Literature Review

Machine Learning-Based Wheat Yield Prediction (የስንዴ ምርት ትንበያ)Using Environmental Factors

GROUP 15

Name **Github Username**

1. Lielina Akele—-------------- @Lielina19

2. Martha Egigu—------------- @marthaoo

3. Melat Kebede—------------ @MelatKebedeAbraham

4. Fikremariam Yalew—----- @Fikiremariyam

5. Sintayehu Mandefro—--- @Sintayehu-M

Submission date:April 10 ,2025

1.1 Introduction

Wheat is a staple crop that significantly contributes to food security and economic stability in Ethiopia. With the country's dependence on agriculture and increasing climate variability, ensuring optimal wheat yield is essential. However, traditional farming practices and limited access to timely agronomic information have posed significant challenges. In this context, yield prediction using machine learning and environmental data becomes crucial to support informed decision-making by farmers and policymakers.

A review of existing literature is necessary to understand the current state of knowledge, identify effective methodologies, and discover gaps this capstone project can fill. It provides a foundation to compare traditional and modern techniques, analyze the use of various datasets, and assess the feasibility of adopting data-driven tools in the Ethiopian context.

1.2 Organization of Literature

This literature review is organized **thematically**, covering traditional yield prediction models, the rise of machine learning in agriculture, the role of environmental factors, data challenges specific to Ethiopia, explainable ML, scalability, and usability.

1.3 Traditional Approaches to Yield Prediction

Historically, yield prediction has been conducted using statistical and process-based crop simulation models such as the **Decision Support System for Agrotechnology Transfer (DSSAT)** and **Agricultural Production Systems sIMulator (APSIM)**. These tools simulate crop growth by modeling plant–environment interactions (Jones et al., 2003). Although effective in controlled environments, they require detailed field-level data, limiting their applicability in regions like Ethiopia where such data is sparse and fragmented.

1.4 Machine Learning in Yield Prediction

Machine learning offers powerful alternatives to traditional methods by uncovering complex nonlinear relationships in agricultural datasets. **Jeong et al. (2016)** applied Random Forests to predict crop yields in the U.S. Midwest and found significantly improved accuracy over multiple linear regression. Similarly, **Khaki and Wang (2019)** developed deep learning models to predict corn yield using satellite and weather data, outperforming traditional techniques.

A variety of ML algorithms—such as Random Forest, Gradient Boosting, Support Vector Machines, and Deep Neural Networks—have been adopted to improve prediction accuracy and generalizability (You et al., 2017).

1.5 Role of Environmental and Remote Sensing Data

Environmental parameters like **temperature, precipitation, solar radiation, cloud cover**, and **soil moisture** are widely acknowledged to influence crop yield (Lobell & Burke, 2010). The **Normalized**

Difference Vegetation Index (NDVI), derived from satellite imagery, is a key indicator of plant health and biomass, making it a valuable input in yield modeling.

Remote sensing data from platforms like **NASA POWER**, **FAO GeoNetwork**, and **Sentinel-2** are increasingly used due to their accessibility and spatial coverage (Brown et al., 2020). These datasets enable large-scale analysis, particularly in regions lacking consistent ground-truth data.

1.6 Data Limitations in Ethiopia

Local agricultural data in Ethiopia is often fragmented, unstructured, or unavailable. This creates barriers to high-resolution, localized modeling. To address this, **data fusion**—combining satellite, meteorological, and limited field data—has become a standard approach (Tadesse et al., 2021). Open-source satellite data helps fill gaps, while historical government datasets, though limited, provide useful baselines.

1.7 Explainable AI and Trust Building

Interpretability in ML is crucial for adoption. Techniques such as **SHAP (SHapley Additive exPlanations)** enable users to visualize feature importance and understand which variables influence predictions (Lundberg & Lee, 2017). This transparency builds trust among end-users like farmers, researchers, and policymakers who may otherwise hesitate to rely on "black box" systems.

1.8 Scalability and Multi-Crop Modeling

Several recent studies have advocated for **multi-target regression** and crop-specific modeling to extend prediction tools to multiple crops. For instance, **Kamilaris and Prenafeta-Boldú (2018)** emphasized the potential of using the same frameworks across crops with only minimal adjustments to input features. Such adaptability is especially valuable for national-level agricultural planning.

1.9 Usability, Localization, and Inclusion

User-centric design is essential for practical implementation. **Zhang et al. (2020)** highlight the need for interfaces that support local languages, mobile devices, and visual outputs tailored for non-expert users. In Ethiopia, where digital literacy and infrastructure vary, such design principles are vital to ensure adoption by smallholder farmers.

1.10 Summary and Contribution

To summarize, existing research confirms the efficacy of ML-based systems in predicting crop yields using environmental and remote sensing data. However, challenges remain in data availability, model explainability, and user accessibility—particularly in the Ethiopian context.

Our project addresses these gaps by:

- Integrating local and satellite data to improve prediction accuracy,
- Using explainable ML (e.g., SHAP) to increase transparency,
- Providing a web-based application with support for local languages.

This contributes to the body of knowledge by demonstrating a scalable, interpretable, and inclusive approach to crop yield prediction tailored for Ethiopia. It supports national agricultural resilience and aligns with **Sustainable Development Goal 2: Zero Hunger**.

References

- Brown, M. E., et al. (2020). The use of remote sensing in crop yield prediction. *Remote Sensing in Agriculture*, 12(2), 224.
- Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, 11(6), e0156571.
- Jones, J. W., Hoogenboom, G., Porter, C. H., et al. (2003). DSSAT Cropping System Model. *European Journal of Agronomy*, 18(3-4), 235–265.
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90.
- Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, 621.
- Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11), 1443–1452.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Tadesse, M., Ayalew, Z., & Degu, A. (2021). Integrating satellite and ground-based data for crop yield estimation in Ethiopia. *Journal of Agriculture and Food Research*, 3, 100120.
- You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. *AAAI Conference on Artificial Intelligence*.
- Zhang, D., et al. (2020). Designing agriculture decision support systems for low-resource users. *Information Processing in Agriculture*, 7(2), 230–240.