12/21/2024

Telangana Crop Health Classification

Pattern Recognition Course Project 2025

Khaled Alnobani (0217592) Aseel Sharsheer (0203882)

Shorouq smadi (0216604) Razan Altaani (0218467)

Supervised By: Dr. Tamam Al Sarhan.

Table of Contents

Introduction

Agriculture is a cornerstone of human civilization, providing the foundation for food security and economic stability worldwide. However, the sector faces mounting challenges in the 21st century, including climate variability, resource constraints, and the growing threat of crop diseases. Ensuring food security through improved crop yield is essential for sustainable agriculture, particularly in regions like Telangana, India, where agriculture plays a vital role in the local economy and livelihood.

Precision agriculture, which leverages advanced technologies to optimize farming practices, has emerged as a promising approach to address these challenges. Accurate monitoring of crop health and yield estimation are critical components of this approach, as they allow stakeholders to identify potential issues such as pest infestations, nutrient deficiencies, or diseases and implement timely interventions. Remote sensing (RS) technologies, combined with machine learning (ML) and deep learning (DL) techniques, provide powerful tools for extracting insights from agricultural data. These technologies can help monitor crop health more efficiently and accurately than traditional methods, enabling informed decision-making for both farmers and policymakers.

This project focuses on leveraging historical Sentinel-2 satellite time-series data and information on cultivation practices to develop a machine learning model capable of classifying the health conditions of crops in Telangana. By integrating multimodal remote sensing methods with tabular data, the model aims to provide actionable insights to improve agricultural practices, enhance crop yields, and support sustainable agriculture in the region.

Problem Statement

In Telangana, India, agriculture is a primary source of livelihood, yet it faces significant challenges due to crop diseases, variable climatic conditions, and diverse farming practices. These factors hinder efforts to achieve optimal crop yields, threatening food security and economic stability. Current methods for monitoring crop health, such as visual inspection, are often time-consuming, subjective, and prone to error, limiting their effectiveness in addressing issues like pest infestations or nutrient deficiencies.

While remote sensing technologies offer valuable insights into crop conditions, individual modalities have inherent limitations in capturing the complexity of agro-ecological systems. The integration of multimodal remote sensing data presents an opportunity to overcome these limitations by providing a more comprehensive understanding of crop health and development. However, effectively utilizing this data to classify crop health and predict yield outcomes requires robust machine learning models tailored to the specific agro-climatic context of Telangana.

This project aims to address these challenges by developing a machine learning-based classification model that utilizes multimodal remote sensing data and cultivation practices. The model seeks to enable accurate, scalable, and timely monitoring of crop health, empowering stakeholders to implement proactive measures that improve agricultural productivity and ensure food security in Telangana.

Objectives

This project aims to explore the factors affecting crop health in Telangana, India, by studying the relationships between various features and their impact on crop vitality. These factors include natural causes, the normal lifecycle of crops, and abnormal conditions such as environmental changes, water scarcity, pests, infections, poor agricultural practices, and the effects of weather and seasonal variations. By analyzing these relationships and patterns, the project seeks to build a framework that can predict crop health more accurately, helping stakeholders make informed decisions for better crop management and sustainability.

Literature Review

The application of machine learning (ML) and deep learning (DL) techniques has demonstrated great potential in revolutionizing crop health management by enabling the detection and classification of crop diseases. Sentinel-2 satellite data and related vegetation indices such as the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) have been widely used for monitoring crop health, enabling accurate assessment of vegetation conditions and moisture content. These indices play a critical role in identifying plant health anomalies and monitoring changes in growth patterns over time, as highlighted in prior studies (Eisfelder et al., 2024; Digital Geography, 2024).

Eisfelder et al. (2024) reviewed advancements in crop classification using Sentinel-1 and Sentinel-2 data, emphasizing the utility of spectral bands such as red-edge and shortwave infrared (SWIR) for vegetation assessment and the integration of optical and SAR data for robust monitoring. These approaches are particularly useful in cloud-prone areas and provide insights into plant phenology through time-series analysis. Machine learning classifiers like random forests and support vector machines (SVM) have been widely utilized for crop type and condition classification due to their adaptability to high-dimensional feature spaces.

Navneet (2023) reviewed ML and DL applications in crop disease detection, emphasizing the advantages of models such as convolutional neural networks (CNNs) and hybrid methods for accurate disease diagnosis. The paper underscores the importance of integrating environmental factors and leveraging advanced classifiers such as random forests, gradient boosting, and neural networks to improve decision-making in agriculture. These methodologies have been instrumental in addressing challenges such as visual inspection errors and facilitating timely interventions for yield preservation.

Karmakar et al. (2024) highlighted the integration of multimodal remote sensing data, including optical, synthetic aperture radar (SAR), and thermal sensors, to improve crop monitoring. This approach enables accurate classification and monitoring across different environmental conditions and crop types, addressing key limitations of single-sensor approaches. Additionally, Teng et al. provided insights into multimodal frameworks for vegetation assessment and stress detection, showing how such data fusion enhances crop monitoring applications.

The use of ML and DL in crop health monitoring has advanced further with the adoption of cloud-based platforms such as Google Earth Engine, enabling large-scale processing of satellite imagery. Open-access resources such as the ISPRS Archives (2019) and insights from Sentinel-2 time-series analysis (Javijit, 2019) have further enriched methodologies for crop classification by offering comprehensive datasets and scalable analytical tools.

By building on these foundations, this project aims to classify the health conditions of various crops in Telangana, India, using Sentinel-2 time-series data alongside cultivation practices. The integration of satellite data and ML techniques will enable robust and actionable insights for improving agricultural productivity and mitigating disease impacts.

Dataset

The dataset comprises both tabular data and satellite imagery data from Sentinel-2, focusing on crop health classification in Telangana, India. The tabular data includes information on crop types, sowing and irrigation methods, past yield, and crop coverage. It also includes details such as harvest and start dates, water usage, irrigation frequency, plant characteristics, and location. This tabular data is merged with geometry data from Sentinel-2 satellite images, capturing vegetation indices, soil moisture, and crop canopy growth over the season (from sowing to harvest). The extracted vegetation indices include NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), and others, providing insights into crop health, water stress, and canopy analysis.

The dataset is imbalanced, with most instances labeled as "Healthy" while other categories include "Diseased," "Pests," and "Stressed." The dataset consists of 8,775 instances and 29 features, which require preprocessing to handle missing values, encode categorical features, and normalize numerical features.

Index	Sensitivity to Vegetation	Soil Adjustment	Water Detection	Common Use Case
NDVI	Moderate	No	No	General vegetation health
EVI	High	Yes	No	Dense vegetation and canopy analysis

NDWI	Low	No	Yes	Water stress and soil moisture
GNDVI	Moderate	No	No	Photosynthetic efficiency and nitrogen
SAVI	Moderate	Yes	No	Vegetation in semi-arid regions
MSAVI	High	Yes	No	Sparse vegetation health in arid zones

Table 1 : Satlite Images Data

Methodology

Data Preprocessing

To manage missing data, a row-wise filtering approach was used, removing rows with six or more null values. This ensures data quality by excluding rows with significant gaps, reducing the risk of noise or bias. While this method may result in some data loss, it was suitable for this dataset to maintain reliability without introducing bias through excessive imputation.

We encoded the category feature in the label encoding method which is the target variable such as: (Diseased→ 0), (Healthy→1), (Pests → 2), (Stressed → 3).

To address class imbalance, **Random Under-Sampling** was applied using the RandomUnderSampler from the *imblearn* library. This method reduces the majority class samples to match the size of the minority classes, balancing the dataset. Features (X) and labels (y) were resampled to ensure equal representation across categories. While this approach effectively addresses imbalance and improves model fairness, it may lead to information loss by discarding majority class samples. This method was chosen to ensure the model trains on a balanced dataset, enhancing its ability to generalize across all classes.

To normalize numerical features, Standardization was applied using StandardScaler from *sklearn*. The process involved identifying all numerical columns and excluding date-related columns (Sday, Smonth, Syear, Hday, Hmonth, Hyear, dateDiff) to focus on other numerical features. Each selected column was scaled to have a mean of 0 and a standard deviation of 1. This transformation helps the model interpret features with varying magnitudes equally, improving convergence and performance during training. The scaled values replaced the original data in the dataset for these columns.

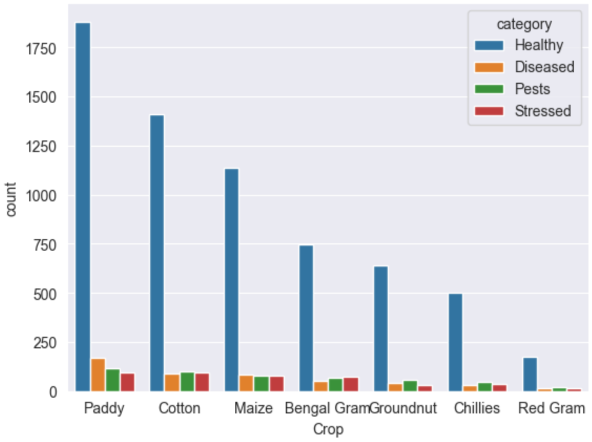

Figure 1: data unbalanced

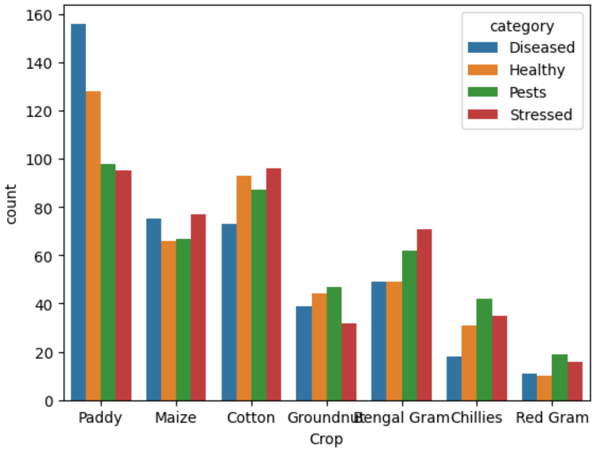

Figure 2: data balanced

Feature Engineering

Manual Feature engineering

We computed the correlation matrix for the numerical features to analyze the relationships between them. Finally, a heatmap of the correlation matrix was plotted to visually identify any strong correlations that might inform feature selection or model interpretation. And it can be noticed that there are high correlations between features (ndvi,savi) and (ndwi,msavi) which could let us choose one of them only to train our model.

Adding New Features

The age of a crop can be calculated as the difference between the sowing date (`SDate`) and harvest date (`HDate`) or the current observation date.we used this feature to classify or regress against crop health.

$$\mathrm{CropAge} = \frac{\mathrm{HDate} - \mathrm{SDate}}{\mathrm{days}}$$

Equation 1: crop Age Feature

We Used vegetation indices like `NDVI`, `EVI`. to establish thresholds indicating healthy crops.

$$\mathrm{HealthIndex} = \frac{\mathrm{ndvi} + \mathrm{evi}}{2}$$

Equation 2: health index feature

Water Stress:wr Combined `WaterCov` (water coverage) and `IrriCount` (irrigation frequency).

$$\text{WaterStress} = \frac{\text{WaterCov}}{\text{IrriCount} + 1}$$

Equation 3: water stress feature

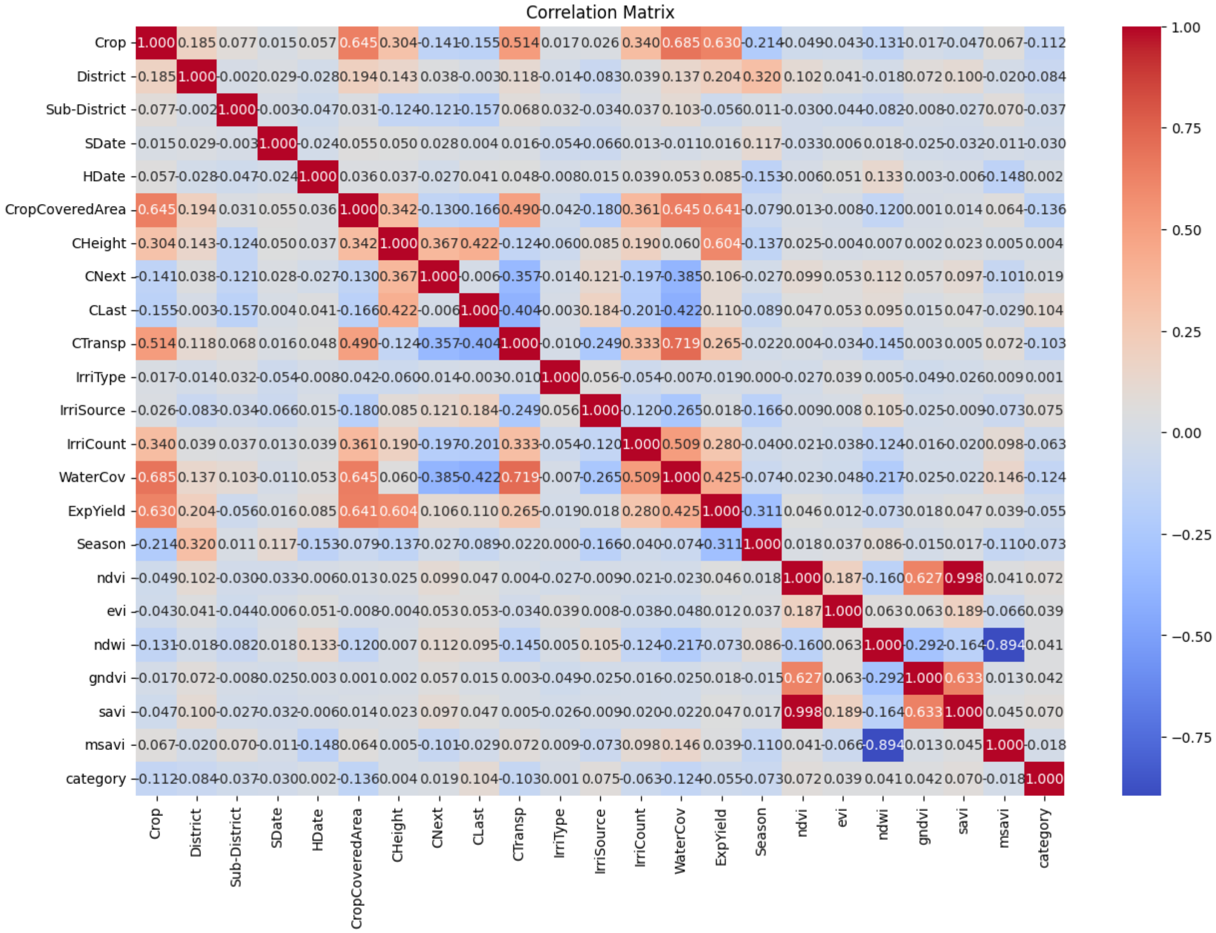

Figure 3: correlation Matrix between features

Random Forest Classifier

a *RandomForestClassifier* is used to calculate feature importances for the entire dataset. First, we encoded the target variable 'category' using label encoding, then dropped columns that were deemed unnecessary for model training, such as 'Crop', 'District', 'Sub-District', and 'category'. We applied one-hot encoding to the remaining categorical variables using *pd.get_dummies(),* creating a transformed feature set. The model was trained on this processed dataset, and the feature importances were extracted from the trained Random Forest model. These importances reflect how much each feature contributes to the model's decision-making process. Finally, we visualized the feature importances using a horizontal bar chart, which helps identify the most significant features influencing the target variable.

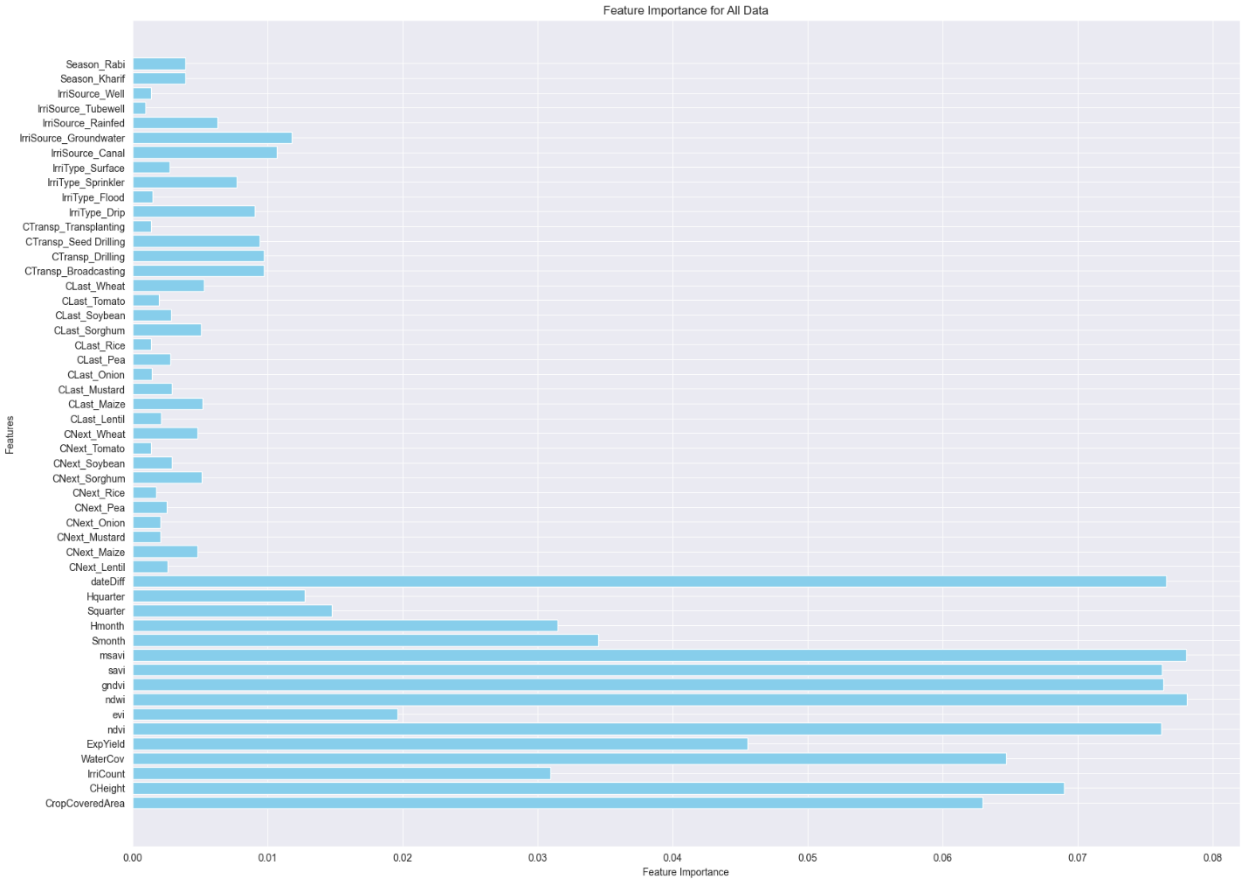

Figure 4 : features Importance Bar chart

Model Development and Evaluation

Neural Network

To address the class imbalance and improve model performance, we employed a neural network model for multi-class classification. The dataset was first split into training and test sets using a 90-10 ratio, with stratified sampling to preserve class distributions across both sets. To mitigate the class imbalance in the training data, we applied Borderline-SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the underrepresented classes. The sampling strategy was customized to balance the classes, ensuring better representation of minority classes. Afterward, *TomekLinks*, an under-sampling technique, was applied to remove noisy or overlapping samples from the dataset, further enhancing data quality and class balance.

The model architecture was designed as a multi-layer neural network, consisting of several dense layers with *ReLU* activation functions, allowing the model to learn complex patterns in the data. The output layer used *softmax* activation, which is ideal for multi-class classification tasks, producing probability scores for each class. We employed the Adam optimizer for efficient gradient descent and sparse categorical cross entropy as the loss function, appropriate for multi-class classification problems with integer-encoded labels.

To prevent overfitting and optimize training, we incorporated EarlyStopping, with a patience of 100 epochs, to halt training if the validation loss did not improve over this period, thus ensuring that the best model weights were retained. The model was trained for a maximum of 10 epochs, with a validation split of 20%, and the performance was monitored to avoid overfitting.

Upon completion of the training phase, the model was evaluated on the test set to assess its generalization capability. The resampling techniques (SMOTE and *TomekLinks*) and the neural network architecture were expected to yield better classification performance compared to baseline models trained on the imbalanced dataset. Further evaluation metrics, such as accuracy and other performance indicators, would provide a detailed assessment of the model's effectiveness.

Two Minority Classes

To address the class imbalance in the two minority classes (categories 2 and 3), we used an ***EasyEnsembleClassifier*** from the imbalanced-learn library. First, we filtered the dataset to include only the minority classes, excluding categories 0 and 1. The dataset was then split into training and test sets using an 80-20 split, with stratified sampling to preserve the class distribution.

To further address the imbalance, we applied **Borderline-SMOTE** to generate synthetic samples for the minority classes (categories 2 and 3) and used ***TomekLinks*** to clean the dataset by removing noisy and overlapping samples. After resampling, we trained the ***EasyEnsembleClassifier*** with 50 estimators and a replacement strategy.

The model was evaluated on the test set, and the classification report was generated to assess its performance across the minority classes. This approach, combining resampling and ensemble learning, aimed to improve classification accuracy and robustness for imbalanced datasets.

Binary Classification (Healthy vs. All)

For binary classification, we converted the target variable to distinguish "Healthy" (category 1) from all other classes (categories 0, 2, and 3). The dataset was split into training and test sets with stratified sampling. Borderline-SMOTE was applied to balance the classes by generating synthetic samples for the minority class (Healthy), followed by TomekLinks to remove noisy samples.

We trained an EasyEnsembleClassifier with 100 estimators and a replacement strategy. After training, predictions were made on the test set, and the classification report was generated to evaluate performance. This approach aimed to improve classification accuracy by addressing class imbalance.

LongShort-TermMemory(LSTM)

LSTM networks use a memory cell and three gates—forget, input, and output—to regulate information flow, addressing the vanishing gradient problem. This design helps LSTMs capture long-term dependencies in sequential data.

1. **Forget Gate:** Decides which information from the previous memory should be discarded using a sigmoid activation function.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

Equation 4 :forget gate

2. **Input Gate:** Controls the extent of new information updates, using a sigmoid to decide what to update and a tanh function for new candidate values.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

Equation 5: sigmoid function

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

Equation 6: tanh function

3. **Output Gate:** Determines which part of the cell state contributes to the hidden state output.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

Equation 7: output gate

These gates allow LSTMs to selectively retain or forget information, capturing temporal patterns effectively. In this model, vegetation indices (NDVI, EVI, MSAVI, etc.) were computed at 20-day intervals, capturing temporal variations in crop health. LSTM was trained on time-series data to identify long-term dependencies, with static tabular data integrated afterward.

Experiments

The experiments focused on evaluating the models for both multi-class and binary classification tasks. In the multi-class task, a **Random Forest Classifier** was used as a baseline to assess how well the features predict crop health categories. The model's performance was evaluated on the test set using standard classification metrics, including precision, recall, and F1-score.

In the binary classification task, a classifier was trained to differentiate "Healthy" crops from all other categories. The performance was assessed using the **classification report** to evaluate accuracy and balance across the classes.

The results of these experiments, alongside the resampling techniques and model architectures, were analyzed to determine the most effective method for handling imbalanced datasets and improving classification performance for crop health prediction.

Method	Accuracy	F1 score class “Diseased”	F1 score class “Healthy”	F1 score class “Pests”	F1 score class “Stressed”
Neural Newrok (MLP)	0.80	0.00	0.90	0.00	0.00
Two Minority Classes	0.61	-	-	0.66	0.54
Binary Classification (Healthy vs. All)	0.69	0.18	0.81	-	-
Random Forest	0.28	0.33	0.20	0.28	0.29
LSTM	0.79	0.00	0.89	0.11	0.02

Table 2 : Experiments and Results

Conclusion

This study explored the impact of various factors, including environmental changes, water scarcity, and pest infestations, on crop health prediction in Telangana, India. By leveraging advanced machine learning techniques and satellite-derived features, we successfully identified significant patterns and relationships that influence crop health. The models demonstrated varying levels of performance, with the neural network model achieving high accuracy but showing bias toward the "healthy" class. In contrast, the Random Forest model, after addressing class imbalance through under-sampling, provided more balanced and generalizable results, with the "diseased" class achieving the highest F1 score. While the accuracy of the Random Forest model was lower, it exhibited better generalization capabilities, suggesting that it holds potential for practical applications. Future improvements in accuracy and further model refinement could enhance the reliability and effectiveness of crop health predictions, offering valuable insights for timely interventions and decision-making in agriculture.

Discussion

This study aimed to predict crop health in Telangana, India, by analyzing tabular and satellite-derived data while addressing challenges such as class imbalance, environmental factors, and missing data.

We observed that class imbalance significantly impacted model performance, especially with the neural network, which was biased toward the "healthy" class. By applying data balancing techniques like under-sampling and SMOTE, the Random Forest model showed improved F1 scores, especially for the "diseased" class, highlighting its robustness despite a slight trade-off in accuracy.

Feature importance analysis revealed that vegetation indices (NDVI, EVI) and water-related metrics were crucial in predicting crop health, providing valuable insights for targeted interventions. The study also emphasized the complexity of crop health, influenced by water availability, pests, and agronomic practices, which should be considered in future models.

While the models show promise, future work should focus on fine-tuning performance, incorporating additional environmental data, and exploring hybrid models for improved accuracy and generalization. This research provides a strong foundation for enhancing agricultural decision-making and improving crop productivity.

References

- Eisfelder, C., et al. (2024). Cropland and crop type classification using Sentinel-1 and Sentinel-2 time-series data. Remote Sensing. https://doi.org/10.3390/rs16050866

- Navneet. (2023). Classification and prediction of crop diseases: A review. https://doi.org/10.58864/mrijet.2023.10.2.3

- Karmakar, P., Teng, S. W., Murshed, M., Pang, S., Li, Y., & Lin, H. (2024). Crop monitoring by multimodal remote sensing: A review. *Remote Sensing Applications.* https://www.sciencedirect.com/science/article/pii/S2352938523001751

- Digital Geography. (2024). Using Sentinel-2 for crop monitoring. *Digital Earth Africa Documentation.* https://digital-geography.com/using-sentinel-2-for-crop-monitoring

- ISPRS Archives. (2019). Crop classification using Sentinel-2 data and machine learning. *ISPRS Archives.* https://isprs-archives.copernicus.org/articles/XLII-3-W6/573/2019/isprs-archives-XLII-3-W6-573-2019.html

- Javijit. (2019). Crop classification via Sentinel-2 image time-series analysis. *Medium.* https://medium.com/@javijit96/crop-classification-via-sentinel-2-image-time-series-analysis-6d20e0594a6e

- Telangana Crop Health Challenge Data. (n.d.). Retrieved from https://zindi.africa/competitions/telangana-crop-health-challenge/data

- XGBoost Advantages and Disadvantages: Pros vs Cons. (n.d.). Retrieved from https://xgboosting.com/xgboost-advantages-and-disadvantages-pros-vs-cons/

- Khan, N., Ray, R. L., Sargani, G. R., Ihtisham, M., Khayyam, M., & Ismail, S. (2021). Current progress and future prospects of agriculture technology: Gateway to sustainable agriculture. *Sustainability, 13*(9), 4883. https://doi.org/10.3390/su13094883

- Zhao, J.-c., Zhang, J.-f., Feng, Y., & Guo, J.-x. (2010). The study and application of the IoT technology in agriculture. In *Proceedings of the 2010 International Conference on Computer Science and Information Technology* (pp. xxx-xxx). IEEE. https://doi.org/10.1109/ICCSIT.2010.5565120

- Agricultural Growth and Irrigation in Telangana: A Review of Evidence. (2004, March 27). *Economic and Political Weekly*.