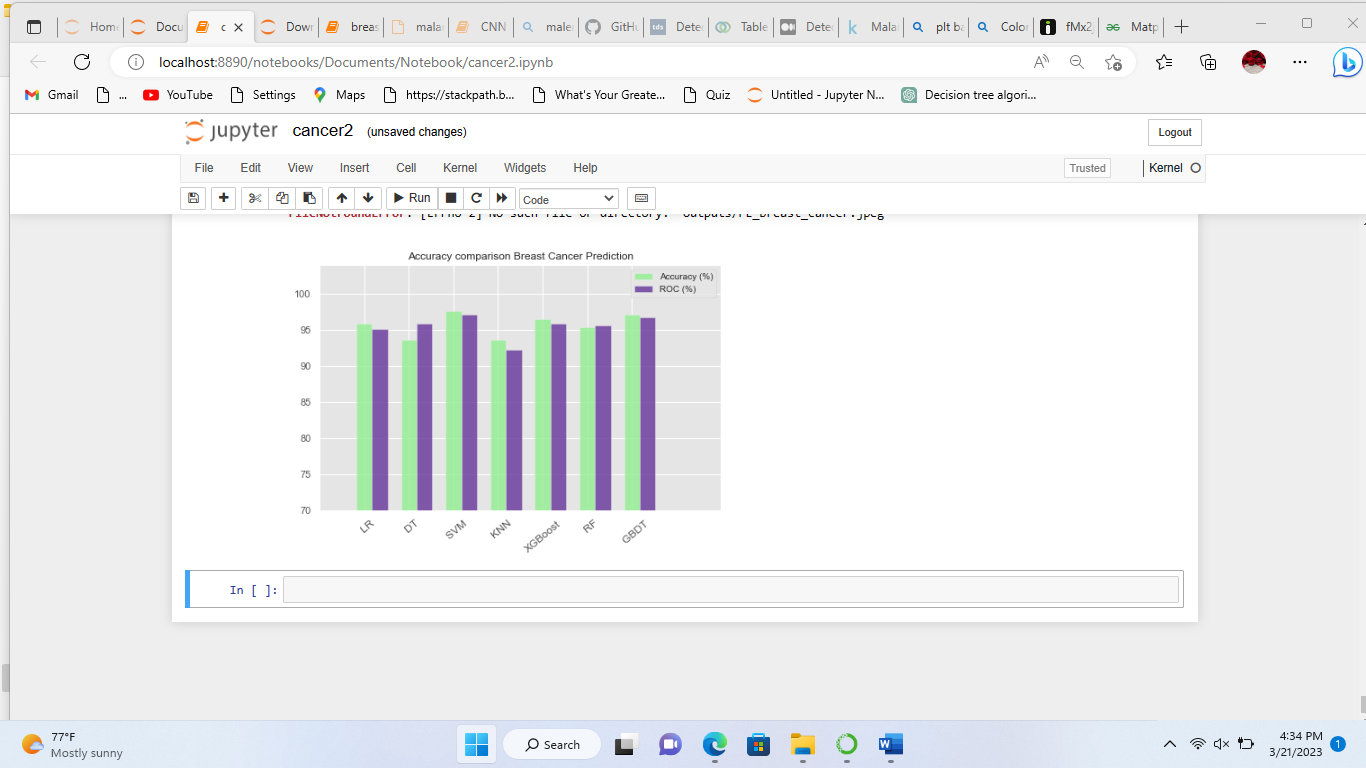
**Transfer Learning:**

Transfer learning is a machine learning strategy that entails utilizing a pre-existing model trained for a similar task as a foundation for a fresh task. This approach can diminish the data and computational demands for training a new model, consequently enhancing its overall performance.

**Accuracy comparison for each disease:**

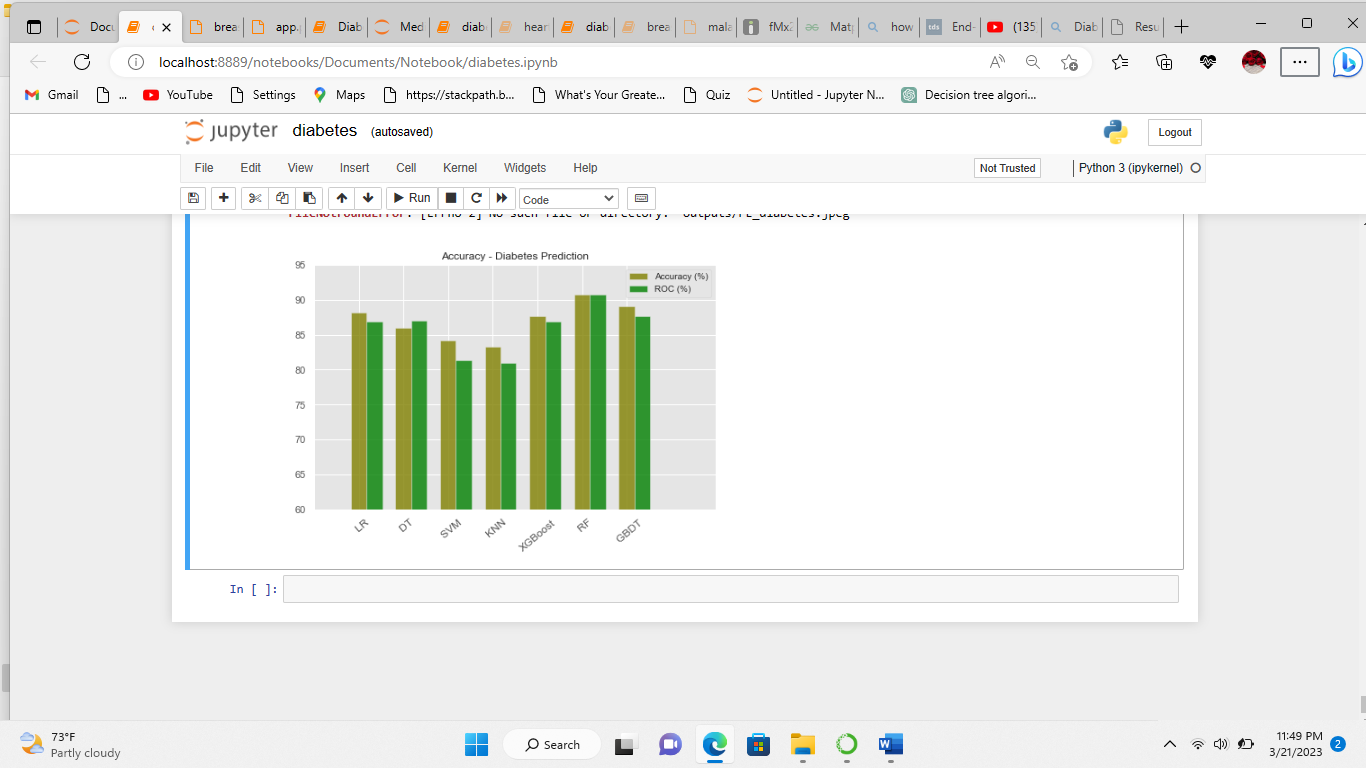
**Breast Cancer:**

In the context of Breast cancer, we have employed several algorithms to assess and contrast accuracy scores. The results are as follows: SVM - 97.66, Gradient Boosting Classifier - 97.66, Random Forest Classifier - 96.49, XGBoost - 96.49, K-Nearest Neighbors (KNN) - 93.57, Decision Tree - 95.32, and Logistic Regression - 95.51.

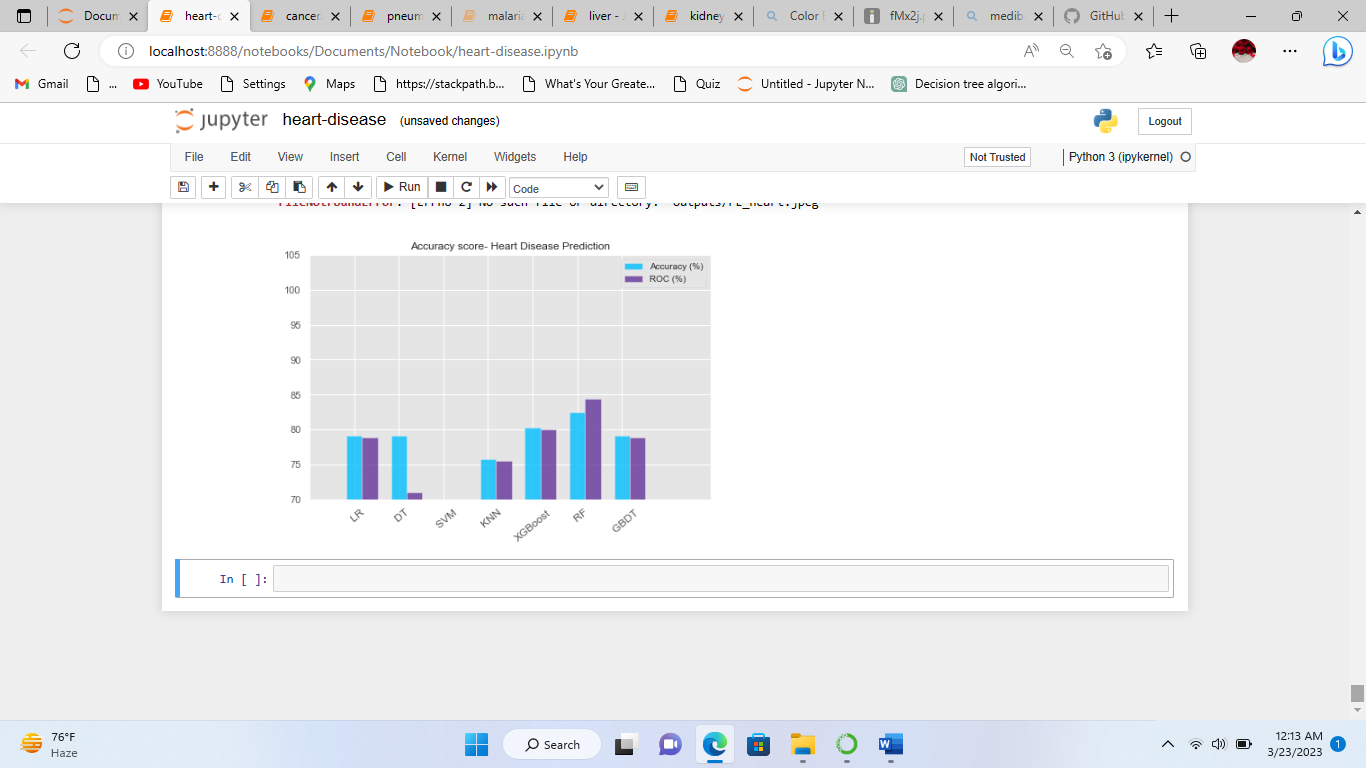


**Diabetes:**

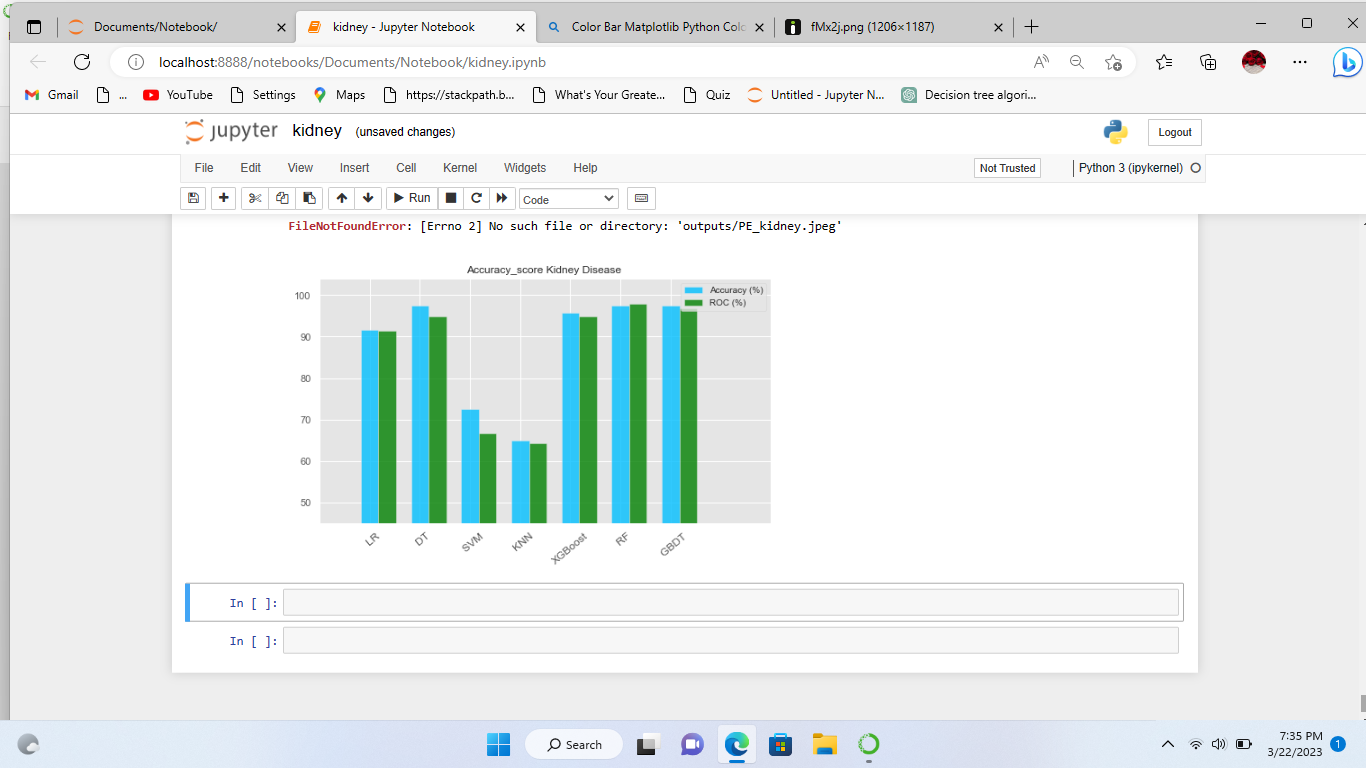
For Diabetes disease, after comparison, we get, Random Forest Classifier- 90.79 with highest accuracy score, SVM -84.21, Gradient boosting classifier-89.04, Xgboost- 87.72, KNN-83.33 Decision tree-85.96 logistic-regression-88.16.

****

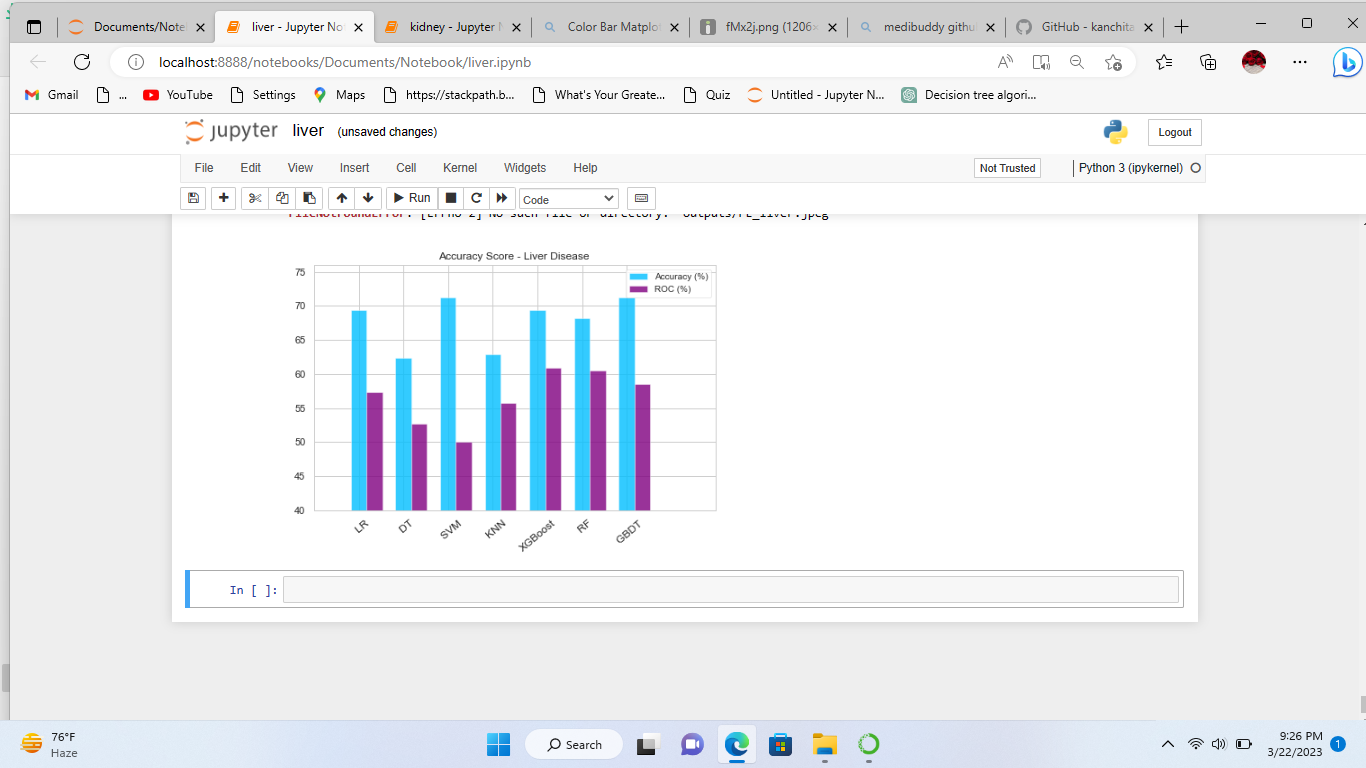
**Heart disease**: After comparison, we get, Random Forest Classifier-0.824176 with highest accuracy score, SVM -0.516484, Gradient boosting classifier- 0.791209, Xgboost- 0.802198, KNN-83.33, Decision tree-0.791209, logistic-regression-0.791209

****

**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

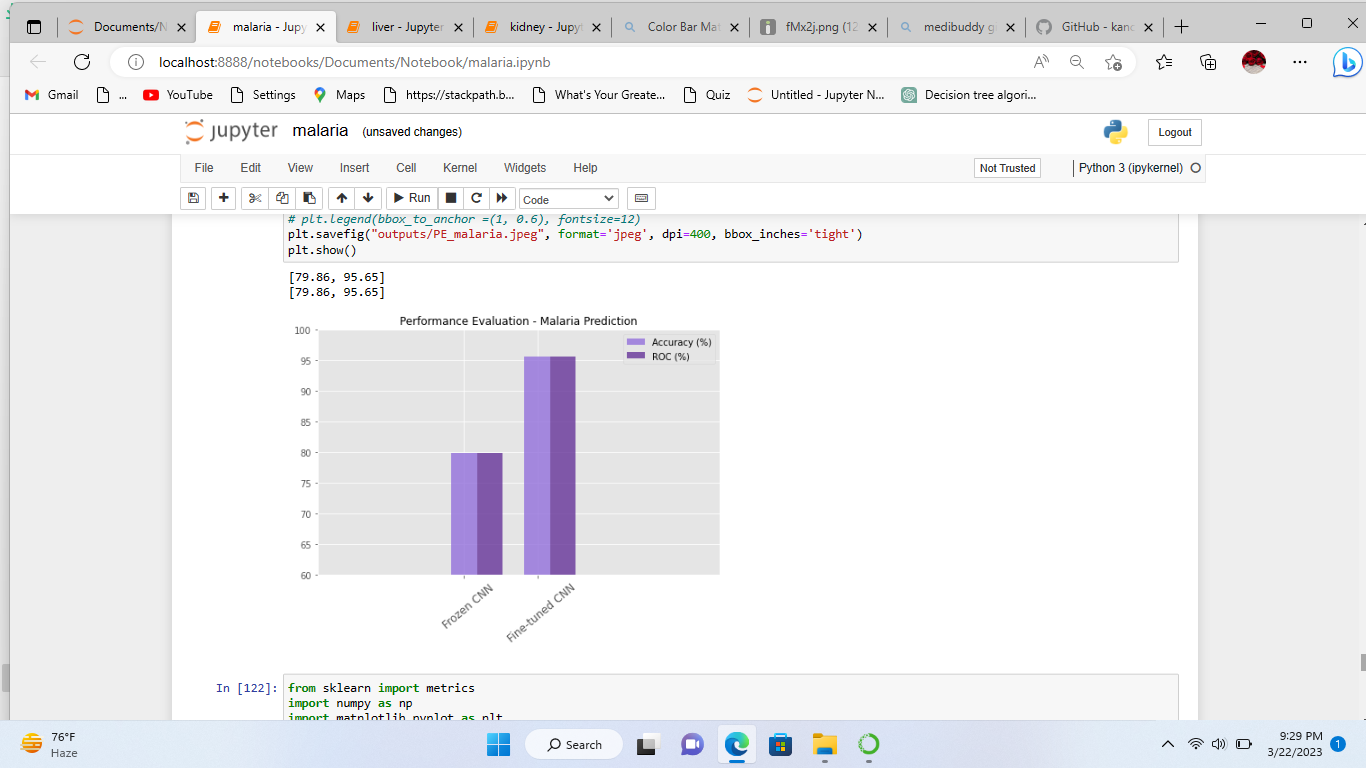
****

**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

****

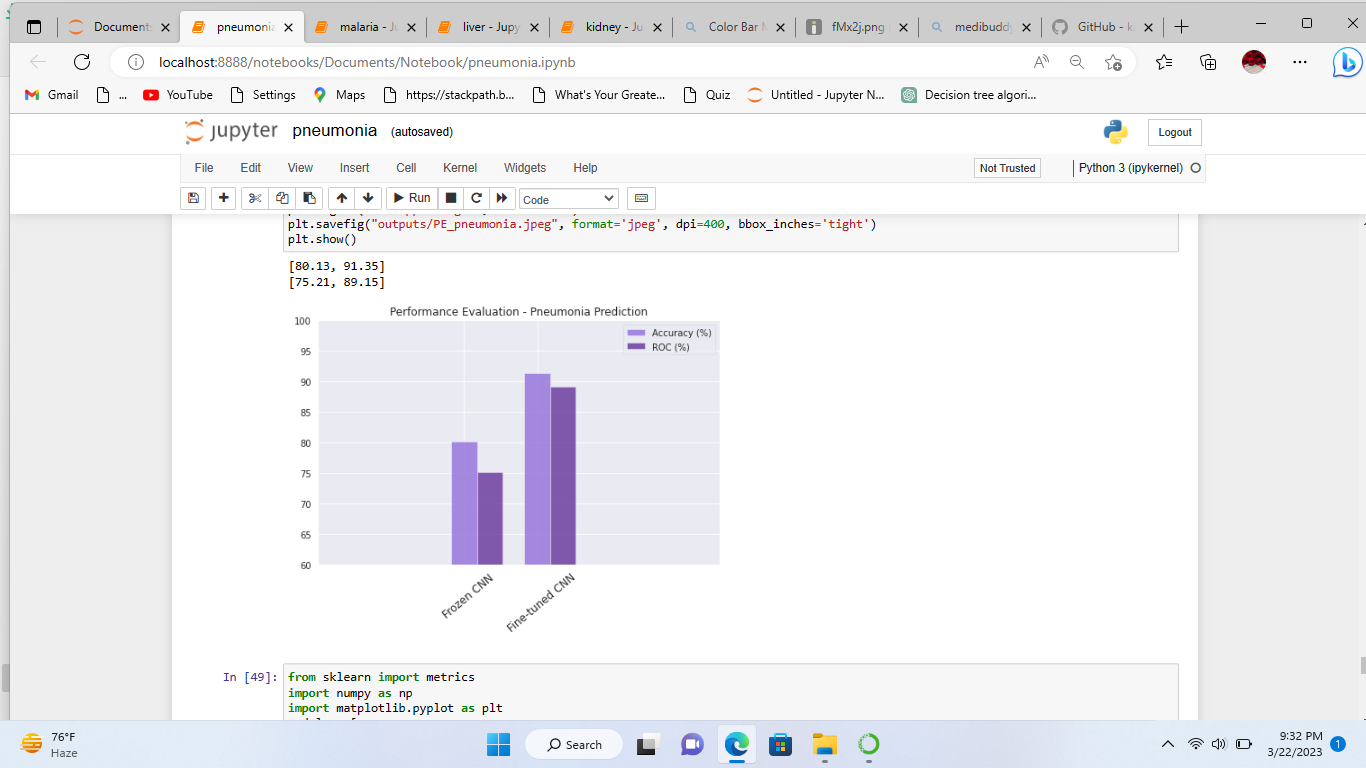
**Malaria:**

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%.

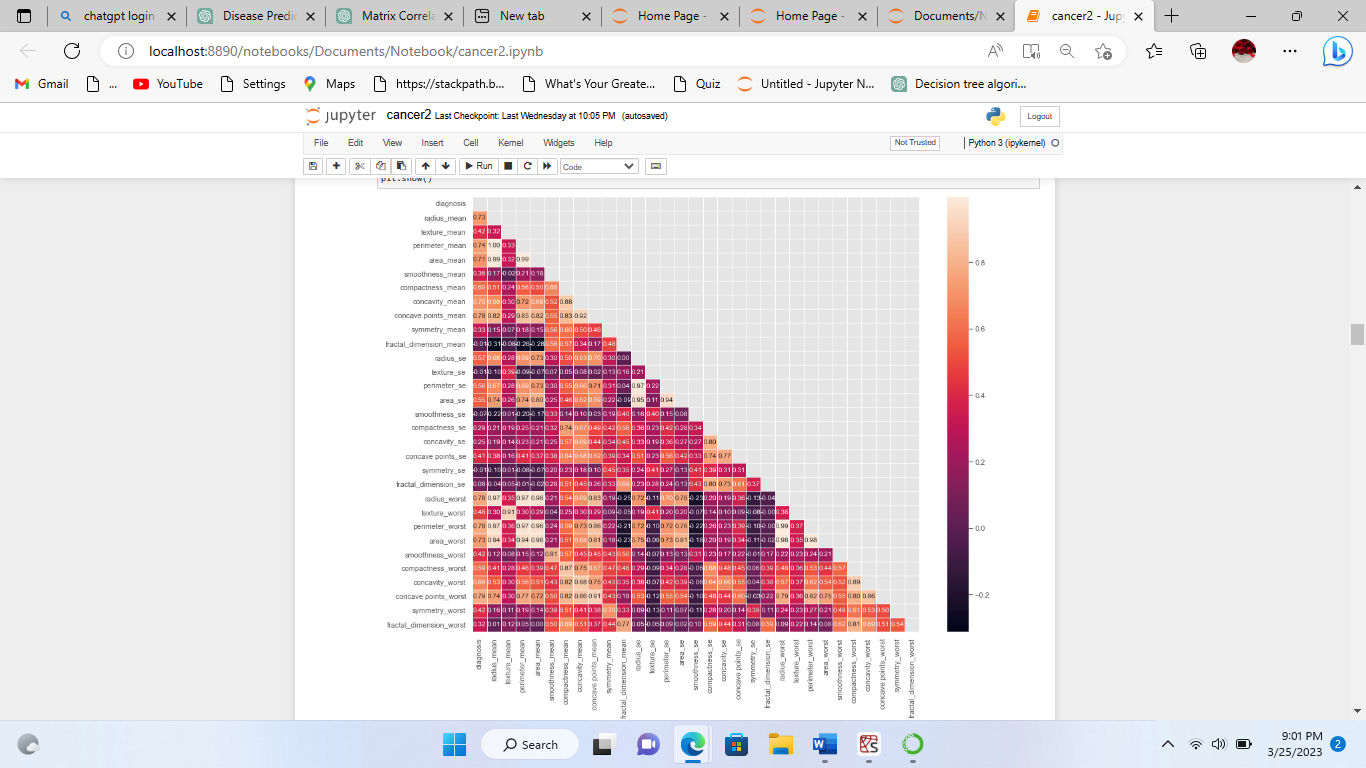
****

**Result**

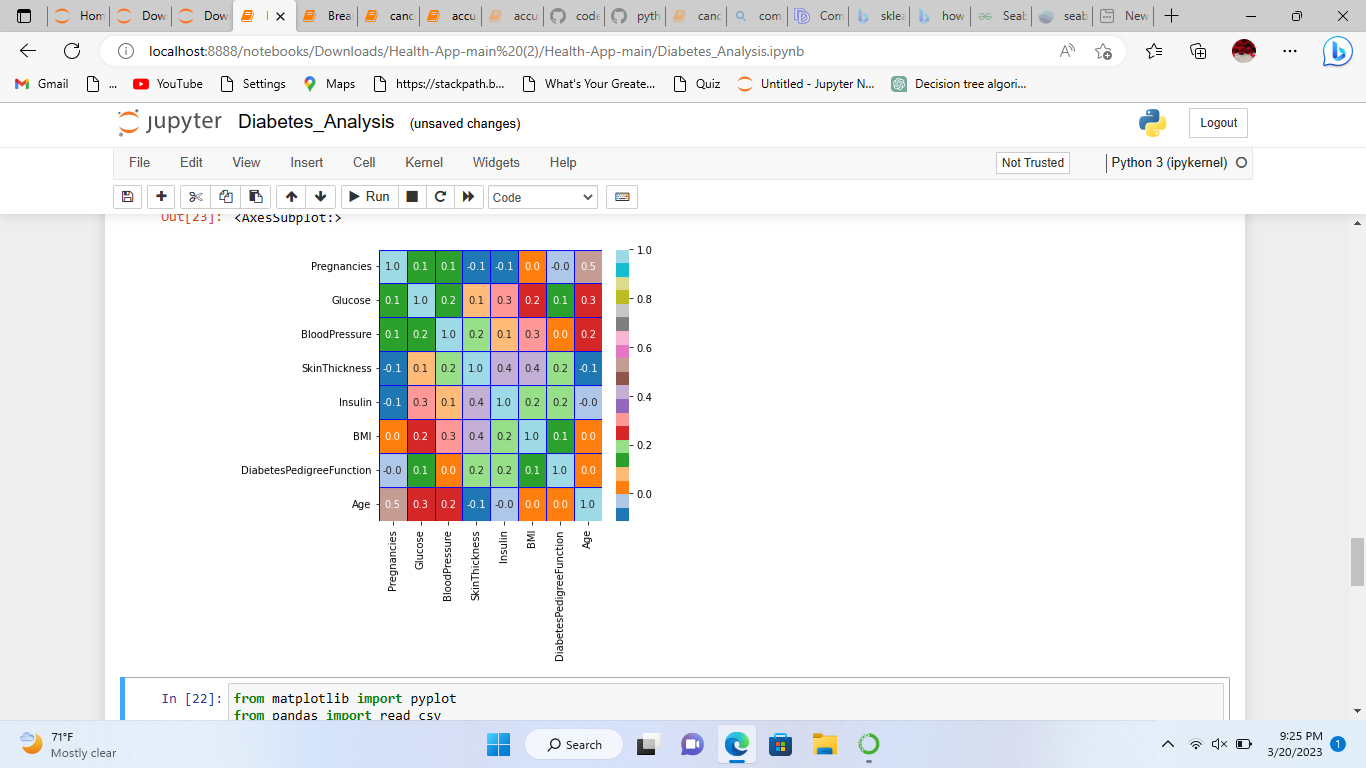
Within the Multiple Disease Prediction System, distinct models are employed for predicting diabetes, heart disease, kidney issues, liver conditions, and breast cancer. These models are built utilizing algorithms such as SVM, Decision Tree, and Random Forest. Conversely, for Malaria and Pneumonia, Convolutional Neural Networks (CNN) and Transfer Learning are leveraged due to their superior accuracy. When a patient provides relevant parameters for a specific disease (such as breast cancer, diabetes, liver, kidney, or heart disease), the system assesses whether the patient exhibits the disease. The parameters furnished by the patient are validated within predefined ranges; any out-of-range or invalid values trigger a warning, prompting the patient to input accurate data. For Malaria and Pneumonia, the system processes uploaded images to predict whether they portray an infected or uninfected state, delivering predictions with high accuracy.

Matrix correlation: In the context of machine learning, the matrix correlation metric employs the Pearson correlation coefficient. This coefficient gauges the degree of linear correlation between two continuous variables. Its values span from -1 to 1, with -1 signifying a complete negative correlation.

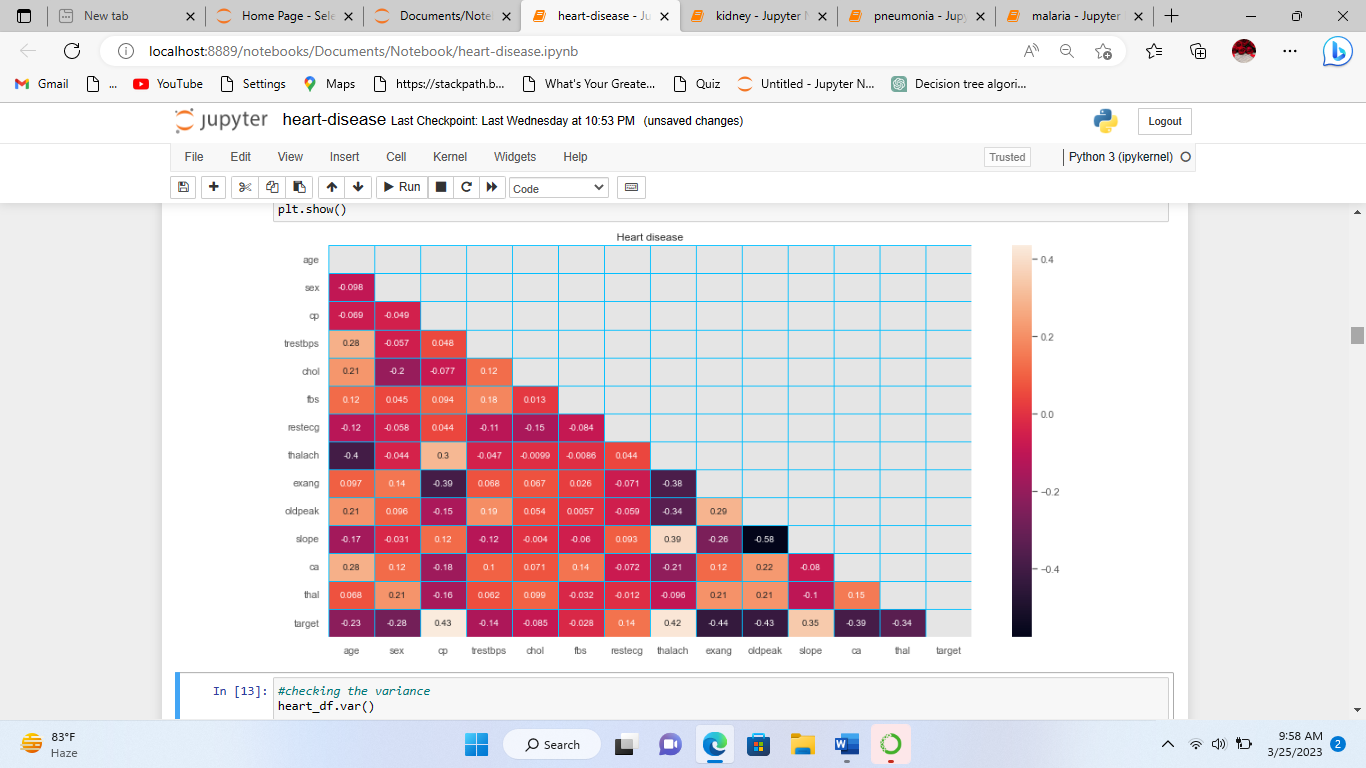
**Matrix correlation of Breast Cancer disease:**

****

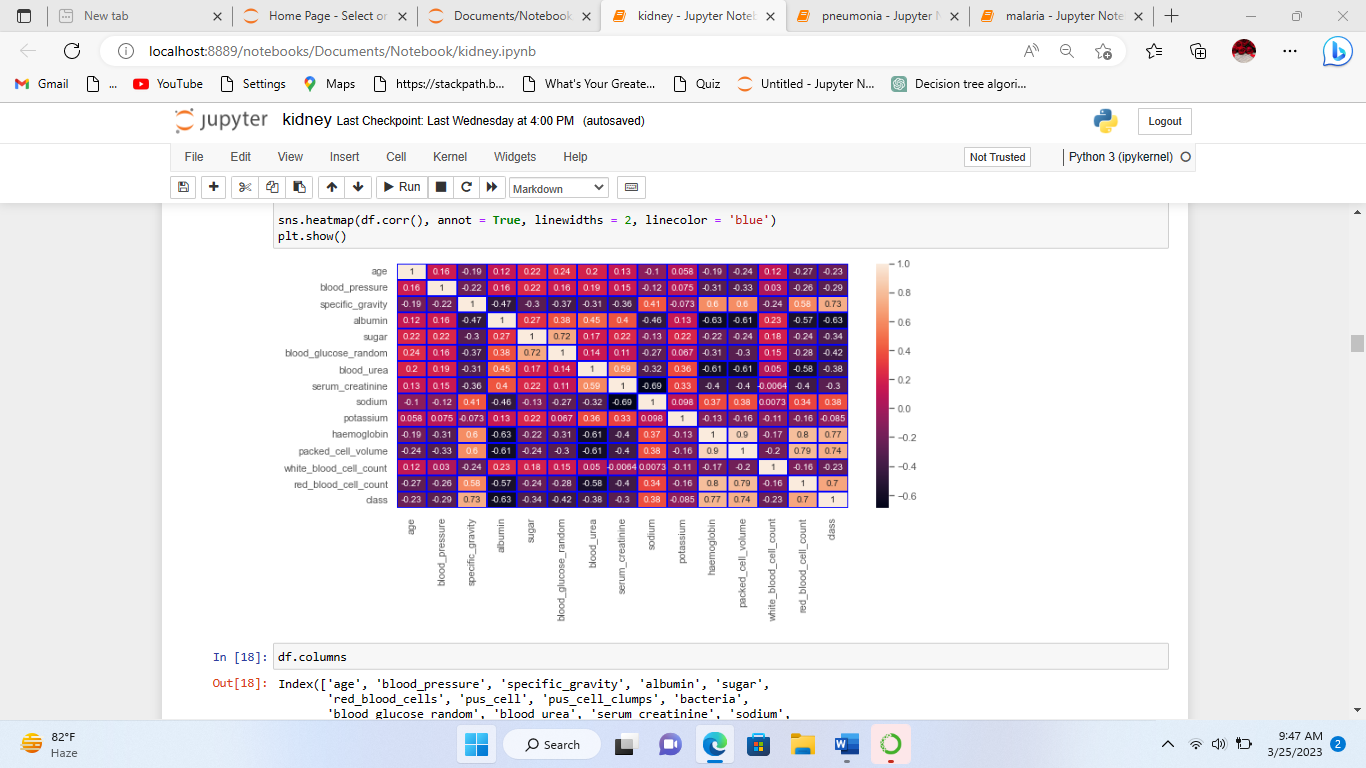
**Matrix correlation of Diabetes disease:**

****

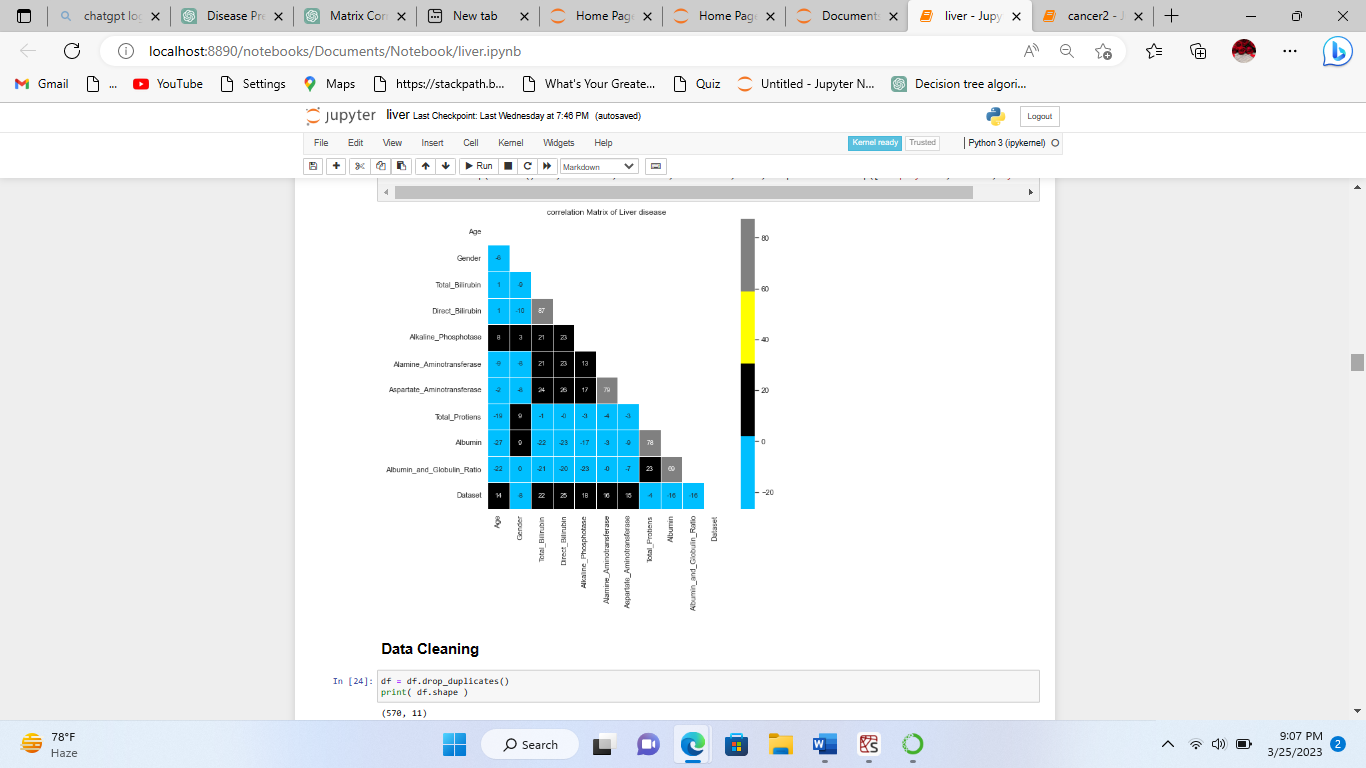
**Matrix correlation of Heart disease:**

****

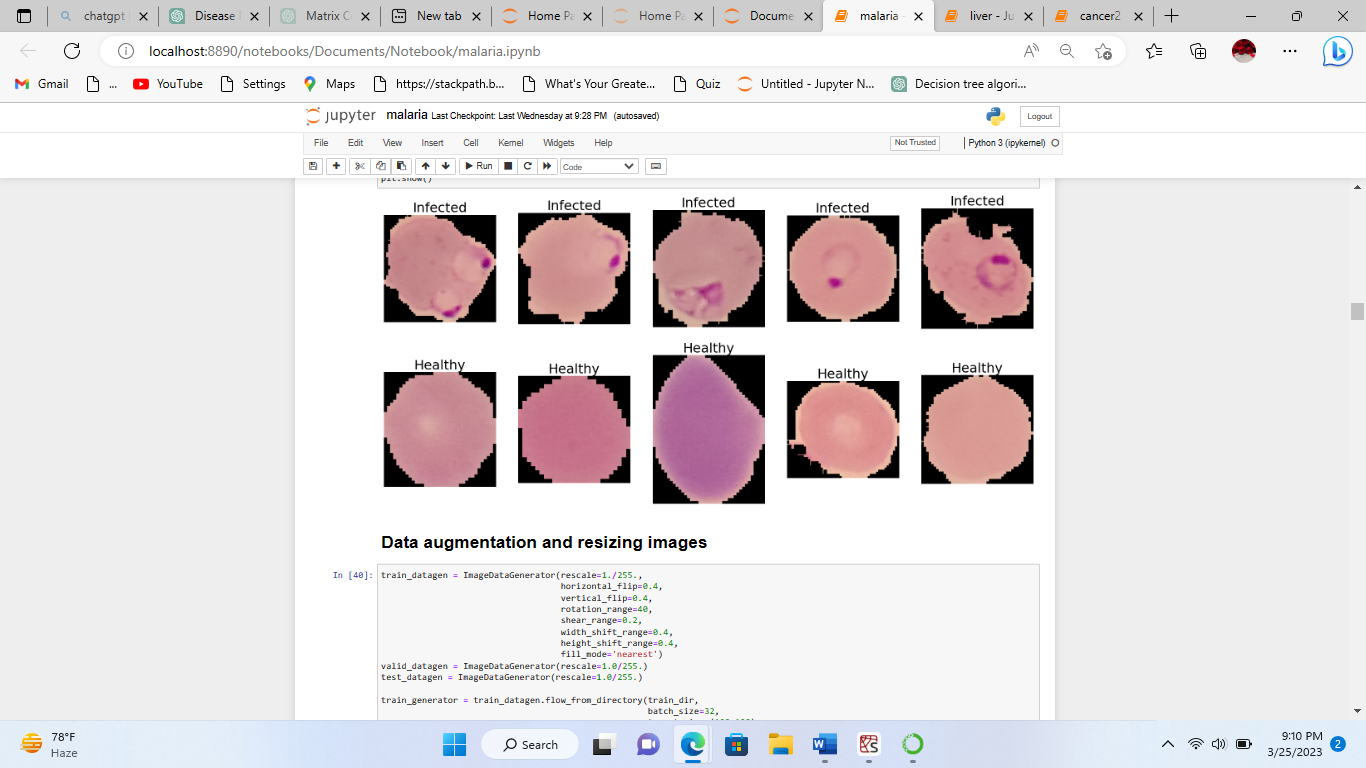
**Matrix correlation of kidney disease:**

****

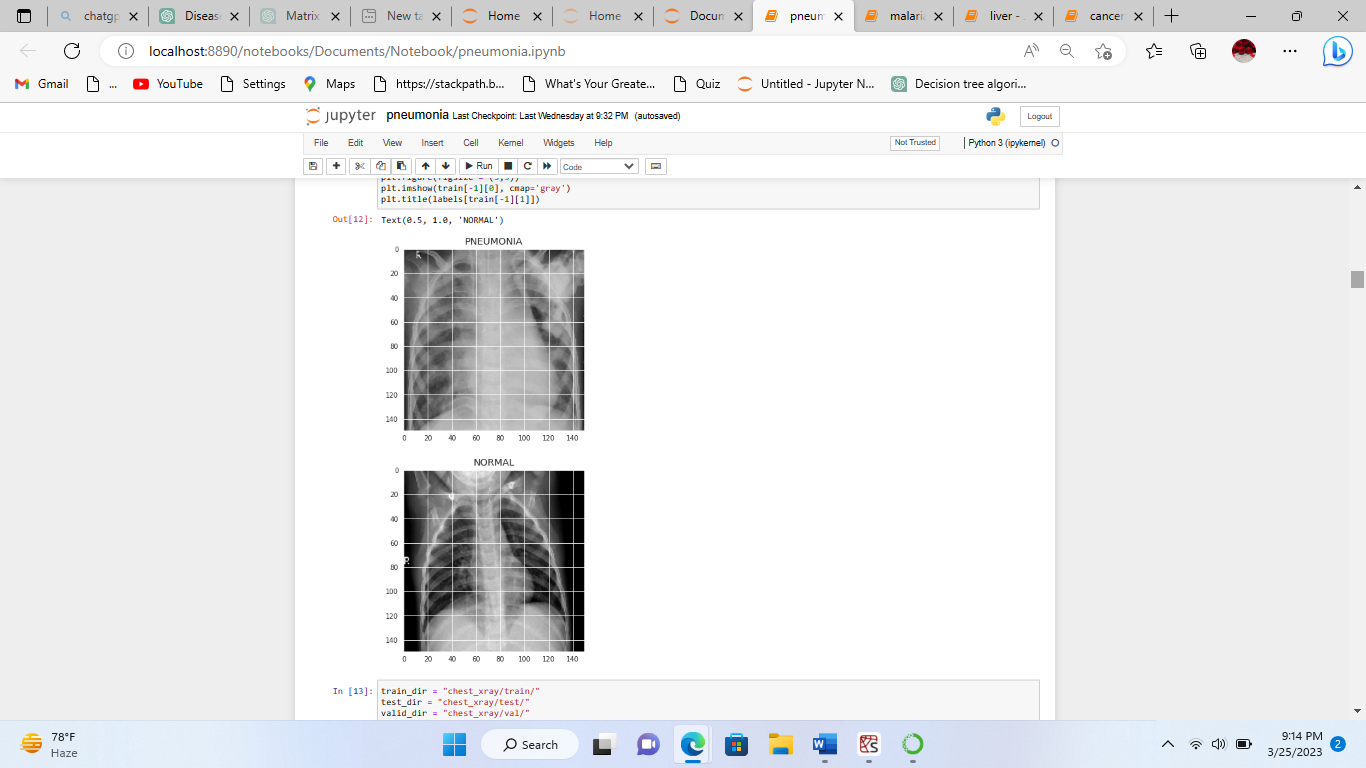
**Matrix correlation of Liver disease:**

****

**Result of Malaria:**

****

**Result of Pneumonia:**

****

**Conclusion:**

A multiple disease prediction system holds significant value by aiding both healthcare professionals and individuals in proactively identifying potential health concerns and taking appropriate measures for prevention or treatment. By scrutinizing a patient's symptoms, medical history, and pertinent information, these systems can furnish predictions regarding potential diseases or conditions that an individual might be susceptible to. This approach can contribute to enhanced patient outcomes, diminished healthcare expenses, and optimized resource utilization within the healthcare sector. Nevertheless, it's crucial to emphasize that these systems should not substitute the expertise of healthcare practitioners; rather, they should complement their judgment by assisting in diagnosis and treatment planning. Continued research and advancements in this field could result in more precise and efficacious disease prediction systems, ultimately benefiting public health and individual well-being.

**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>