Capstone Project Concept Note and Implementation Plan

Project Title: Crop Yield Prediction for Farmers

Team Members:

1. Natan Amanuel Mamo

Concept Note

1. Project Overview

As it is known, Crop yield directly determines the ability to ensure food security. Hence, Predicting crop yield is critical for optimizing resource usage, supporting farmers and making better decisions to have a better food security. This project aims to develop a hybrid machine learning (ML) and deep learning (DL) model to predict crop yields accurately by integrating environmental data, agricultural practices and satellite imagery. The solution addresses Sustainable Development Goal (SDG) 2: Zero Hunger by optimizing resource allocation and improving food security, and SDG 13: Climate Action by helping farmers adapt to climate variability and optimize agricultural productivity.

Problem Statement:

Smallholder farmers struggle with unpredictable yields due to climate change, soil degradation, and lack of data-driven decision-making tools. In addition, traditional methods are labor-intensive. The potential Impact of our model could be:

- Empower farmers with better insights to maximize yields.
- Reduce resource waste.
- Contribute to global food security efforts.

2. Objectives

1. Develop a hybrid Machine learning and Deep learning model to predict crop yields.

2. Integrate multi-source data (satellite imagery, weather, soil data) for accurate predictions.

3. Validate the model using real-world datasets from Ethiopia and other regions.

4. Build a prototype tool to assist farmers in making better planting and resource decisions.

5. Contribute to improving food security and climate adaptation strategy.

6. Design a scalable solution accessible to farmers in low-bandwidth regions.

3. Background

Random assessment of crop yields uses both human surveys supported by basic statistical models yet these methods produce low precision combined with high cost and extended timescales. Current research utilizing ML and DL approaches demonstrates successful yield estimation capabilities by processing satellite data alongside environmental variables through predictive modeling systems.

The applications face obstacles regarding data quality difficulties together with interpretability challenges and limited suitability for smallholder farmers. A hybrid model utilizing Random Forest/XGBoost among CNN-LSTM architectures forms the foundation of our project to overcome current data limitations and achieve optimal precision across different datasets.

4. Methodology

- Data: Satellite imagery (Sentinel-2), weather data (NASA POWER), soil data (SoilGrids), and historical yields (FAO).

- ML Techniques: For structured data

  - XGBoost
  - Random Forest

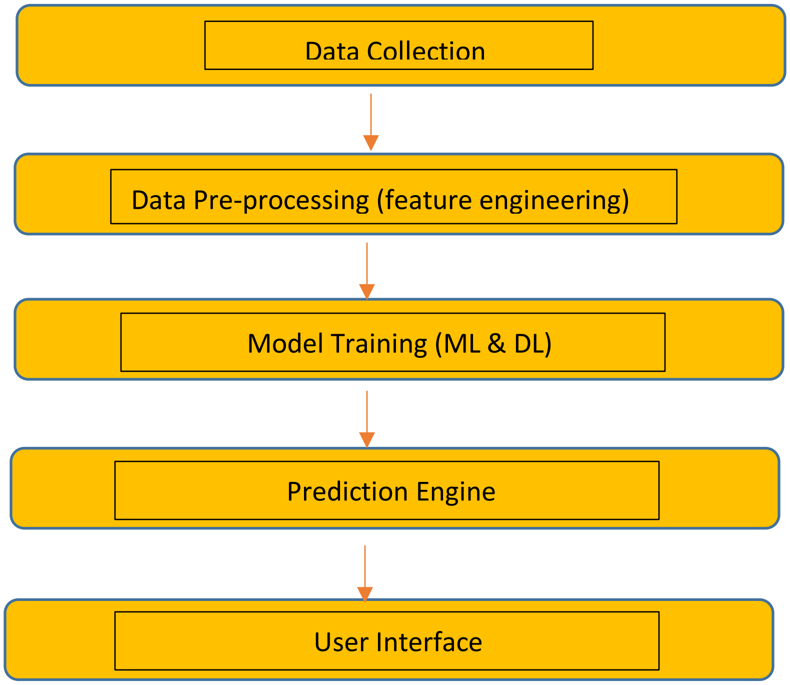
- Deep Learning:

  - CNN: Process satellite imagery.
  - LSTM: Model time-series weather patterns.

- Hybrid Approach: Combine ML (feature importance) and DL (spatial-temporal patterns).

- Frameworks: TensorFlow (for Deep learning models), Scikit Learn (for Machine Learning model).

5. Architecture Design Diagram



Components:

1. Data Collection: Collect data from diverse sources.

2. Feature Engineering: Normalize, impute, and encode features.

3. Model Training: Hybrid XGBoost-CNN-LSTM pipeline.

4. Prediction Engine: Generate yield forecasts.

5. Usable user interface

6. Data Sources

Data Type	Source	Format
Historical Yields	FAO Crop Production Data	CSV
Weather	NASA POWER	CSV
Soil Properties	SoilGrids	CSV
Satellite Imagery	Google Earth Engine	Images

7. Literature Review

Studies have demonstrated that hybrid ML-DL models (e.g., CNN + XGBoost) outperform standalone models in crop yield prediction tasks. Existing works focus heavily on environmental and satellite data integration, and our project builds upon these techniques by tailoring the solution to smallholder farmers in developing regions.

Implementation Plan

1. Technology Stack

- **Programming Languages**: Python

- **Libraries**: TensorFlow, PyTorch, Scikit-learn, XGBoost

- **Tools**: Google Earth Engine, GDAL

- **Visualization**: Matplotlib, Seaborn

- **Deployment**: Flask/Django

2. Milestones

1. Data collection and preprocessing.

2. ML model

3. DL model

4 hybrid model integration

5. Model achieve 85% accuracy.

6. Functional Prototype.

3. Challenges and Mitigations

- **Data Scarcity**

- **Model Bias**: Validate model across diverse regions.

4. References

- Mekecha & Gorbatov (2024), Khaki et al. (2020), FAO (2021).