Atal Bihari Vajpayee Indian Institute of Information Technology and Management Gwalior

INFORMATION TECHNOLOGY

Minor Project

Saathi

An ML based crop recommendation and plant disease identification website

---

STUDENT ID

SAHIL GANGURDE (2019IMT-034)

UNDER THE SUPERVISION OF

DR. PINKU RANJAN

Contents

1 ABSTRACT

Since there have been climate changes which have resulted in an increasing amount of unexpected rainfalls, par below temperatures and heatwaves in the region, which result in significant loss of ecosystem. Machine learning has helped develop various utility tools to tackle world problems. This problem of agriculture can be solved by using various ML algorithms. This project aims to create two things - a)A crop recommendation system and b) a Plant disease identification system embedded into a single website. The datasets were publicly available over the internet. Once the features for task one are extracted, then the dataset is trained on five different algorithms - logistic regression, decision tree, support vector machine(SVM), multi-layer perceptron and random forest. For the second task, three CNN architectures, VGG16, ResNet50 and EfficientNetV2, are trained, and a comparative study is done between them. For the task one, random forest achieved an accuracy of 99.31%, and for the second task, EfficientNetV2 achieved an accuracy of 96.06%.

2 INTRODUCTION

Machine learning can help humans solve problems which are not easily solvable by humans. Machine learning can be used to solve tasks like classification, prediction, identification, etc. and can be applied to a wide range of other fields such as agriculture, sports, trade and business, and so on. This project aims at building a website focused on the agriculture sector, solving two significant issues crop recommendation and crop disease identification. The method used to solve these problems is by training models on datasets available over the internet and comparing them. Models with reasonable accuracy are embedded into the website, which can be then deployed on the cloud.

2.1 Motivation

Climate change has been quite effective over the past five years. Less knowledge about scientific ways of farming also leads to wrong decisions in the selection of crops. Farmers often tend to rely on experiences which are limited and also full of errors. Overall the agriculture industry suffers a huge loss due to improper usage of knowledge such as soil constituents present, pH of the soil, and early detection of diseases of plants. This problem can be solved by making proper use of technology. With the help of machine learning and the web, this solution can reach every individual having access to a mobile phone with an internet connection on it.

2.2 Literature Survey

There are many attempts made to tackle the problem of crop recommendation and plant disease classification. G Chauhan and A. Chaudhary proposed the ways to recommend crops based on soil type and used random forest and decision trees to make the prediction(2). This showed that the task of predicting crops based on land patterns would be helpful via a random forest classifier.

For plant disease detection, S.P. Mohanty has an open-source dataset of plant leaf images along with their grayscale and segmented part. The paper published by him talks about the classification task done by using AlexNet(3). This paper(4) talks about using the above-mentioned dataset on various DL classification models, and Inspired by this state of the art result, and I decided to use the CNN approach for plant disease detection.

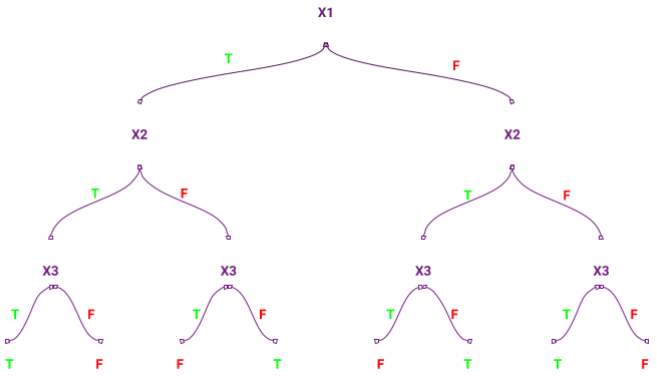

Figure 1: Pictorial representation of the decision tree

2.3 Project Objectives

The project is divided into 3 parts crop recommendation, disease identification and deployment. The first two tasks are related to creating ML models with higher accuracy and the last task is related to creating a web application using flask and also create a neat and robust system design which will help us to serve larger requests over the network.

2.4 Classification Algorithms

This section gives a brief overview of the classification algorithms used in this project.

Logistic Regression

Logistic regression is similiar to linear regression but it is used to solve classification tasks. The function used in the logistic regression is a signmoid function which gives a value between 0 and 1.

Decision Tree

Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves. Entropy is the uncertainty in our dataset or measure of disorder. Information gain measures the reduction of uncertainty given some feature and it is also a deciding factor for which attribute should be selected as a decision node or root node. See figure:1 for better understanding.

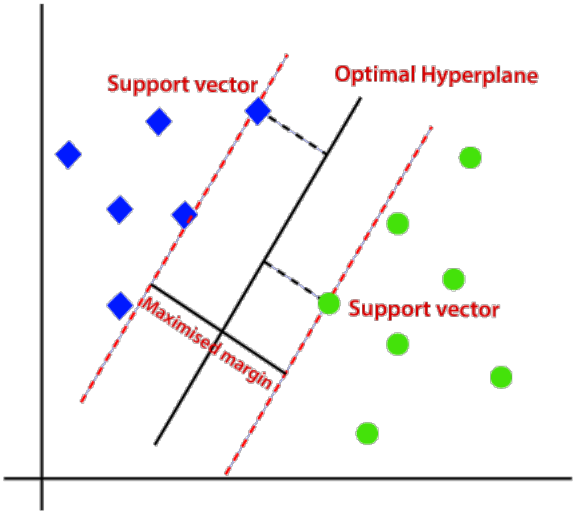

Figure 2: Pictorial representation of the decision tree

Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. See figure:2 for better understanding.

Multi-Layer Perceptron

MLPClassifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLPClassifier relies on an underlying Neural Network to perform the task of classification. The multilayer perceptron (MLP) is a feedforward artificial neural network model that maps input data sets to a set of appropriate outputs. See figure:3 for better understanding.

Random Forest

random forest is build using multiple decision trees. The decisions from each tree is used and then the maximum outcome from all is chosen to give as the output. Due to this most of the time Random forest works as the best classifier model. See figure:4 for better understanding.

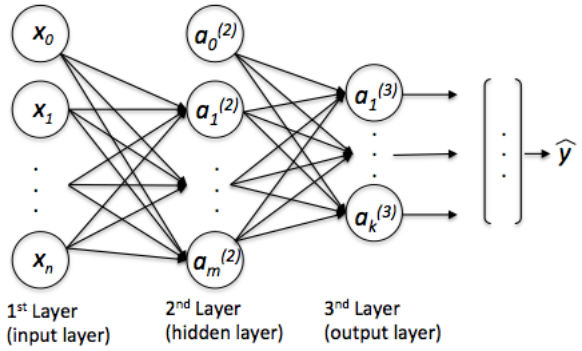

Figure 3: Multi-Layer Perceptron Classifier

Figure 4: Random forest

2.5 Convolutional Neural Networks

While handling image data CNNs are the best type of approach to perform classification. The below mentioned models are used in plant disease detection task.

VGG16

The VGG-16 is one of the most popular pre-trained models for image classification. Introduced in the famous ILSVRC 2014 Conference, it was and remains THE model to beat even today. Developed at the Visual Graphics Group at the University of Oxford, VGG-16 beat the then standard of AlexNet and was quickly adopted by researchers and the industry for their image Classification Tasks. Figure:5 shows the VGG16 architecture.

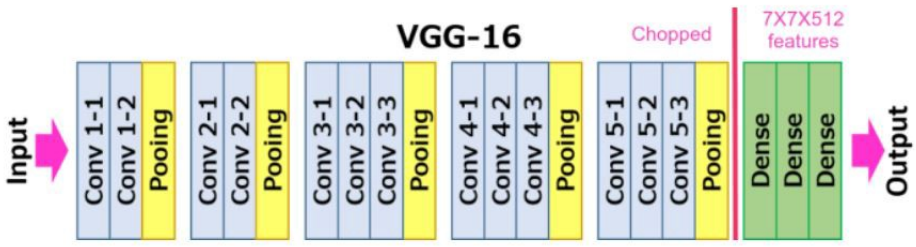

Figure 5: VGG 16 architecture

ResNet50

ResNet50 is a variant of ResNet model which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer. It has 3.8 x $10\hat{9}$ Floating points operations. It is a widely used ResNet model and we have explored ResNet50 architecture in depth. This architecture can be used on computer vision tasks such as image classififcation, object localisation, object detection. Figure:6 shows the ResNet50 architecture.

EfficientNet

EfficientNet is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image. The base EfficientNet-B0 network is based on the inverted bottleneck residual blocks

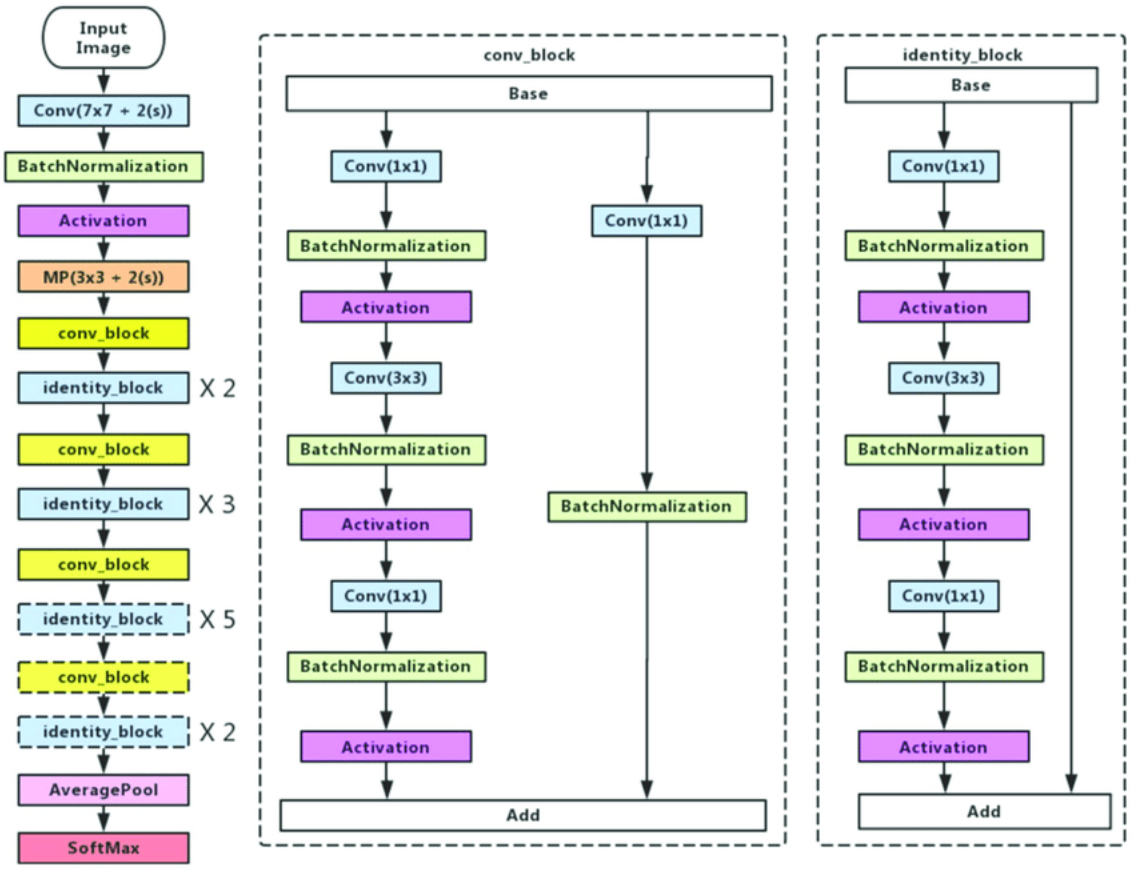

Figure 6: ResNet50 architecture

of MobileNetV2, in addition to squeeze-and-excitation blocks. In this project EfficientV2S is used which the current SOTA of Google on the ImageNet dataset. Figure:14 shows the EfficientNet architecture.

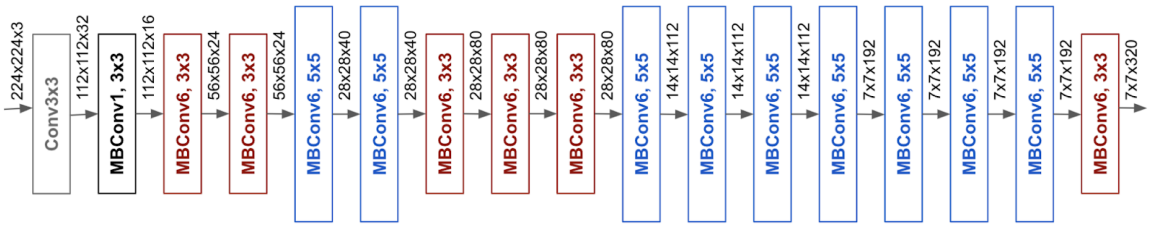

Figure 7: EfficientNet architecture

3 METHODOLOGY

3.1 System Architecture

The overall system can be divided into two parts a) Web server and b) a ML container. The whole web server can be connected to a database most probably a SQL database. The architecture is designed in a way to allow large number of requests. It is easily scalable and deployable.

The technologies used for this whole project are:

- Docker: Docker helps in containarization process. It creates a whole development environment which is easy to use and easily deployable across any server
- NginX: Nginx is used to create a load balancer. A load balancer help distrubute the traffic among multiple available servers using different algorithms.
- Tensorflow: It is an open-source machine learning framework provided by Google.
- Keras: Keras is a deep learning library which provides powerful deep learning solutions
- MySQL: A relational database system
- Flask: A light-weight python web server

See figure:8 for a better understanding of the architecture

3.2 Datasets

The datasets for both the tasks were available online and were open sourced under open license to be used for anybody. The dataset for the first task consists of a CSV file which contains 2200 entries of the various factors such as soil condition, temperature, pH, humidity and rainfall and label output as the type of crop well produced in that type of conditions. Refer figure: 9 for peak into dataset.

There are in total 22 types of crops available in the dataset and they are 'rice' 'maize', 'chickpea', 'kidneybeans', 'pigeonpeas', 'mothbeans', 'mungbean', 'blackgram', 'lentil', 'pomegranate', 'banana', 'mango', 'grapes', 'watermelon', 'muskmelon', 'apple', 'orange', 'papaya', 'coconut', 'cotton', 'jute' and 'coffee'.

The second dataset consists of 70,000 plant images having various diseases. It was a 5GB data all of resolution 256x256. There are in total 38 classes available where

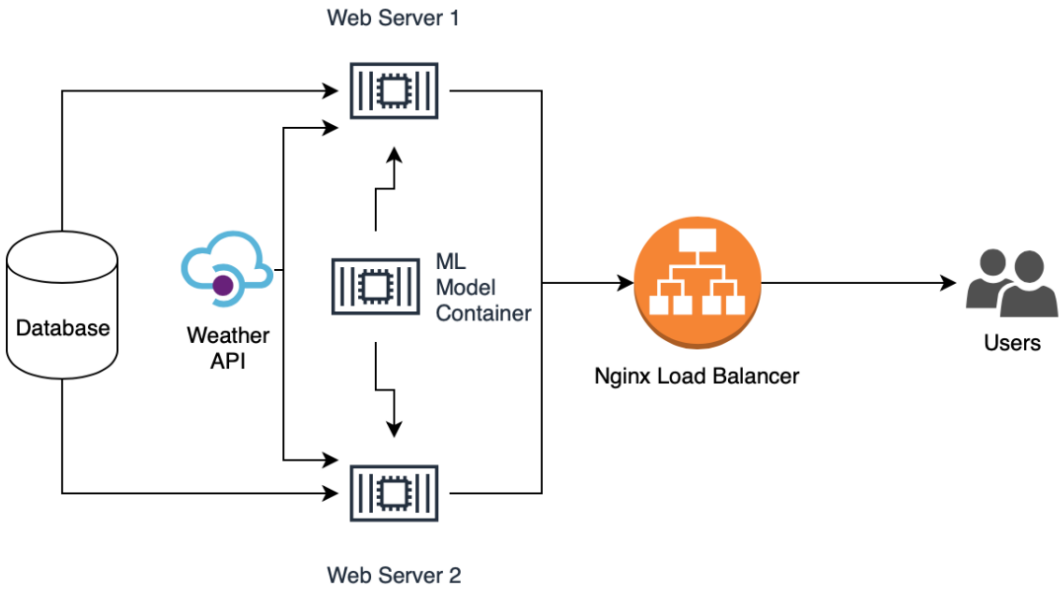

Figure 8: System Architecture

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

Figure 9: Random data from the dataset for crop recommendation

there are 14 different plants and 26 diseases to be identified. Refer figure: 10 for peak into dataset.

These datasets were used to train all the further mentioned ML algorithms and the one with best accuracy was chosen.

3.3 Crop Recommendation

The crop recommendation is basically a classification task. Standard ML algorithms were used to classify various plants. The features trained for the classification were

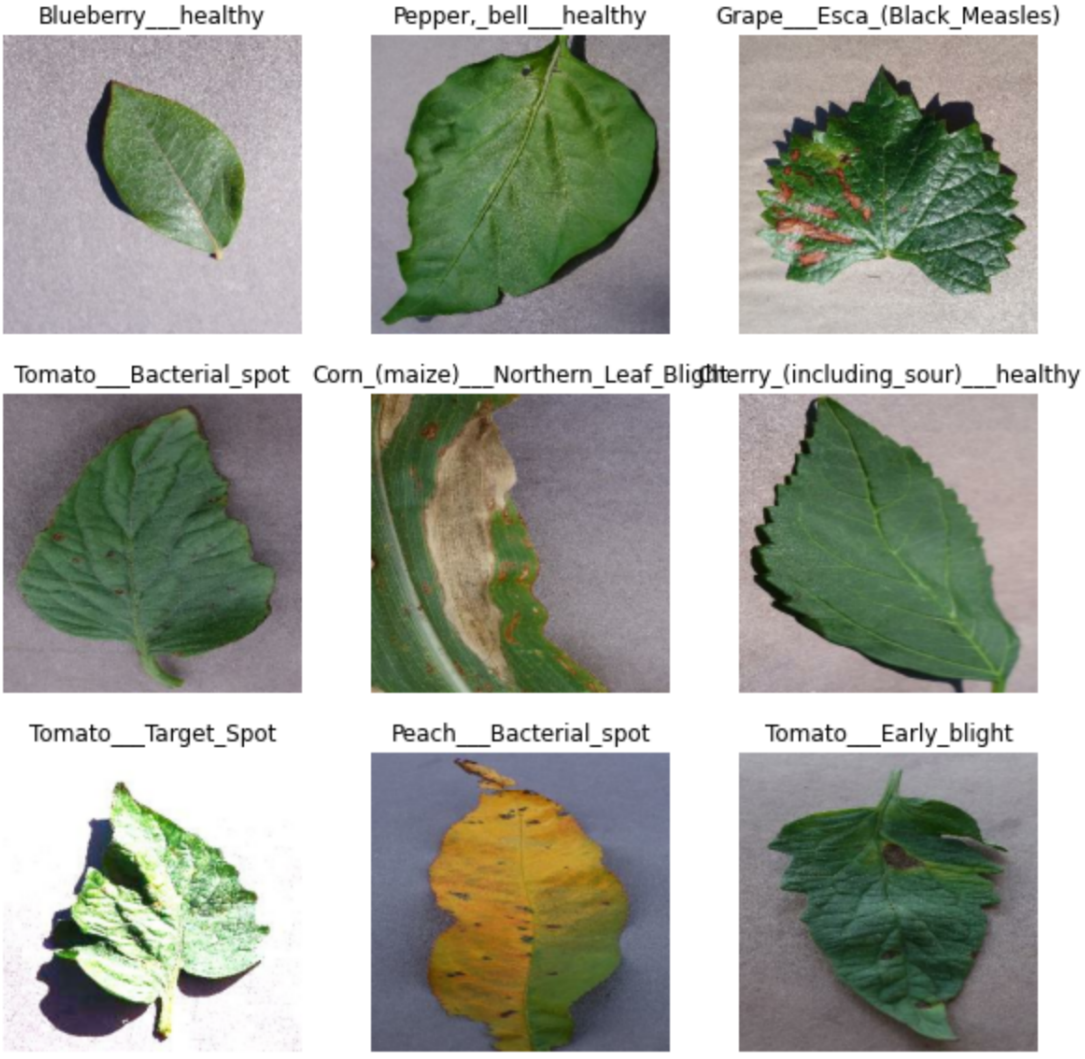

Figure 10: Random data from the dataset for plant disease identification

the NPK value of the soil, temperature of the surroundings, pH of the soil and the rainfall in a particular area.

This dataset was trained on the following algorithms:

- Logistic Regression
- Decision Tree
- Support Vector Machine (SVM)
- Multi Layer Perceptron
- Random Forest

These five algorithms were chosen because given the features and labeled dataset these algorithms can run faster and provide good result on labeled dataset for classification problem.

The accuracy of the mentioned algorithms is given in the observations section.

3.4 Plant Disease Identification

The dataset available for this task is a collection of 70,000 images in total. Image classification is a hard task for normal ML algorithms since the feature detection is not easy. Hence for major image classification problems convolutional neural networks(CNNs) are used which can detect features while training and provide better accuracy over the traditional ML algorithms.

There are 3 different state-of-the-art CNN architecture available for the image classification task.

- VGG16
- ResNet50
- EfficientNet

The discussion about the various model accuracy is given in the observations section.

3.5 Tools and Technologies

As the project had two parts, web and ML two spectrums of tools are been used. Those are listed below:

- Python
- Tensorflow

- Keras
- Flask
- Docker
- Nginx
- HTML, CSS, JS

4 RESULT

4.1 Model results

For the first task, 5 classification algorithms were taken and then the accuracy over each of them were measured. For the task 2 the best image classification CNN architectures were trained. VGG16, ResNet50 and EfficientNetv2 are trained. The tables below show the results 2

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	94.54	0.95	0.95	0.94
Decision Tree	97.72	0.98	0.98	0.98
Support Vector Machine (SVM)	9.09	0.59	0.09	0.11
Multi-Layer Perceptron	95.22	0.96	0.95	0.95
Random Forest	99.31	0.99	0.99	0.99

Table 1: Crop recommendation task accuracy over various algorithms. This table concludes that the random forest outperforms every other algorithm achieving 99.92% accuracy

A histogram representation of the accuracy vs models is also shown for the task 1 of the project. See figure:11 for more information.

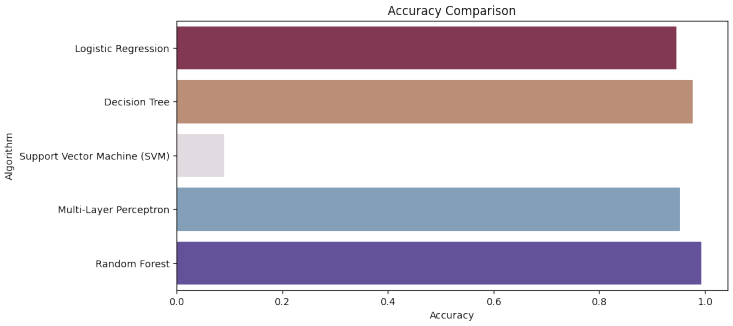

Figure 11: Histogram representation of the table 1

For the task 2 i.e. the plant disease identification task three models are trained VGG16, ResNet50 and EfficientNetV2S. The accuracy and loss function graph over various epochs of both the models is shown below.

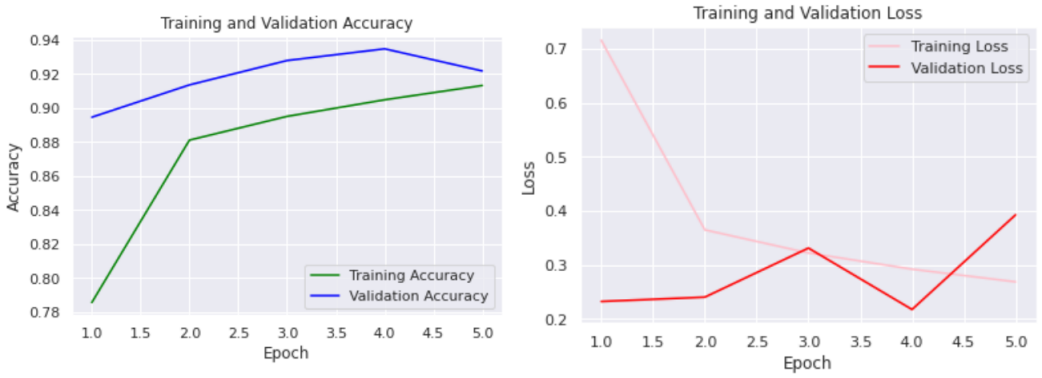

Figure 12: Accuracy and Loss graphs of the VGG model

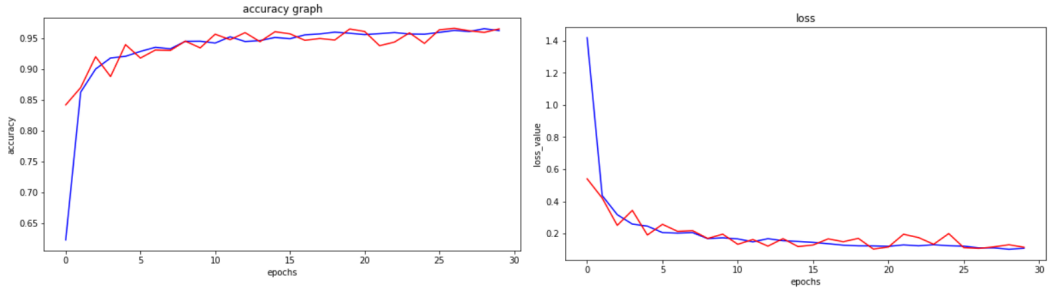

Figure 13: Accuracy and Loss graphs of the ResNet50 model

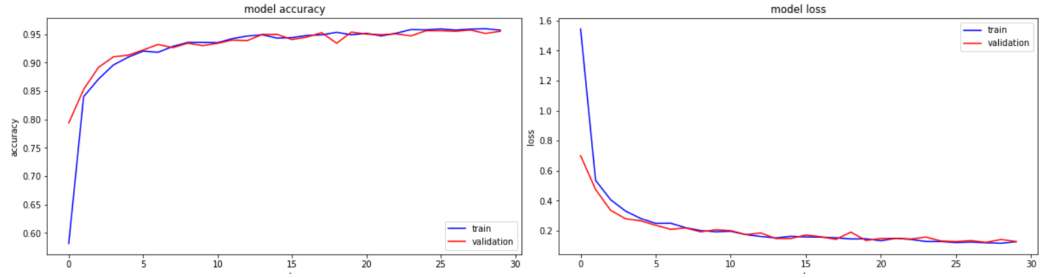

Figure 14: Accuracy and Loss graphs of the EfficientNetV2S model

Architecture	Training Acc.	Validation Acc.	Test Accuracy.
VGG16	92.18	91.33	91.78
ResNet50	96.02	95.41	95.53
EfficientNetV2	96.06	95.53	95.83

Table 2: Plant Disease Classification task accuracy over various architectures. This table concludes that EfficientNetV2 is efficient.

4.2 Deployed website

This section shows the various outputs of the deployed application. The web application is created in Flask and has a ML backend to it. It will later be deployed on cloud service platform.

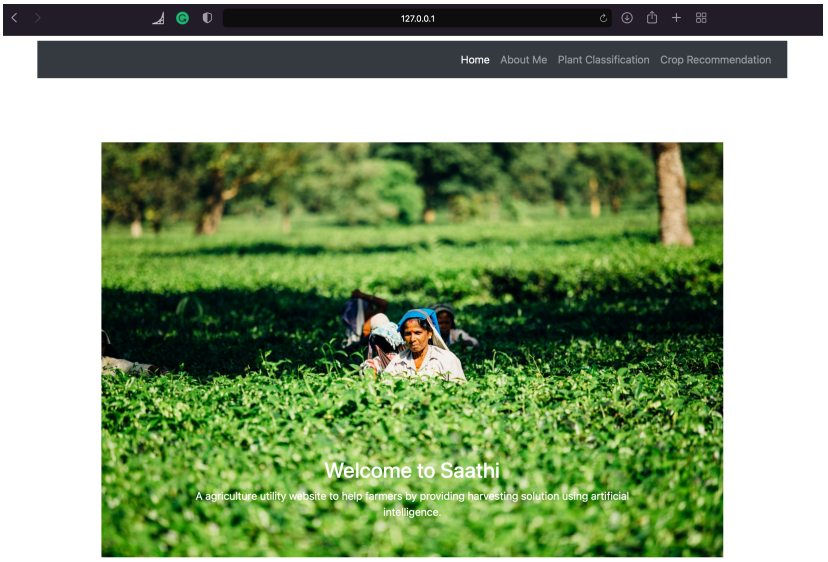

Figure 15: Landing page of the website

Figure 15 shows the landing page of the website. The pages are build using bootstrap and flask. Figure 16 shows the landing page of the plant disease classification and figure 17 shows the result of the plant disease classification. Figure 18 shows the landing page of crop recommendation system and figure 19 shows the corresponding result.

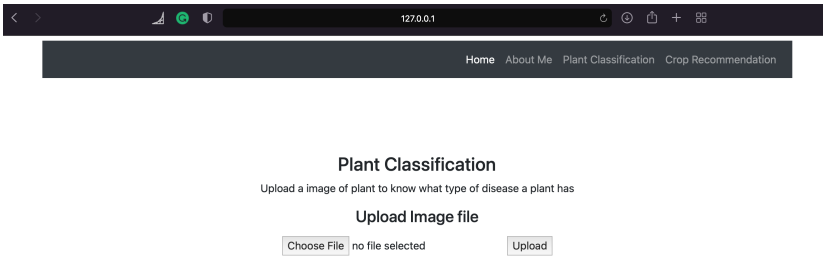

Figure 16: Plant disease classification landing page

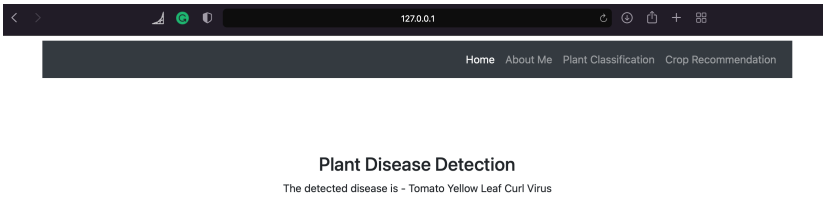

Figure 17: Result of the plant disease classification

5 FUTURE SCOPE

CNN architectures have been replaced by Transformer Networks. Models like ViT, CoAtNet which uses transformer neural networks which are the current state-of-the-art image classification neural nets. If these networks are used accuracy can

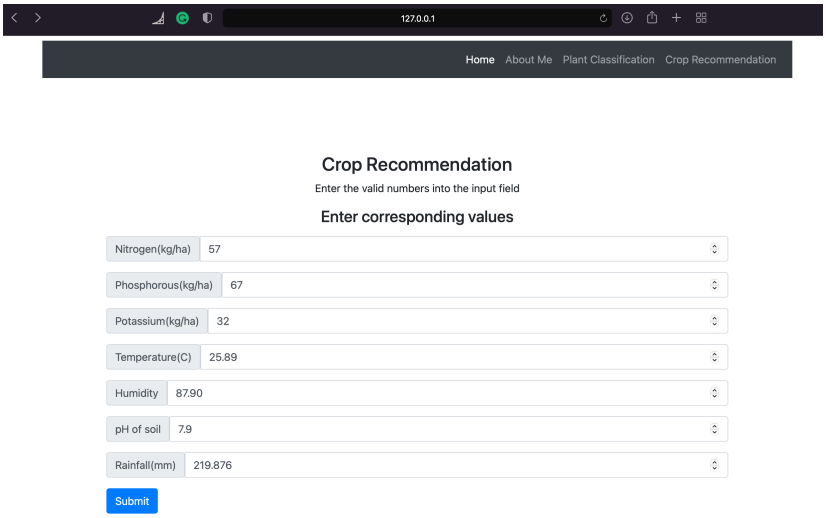

Figure 18: Crop recommendation system

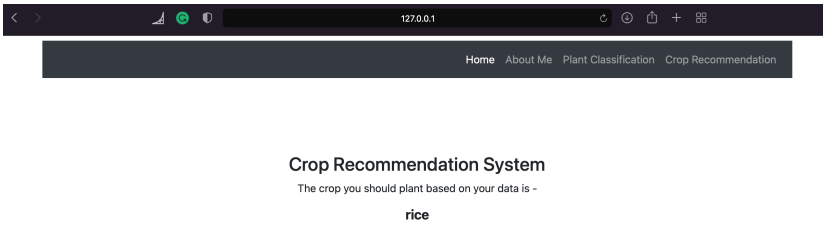

Figure 19: Result of crop recommendation

be boosted by 2 to 3 %. Also in terms of datasets, for the crop recommendation task, more data from different regions should be collected to improve the accuracy further. For plant disease classification more classes should be covered in terms of diseases as well as the plants.

For the deployment of the project, as currently there is no efficient load balancing solution given to the deployed website. Efficient load balancing algorithms can be used to properly distribute the traffic over available servers. The frontend can be improved for better user experience using modern web frameworks such as ReactJS or VueJS.

6 CONCLUSION

This project solves the problem of agricultural industry by providing the solution to a major problem of harvesting. We studied 5 different algorithms for the task 1 and reached to a conclusion that Random Forest is the best suited for the selected dataset.Random Forest achieved an overall accuracy of 99.3%. For the task 2 a comparative study was shown between VGG16, ResNet50 and EfficientNetV2S. EfficientNetV2 outperformed VGG16 and ResNet50 by achieving an overall accuracy of 96.06%. ResNet50 performed better than VGG16 gaining an overall accuracy of 95.53%. These two models were then deployed on web to be accessed by people.

References

[1] Kulkarni, P., Karwande, A., Kolhe, T., Kamble, S., Joshi, A. and Wyawahare, M., 2021. Plant Disease Detection Using Image Processing and Machine Learning. arXiv preprint arXiv:2106.10698.

[2] G. Chauhan and A. Chaudhary, "Crop Recommendation System using Machine Learning Algorithms," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), 2021, pp. 109-112, doi: 10.1109/SMART52563.2021.9676210.

[3] Sharada Prasanna Mohanty, David Hughes, Marcel Salathe, 2016, Using Deep Learning for Image-Based Plant Disease Detection. arXiv preprint arXiv:2106.10698.

[4] Hassan SM, Maji AK, Jasiński M, Leonowicz Z, Jasińska E. Identification of Plant-Leaf Diseases Using CNN and Transfer-Learning Approach. Electronics. 2021; 10(12):1388