SOMAIYA VIDYAVIHAR UNIVERSITY

Mini Project Report

On

Crop Identification Using Machine Learning

Submitted in the partial fulfillment of the requirements

for the degree

Masters in Technology

by

Darshan Sunil Shirsat

Roll No: 16034424028

Guide

Ms. Sujata Pathak

Department of Information Technology (AI & DS)

K. J. Somaiya School of Engineering

Batch 2024 -2026

Somaiya Vidyavihar University

K. J. Somaiya School of Engineering

Certificate

This is to certify that the Mini project report entitled "Crop Identification Using Machine Learning" is bonafide record of the work done by Darshan Sunil Shirsat in the year 2024-25 under the guidance of Ms Sujata Pathak in partial fulfillment of requirement for the Masters in Technology degree in Artificial Intelligence and Data Science of Somaiya Vidyavihar University.

_____ _____

Guide / Co-Guide Head of the Department

_____

Principal

Date:

Place: Mumbai-77

Somaiya Vidyavihar University

K. J. Somaiya College of Engineering

Certificate of Approval of Examiners

This is to certify that the Mini project report entitled "Crop Identification Using Machine Learning" is bonafide record of the work done by Darshan Sunil Shirsat in partial fulfillment Ms Sujata Pathak of requirement for the Masters in Technology degree in Artificial Intelligence and Data Science of Somaiya Vidyavihar University.

_____ _____

Expert / External Examiner Internal Examiner / Guide

Date:

Place: Mumbai-77

Somaiya Vidyavihar University

K. J. Somaiya College of Engineering

DECLARATION

I declare that this written Mini Project report submission represents the work done based on my and / or others' ideas with adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity as per norms of the Somaiya Vidyavihar University. I have not misinterpreted or fabricated or falsified any idea/data/fact/source/original work/matter in my submission.

I understand that any violation of the above will be cause for disciplinary action by the university and may evoke the penal action from the sources which have not been properly cited or from whom proper permission is not sought.

_____

Signature of the Student

_____

Name of the Student

16034424028

Roll No.

Date:

Place: Mumbai-77

Abstract

This project presents a machine learning pipeline for automated crop classification based on image data. A balanced dataset comprising five different crop types was utilized. To effectively capture important visual characteristics, handcrafted feature extraction techniques such as **color histograms** were applied, alongside flattening of image data. Subsequently, **Principal Component Analysis (PCA)** was employed to reduce dimensionality while preserving critical variance.

A lot of literature survey is done on this topic which shows that a lot of work is being carried out in Agriculture Domain regarding crop recognition using Machine learning.

Multiple baseline classifiers—including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Decision Tree—were developed, with regularization and hyperparameter tuning techniques implemented to mitigate overfitting. Ensemble methods, including hard voting, bagging (using multiple models, Random Forest), boosting (AdaBoost, Gradient Boosting, XGBoost, were further incorporated to enhance overall model robustness and predictive accuracy. A unique architecture of ensembles was designed to combine the strengths of individual models. This project showcases the integration of feature extraction, dimensionality reduction, ensemble modeling for solving real-world agricultural classification challenges.

***Key words*:** Principal Component Analysis (PCA), color histograms (feature extraction), Ensemble Learning (Bagging, Boosting, Voting)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents an overview of the mini-project work titled "Crop Identification using Machine Learning." It explains the motivation behind the project, highlights the scope of the work undertaken, and sets the foundation for the following chapters.

1.1 Background And Motivation Agriculture plays a critical role in the economy and food security of a nation. Accurate and timely classification of crop types is essential for various applications such as yield prediction, land use analysis, and resource allocation. Traditional methods of crop monitoring are labor-intensive and error-prone, making them inefficient for large-scale implementation.The motivation behind this project arises from the need to enhance the efficiency and accuracy of crop classification using machine learning techniques. With the availability of large-scale image data and the advancement of computational methods, there is a significant opportunity to automate crop classification using AI-based approaches.

1.2 Problem Statement

Accurate identification of crop types is fundamental to effective agricultural management, influencing decisions related to crop monitoring, yield estimation, and resource allocation. Traditional methods, such as manual field surveys and visual inspections, are often labor-intensive, time-consuming, and susceptible to human error. These limitations can hinder timely decision-making, especially in regions with vast agricultural landscapes.The advent of machine learning (ML) offers a promising avenue to automate and enhance crop identification processesHowever, challenges persist, including the need for large, high-quality labeled datasets, handling data variability due to seasonal changes, and ensuring model generalizability across different geographic regions.

This project aims to address these challenges by developing a machine learning-based system for crop identification that utilizes dimensionality reduction techniques, such as Principal Component Analysis (PCA), to manage high-dimensional data. The goal is to create a robust, scalable, and efficient model capable of accurately classifying various crop

types, thereby supporting precision agriculture initiatives and contributing to sustainable farming practices.

1.3 Scope The scope of the Project This mini-project aims to develop a machine learning model that classifies crops based on input images. The model addresses challenges like class imbalance, high-dimensional data, and varying environmental conditions. The project also includes the deployment of the model through a web interface using Flask and containerization using Docker.

1.4 Objectives

- To collect and preprocess an image dataset of various crop types
- To reduce dimensionality using Principal Component Analysis (PCA)
- To build and compare multiple machine learning models
- To enhance accuracy using ensemble techniques
- To use feature extraction techniques

1.5 Organization of the Report

This report is organized into seven chapters, each describing a key aspect of the mini project titled *"Crop Identification Using Machine Learning"*:

Chapter 1 – Introduction:
Provides the background, motivation, scope, objectives, and an overview of the report's structure.

Chapter 2 – Literature Survey:
Reviews prior work related to crop classification using machine learning and remote sensing, including dimensionality reduction, ensemble learning, and deep learning approaches.

Chapter 3 – Dataset and Preprocessing:
Describes the dataset used in the study, its structure, and the preprocessing techniques applied to prepare the data for model training.

Chapter 4 – Proposed Methodology:

Details the feature extraction methods, dimensionality reduction using PCA, machine learning algorithms employed, and ensemble techniques used for classification.

Chapter 5 – Implementation:

Outlines the implementation workflow, including model training, hyperparameter tuning, evaluation strategies, and comparison of models with and without PCA.

Chapter 6 – Results and Discussion:

Presents the evaluation results, performance metrics (accuracy, precision, recall, F1-score), and discusses the impact of various methods and models on performance and computational time.

Chapter 7 – Conclusions and Future Work:

Summarizes the findings of the project, highlights the robustness and efficiency of the proposed system, and suggests directions for future enhancement using advanced feature extraction and deep learning techniques.

Chapter 2

Literature Review

This chapter presents a review of existing research and techniques related to crop classification using machine learning. It discusses various methodologies and identifies gaps that the current project aims to fill.

The advancement in remote sensing and machine learning (ML) has paved the way for highly accurate crop classification techniques. This chapter provides a comprehensive review of recent developments in the domain, focusing on remote sensing imagery, dimensionality reduction, class balancing strategies, machine learning models including ensemble methods, and deep learning techniques.

2.1 Remote Sensing and Crop Classification

Remote sensing imagery, particularly from multispectral and hyperspectral satellites like Sentinel-2, has proven indispensable for crop type identification. As discussed in [1], such imagery enables the extraction of critical vegetation indices like NDVI and EVI, which serve as robust predictors in classification tasks. Support Vector Machines (SVM), Random Forests (RF), and Gradient Boosting Machines (GBM) are among the most commonly employed ML classifiers owing to their strong performance in high-dimensional feature spaces. For instance, [2] and [3] highlight the synergistic use of optical and radar data, where radar backscatter complements the spectral information from optical bands, leading to higher classification accuracies.

2.2 Dimensionality Reduction Techniques

Given the high-dimensional nature of hyperspectral imagery, dimensionality reduction is essential. Principal Component Analysis (PCA) is one of the most widely adopted techniques in the field. Studies like [4] and [5] demonstrate that PCA not only enhances classifier performance but also significantly reduces training times. Alternative methods such as Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), and manifold learning techniques have also been explored. For instance, [6] compared multiple dimensionality reduction methods and found ICA to be effective in preserving non-Gaussian features relevant to crop discrimination.

2.4 Ensemble Learning Methods

Ensemble learning has emerged as a cornerstone technique in the domain of crop classification, offering a means to improve both accuracy and model robustness by aggregating predictions from multiple base learners. Rather than relying on a single model, ensemble methods combine the outputs of various classifiers to produce a more reliable and generalizable result. A notable trend in recent research is the development of hybrid ensemble models, which combine diverse algorithms such as Random Forest (RF) and k-Nearest Neighbors (k-NN). As demonstrated in [10], such combinations can lead to significant performance gains, particularly when dealing with complex and heterogeneous agricultural datasets.

In parallel, boosting-based ensemble methods, such as XGBoost and LightGBM, have gained traction due to their superior capability in modeling non-linear relationships and capturing subtle feature interactions. These methods work by sequentially training weak learners, with each subsequent model focusing on the errors of its predecessor. This iterative refinement leads to highly accurate and low-bias models that often outperform standalone classifiers. For instance, [11] highlights that boosting frameworks not only improved predictive accuracy but also enhanced generalization across diverse geographic regions, Furthermore, ensemble models often exhibit greater resilience to overfitting, especially when properly tuned and validated using techniques such as spatial or stratified cross-validation.

2.5 Deep Learning Approaches

Deep learning, particularly Convolutional Neural Networks (CNNs), is emerging as a transformative approach. In [12], a CNN trained on time-series NDVI and EVI indices achieved significant accuracy improvements over traditional ML models. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, as used in [13] and [14], are particularly suited for modeling temporal dynamics of crop growth over a season. However, these models are data-hungry and computationally intensive. Transformer-based models are gaining traction due to their ability to model long-range dependencies and contextual relationships. A Vision Transformer (ViT), explored in [15], outperformed standard CNNs on multi-temporal classification tasks. Similarly, attention mechanisms in [16] enabled the model to focus on the most informative spectral and temporal features,

enhancing interpretability and accuracy. Transfer learning has shown promise in addressing the issue of limited training data and generalization to unseen regions. As reported in [19], pretraining models on large-scale agricultural datasets followed by fine-tuning led to improved performance in new environments. Domain adaptation techniques further enhance cross-region model transferability.

2.10 Summary and Gaps

Recent literature on crop classification showcases a diverse and evolving set of methodologies that leverage advanced technologies to improve accuracy, scalability, and interpretability. One major advancement is the use of high-resolution satellite imagery, which provides detailed spatial and spectral information crucial for distinguishing between different crop types, growth stages, and land use patterns. The field has also seen a surge in the adoption of Transformers, which, originally developed for natural language processing, are now being adapted for image-based tasks. These models excel at capturing long-range dependencies and context in data, which is particularly useful in analyzing large-scale spatial information from satellite images.

To handle the high dimensionality and noise in remote sensing data, preprocessing techniques like Principal Component Analysis (PCA) are widely adopted. These techniques help reduce redundancy, enhance relevant features, and optimize computational efficiency, forming a strong foundation for downstream classification tasks.

A broad range of classification models have been explored. Traditional machine learning methods such as Support Vector Machines (SVM), Random Forests, and Gradient Boosting have demonstrated solid performance, especially when combined into ensemble methods that harness the strengths of multiple models. However, deep learning models, particularly Convolutional Neural Networks (CNNs), have emerged as powerful tools due to their ability to automatically extract hierarchical spatial features from raw imagery.

The field has also seen a surge in the adoption of Transformers, which, originally developed for natural language processing, are now being adapted for image-based tasks. These models excel at capturing long-range dependencies and context in data, which is particularly useful in analyzing large-scale spatial information from satellite images.

Chapter 3

Dataset and Preprocessing

This chapter gives you an overview of dataset used in this problem and the preprocessing techniques. It was a readily available Kaggle dataset of crop images. The preprocessing was done using OpenCV library.

3.1 Dataset Overview:

The dataset was Kaggle dataset consisting of images of the crop . It was a balanced dataset with each variety of crops equally distributed . The variety of crops were 5 in total viz, wheat ,rice ,maize, sugarcane, jute, Overall it consists of 1000 rgb images of crop of shape (224,224,3). The dataset was having a csv file in which all urls of images of crop and the and the corresponding name and numerical labels were given.

The folder structure of the dataset used is given below:

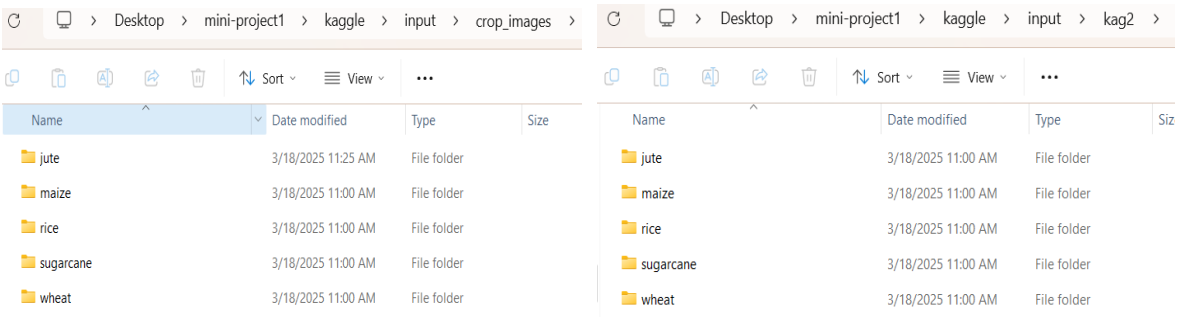

Figure 3.1 Folder structure

Actually there were two folders in which images of crop were stored (crop_images and kag2). Each of them were having each folder for each crop

3.2 Dataset Preprocessing:

The dataset was first splitted into train_test_val split. The training data consist of 60% of overall data. The test contribute to 20%of data and val contribute to 20% of data,thus making it to (60-20-20) spit of overall dataset. The images in every set were first fetched through a csv file. The images were flattened using OpenCV to convert it into 1d array so that it can be given to ML models. The shape of single image was (1,150,528). For overall

dataset it become (1000,150528).This was computationally very heavy in terms time and memory. So, before flattening, we resized it to (64,64,3) and then flattened but this was losing most information of data, thus affecting performance of model. Hence , these approaches were not the appropriate for training models.

We , then flattened our overall dataset, then applied PCA to reduce number of features while maintain variance. Features which were150,528 initially were dramatically reduced to 250 . This technique was found to be useful but was only focusing on variance of data. As, this was image dataset, focusing only on variance was not sufficient , Hence we deep dive into feature engineering techniques like color histograms, Haralick features, lbp, hog, etc. to extract meaningful features. Different features fusion combination was also studied during this stage. Color Histogram was found to be very helpful to distinguish between crops. After using the above technique the final shape of dataset was (1000,512) .It has more features compared to earlier ones but were considered as they significantly improve performance.

Chapter 4

Proposed Methodology

This chapter presents the proposed methodology, techniques, and algorithms used in developing the crop classification system. It provides details on PCA, feature extraction, ML algorithms and the ensemble learning techniques adopted.

4.1 Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space by identifying the principal components. It helps in reducing computational cost while retaining the significant features. Mathematically, PCA identifies eigenvectors of the covariance matrix of the data and projects the data onto these vectors as it is proven that eigenvector with maximum magnitude (eigen value) has the most variance along its direction.

4.2 Feature Extraction:

Various feature extraction techniques were used for further performance improvement The reason behind using these techniques is image dataset used in this problem . As for image dataset only PCA cannot help as it only explains variance. So, we have to extract unique features using computer vision techniques.

4.2.1 Local Binary Pattern (LBP) is a texture descriptor. It helps a machine "describe" the of an image by comparing texture each pixel to its neighbors.

4.2.2 Histogram of Oriented Gradients (HOG) is a feature descriptor — like LBP, but it captures shapes and edges instead of just texture. It tells the machine where the edges are and which direction they are pointing.

4.2.3 Color Histogram is a graph that shows how frequently each color appears in an image. It purely looks at distribution of color using bins.

4.2.4 Haralick features are a set of statistical features that describe the texture of an image. They are calculated from something called a Gray-Level Co-occurrence Matrix (GLCM).

4.2.5 Fourier Transform converts a signal (or image) from the time/space domain into frequency domain. It counts frequency of distinct pixels

4.3 Machine Learning Algorithms:

Various machine learning techniques were used to solve the problem. The idea behind this is to study different ML algorithms across our dataset. This was done to make our classifier robust and also to get intution so that it is useful further in framing ensemble strategy. It makes our strategy robust

4.2.1 Decision Tree: A non-liner tree based model using some programming logic(if-else) to capture pattern in data for making predictions.

4.2.2 Logistic Regression A linear model that predicts the probability of a class using the sigmoid function. It is useful for binary and multi-class classification with interpretable coefficients.

4.2.3 Support Vector Machine (SVM) SVMs are powerful classifiers that find the optimal hyperplane that maximizes the margin between classes. With kernels, SVM can also handle non-linear data.

4.2.4 K-Nearest Neighbors (KNN) KNN is a simple, instance-based learning algorithm that classifies a sample based on the majority class among its K-nearest neighbors in the feature space.

4.25 Naïve Bayes: A probabilistic model which classifies data based on conditional probabilities of class labels with features. Internally it uses Bayes Theorem assuming conditional independence.

4.6 Ensemble Techniques

Bagging: Combines predictions from multiple models trained on bootstrapped subsets to reduce variance. Helps in running models parallelly improving speed.

Boosting: Builds models sequentially, each correcting errors of the previous one. These are little slow as it follows a proper order.

Voting Classifier: Aggregates predictions from different models to make the final decision based on majority voting.

The ensemble strategy used was literally unique having unique architecture. It comprises of combining bagging and boosting algorithm. A voting ensemble (vt_clf1) was initially used to classify crops which was based purely based on machine learning model

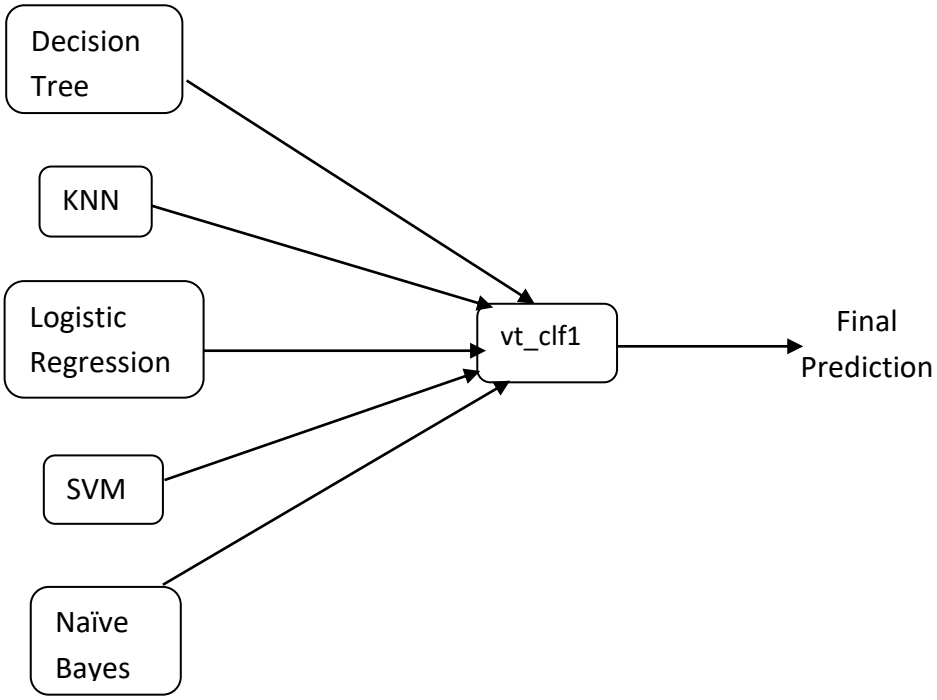

Fig 4.1 Voting Classifier1

Then finally a voting classifier of all ensembles(bagging(4)+boosting(3)) was implemented to make classifier more robust. Also weights were assigned to each base estimator in voting depending on its accuracy so that it prioritises the right base model while decision making the idea is to drop the conventional majority vote class method.

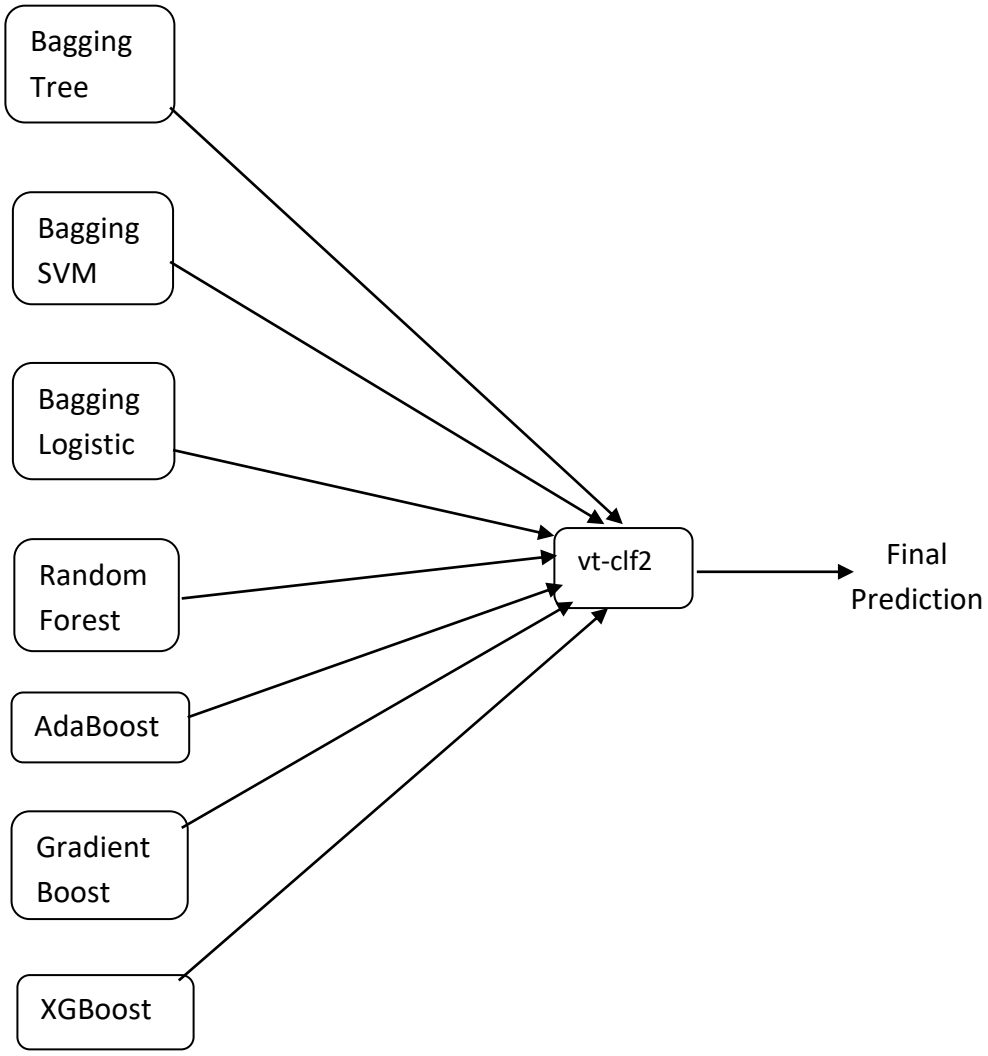

Fig 4.2 Voting Classifier2

Chapter 5

Implementation

This chapter presents the implementation of the crop identification system. It covers dataset preprocessing, model training, hyperparameter tuning, It show emphasis on classification of crop to its identification. It provide detail on entire ml pipeline.

5.1 Dataset Description

The data set was Kaggle dataset consisting of images of the crop . The dataset was having a csv file in which all urls of images of crop and the and the corresponding name and numerical labels were given. The labels of each crop in dataset was given below.

Table 4.1 Crop labels

Sugarcane	Rice	Wheat	Jowar	Maize
3	1	4	0	2

5.2 Preprocessing

As the dataset was a image dataset, ML algorithms don't accept such data. So we neec to convert these into tabular/2d numpy array form. So we flattened our images So each image which was having shape (224,224,3) initially was converted into (,150000) shape As we can see there are 1.5 million feature for each row so it is high dimensional dataset. So we need to reduce it so that our ML algorithm perform well on data.

There were two techniques to reduce dimensions resizing and or applying PCA. The problem with resizing was loss of pixels/information thus reducing performance. This technique is similar to compressing technique where pixels are hammered leading to low quality image while reducing dimensions. But PCA when directly applied to image transforms features into new features which are more useful than previous. These transformations are applied based on variance A threshold of 94-95% total variance is kept. Only those features were kept which explained threshold variance Rest all were discarded.

Hence features were reduced by significant amount . Initially the shape of image was (,150000). It was reduced to (,200). It made the model less computational in terms of time and resources.

Another approach was feature extraction using computer vision techniques such as LBP , HOG, color histograms (color-based), haralick features fourier transform(frequency-based), These technique were usually used on each of the ml model. But only three of them were giving good results. The three were then combined in order to improve accuracy, but that attempt also failed. Besides principal components analysis (PCA) were used individually and then combine d features to reduce noise in features but then too it was not that beneficial. So color histograms were selected as final features as the input to model. This also boosted our performance by high amount and also reduced training and test time. The details of which are given below in table.

5.3 Model Training

The idea is to use two dataset. One without applying PCA and another with PCA. So that we don't lose important features(pixels). To compare time, we have created a pair of every model as clf1 and clf2

clf1:(Without PCA)

clf2: (With PCA and extracted features)

List of algorithms used for both clf1 and clf2 classifier

- Decision Tree
- K-Nearest Neighbors
- Support Vector Machine
- Logistic Regression
- Naïve Bayes

All these algorithms were sample execution of the project .Just to get an intuition so as to take further effective steps. The main thing was also to study time comparision between overall data and reduced data.

Ensemble Strategy: This is where the main approach of the project lies.As said earlier bagging boosting and voting were primarily used Bagging was implemented using four bagging classifier. Besides two separate voting classifier were used. First one was on traditional ML models and second one was Ensemble models. Following are th ensembles used in voting classifier

- Bagging with Decision Tree
- Bagging with Support Vector Machine
- Bagging with Logistic Regression
- A special Bagging technique called Random Forest

Boosting was implemented using three boosting classifier

- Ada boost
- Gradient Boost
- Extreme Gradient Boost

5.4 Hyperparameter Tuning: A special care was taken of even individual models . If not handled can cause overfitting like fully grown Decision Tree, KNN with low k value high,

Parameters Chosen:

Decision Tree:	SVM:	KNN
max_depth=15	C=2.5	k=5

All the ensembles were tuned at their best parameters so that the performance is maximum. The hyperparameters used for tuning were different for bagging and boosting. They are as follows.

Bagging: Boosting:

n_estimators=50	n_estimators=100
max_samples=0.9	learning_rate=0.1
max_features=0.9	subsample=0.9
	max_features=0.9

5.5 Evaluation and Prediction: All the models were evaluated for their accuracy and other classification metrics for test train validation set. Also the time for traditional models were calculated (with and without PCA). Not only accuracy, but for final models, several other metrics like precision,recall,f1_score and confusion metric were calculated..The final model ie vt_clf2 was used for prediction. A sample dataset was initially created which was then used for making prediction. It was then processed similarly as the way test train was validation sets were processed. Then finally given to our voting ensemble.

Chapter 6

Results and Discussion

This chapter presents the results of the experiments carried out in project work. It displays values of accuracy and time of ML models used in the project. Other classification metrics are also given equal importance in evaluation to address class imbalance. Th

Table 6.1 Accuracy on test data.

Algorithm	Without PCA	With PCA
Decision Tree	0.83	0.88
KNN	0.78	0.79
SVM	0.9	0.92
Logistic	0.86	0.88
Naïve Bayes	0.56	0.58

Table 6.2 Accuracy based on color histograms

Algorithm	Training Accuracy	Test Accuracy	Validation Accuracy
Decision Tree	1	0.96	0.95
KNN	0.801	0.696	0.784
SVM	0.996	0.96	0.97
Logistic	0.96	0.83	0.84
Naïve Bayes	0.72	0.68	0.66

The SVM performed very well (0.96) when compared to other models performing individually with Naïve Bayes performing poor(0.68). This model cant handle high dimensional data very well.

Table 6.3 Ensemble Accuracy based on Color Histograms

Algorithm	Training Accuracy	Testing Accuracy	Validation Accuracy
Bagging Tree	1	0.98	0.952
Bagging SVM	0.93	0.87	0.86
Bagging Logistic	0.99	0.97	0.98
Random Forest	1	0.99	0.97
AdaBoost	0.74	0.67	0.62
Gradient Boost	1	0.99	0.98
XG Boost	1	0.99	0.97

Out of Ensembles, Random Forest, Gradient Boost and XG Boost performed outperformed all the other algorithms with accuracy of 0.99 on testing data. Adaboost was showing lowest accuracy on test data

Table 6.4 Final Model metrics based on color histogram

Phase	Algorithm	Accuracy	Precision	Recall	F1_score
Training	Voting Classifier1	0.99	0.99	0.99	0.99
	Voting Classifier2	1	1	1	1
Validation	Voting Classifier1	0.96	0.96	0.96	0.96
	Voting Classifier2	0.98	0.98	0.98	0.98
Testing	Voting Classifier1	0.97	0.97	0.97	0.97
	Voting Classifier2	0.98	0.98	0.98	0.98

From the above tables it is clearly visible that use of ensembles have great impact on accuracy along with other metrics. Feature engineering is found to be useful in enhancing models overall performance. Besides Voting Ensembe strategy was making classifier more robust and high in accuracy. Voting classifier 2 was better than than voting classifier 1 in terms of every classification metric. This implies our technique was very useful in classifying crops.

Table 6.5 Training Time

Algorithm	Without PCA	With PCA	Features Extracted
Decision Tree	116	0.23	0.115
KNN	0.01	0.01	0.002
Logistic	39	0.024	0.113

SVM	39	0.011	0.065
Naïve Bayes	2.64	0.01	0.005

Table 6.6 Testing Time

Algorithm	Without PCA	With PCA	Features Extracted
Decision Tree	0.2	0.002	0.002
KNN	6	0.36	0.024
Logistic	0.45	0.004	0.009
SVM	0.25	0.037	0.055
Naïve Bayes	3.72	0.006	0.003

As the number of features are significantly reduced, it had impacted the time required for training and testing. However, testing has been impacted less than training time for all models except KNN and NaiveBayes since it does nothing in training phase.

Chapter 7

Conclusions and Future work

This chapter presents the implementation of the crop identification system. It covers dataset preprocessing, model training, hyperparameter tuning, It show emphasis on classification of crop to its identification. It provide detail on entire ml pipeline.

7.1 Conclusions:

7.1.1 Robust Classifier

The individual models performed well but ensemble strategy was great fit for this data. They simply helped to reduced overfitting and make models more generalizable on unseen data. The voting ensembles on normal models has also showed great performance. But the performance of main voting classifier trained on ensemble models was the highest achieved of all. It was due to the amount of ensembles used and different strategy weighted based on the accuracy of every model.

7.1.1 Performance Enhancement

Besides this performance increased due to feature extraction techniques like PCA. It aimed at extracting important features explaining maximum variance of the data thus reducing noise, preforming dimensionality reduction. PCA reduced time by 70 times the time required to train a model with just flattening it. Besides PCA, several feature extraction techniques like HOG, LBP,etc .The color histogram technique was the best even better than PCA as it showed much higher accuracy on almost every model. The reason why color histograms performed well was the models where able to distinguish distinct datapoints(crops) based on colors.

7.2 Future Work

7.2.1 Feature Engineering:

There are many techniques explored in this context but color histogram finds a best fit. But there are several techniques yet to be explored like Gabor Filters and Canny Features.

Besides, the features which were not useful in our study might prove to be very useful on some other dataset. Advanced feature extraction like CNN can also give promising results.

7.2.2 Ensemble Methods:

Various boosting algorithms are still missing in this study e.g. CatBoost, LightBGM. Bagging classifiers with different machine learning algorithms can be examined . Stacking and Blending, one of the finest Ensemble models are yet to executed on data. Multilevel stacking with k-fold cross validation is way stronger approach.

7.2.3 Deep Learning:

As the problem we are dealing with is full of images, it is deep learning specific problem especially when it comes to large datasets. When dataset becomes very large i.e. more than 10000 images.ml models can suffer even when properly feature engineered. There can be cases where this can be handled when engineered very carefully, but still Dep learning has its own benefit.

References

[1] A. Sharma et al., "Crop type classification using Sentinel-2 and machine learning," *Remote Sensing of Environment*, vol. 229, pp. 111-123, 2020.

[2] J. Wang et al., "Fusion of optical and radar data for crop mapping using ensemble learning," *IEEE Access*, vol. 9, pp. 78921-78933, 2021.

[3] S. Zhao and Y. Liu, "Improved classification accuracy by combining multi-source satellite data," *Sensors*, vol. 20, no. 3, pp. 600-612, 2020.

[4] B. Singh and R. Kumar, "Dimensionality reduction of hyperspectral data using PCA for crop classification," *International Journal of Remote Sensing*, vol. 41, no. 12, pp. 4556-4574, 2020.

[5] T. N. Tran et al., "Feature reduction techniques in agricultural remote sensing," *Computers and Electronics in Agriculture*, vol. 178, 105736, 2020.

[6] M. Becker et al., "Comparative study of dimensionality reduction methods in hyperspectral analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 88-97, 2020.

[7] K. Patel et al., "Addressing class imbalance with SMOTE in crop classification," *Computers and Electronics in Agriculture*, vol. 175, 105603, 2020.

[8] F. Lin et al., "ADASYN-based oversampling for imbalanced crop data classification," *Agricultural Systems*, vol. 185, 102937, 2020.

[9] J. R. Gomez and D. Perez, "Bagging ensemble with stratified sampling for imbalanced datasets," *Remote Sensing Letters*, vol. 11, no. 9, pp. 832-841, 2020.

[10] M. Liu and X. Wang, "Hybrid ensemble learning for accurate crop type classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 23-34, 2020.

[11] A. Desai et al., "Boosted decision trees for spatially generalized crop classification," *IEEE GRSL*, vol. 17, no. 8, pp. 1390-1394, 2020.

[12] Y. Chen et al., "Deep learning-based crop classification using NDVI time series," *IEEE JSTARS*, vol. 13, pp. 1234-1245, 2020.

[13] Z. Yang et al., "Crop classification with LSTM networks using time-series Sentinel data," *Remote Sensing*, vol. 12, no. 3, 456, 2020.

[14] M. Suresh et al., "Temporal feature modeling for crop mapping using RNNs," *Computers and Electronics in Agriculture*, vol. 174, 105465, 2020.

[15] D. Kwon et al., "Vision Transformer for crop classification from remote sensing imagery," *IEEE Access*, vol. 10, pp. 53290-53300, 2022.

[16] P. Singh et al., "Self-attention models for satellite-based agricultural monitoring," *Remote Sensing of Environment*, vol. 263, 112543, 2021.

[17] D. Roy et al., "Evaluating classification performance in multi-crop systems," *Agricultural Systems*, vol. 180, 102763, 2020.

[18] P. Zhang and S. Li, "Spatial cross-validation for agricultural remote sensing," *Remote Sensing*, vol. 12, no. 4, pp. 765-778, 2020.

[19] H. Javed et al., "Transfer learning for seasonal crop type mapping," *ISPRS International Journal of Geo-Information*, vol. 9, no. 11, 645, 2020.

[20] R. Mehta and L. Banerjee, "Edge AI for real-time crop classification in resource-limited settings," *IEEE IoT Journal*, vol. 7, no. 6, pp. 5434-5445, 2020.