LifeAI

Multiple Disease Predictor Software

ANANYA GHOSH

20MIC0063

Submitted to
Prof. Boominathan P., SCOPE

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

CONTENTS

1. Abstract

LifeAI- Multiple Disease Detection Software will be a platform for the users to fill their data and get to know the predictions of them being positive or negative to certain chronic diseases. As we know, due to the large population, getting healthcare access is also becoming difficult in various places. There is an increased pressure on the healthcare infrastructure especially due to the pandemic. Also, most people die of these chronic diseases due to late checking. This platform will help users in all these aspects. We use Machine Learning and Deep Learning Algorithms and BioInspired Algorithms along with Django to build this software.

2. Introduction

Alzheimer's disease is a brain disorder that gets worse over time. It's characterized by changes in the brain that lead to deposits of certain proteins. Alzheimer's disease causes the brain to shrink and brain cells to eventually die. Alzheimer's disease is the most common cause of dementia — a gradual decline in memory, thinking, behavior and social skills. These changes affect a person's ability to function. About 6.5 million people in the United States age 65 and older live with Alzheimer's disease. Among them, more than 70% are 75 years old and older. Of the about 55 million people worldwide with dementia, 60% to 70% are estimated to have Alzheimer's disease. The early signs of the disease include forgetting recent events or conversations. Over time, it progresses to serious memory problems and loss of the ability to perform everyday tasks. Medicines may improve or slow the progression of symptoms. Programs and services can help support people with the disease and their caregivers. There is no treatment that cures Alzheimer's disease. In advanced stages, severe loss of brain function can cause dehydration, malnutrition or infection. These complications can result in death.

Dementia is a term used to describe a group of symptoms affecting memory, thinking and social abilities severely enough to interfere with your daily life. It isn't a specific disease, but several diseases can cause dementia. Though dementia generally involves memory loss, memory loss has different causes. Having memory loss alone doesn't mean you have dementia, although it's often one of the early signs of the condition.

Breast cancer arises in the lining cells (epithelium) of the ducts (85%) or lobules (15%) in the glandular tissue of the breast. Initially, the cancerous growth is confined to the duct or lobule (“in situ”) where it generally causes no symptoms and has minimal potential for spread (metastasis). Over time, these in situ (stage 0) cancers may progress and invade the surrounding breast tissue (invasive breast cancer) then spread to the nearby lymph nodes (regional metastasis) or to other organs in the body (distant metastasis). If a woman dies from breast cancer, it is because of widespread metastasis. Breast cancer treatment can be highly effective, especially when the disease is identified early. Treatment of breast cancer often consists of a combination of surgical removal, radiation therapy and medication (hormonal therapy, chemotherapy and/or targeted biological therapy) to treat the microscopic cancer that has spread from the breast tumor through the blood. Such treatment, which can prevent cancer growth and spread, thereby saves lives.

Diabetes mellitus refers to a group of diseases that affect how the body uses blood sugar (glucose). Glucose is an important source of energy for the cells that make up the muscles and tissues. It's also the brain's main source of fuel. The main cause of diabetes varies by type. But no matter what type of diabetes you have, it can

lead to excess sugar in the blood. Too much sugar in the blood can lead to serious health problems. Chronic diabetes conditions include type 1 diabetes and type 2 diabetes. Potentially reversible diabetes conditions include prediabetes and gestational diabetes. Prediabetes happens when blood sugar levels are higher than normal. But the blood sugar levels aren't high enough to be called diabetes. And prediabetes can lead to diabetes unless steps are taken to prevent it. Gestational diabetes happens during pregnancy. But it may go away after the baby is born.

Heart/ Coronary artery disease is a common heart condition that affects the major blood vessels that supply the heart muscle. Cholesterol deposits (plaques) in the heart arteries are usually the cause of coronary artery disease. The buildup of these plaques is called atherosclerosis (ath-ur-o-skluh-ROE-sis). Atherosclerosis reduces blood flow to the heart and other parts of the body. It can lead to a heart attack, chest pain (angina) or stroke. Coronary artery disease symptoms may be different for men and women. For instance, men are more likely to have chest pain. Women are more likely to have other symptoms along with chest discomfort, such as shortness of breath, nausea and extreme fatigue.

Lung cancer is a type of cancer that begins in the lungs. Your lungs are two spongy organs in your chest that take in oxygen when you inhale and release carbon dioxide when you exhale. Lung cancer is the leading cause of cancer deaths worldwide. People who smoke have the greatest risk of lung cancer, though lung cancer can also occur in people who have never smoked. The risk of lung cancer increases with the length of time and number of cigarettes you've smoked. If you quit smoking, even after smoking for many years, you can significantly reduce your chances of developing lung cancer.

Parkinson's disease is a brain disorder that causes unintended or uncontrollable movements, such as shaking, stiffness, and difficulty with balance and coordination. Symptoms usually begin gradually and worsen over time. As the disease progresses, people may have difficulty walking and talking. They may also have mental and behavioral changes, sleep problems, depression, memory difficulties, and fatigue. Older woman and her caregiverWhile virtually anyone could be at risk for developing Parkinson's, some research studies suggest this disease affects more men than women. It's unclear why, but studies are underway to understand factors that may increase a person's risk. One clear risk is age: Although most people with Parkinson's first develop the disease after age 60, about 5% to 10% experience onset before the age of 50. Early-onset forms of Parkinson's are often, but not always, inherited, and some forms have been linked to specific alterations in genes.

LifeAI- Multiple Disease Detection Software will be a platform for the users to fill their data and get to know the predictions of them being positive or negative to certain chronic diseases. As we know, due to the large population, getting healthcare access is also becoming difficult in various places. There is an increased pressure on the healthcare infrastructure especially due to the pandemic. Also, most people die of these chronic diseases due to late checking. This platform will help users in all these aspects. We use Machine Learning and Deep Learning Algorithms and BioInspired Algorithms along with Django to build this software.

With the popularity of artificial intelligence (AI) and its integration into every field, it has yielded positive results that have increased productivity and aided us in solving complex problems. Deep learning (DL) is a subset of AI developed to mimic the human brain. It is a way for a computer to perform actions that come naturally to humans. It is a tool extensively used to organize unsupervised or unlabeled data and find patterns in them. DL has had a major impact on the field of medical science owing to its applications in medical drug

discovery, medical imaging, genome synthesis, disease diagnosis, and much more. The growth of DL in this field has increased significantly owing to the processing and type of data used in the models. Focusing on the type of data, preexisting or curated, can drastically affect the success rate of a DL model.

3. Proposed Model/ Solution

LifeAI- A multiple disease predictor software is the proposed solution to the healthcare industry for disease prediction. It predicts 6 diseases, namely Alzheimer's Dementia, Heart Disease, Breast Cancer, Diabetes, Parkinson's and Lung Cancer. I used several different algorithms to predict different diseases, so we can compare the working of the algorithms and their accuracies. LifeAI- Multiple Disease Detection Software will be a platform for the users to fill their data and get to know the predictions of them being positive or negative to certain chronic diseases.

Alzheimer's Dementia Prediction- 2D CNN for Dementia Architecture:

The preprocessed data has been taken from Kaggle Dataset. The dataset is in 2D format. The train and test split is done and data is segregated into 2 folders 'train' and 'test'. There are 4 classes for prediction - 'Very Mild Demented', 'Mild Demented', 'Moderate Demented' and 'Non Demented'. The model is a Sequential model, meaning the layers will be stacked in the order they are added to the model. The first layer, "Rescaling", is a preprocessing layer that scales the input data by dividing it by 255. This normalizes the input data, so it falls in the range of 0 to 1. The input shape of this layer is defined as (IMG_HEIGHT, IMG_WIDTH, 3), meaning the input images are expected to have three color channels (red, green, and blue). The next several layers are Convolutional Neural Network (CNN) layers. The first layer is a Conv2D layer with 16 filters and a kernel size of (3,3). The padding is set to 'same', meaning the padding around the input is such that the spatial dimensions of the output remain the same. The activation function is 'relu', meaning rectified linear unit activation. The kernel initializer is set to "he_normal", meaning the weights are initialized using the He normal distribution.

After the Conv2D layer, there is a MaxPooling2D layer, which performs pooling to reduce the spatial dimensions of the feature maps. This helps to reduce the computational cost of the model and prevent overfitting. This pattern of Conv2D and MaxPooling2D layers is repeated two more times, with the number of filters in the Conv2D layer increasing to 32 and then 64. Between the Conv2D and MaxPooling2D layers, there are Dropout layers that randomly set a portion of the neurons in the layer to 0 during training, helping to prevent overfitting. After the final MaxPooling2D layer, there is a Flatten layer that flattens the feature maps into a single vector. This is followed by two dense layers, or fully connected layers, with 128 and 64 neurons, respectively. The activation functions for these layers are 'relu', meaning rectified linear unit activation. The final layer is a Dense layer with 3 neurons, which corresponds to the number of classes in the target variable. The activation function for this layer is 'softmax', meaning the outputs of this layer will be normalized probabilities that sum to 1. This makes the final layer suitable for multi-class classification problems.

The model.compile method is used to specify the loss function and optimizer for the model. The loss function chosen is "sparse_categorical_crossentropy". This is a commonly used loss function for multi-class classification problems when the target classes are represented as integer values. The optimizer used is "Adam". This is a gradient-based optimization algorithm that adapts the learning rate of each parameter during

training. The metrics argument is used to specify the evaluation metric for the model. In this case, the accuracy of the model is used as the evaluation metric.

They perform convolutions on the input image with a specified number of filters, kernel size, and padding. The activation function used in these layers is the rectified linear unit (relu). The MaxPooling2D layers are used for down-sampling the input image by taking the maximum value in a specified pool size. The Dropout layers are used for regularization to prevent overfitting. The Flatten layer is used to flatten the output from the last Conv2D layer into a 1D vector. The last three layers are Dense layers, which are fully connected layers in a neural network. They are used to make predictions based on the output from the Flatten layer. The activation function used in the last layer is "softmax", which is used for multi-class classification. The output from this layer is a probability distribution over the different classes. Finally, the total number of parameters in the model is 2,129,380 and all of them are trainable.

Then we trained the neural network model on the training dataset "train_ds" and validated it on the validation dataset "val_ds" using the fit() method of the model. The training process is performed for 100 epochs, with a batch size of 64, and verbosity level set to 1. The verbosity level 1 means that the model will show a progress bar during the training process. The output of the training process is stored in the "hist" variable.

The accuracy and loss of the model during training and validation is plotted. The 'hist' object is a history object returned by the 'fit' method of the model that stores the history of training and validation accuracy and loss. The first plot shows the accuracy and loss of the training data. The second plot shows the accuracy of both training and validation data. The third plot shows the loss of both training and validation data. The 'get_ac' and 'get_los' variables are the accuracy and loss of the training data respectively. In the plots, the green line represents the accuracy or loss of the training data and the red line represents the accuracy or loss of the validation data.

Then the performance of the trained model is evaluated on the test data set. The model.evaluate() method takes in the test data set and returns the loss and accuracy of the model on the test data. The returned values are stored in the variables loss and accuracy respectively. The loss value represents how well the model is able to make correct predictions on the test data and the accuracy value represents how well the model is able to classify the test data into the correct categories. Then we test the model on testing data, save the model and form the confusion matrix.

The web application has been created for predicting very mild dementia, mild dementia, moderate dementia and no dementia from the MRI scans uploaded by users. The user interface is designed using HTML, CSS and the backend has been developed using Django, a Python Framework. The user interface is intuitive and easy to use, with clear instructions on how to use the application. The uploaded image is in 2D format jpg format. The model integrated predicts the status of the MRI as very mildly demented, mildly demented, moderately demented and non demented.

Breast Cancer Prediction- Hyperparameter Tuning and Genetic Algorithm:

I used the Ensemble Learning algorithm ExtraTreesRegressor, RandomForest Classifier, GridSearchCV to predict at the beginning and conduct the hyperparameter tuning. Then we use the keras tensorflow Sequential

Model. We then use Principal Component Analysis for dimensionality reduction and apply MLPClassifier along with Genetic Algorithm for 50 Generations. We receive an accuracy of 96.5%.

Genetic Algorithm (GA) is a search-based optimization technique based on the principles of Genetics and Natural Selection. It is frequently used to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifetime to solve. It is frequently used to solve optimization problems, in research, and in machine learning. In GAs, we have a pool or a population of possible solutions to the given problem. These solutions then undergo recombination and mutation (like in natural genetics), producing new children, and the process is repeated over various generations. Each individual (or candidate solution) is assigned a fitness value (based on its objective function value) and the fitter individuals are given a higher chance to mate and yield more "fitter" individuals. This is in line with the Darwinian Theory of "Survival of the Fittest".

GA()
initialize population
find fitness of population
while (termination criteria is reached) do
parent selection
crossover with probability pc
mutation with probability pm
decode and fitness calculation
survivor selection
find best
return best

Heart Disease Prediction- Ensemble Learning with Ant Colony Optimization:
In Ensemble Learning we use 5 algorithms, namely, Adaboost, Bagging, Random Forest, Gradient Boosting and Extra Trees.

- Adaboost is to create a strong classifier by combining multiple weak classifiers. A weak classifier is a model that performs slightly better than random guessing, i.e., its accuracy is slightly better than 50%. The weak classifier is typically a decision tree with a depth of one, called a decision stump.
- Bagging (Bootstrap Aggregation) is a specific type of bagging algorithm that is used for classification tasks. It works by training multiple decision tree models on randomly sampled subsets of the training data, where each subset is selected with replacement (known as bootstrapping). The predictions of the individual models are then combined by taking a simple majority vote.
- Random Forest is a popular ensemble learning algorithm for classification and regression tasks. It is an extension of the bagging algorithm that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or the average prediction (regression) of the individual trees.
- Gradient Boosting is an iterative algorithm that starts with a weak learner (often a decision tree with a small depth) and then iteratively adds new trees to the ensemble by focusing on the data points that were misclassified by the previous trees.
- Extra Trees (short for Extremely Randomized Trees) is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time

and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Then we combine these classifiers with Ant Colony Optimization.

I applied the Ant Colony Optimization on each of the ensemble learning algorithms separately on the Heart Disease/ Breast Cancer dataset. Accuracy is the class to store the accuracy of a solution. Edge is a class to represent an edge in a graph, with its origin, destination, cost, and pheromone. Graph is a class to represent a graph, with its number of vertices, edges, neighbors, and vertices. It has methods to add edges, obtain the cost of an edge, obtain the pheromone of an edge, set the pheromone of an edge, and obtain the cost of a path. GraphComplete is the subclass of Graph that generates a complete graph from a matrix of weights. Ant is the class to represent an ant, with its current city, solution, cost, and accuracy. It has methods to obtain the current city, set the current city, obtain the solution, set the solution, obtain the cost of the solution, and obtain the accuracy. ACO or Ant Colony Optimization is the class to implement the ACO algorithm for attribute selection. It takes a graph, number of ants, alpha, beta, iterations, evaporation, and number of FS (feature subsets). It has methods to initialize the ants, initialize the pheromone, update the pheromone, select the next city, construct the solution, and run the ACO algorithm. It also has a method to print the parameters of the ACO. The ACO algorithm initializes a graph, with the number of vertices equal to the number of attributes. It then initializes the ants, by randomly assigning a starting city to each ant. It initializes the pheromone of each edge in the graph to 1 divided by the number of vertices times the cost of a greedy solution, which is computed as follows: starting from a random vertex, the ant selects the next vertex with the lowest cost, until all vertices have been visited. The cost of the greedy solution is the sum of the costs of the edges traversed, plus the cost of the edge from the last visited vertex to the starting vertex. The number of FS is the number of ants divided by 2. The ACO algorithm then runs for the specified number of iterations, each time constructing a solution for each ant, by selecting the next city with a probability that depends on the pheromone and the heuristic information (in this case, the cosine similarity between attribute pairs), and updating the pheromone of the edges in the solution. The best solution is stored, and the accuracy is computed and stored. The pheromone is updated by evaporating a fraction of it, and then adding to each edge the amount of pheromone deposited by each ant that traversed it, proportional to the accuracy of the ant's solution. Finally, the best feature subset is returned.

Parameter	***Value***	***Importance***
Number of vertices of the graph	*13*	*Number of features in dataset is 13 for Heart Disease and 30 for breast cancer*
Number of Ants	*50*	*50-100 Ants are ideal for this case.*
Rate of Evaporation	*0.2*	*High rate of evaporation can lead to faster evaporation of pheromone hence the following ants might lose their way to the food source. Hence an optimal rate of evaporation*

		is 0.2-0.3.
Alpha Heuristic	*1*	*Importance of pheromone*
Beta Heuristic	*1*	*Importance of heuristic information*
Number of iterations	*100*	*100-200 iterations are optimal as a higher number of iterations means more exploration of paths and finding the most optimal shortest path.*

Ant colony optimization is a metaheuristic optimization algorithm inspired by the foraging behavior of ants. Ants are known for their ability to find the shortest path between their nest and a food source. They achieve this by leaving a trail of pheromones as they move towards the food source, which attracts other ants to follow the same path. This process of pheromone trail laying and following is the basis of ant colony optimization algorithms. Ant colony optimization algorithms have been extensively studied and applied to solve a variety of optimization problems. The algorithm has proven to be highly effective in finding optimal solutions to problems in various fields, including transportation, telecommunications, and engineering.

- There are two paths to reach the food from the colony. At first, there is no pheromone on the ground. So, the probability of choosing these two paths is equal that means 50%. Let consider two ants choose two different paths to reach the food as the probability of choosing these paths is fifty-fifty.
- The distances of these two paths are different. Ant following the shorter path will reach the food earlier than the other..
- After finding food, it carries some food with itself and returns to the colony. When it tracking the returning path it deposits pheromone on the ground. The ant following the shorter path will reach the colony earlier.
- When the third ant wants to go out for searching food it will follow the path having shorter distance based on the pheromone level on the ground. As a shorter path has more pheromones than the longer, the third ant will follow the path having more pheromones.
- By the time the ant following the longer path returned to the colony, more ants already have followed the path with more pheromones level. Then when another ant tries to reach the destination(food) from the colony it will find that each path has the same pheromone level. So, it randomly chooses one.
- Repeating this process again and again, after some time, the shorter path has a more pheromone level than others and has a higher probability to follow the path, and all ants next time will follow the shorter path.

Pheromone Update formula is given by $\tau_{xy} \leftarrow (1-\rho)\tau_{xy} + \sum_k \Delta\tau_{xy}^k$ The left side on the equation indicates the amount of pheromone on the given edge x,y; ρ — the rate of pheromone evaporation; and the last term on the right side indicated the amount of pheromone deposited. $\Delta\tau_{xy}^k = \begin{cases} Q/L_k & \text{if ant } k \text{ uses curve } xy \text{ in its tour} \\ 0 & \text{otherwise} \end{cases}$; where L is the cost of an ant tour length and Q is a constant.

An artificial ant is made for finding the optimal solution. In the first step of solving a problem, each ant generates a solution. In the second step, paths found by different ants are compared. And in the third step, paths value or pheromone is updated.

```
procedure ACO_MetaHeuristic is
   while not_termination do
      generateSolutions()
      daemonActions()
      pheromoneUpdate()
   repeat
end procedure
```

Components of an Ant Colony Optimization Algorithm:

- Problem Representation: The problem is represented as a graph or a network, where the nodes represent the problem variables, and the edges represent the relationships between them.
- Solution Construction: Ants construct a solution to the problem by selecting a path through the graph. At each node, an ant chooses the next node to visit based on a probability that is influenced by the pheromone trail.
- Pheromone Update: As ants move through the graph, they lay pheromone trails on the edges. The pheromone trail evaporates over time, and the pheromone level is updated based on the quality of the solution found.
- Local Search: After constructing a solution, a local search procedure can be used to improve its quality.
- Termination: The algorithm terminates when a stopping criterion is met, such as a maximum number of iterations or a threshold level of pheromone.

Factors Affecting the Performance of Ant Colony Optimization Algorithm:

- The number of ants: Increasing the number of ants can improve the quality of solutions but also increases the computational cost.
- The amount of pheromone: The amount of pheromone deposited on the edges influences the probability of ants choosing a particular path.
- The evaporation rate: The rate at which pheromone evaporates affects the exploration-exploitation balance of the algorithm.
- The local search procedure: The choice of local search procedure can significantly impact the quality of the solutions found.

Parameter	***Value***	***Importance***
Population Size	*50*	*This parameter specifies the size of the population, i.e., the number of candidate solutions. A larger population size can increase the diversity of the solutions, but it may also lead to slower convergence and higher computational costs.*

Awareness Probability	*0.02*	*This parameter controls the probability of a crow selecting a specific feature for optimization. It ranges between 0 and 1 and is used in the update_position() function to decide whether a feature is updated based on the local best or randomly generated values.*
Flight Length	*2*	*This parameter controls the step size of the search process. It is used to balance between local and global search. A higher value of fL increases the global search capability of the algorithm, but it may also decrease the accuracy of the local search.*
Number of Iterations	*100*	*This parameter specifies the maximum number of iterations that the algorithm can run. It determines how long the algorithm will search for the optimal solution.*
Target Function	*fitness*	*This parameter is a user-defined function that calculates the fitness value of each candidate solution. It is used to evaluate the performance of each solution in the population. The goal of the algorithm is to find the solution that maximizes the fitness value.*
Minimum Value	*[5, 1]*	*This parameter is a list that specifies the minimum values of the search space for each feature. It ensures that the search does not go beyond the lower bounds of the problem domain.*
Maximum Value	*[120, 3]*	*This parameter is a list that specifies the maximum values of the search space for each feature. It ensures that the search does not go beyond the upper bounds of the problem domain.*

Diabetes Prediction- Supervised Learning with Crow Search Optimization:

We use K Nearest Neighbours, Random Forest and finally Logistic Regression with Crow Search Algorithm to predict the disease. **K-Nearest Neighbours** is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. **Logistic regression** is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyze the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision-making. For

example email spam or not. **Random Forest** is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Crow Search Algorithm concept of optimization is fundamental in various fields, including engineering, economics, and computer science. Optimization problems aim to find the optimal solution from a set of possible solutions. However, optimization problems can be challenging due to their complexity and the large search space. Metaheuristic optimization algorithms have been developed to tackle such optimization problems. Crow Search Algorithm (CSA) is one such metaheuristic optimization algorithm that was introduced in 2014 by Yang and Deb. CSA is based on the foraging behavior of crows and is used to solve various optimization problems.

The working mechanism of CSA is based on the social behavior of crows, where the foraging behavior of crows is used to search for the optimal solution. CSA starts by initializing a population of crows, where each crow represents a potential solution to the optimization problem. The crows communicate with each other using a combination of local and global information. Local information is obtained by the individual crow's experience, while global information is obtained from the best solution obtained so far. Based on this information, the crows adjust their position in the search space to move towards the optimal solution. CSA uses three main operators, namely exploration, exploitation, and memorization. Exploration aims to search for new solutions by exploring the search space. Exploitation aims to exploit the best solution obtained so far by focusing on the promising areas in the search space. Memorization aims to remember the best solution obtained so far by storing it in the memory of each crow. As per the study on crow behavior, CSA principles are:
(1) Living in flock.
(2) Memorize hiding locations.
(3) Stealing food by following peers.
(4) Guard caches from thief crows.

CSA has shown promising results in various optimization problems, including engineering design, scheduling, and image processing. In engineering design, CSA has been used to optimize the design parameters of a hybrid energy storage system. In scheduling, CSA has been used to optimize the scheduling of tasks in a distributed system. In image processing, CSA has been used to optimize the parameters of a digital image filter. CSA has several strengths that make it a suitable optimization algorithm. It is easy to implement, has a fast convergence rate, and can handle high-dimensional problems. CSA is also robust against noise and can handle multimodal optimization problems. However, CSA has some limitations. It may get stuck in local optima, and the convergence rate may slow down as the search progresses. CSA is a relatively new optimization algorithm, and there is still room for improvement. Future research can focus on developing hybrid algorithms that combine CSA with other optimization algorithms. Furthermore, the optimization performance of CSA can be enhanced by using adaptive parameters and dynamic control strategies. Finally, the application of CSA can be extended to other fields, such as data mining and machine learning.

Parkinson's Disease Prediction- Voice data based Support Vector Machine Algorithm:

90 percent of the people with PD suffer from speech disorders, speech analysis is considered as the most common technique for this aim. Algorithm for diagnosing Parkinson's disease based on voice analysis. In the first step, selecting optimized features from all extracted features. Afterwards a network based on a Support Vector Machine (SVM) is used for classification between healthy and people with Parkinson's.

Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. The main idea behind SVMs is to find a hyperplane that maximally separates the different classes in the training data. This is done by finding the hyperplane that has the largest margin, which is defined as the distance between the hyperplane and the closest data points from each class. Once the hyperplane is determined, new data can be classified by determining on which side of the hyperplane it falls. SVMs are particularly useful when the data has many features, and/or when there is a clear margin of separation in the data.

Lung Cancer prediction using XGBoost:

EXtreme Gradient Boosting (XGBoost) is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.
XGBoost is a fast and efficient algorithm used by the winners of many learning competitions of machine. XG Boost only works with numeric variables.
XgBoost modeling consists of two techniques:

- Bagging: is an approach where you can take random samples of data, build algorithms, learning and using simple means to find bagging probabilities.
- Boosting: It's an approach where the selection of the approach is done more intelligently, that is, more weight is given to sort how to sort.

Types of parameters:

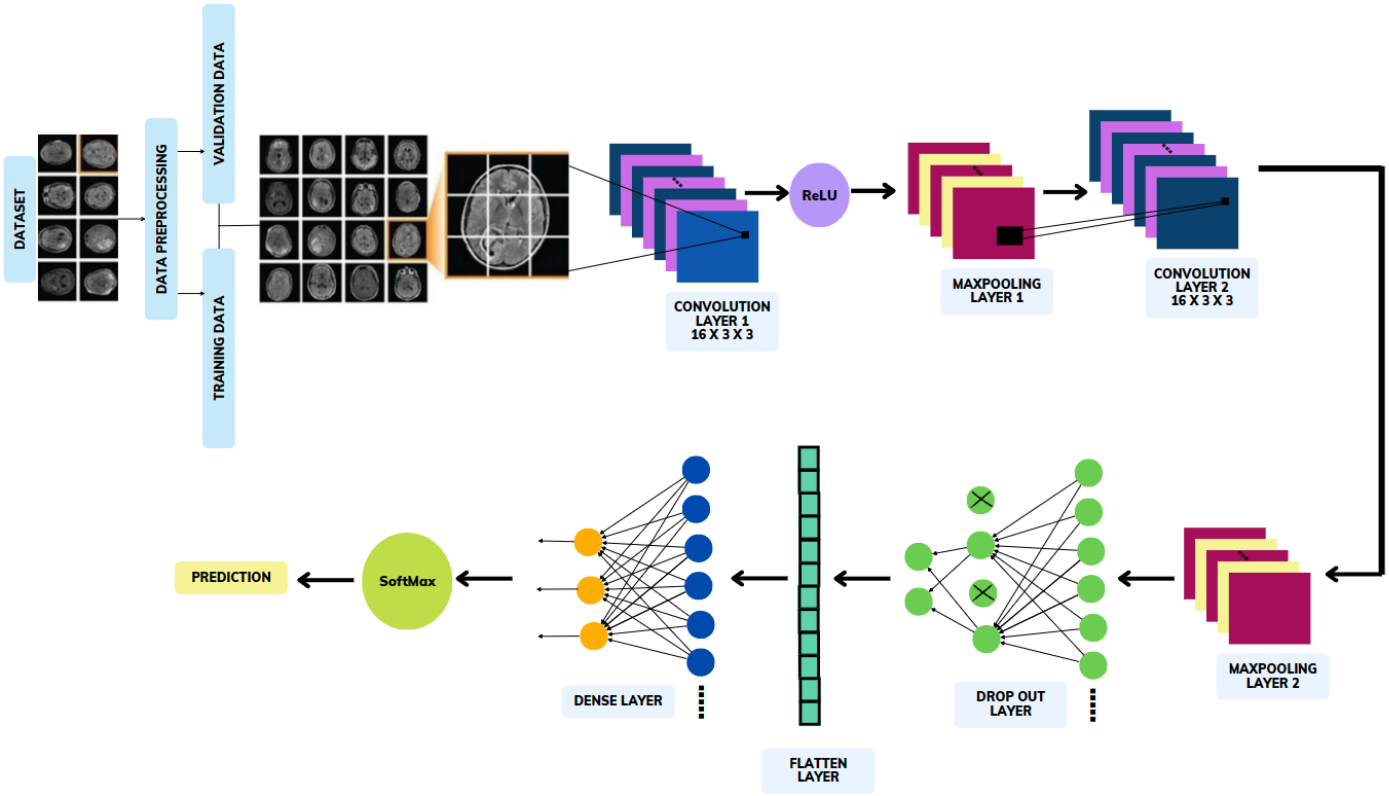
- Booster parameters: It affects the boosting operation
- Learning Task parameters: It guides optmized perfomance
- General Parameters: It affects each Overall function

Some parameters in XGBoost

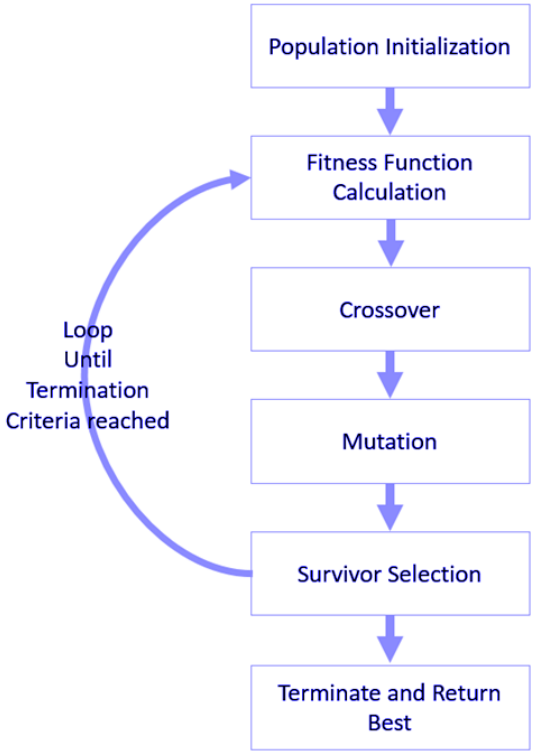
- eta: reduces resource weights to make the boosting process more conservative. The range is from 0 to 1. This is also known as the learning rate or reduction factor. The low eta value means the model is more robust for overfitting.
- gama: the larger the gamma value, the more conservative the algorithm. Its range is from 0 to infinity.
- max_depth: The maximum depth of a tree can be increased using the max_dept parameter.
- subsample: is the proportion of rows the model will randomly select to grow trees.
- colsample_bytree: It is the proportion of variables chosen randomly to build each tree in the model.

4. Flowchart/ Model Workflow

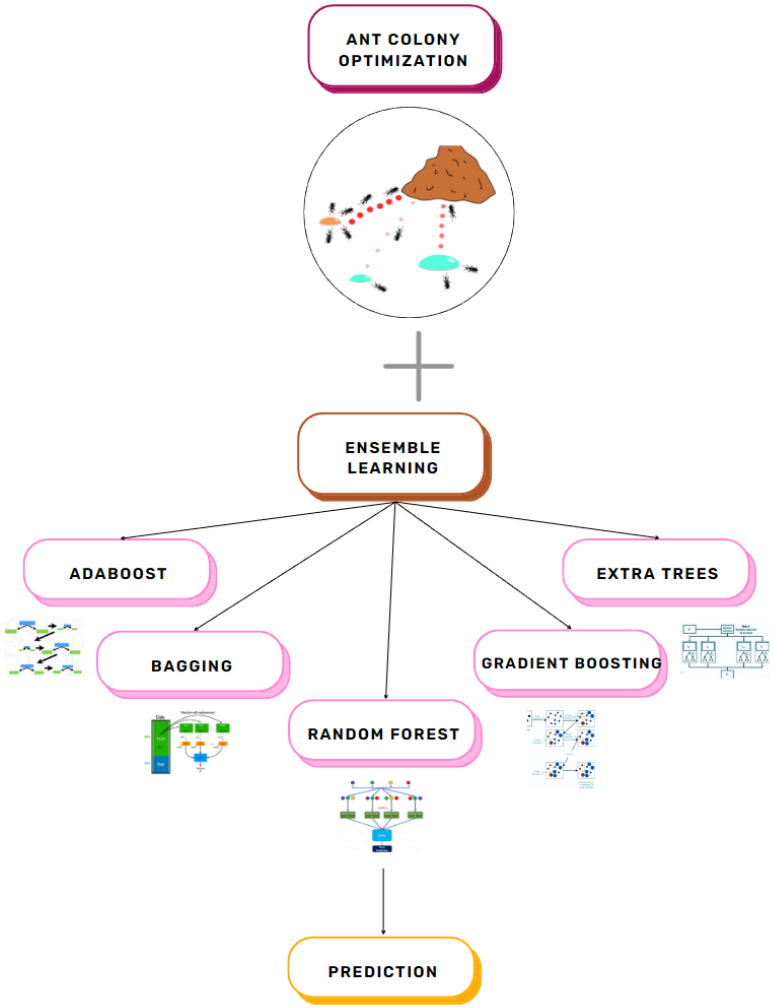
Alzheimer's Dementia Prediction- 2D CNN for Dementia Architecture:



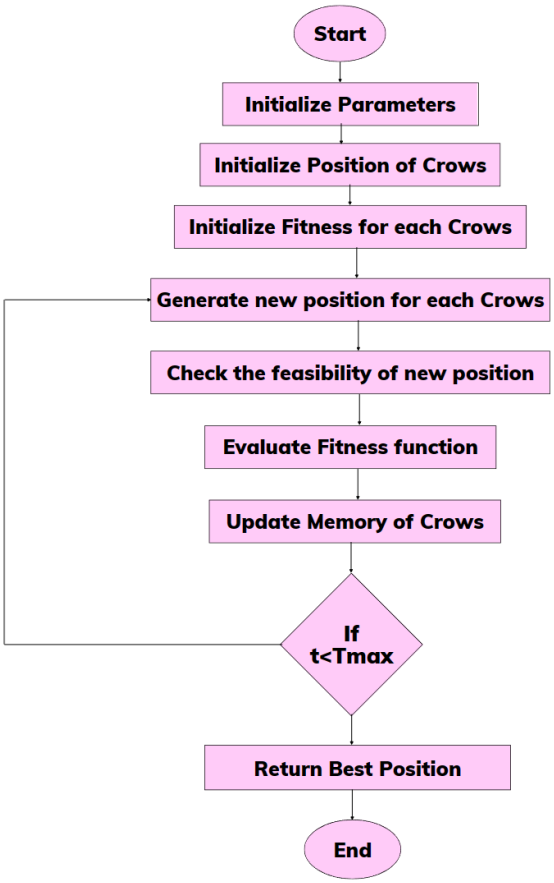
Breast Cancer Prediction- Hyperparameter Tuning and Genetic Algorithm:

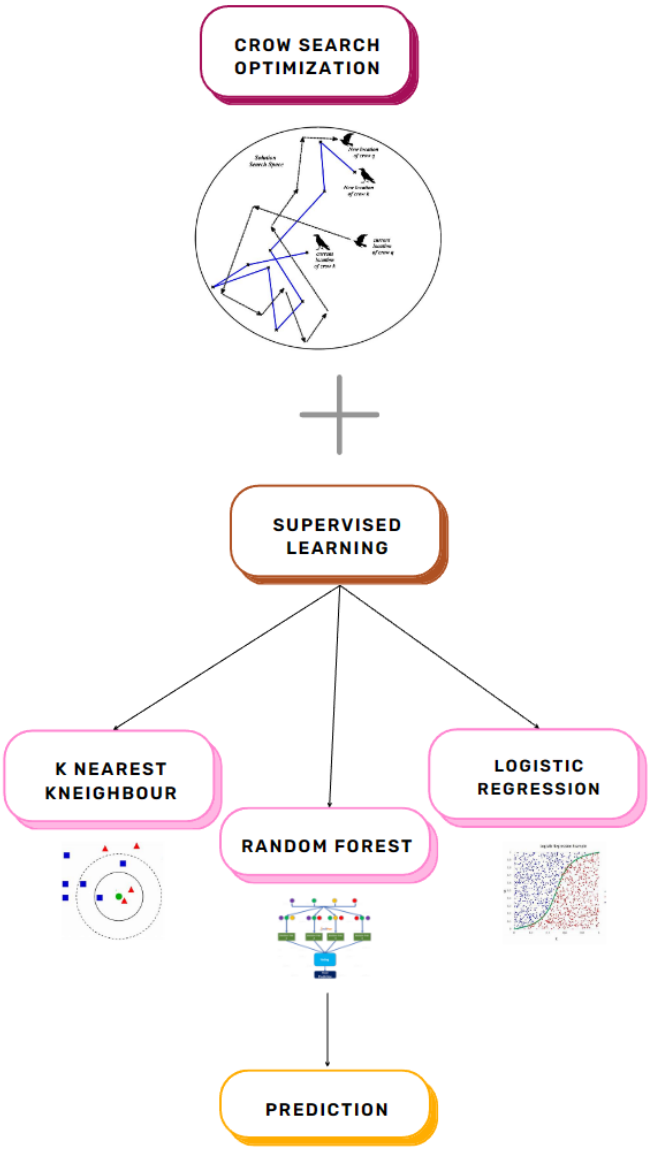


Heart Disease Prediction- Ensemble Learning with Ant Colony Optimization:

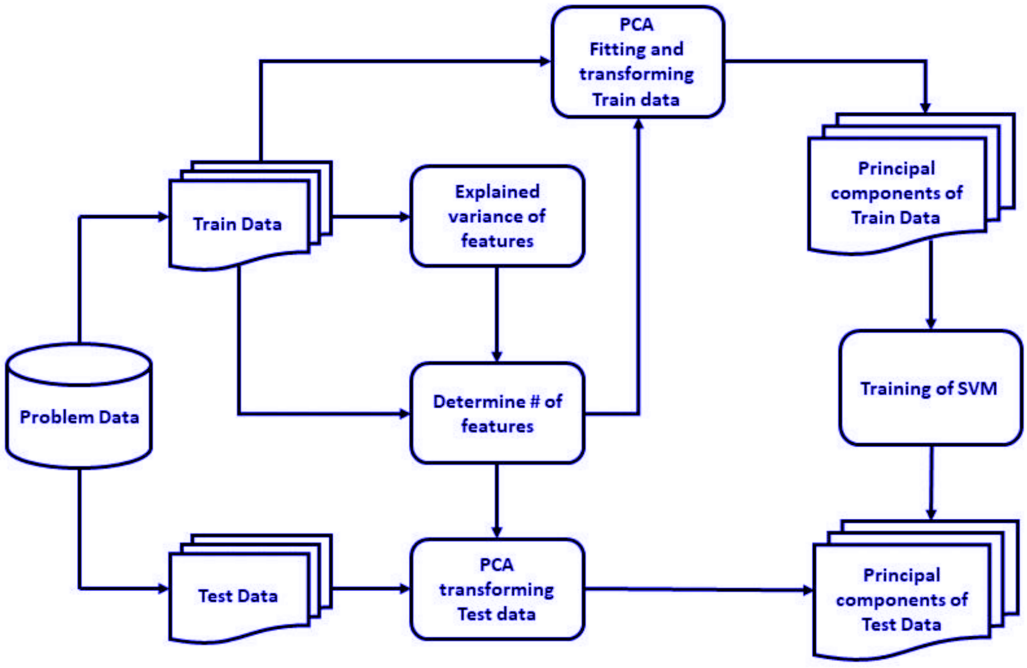


Diabetes Prediction- Supervised Learning with Crow Search Optimization:

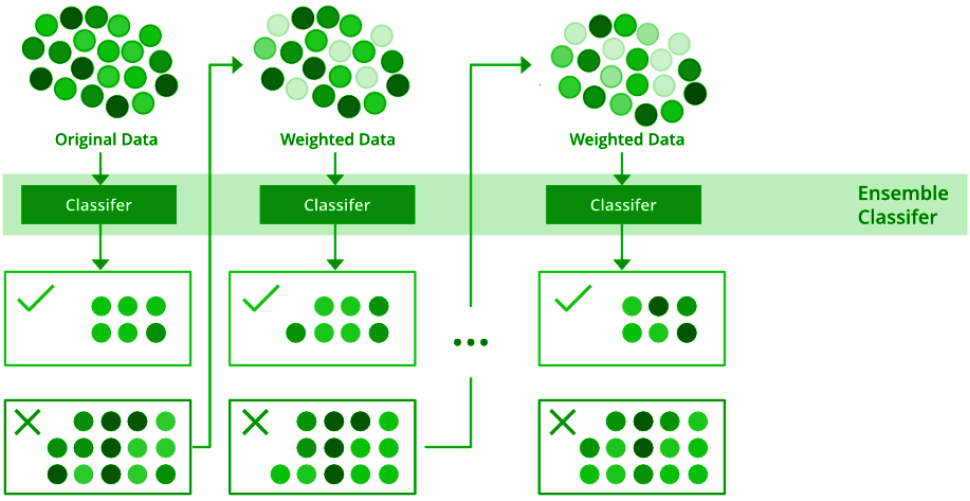



CROW SEARCH OPTIMIZATION
SUPERVISED LEARNING
K NEAREST KNEIGHBOUR
RANDOM FOREST
LOGISTIC REGRESSION
PREDICTION

Parkinson's Disease Prediction- Voice data based Support Vector Machine Algorithm:

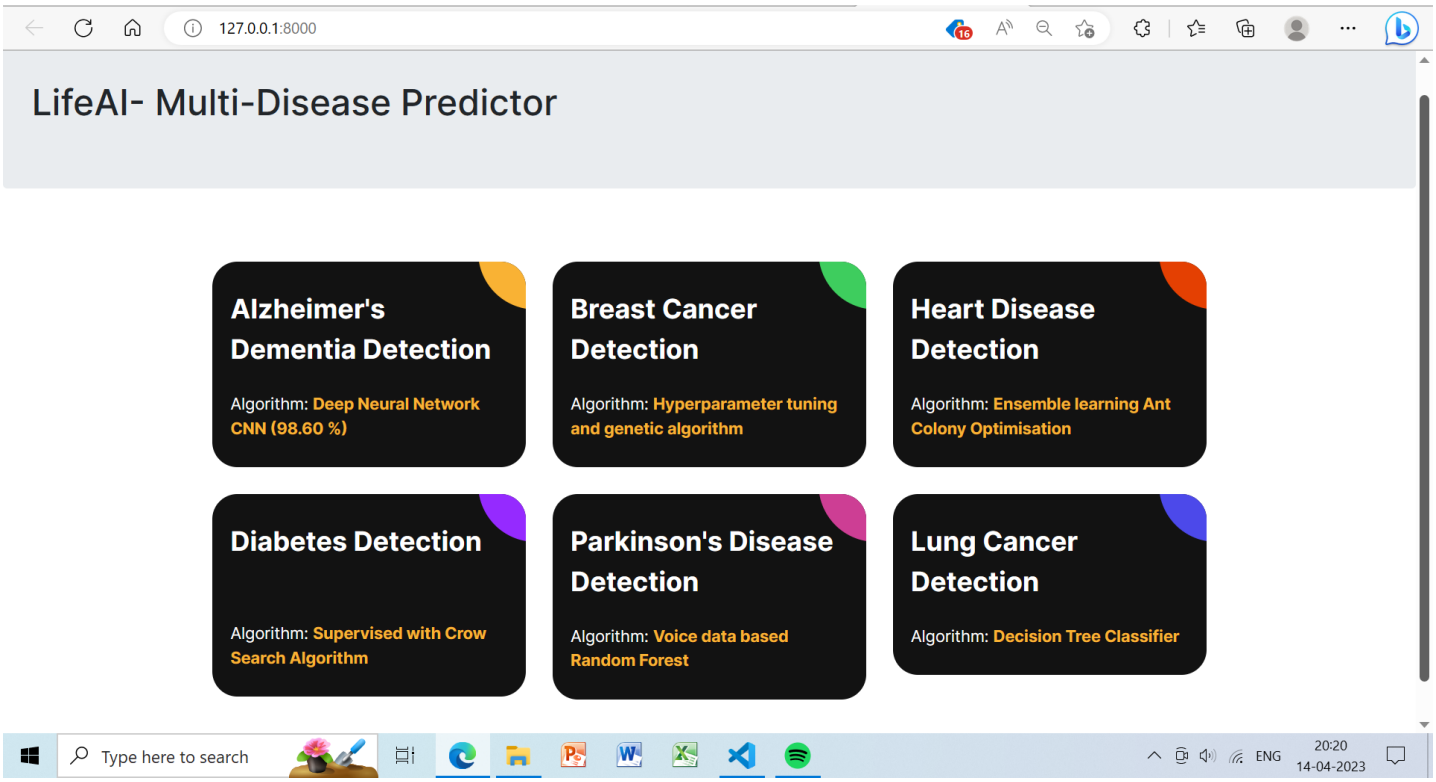


Lung Cancer prediction using XGBoost:



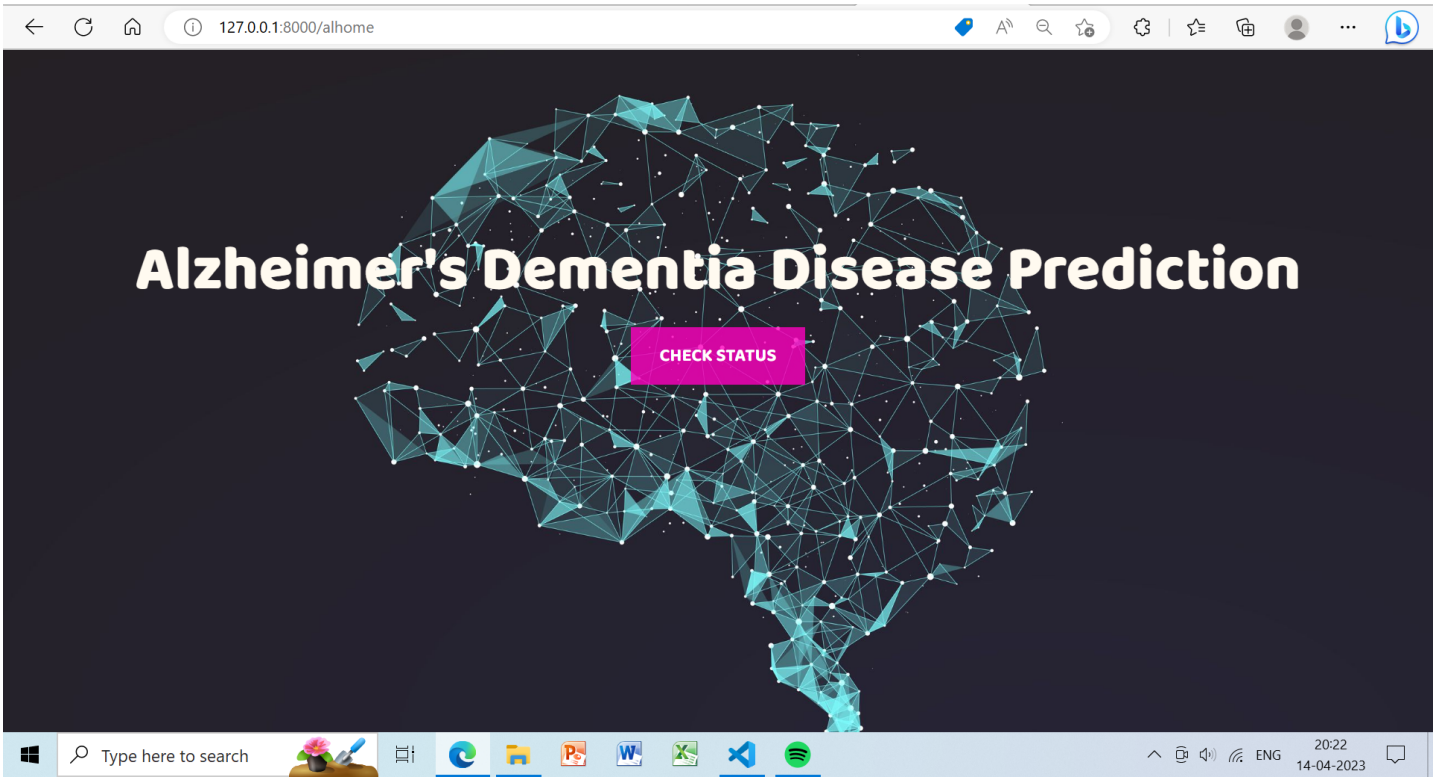
5. Implementation/ Code

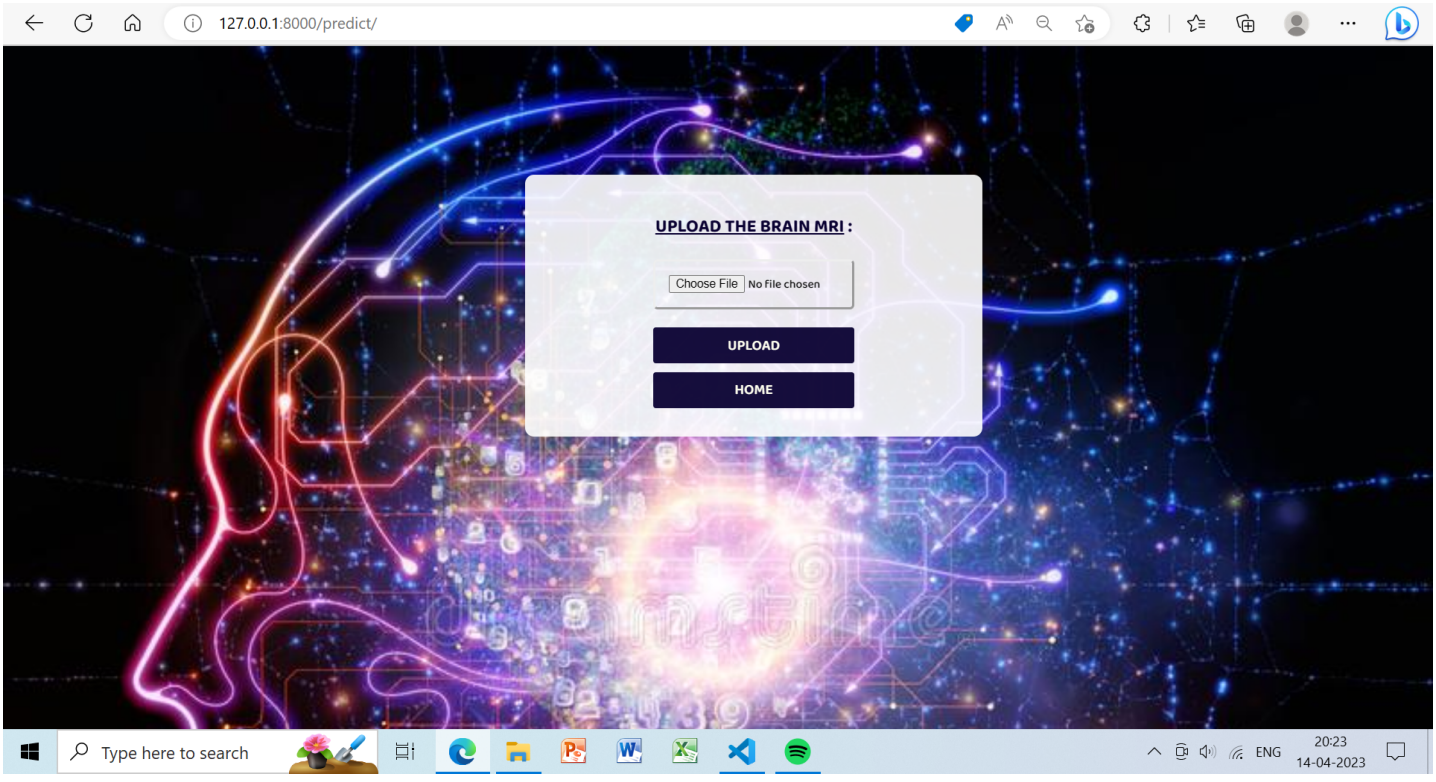
6. Results

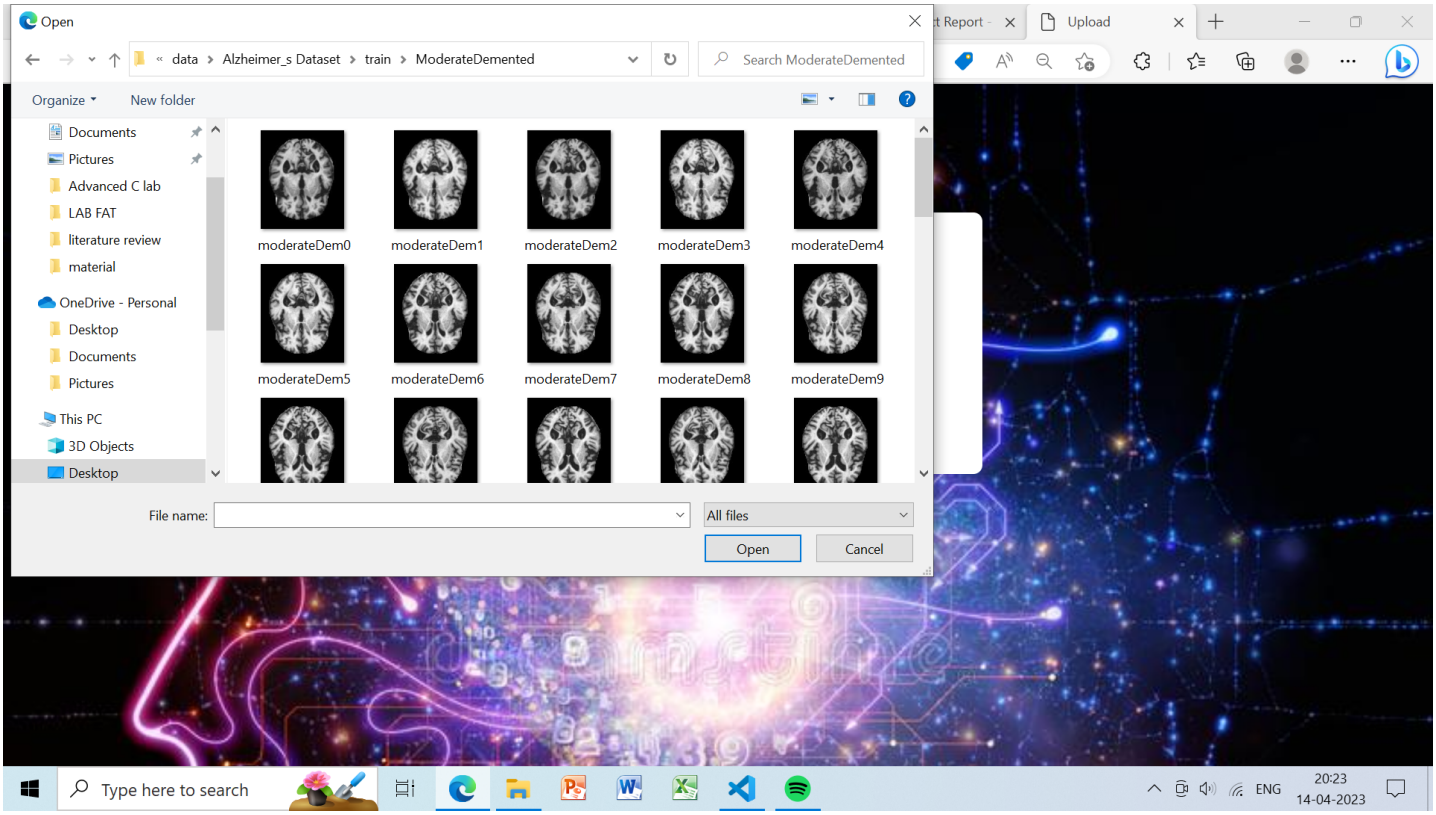

127.0.0.1:8000
LifeAI- Multi-Disease Predictor
Alzheimer's Dementia Detection
Algorithm: Deep Neural Network CNN (98.60 %)
Breast Cancer Detection
Algorithm: Hyperparameter tuning and genetic algorithm
Heart Disease Detection
Algorithm: Ensemble learning Ant Colony Optimisation
Diabetes Detection
Algorithm: Supervised with Crow Search Algorithm
Parkinson's Disease Detection
Algorithm: Voice data based Random Forest
Lung Cancer Detection
Algorithm: Decision Tree Classifier
Type here to search
ENG
20:20
14-04-2023

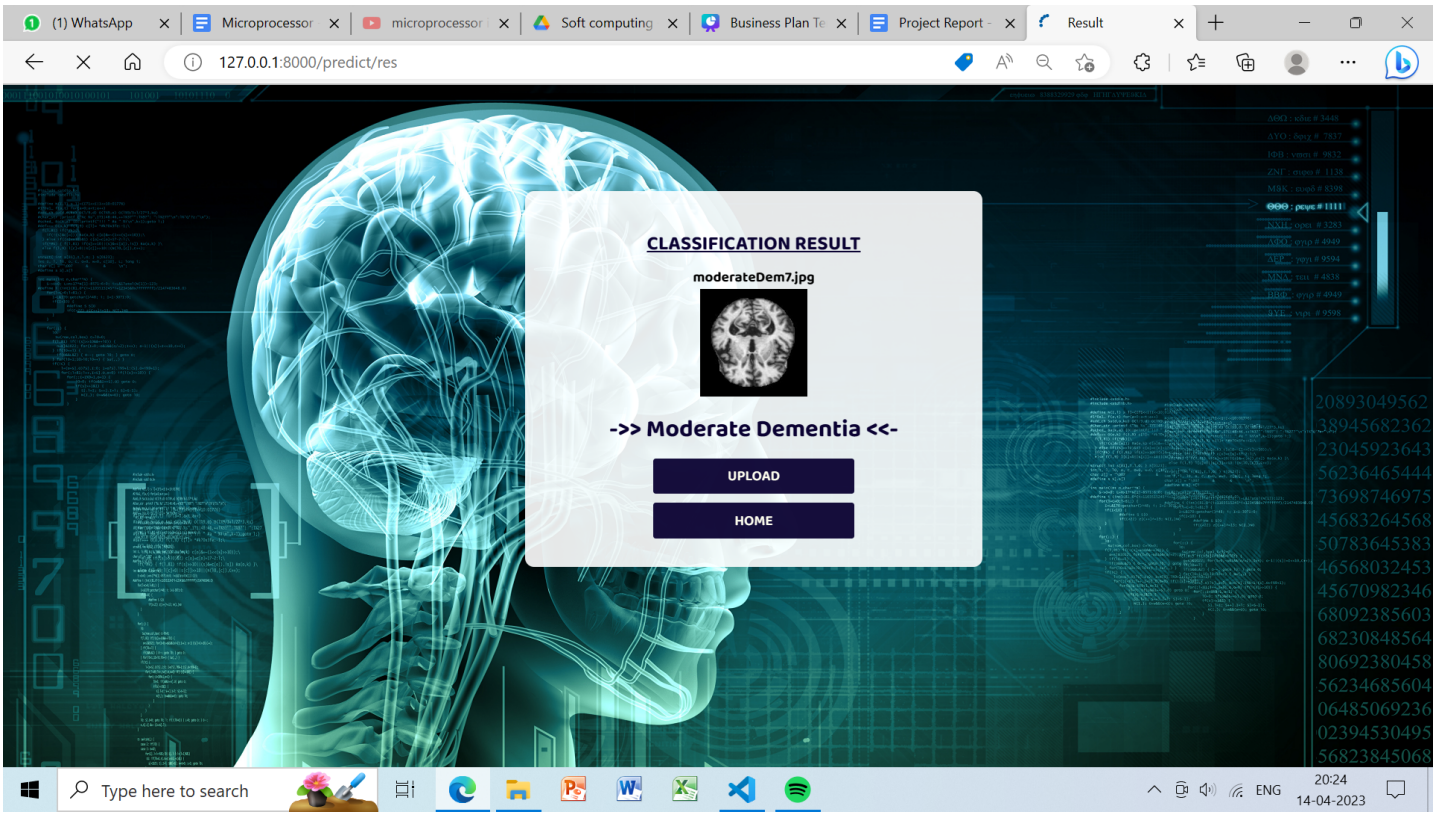
Alzheimer's Dementia Prediction- 2D CNN for Dementia Architecture:

Accuracy= 98.60%



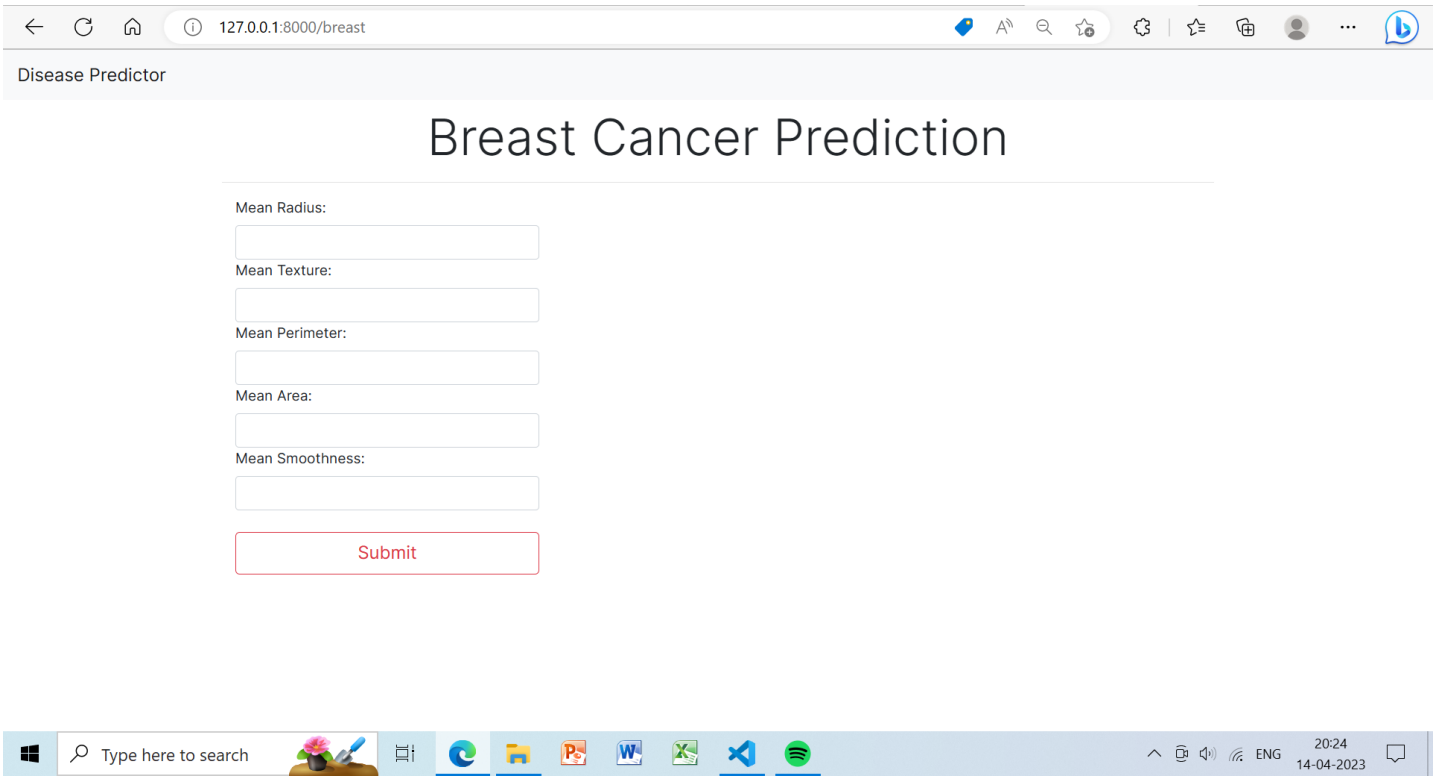


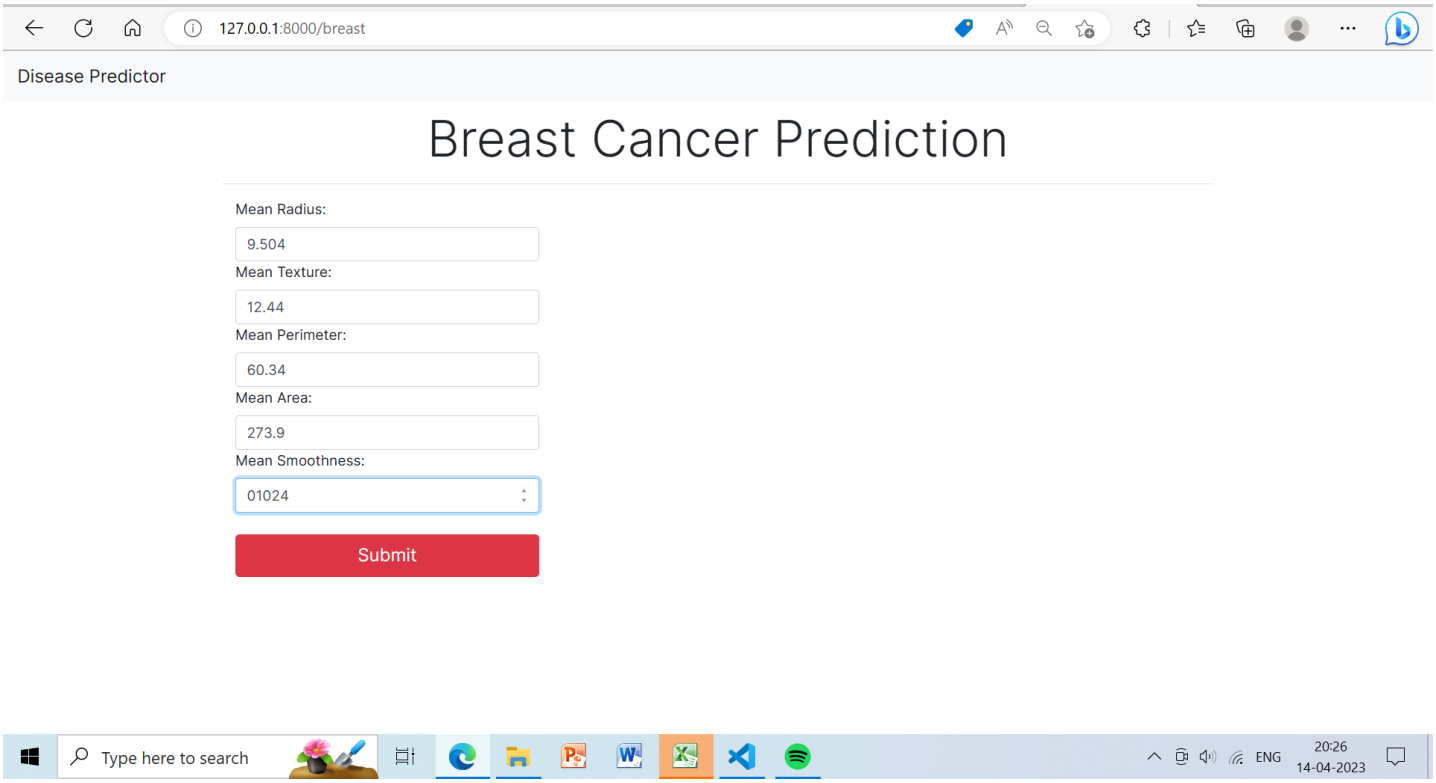

Open
data › Alzheimer_s Dataset › train › ModerateDemented
Search ModerateDemented
Organize
New folder
Documents
Pictures
Advanced C lab
LAB FAT
literature review
material
OneDrive - Personal
Desktop
Documents
Pictures
This PC
3D Objects
Desktop
moderateDem0
moderateDem1
moderateDem2
moderateDem3
moderateDem4
moderateDem5
moderateDem6
moderateDem7
moderateDem8
moderateDem9
File name:
All files
Open
Cancel
Upload
Type here to search
ENG
20:23
14-04-2023

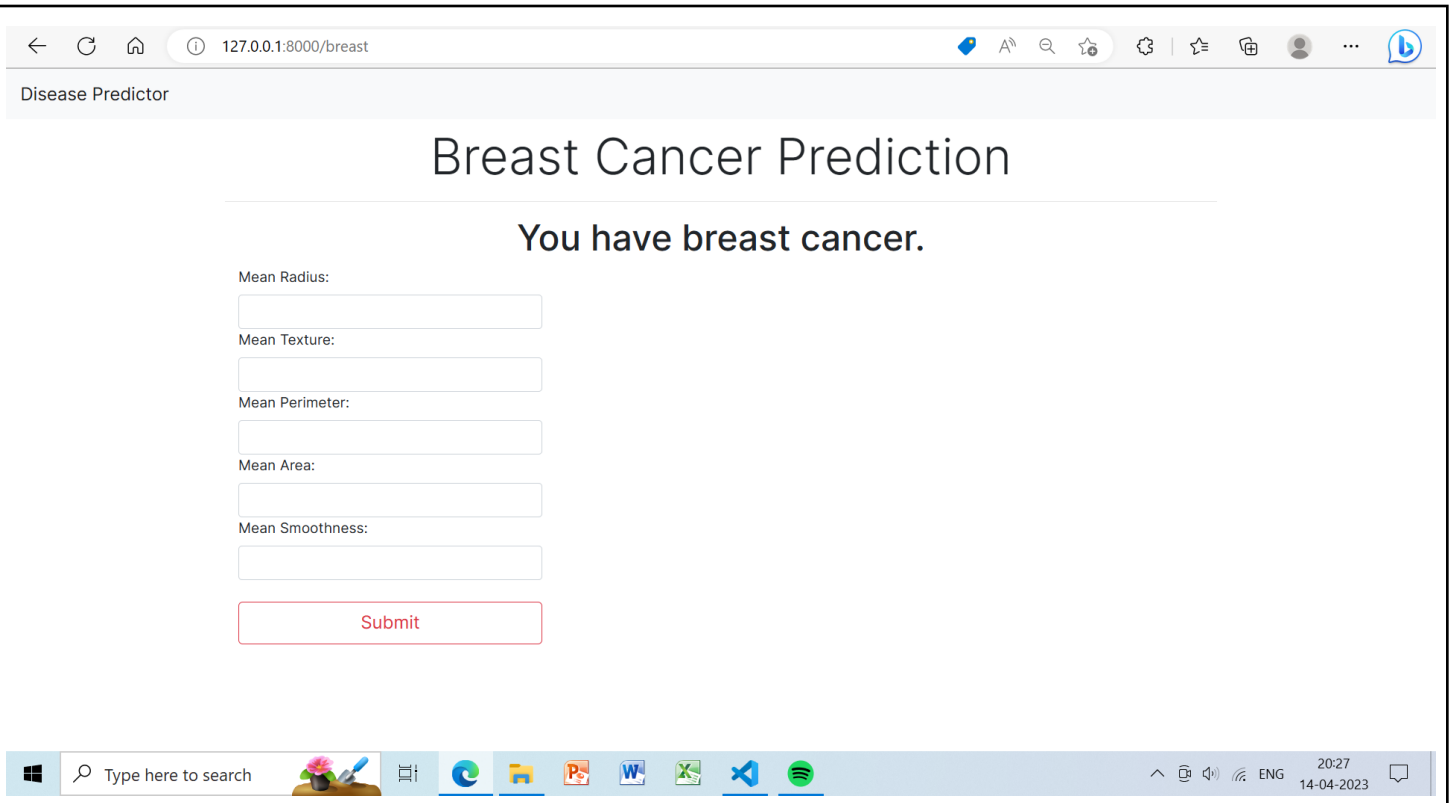

(1) WhatsApp
Microprocessor
microprocessor
Soft computing
Business Plan Te
Project Report
Result
127.0.0.1:8000/predict/res
CLASSIFICATION RESULT
moderateDem7.jpg
->> Moderate Dementia <<-
UPLOAD
HOME
Type here to search
ENG
20:24
14-04-2023

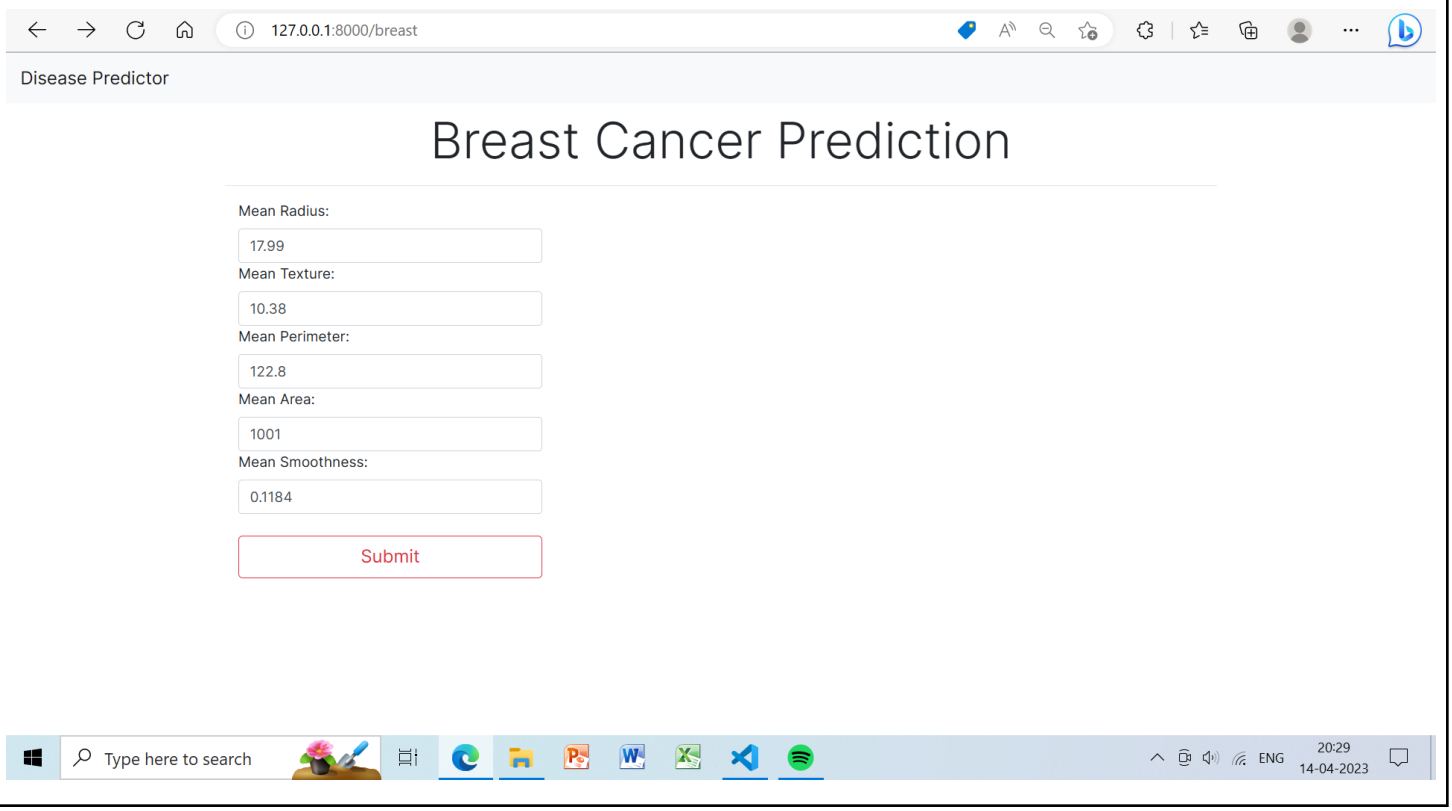
Breast Cancer Prediction- Hyperparameter Tuning and Genetic Algorithm:

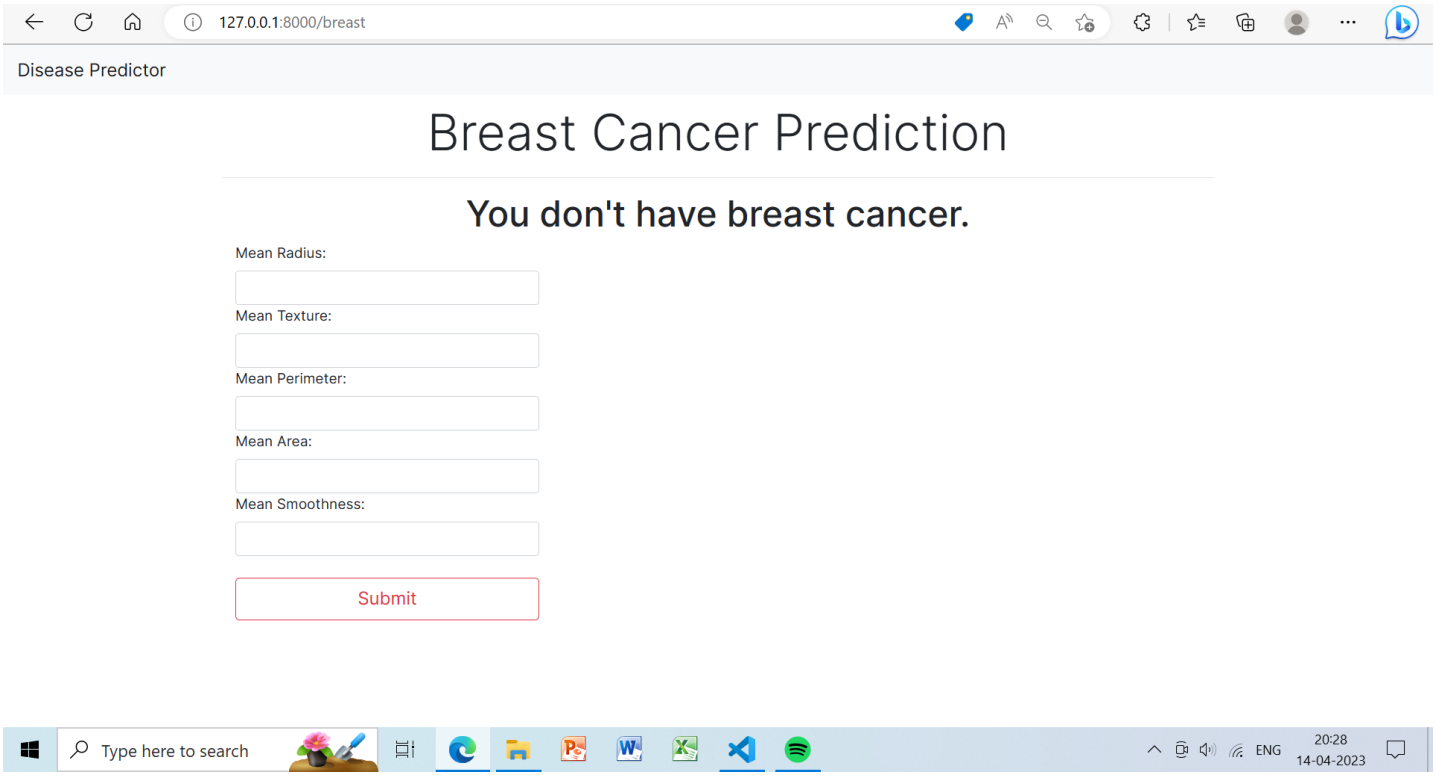
Accuracy= 96.50%






127.0.0.1:8000/breast
Disease Predictor
Breast Cancer Prediction
You have breast cancer.
Mean Radius:
Mean Texture:
Mean Perimeter:
Mean Area:
Mean Smoothness:
Submit
Type here to search
ENG
20:27
14-04-2023


127.0.0.1:8000/breast
Disease Predictor
Breast Cancer Prediction
Mean Radius:
17.99
Mean Texture:
10.38
Mean Perimeter:
122.8
Mean Area:
1001
Mean Smoothness:
0.1184
Submit
Type here to search
ENG
20:29
14-04-2023



Heart Disease Prediction- Ensemble Learning with Ant Colony Optimization:

<table>
<tr><th>Without ACO</th><th>With ACO</th></tr>
<tr><td>Adaboost
<table><tr><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1 score</th></tr><tr><td>78.54</td><td>0.77</td><td>0.82</td><td>0.79</td></tr></table></td><td>Adaboost
<table><tr><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1 score</th></tr><tr><td>82.44</td><td>0.81</td><td>0.86</td><td>0.83</td></tr></table></td></tr>
<tr><td>Bagging
<table><tr><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1 score</th></tr><tr><td>94.64</td><td>0.95</td><td>0.94</td><td>0.95</td></tr></table></td><td>Bagging
<table><tr><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1 score</th></tr><tr><td>99.99</td><td>1.0</td><td>0.97</td><td>0.99</td></tr></table></td></tr>
<tr><td>Random Forest
<table><tr><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1 score</th></tr></table></td><td>Random Forest
<table><tr><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1 score</th></tr></table></td></tr>
</table>

92.20	0.93	0.92	0.92

99.98	1.0	0.98	1.0

Gradient Boosting

Accuracy	Precision	Recall	F1 score
89.27	0.89	0.91	0.89

Gradient Boosting

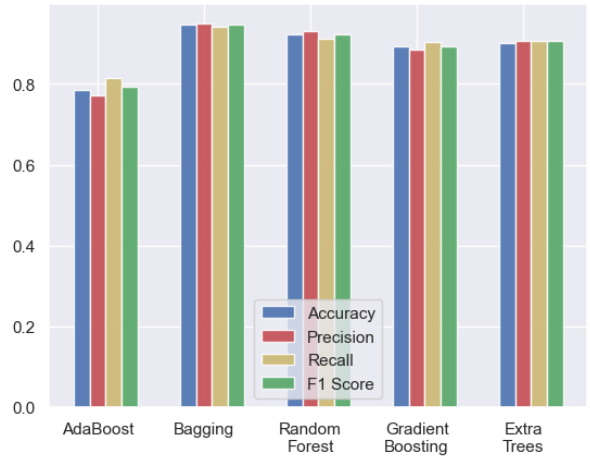
Accuracy	Precision	Recall	F1 score
95.61	0.94	0.94	0.94

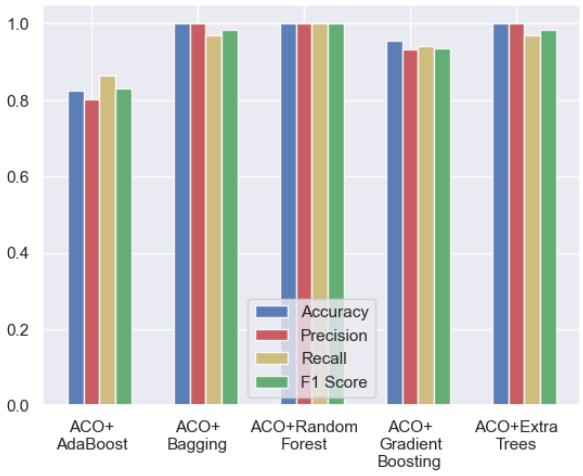
Extra Trees

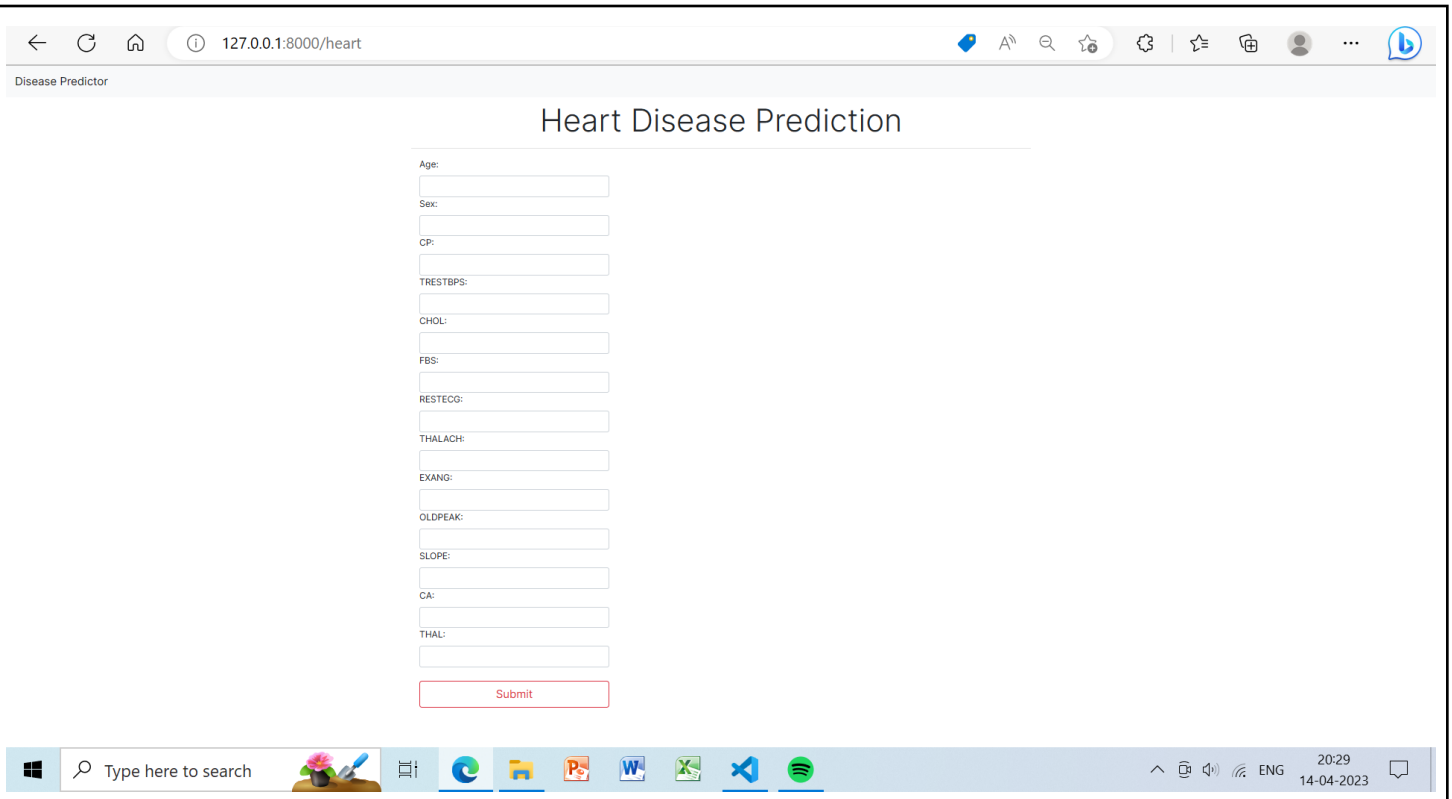
Accuracy	Precision	Recall	F1 score
90.17	0.91	0.91	0.91

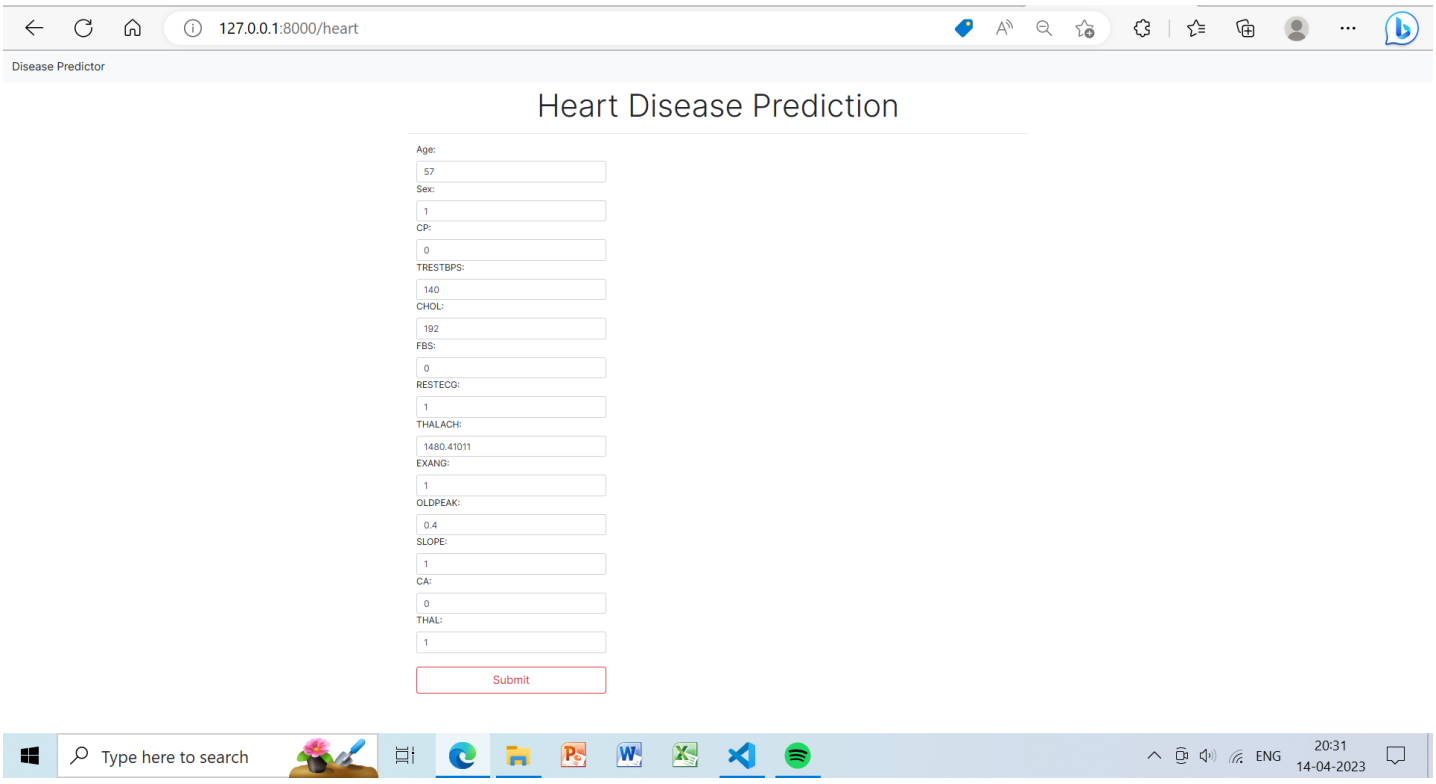
Extra Trees

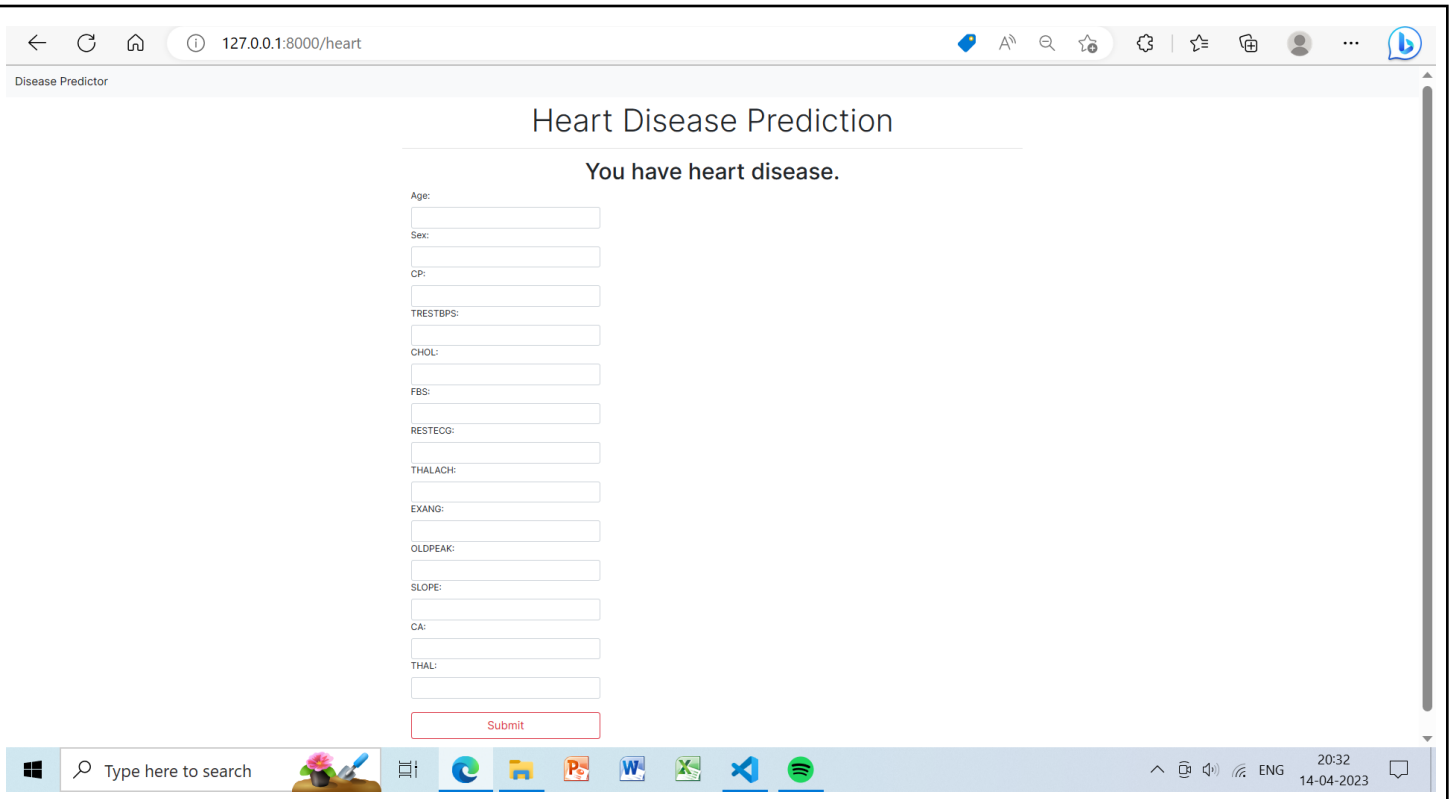
Accuracy	Precision	Recall	F1 score
99.98	1.0	0.97	0.99

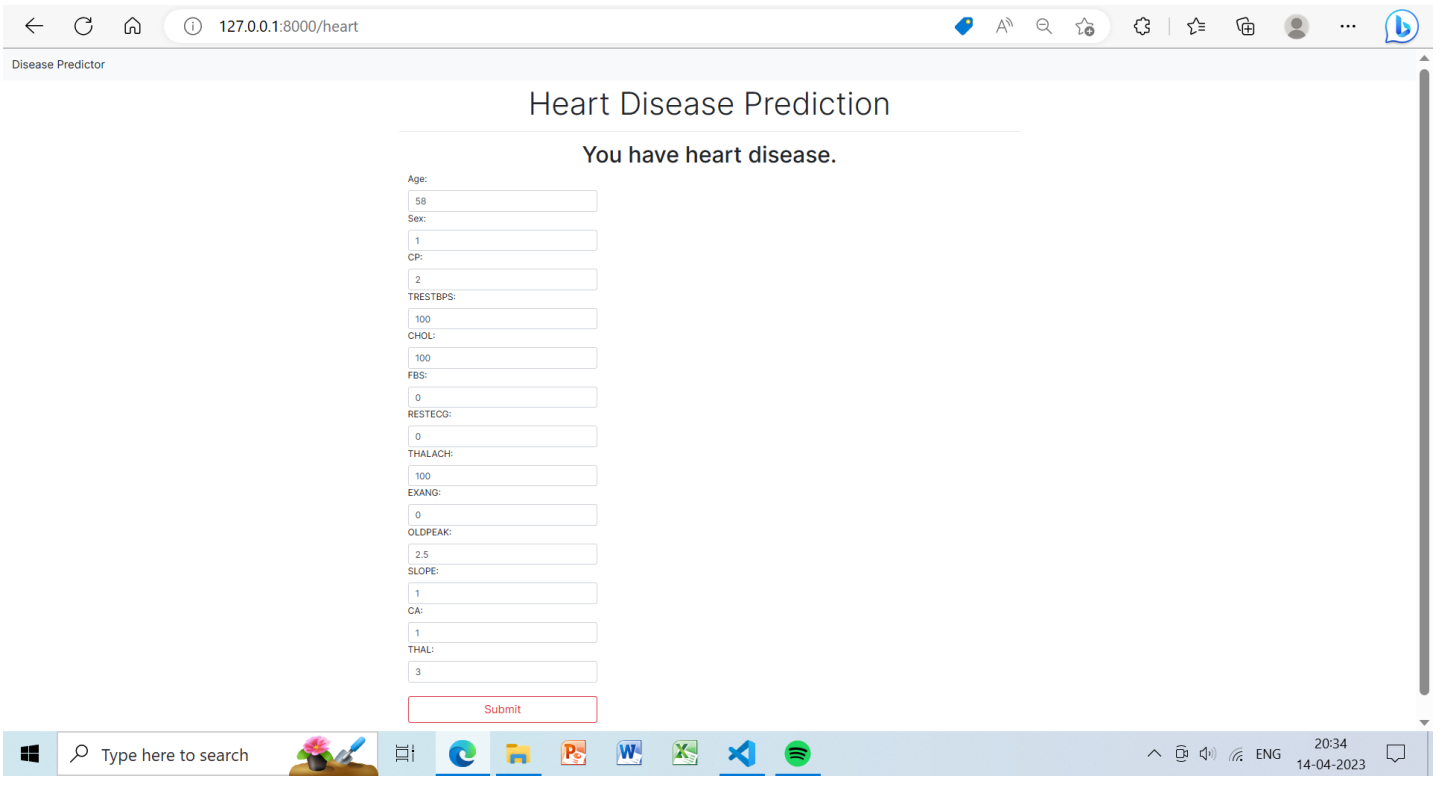


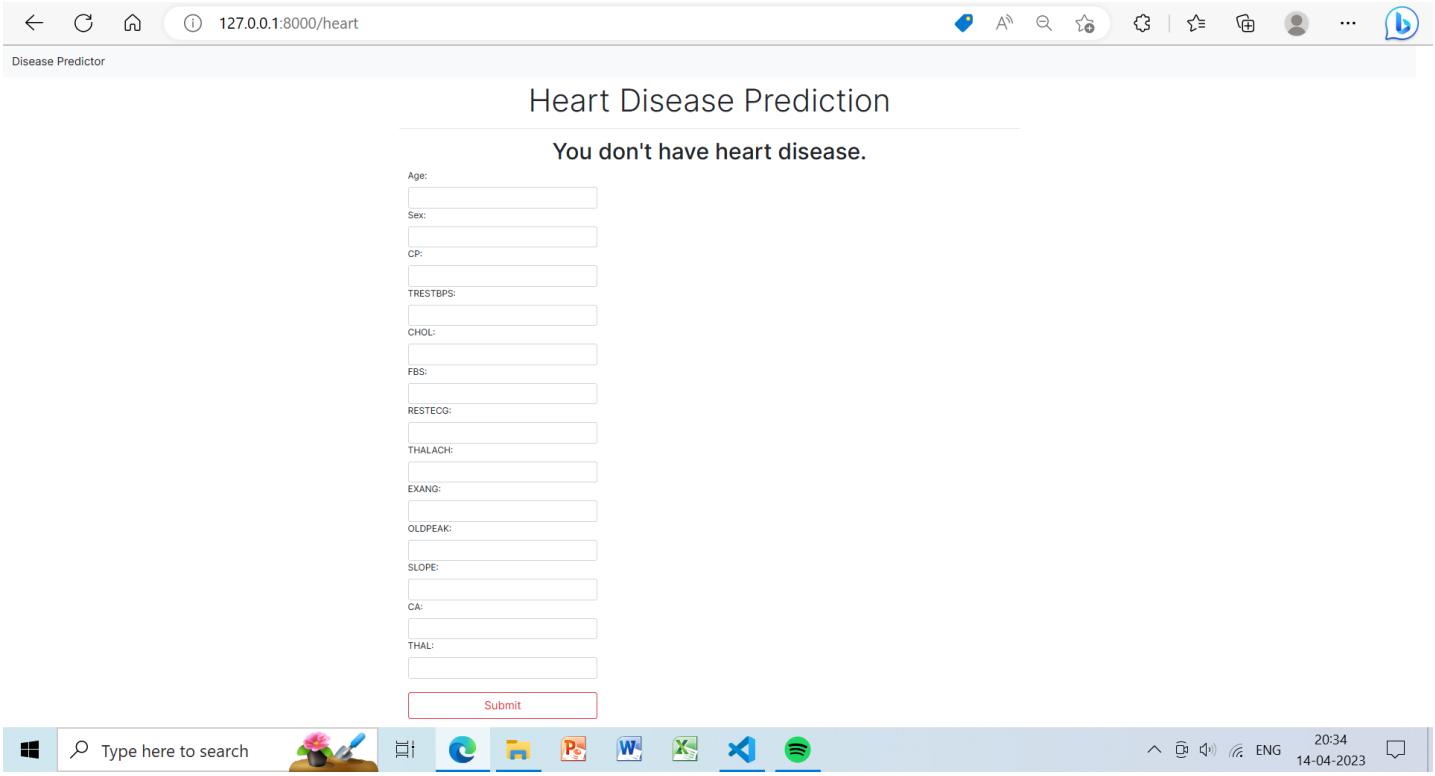



127.0.0.1:8000/heart
Disease Predictor
Heart Disease Prediction
Age:
Sex:
CP:
TRESTBPS:
CHOL:
FBS:
RESTECG:
THALACH:
EXANG:
OLDPEAK:
SLOPE:
CA:
THAL:
Submit
Type here to search
ENG
20:29
14-04-2023


127.0.0.1:8000/heart
Disease Predictor
Heart Disease Prediction
Age:
57
Sex:
1
CP:
0
TRESTBPS:
140
CHOL:
192
FBS:
0
RESTECG:
1
THALACH:
1480.41011
EXANG:
1
OLDPEAK:
0.4
SLOPE:
1
CA:
0
THAL:
1
Submit
Type here to search
ENG
20:31
14-04-2023

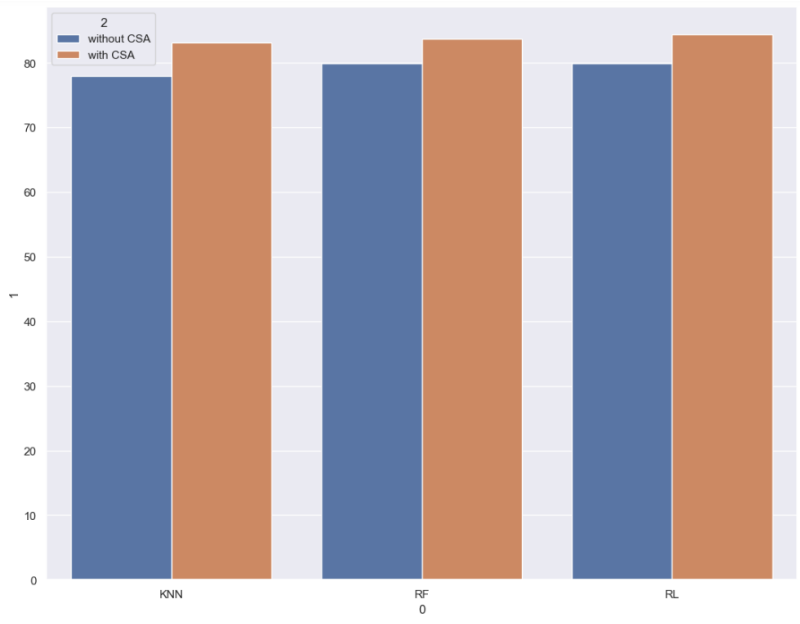

127.0.0.1:8000/heart
Disease Predictor
Heart Disease Prediction
You have heart disease.
Age:
Sex:
CP:
TRESTBPS:
CHOL:
FBS:
RESTECG:
THALACH:
EXANG:
OLDPEAK:
SLOPE:
CA:
THAL:
Submit
Type here to search
ENG
20:32
14-04-2023

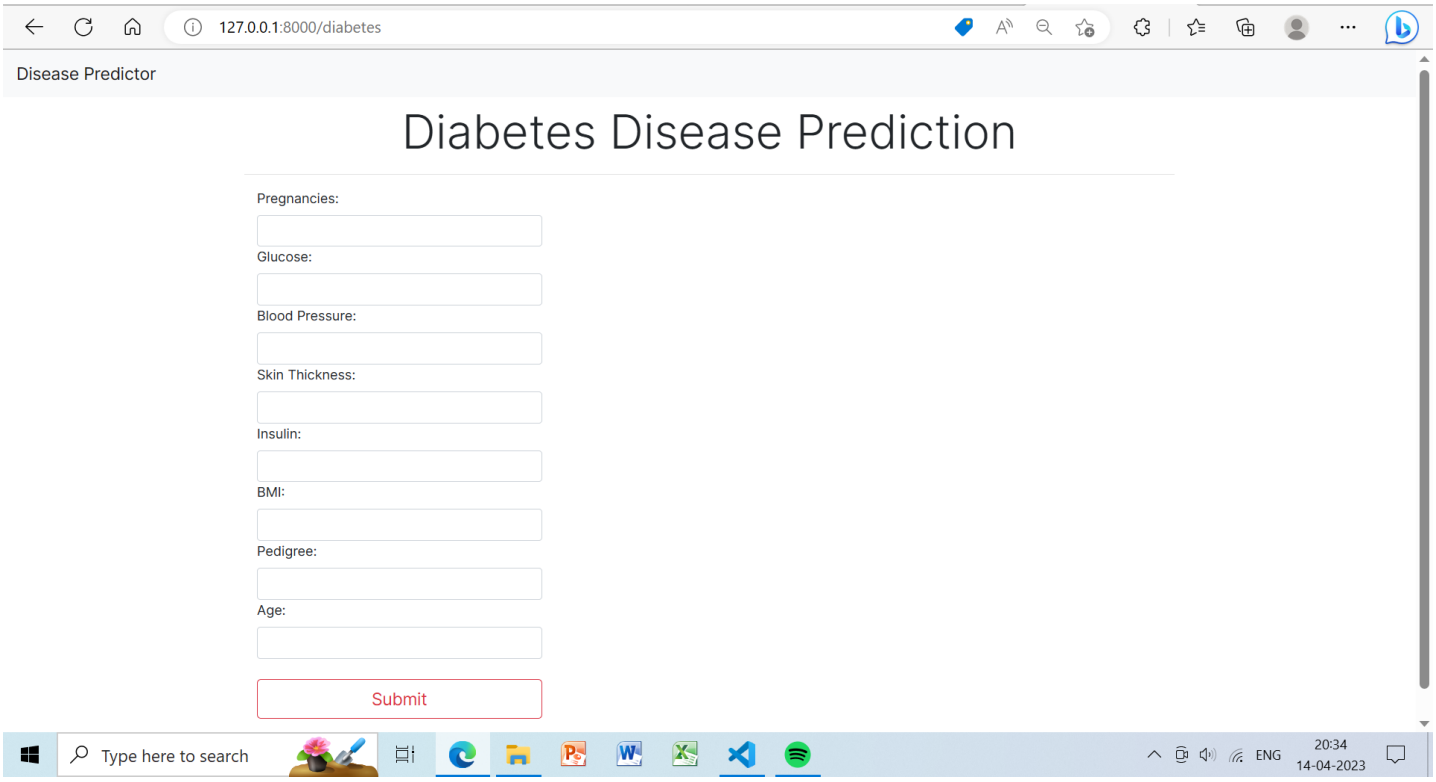

127.0.0.1:8000/heart
Disease Predictor
Heart Disease Prediction
You have heart disease.
Age:
58
Sex:
1
CP:
2
TRESTBPS:
100
CHOL:
100
FBS:
0
RESTECG:
0
THALACH:
100
EXANG:
0
OLDPEAK:
2.5
SLOPE:
1
CA:
1
THAL:
3
Submit
Type here to search
ENG
20:34
14-04-2023

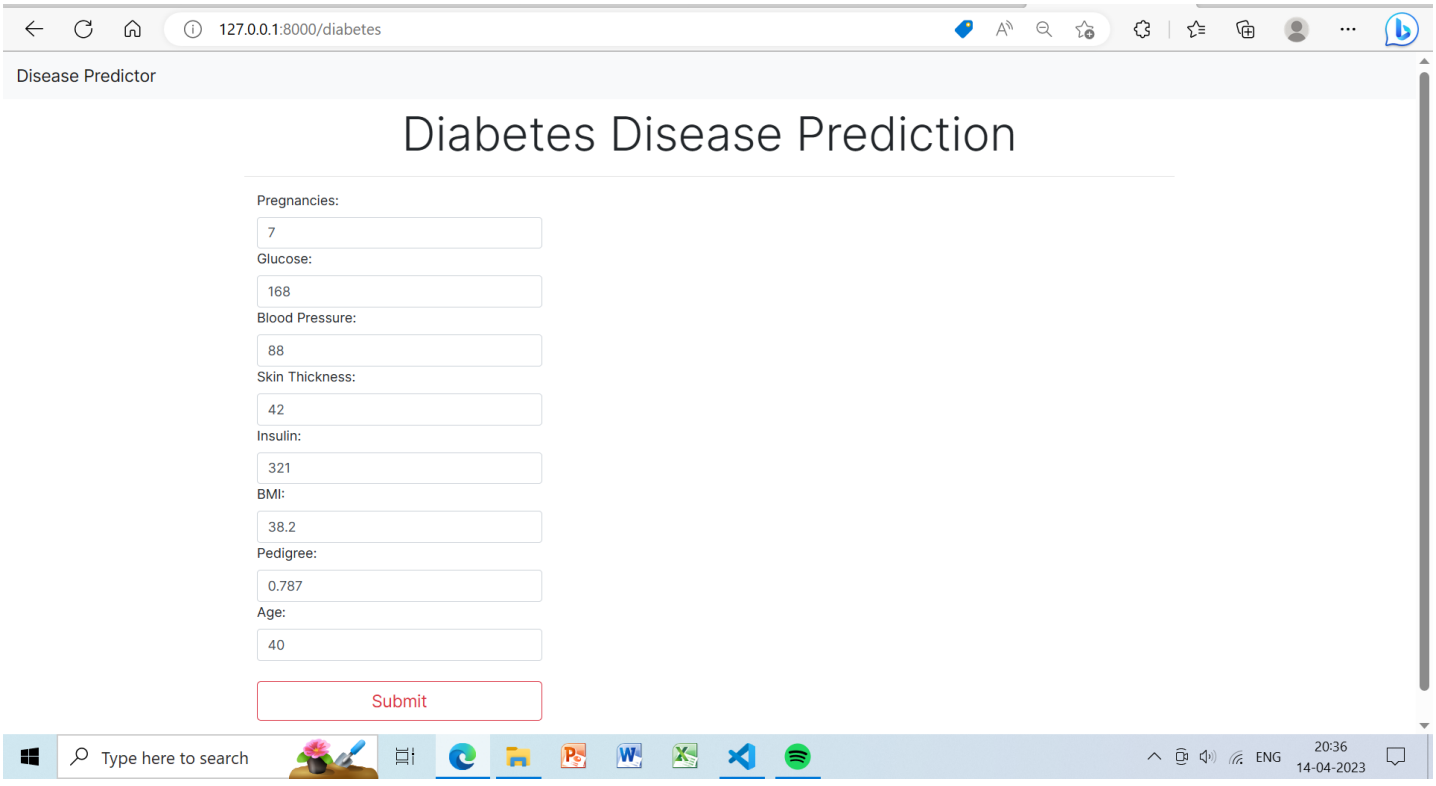


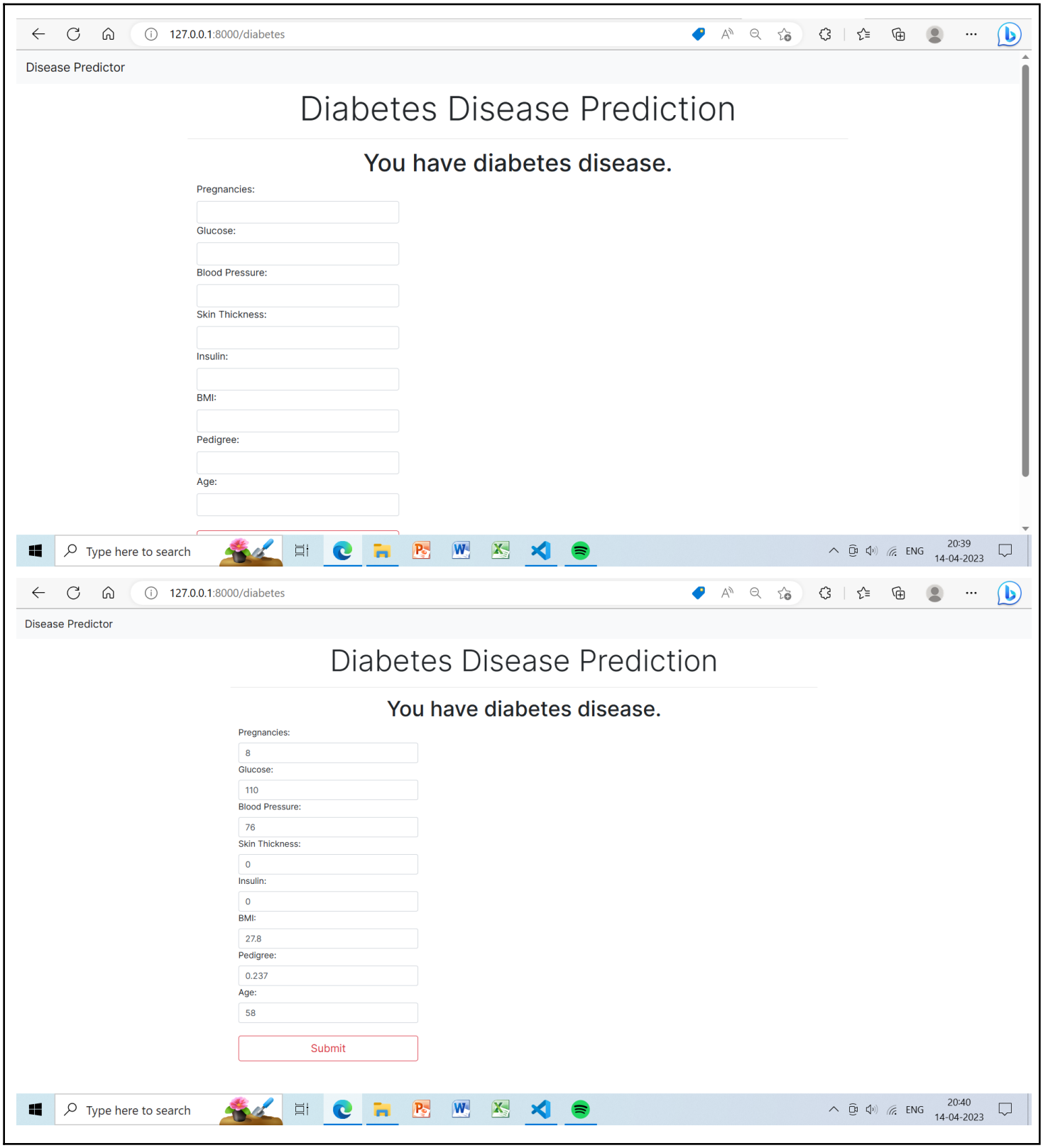
Diabetes Prediction- Supervised Learning with Crow Search Optimization:

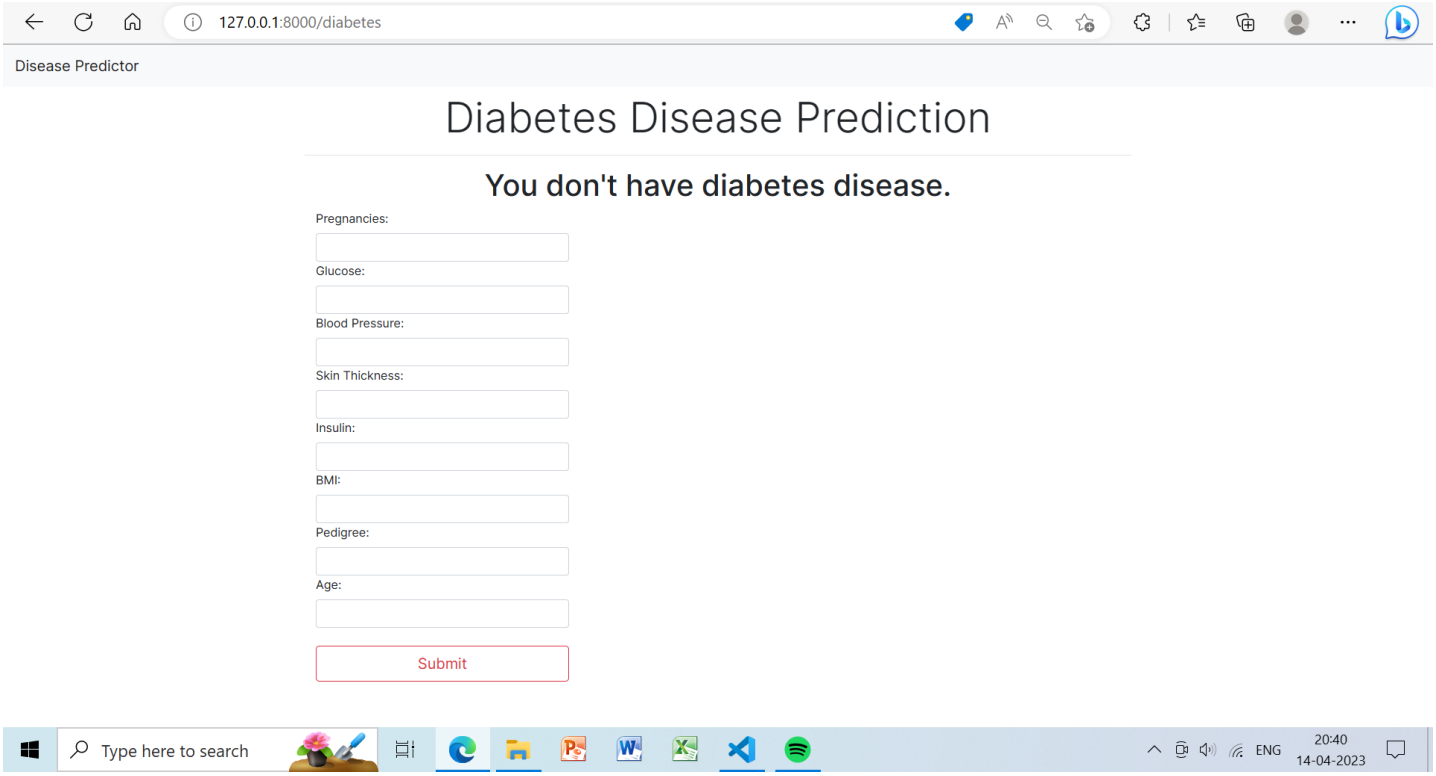
Without CSA	With CSA
KNN = 77.92	KNN = 83.11
RF = 79.87	RF = 83.76
LR= 79.87	LR= 84.41




127.0.0.1:8000/diabetes
Disease Predictor
Diabetes Disease Prediction
Pregnancies:
Glucose:
Blood Pressure:
Skin Thickness:
Insulin:
BMI:
Pedigree:
Age:
Submit
Type here to search
ENG
20:34
14-04-2023

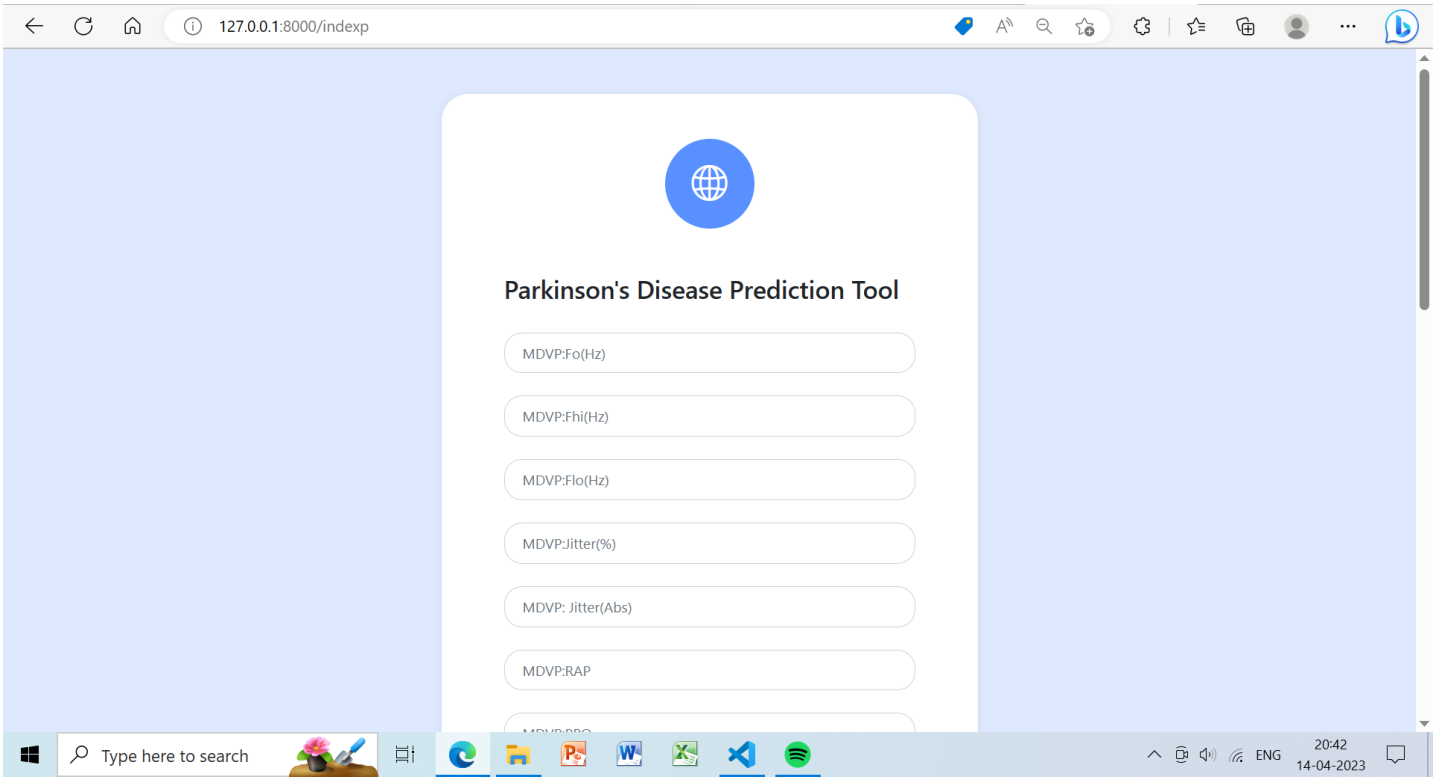

127.0.0.1:8000/diabetes
Disease Predictor
Diabetes Disease Prediction
Pregnancies:
7
Glucose:
168
Blood Pressure:
88
Skin Thickness:
42
Insulin:
321
BMI:
38.2
Pedigree:
0.787
Age:
40
Submit
Type here to search
ENG
20:36
14-04-2023

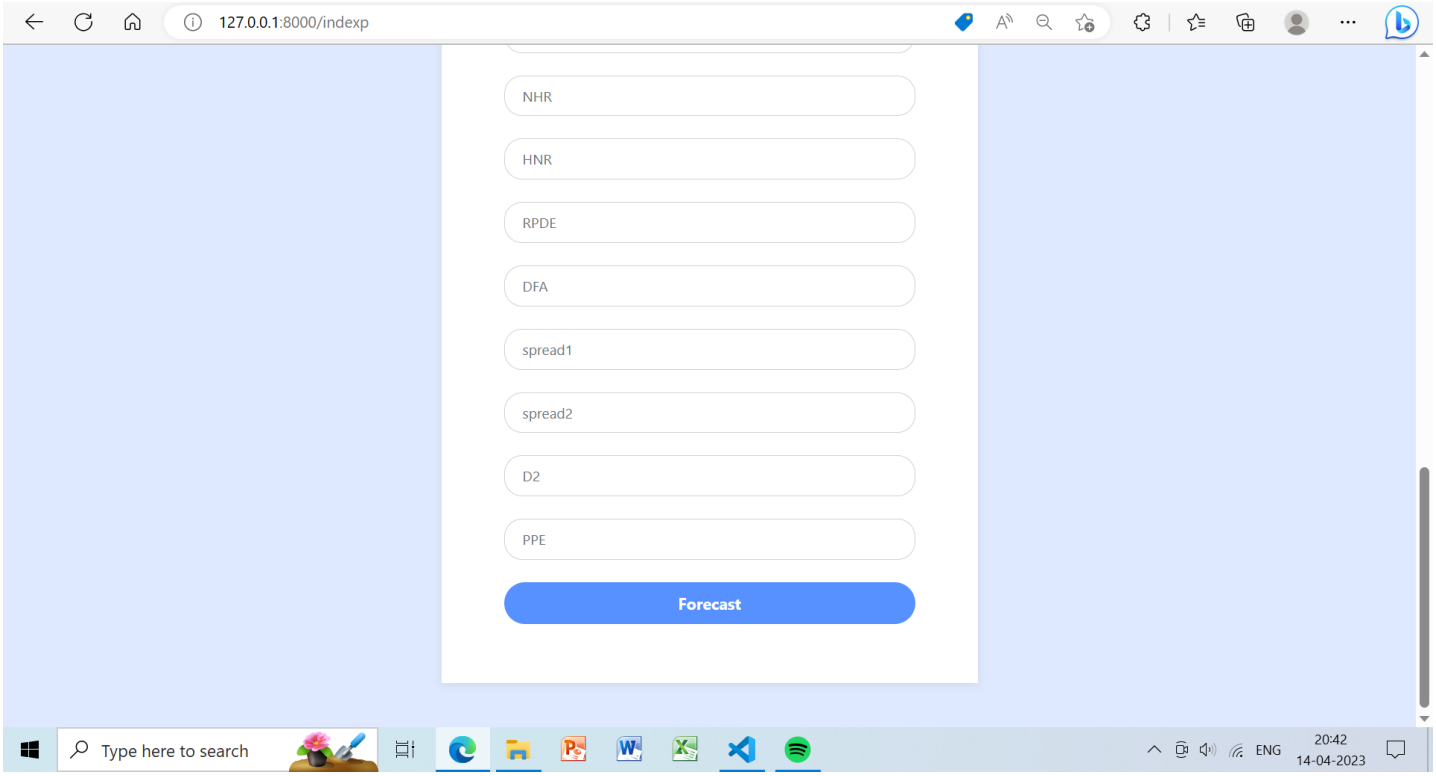

127.0.0.1:8000/diabetes
Disease Predictor
Diabetes Disease Prediction
You have diabetes disease.
Pregnancies:
Glucose:
Blood Pressure:
Skin Thickness:
Insulin:
BMI:
Pedigree:
Age:
Type here to search
ENG
20:39
14-04-2023
127.0.0.1:8000/diabetes
Disease Predictor
Diabetes Disease Prediction
You have diabetes disease.
Pregnancies:
8
Glucose:
110
Blood Pressure:
76
Skin Thickness:
0
Insulin:
0
BMI:
27.8
Pedigree:
0.237
Age:
58
Submit
Type here to search
ENG
20:40
14-04-2023

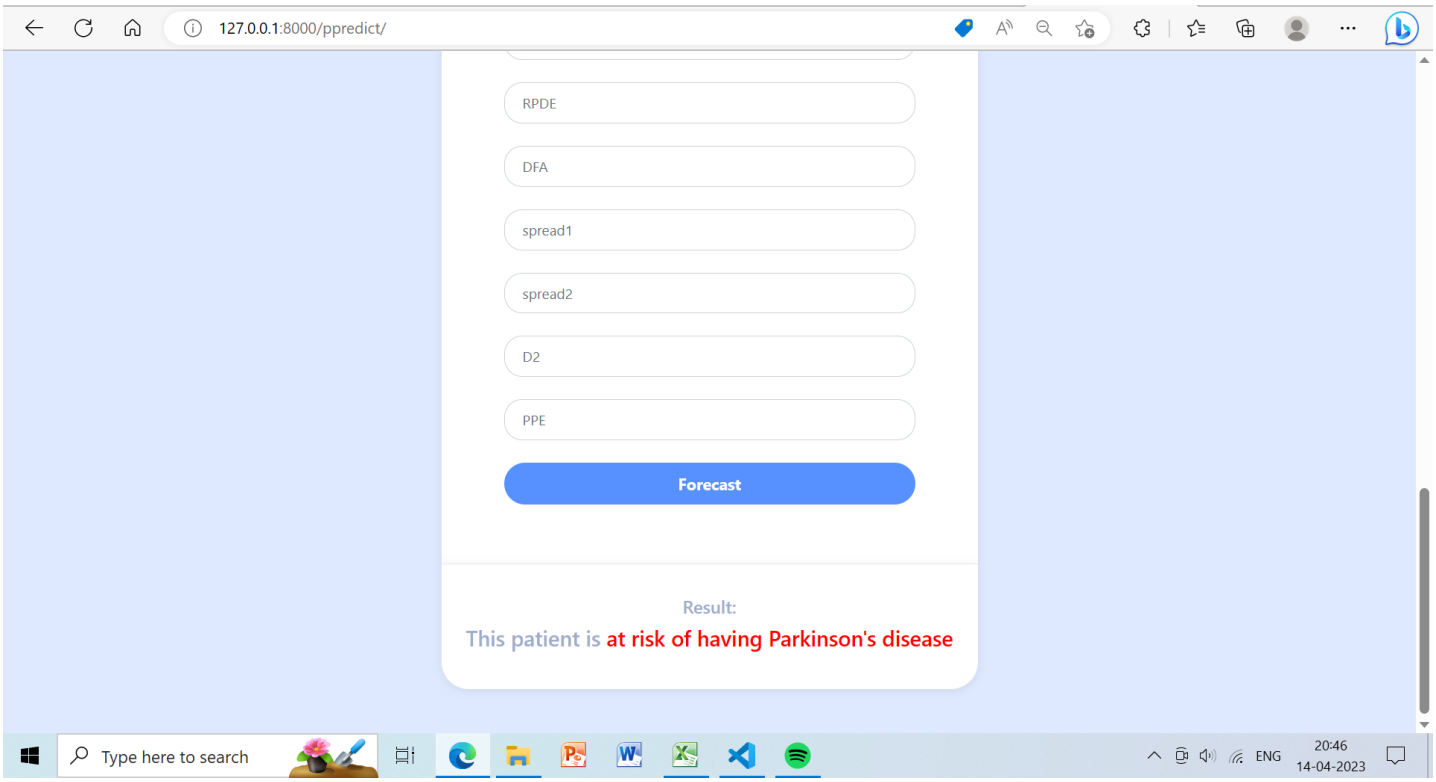


Parkinson's Disease Prediction- Voice data based Support Vector Machine Algorithm:

Accuracy= 88%

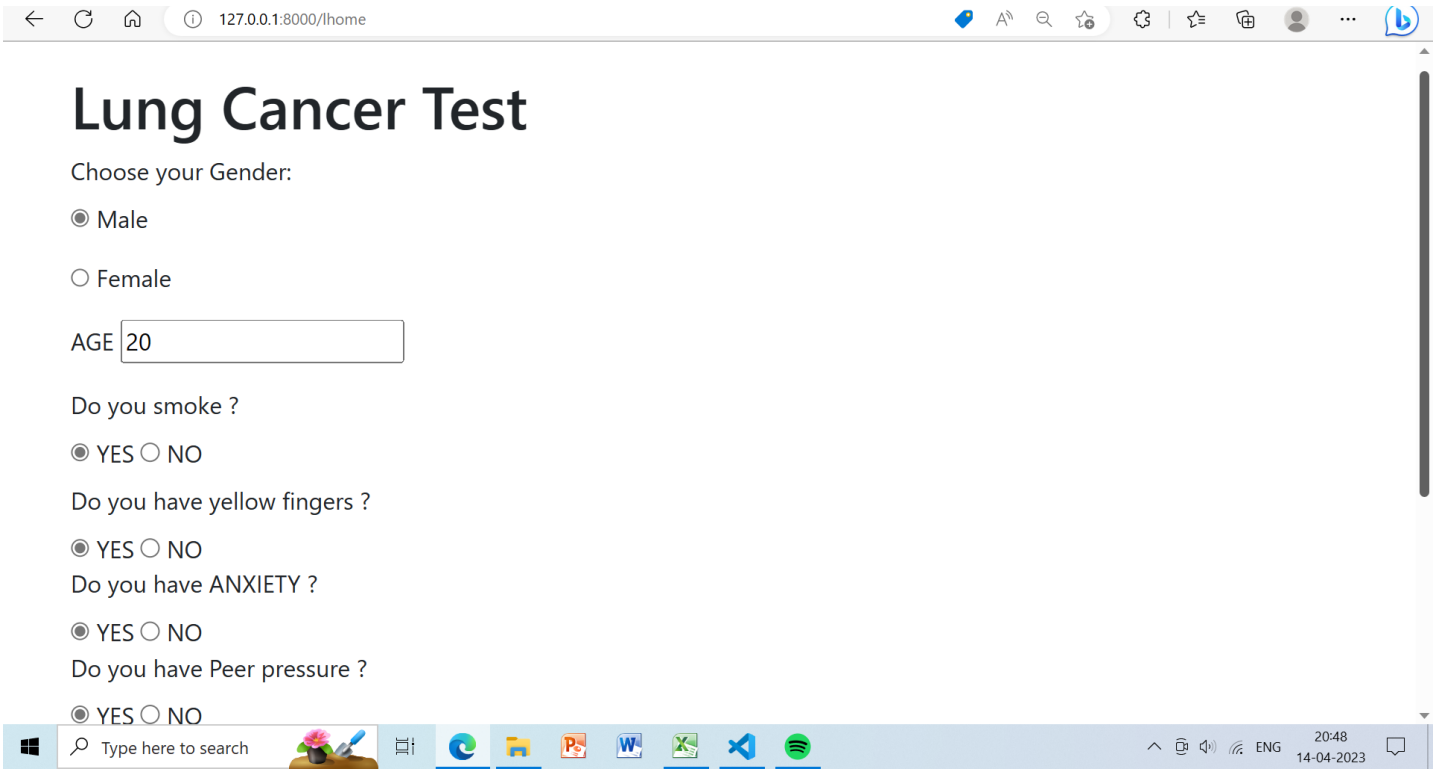


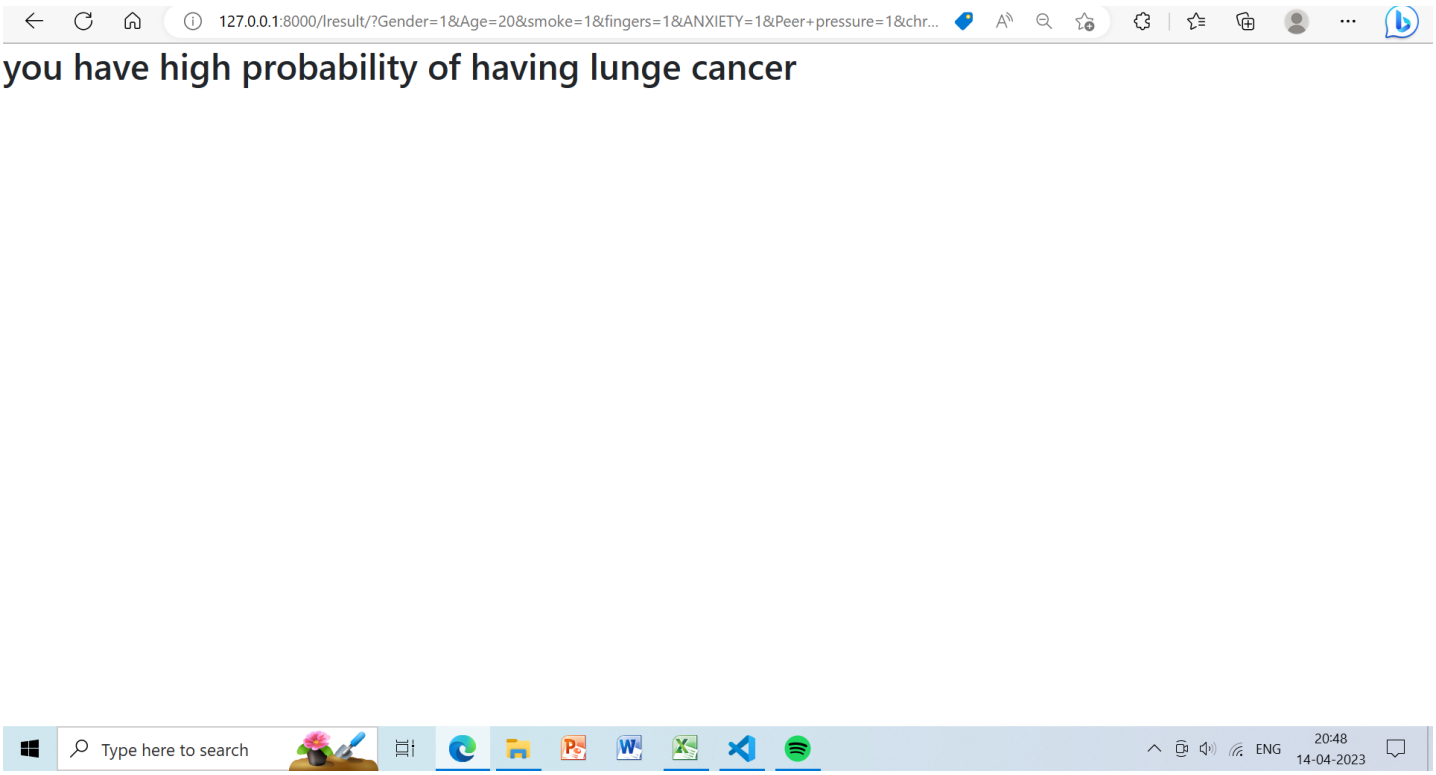

127.0.0.1:8000/indexp
NHR
HNR
RPDE
DFA
spread1
spread2
D2
PPE
Forecast
Type here to search
ENG
20:42
14-04-2023


127.0.0.1:8000/ppredict/
RPDE
DFA
spread1
spread2
D2
PPE
Forecast
Result:
This patient is at risk of having Parkinson's disease
Type here to search
ENG
20:46
14-04-2023

Lung Cancer prediction using XGBoost:

Accuracy= 97%





7. Conclusion

Disease prediction using machine learning, deep learning, and bio-inspired algorithms have shown promising results in recent years. These approaches have enabled the medical community to predict the occurrence of various diseases with high accuracy and efficiency. Machine learning algorithms such as K-Nearest Neighbors, Random Forest, Gradient Boosting, and Extra Trees have been utilized for disease prediction. These algorithms work by learning patterns in the data and predicting the likelihood of disease based on the patterns observed. Deep learning algorithms such as Convolutional Neural Networks and Recurrent Neural Networks have been used to predict diseases from medical images and time-series data, respectively. These algorithms have shown excellent performance in detecting diseases at an early stage. Bio-inspired algorithms such as the Crow Search Algorithm have also been utilized for disease prediction. These algorithms are inspired by the behavior of natural organisms and have been shown to provide better results than traditional algorithms in some cases. Overall, disease prediction using machine learning, deep learning, and bio-inspired algorithms has the potential to revolutionize healthcare by enabling early detection and prevention of diseases. However, further research and development are required to make these approaches more accurate and accessible to the medical community.

9. References

[1] Akhtar, A., 2019. Evolution of Ant Colony Optimization Algorithm--A Brief Literature Review. *arXiv preprint arXiv:1908.08007*.

[2] Hamim, M., El Moudden, I., Pant, M.D., Moutachaouik, H. and Hain, M., 2021. A hybrid gene selection strategy based on fisher and ant colony optimization algorithm for breast cancer classification.

[3] Aldryan, D.P. and Annisa, A., 2018, November. Cancer detection based on microarray data classification with ant colony optimization and modified backpropagation conjugate gradient Polak-Ribiére. In *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)* (pp. 13-16). IEEE.

[4] Anwar, N.H.K., Saian, R. and Bakar, S.A., 2021, July. An enhanced ant colony optimization with Gini index for predicting type 2 diabetes. In *AIP Conference Proceedings* (Vol. 2365, No. 1, p. 020004). AIP Publishing LLC.

[5] Parpinelli, R.S., Lopes, H.S. and Freitas, A.A., 2002. An ant colony algorithm for classification rule discovery. In *Data mining: A heuristic approach* (pp. 191-208). IGI Global.

[6] Dorigo, M., 1992. Optimization, learning and natural algorithms. *Ph. D. Thesis, Politecnico di Milano*.

[7] USHA, S. and KANCHANA, S., 2023. REVIVED ANT COLONY OPTIMIZATION-BASED ADABOOST ALGORITHM FOR HEART DISEASE AND DIABETES (HDD) PREDICTION. *Journal of Theoretical and Applied Information Technology*, *101*(4).

[8] Kavitha, R., Jothi, D.K., Saravanan, K., Swain, M.P., Gonzáles, J.L.A., Bhardwaj, R.J. and Adomako, E., 2023. Ant Colony Optimization-Enabled CNN Deep Learning Technique for Accurate Detection of Cervical Cancer. *BioMed Research International*, *2023*.

[9] Masud, M., Singh, P., Gaba, G.S., Kaur, A., Alroobaea, R., Alrashoud, M. and Alqahtani, S.A., 2021. CROWD: crow search and deep learning based feature extractor for classification of Parkinson's disease. *ACM Transactions on Internet Technology (TOIT)*, *21*(3), pp.1-18.

[10] Wang, Y., Wang, A.N., Ai, Q. and Sun, H.J., 2017. An adaptive kernel-based weighted extreme learning machine approach for effective detection of Parkinson's disease. *Biomedical Signal Processing and Control*, *38*, pp.400-410.
[11] Singh, G., Vadera, M., Samavedham, L. and Lim, E.C.H., 2016. Machine learning-based framework for multi-class diagnosis of neurodegenerative diseases: a study on Parkinson's disease. *IFAC-PapersOnLine*, *49*(7), pp.990-995.
[12] Yang, X.S., 2020. *Nature-inspired optimization algorithms*. Academic Press.
[13] Askarzadeh, A., 2016. A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm. *Computers & structures*, *169*, pp.1-12.
[14] De Souza, R.C.T., dos Santos Coelho, L., De Macedo, C.A. and Pierezan, J., 2018, July. A V-shaped binary crow search algorithm for feature selection. In 2018 IEEE congress on evolutionary computation (CEC) (pp. 1-8). IEEE.
[15] Abdallh, G.Y. and Algamal, Z.Y., 2020. A QSAR classification model of skin sensitization potential based on improving binary crow search algorithm. Electronic Journal of Applied Statistical Analysis, 13(1), pp.86-95.
[16] Sayed, G.I., Hassanien, A.E. and Azar, A.T., 2019. Feature selection via a novel chaotic crow search algorithm. *Neural computing and applications*, *31*, pp.171-188.
[17] Meraihi, Y., Mahseur, M. and Acheli, D., 2020. A modified binary crow search algorithm for solving the graph coloring problem. *International Journal of Applied Evolutionary Computation (IJAEC)*, *11*(2), pp.28-46.
[18] Mandala, J. and Rao, M.C.S., 2019. Privacy preservation of data using crow search with adaptive awareness probability. *Journal of information security and applications*, *44*, pp.157-169.
[19] Gupta, D., Sundaram, S., Khanna, A., Hassanien, A.E. and De Albuquerque, V.H.C., 2018. Improved diagnosis of Parkinson's disease using optimized crow search algorithm. *Computers & Electrical Engineering*, *68*, pp.412-424.
[20] Oliva, D., Hinojosa, S., Cuevas, E., Pajares, G., Avalos, O. and Gálvez, J., 2017. Cross entropy based thresholding for magnetic resonance brain images using Crow Search Algorithm. *Expert Systems with Applications*, *79*, pp.164-180.
[21] Anter, A.M. and Ali, M., 2020. Feature selection strategy based on hybrid crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm for medical diagnosis problems. *Soft Computing*, *24*(3), pp.1565-1584.
[22] Gadekallu, T.R., Alazab, M., Kaluri, R., Maddikunta, P.K.R., Bhattacharya, S. and Lakshmanna, K., 2021. Hand gesture classification using a novel CNN-crow search algorithm. *Complex & Intelligent Systems*, *7*, pp.1855-1868.
[23] Surendar, P., 2021. Diagnosis of lung cancer using hybrid deep neural network with adaptive sine cosine crow search algorithm. *Journal of Computational Science*, *53*, p.101374.
[24] Zhao, S., Wang, P., Heidari, A.A., Zhao, X. and Chen, H., 2023. Boosted crow search algorithm for handling multi-threshold image problems with application to X-ray images of COVID-19. *Expert Systems with Applications*, *213*, p.119095.
[25] Devikanniga, D., Ramu, A. and Haldorai, A., 2020. Efficient diagnosis of liver disease using support vector machine optimized with crows search algorithm. *EAI Endorsed Transactions on Energy Web*, *7*(29), pp.e10-e10.
[26] Alagarsamy, S., Subramanian, R.R., Shree, T., Balasubramanian, M. and Govindaraj, V., 2021, February. Prediction of lung cancer using meta-heuristic based optimization technique: Crow search technique. In *2021*

International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 186-191). IEEE.
[27] Bhoomija, P., Jyothi, A. and Puppala Rani, Y.S., 2022. Prediction of Heart Disease Using Bio-Inspired Algorithms. *JOURNAL OF ALGEBRAIC STATISTICS*, *13*(3), pp.4762-4766.
[28] World Health Organization Fact sheet statistics details of Breast Cancer, 26 March 2021- https://www.who.int/news-room/fact-sheets/detail/breast-cancer
[29] World Health Organization Fact sheet statistics details of Cardiovascular diseases (CVDs), 11 June 2021- https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[30] H. Yu, G. Gu, H. Liu, J. Shen, and J. Zhao, "A modified ant colony optimization algorithm for tumor marker gene selection," Genomics, Proteomics and Bioinformatics, 7, 200–208 (2009).
[31] Y. Marinakis and G. Dounias, "Nature inspired intelligence in medicine: Ant colony optimization for pap-smear diagnosis," International Journal on Artificial Intelligence Tools, 17, 279–301 (2008).
[32] F. M. Ramo, "Diagnosis of heart disease based on ant colony algorithm," International Journal of Computer Science and Information Security, 11 (2013).
[33] N. H. K. Anwar and R. Saian, "Predictive accuracy for two diabetes dataset using ant-miner algorithm," INTERNATIONAL JOURNAL OF SCIENTIFIC AND TECHNOLOGY RESEARCH, 9 (2020).