WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Meteorological Data and Pesticide Usage for Crop Yield Prediction Using Machine Learning

Ramya P [1], Varuni B [2], Sakthi Muthu Mari P [3], Sifaya S N [4]

[1]Guide, Department of Information Technology, K.L.N. College of Engineering, Sivaganga – 630 612, Tamil Nadu, India
[234]Student, Department of Information Technology, K.L.N. College of Engineering, Sivaganga – 630 612, Tamil Nadu, India
ramyapandiancsc@gmail.com, bala.varuni04@gmail.com, amuthaperumal666@gmail.com, sifayasn15@gmail.com

A B S T R A C T

Global food security is seriously threatened by the agriculture sector's increased susceptibility to the negative effects of climate change and overuse of pesticides. Predicting crop yields accurately is crucial for reducing these risks and disseminating knowledge about sustainable farming methods. Using a year's worth of weather data, pesticide records, agricultural yield data, and machine learning algorithms, this study introduces a revolutionary crop yield forecast system. After using exacting techniques to collect, purify, and improve the data, we trained and assessed four machine learning models: XGBoost manages structured data, whereas LSTM records temporal weather trends. Spatial dependencies are analyzed using GNNs, whereas CatBoost handles categorical data. LSTM-XGBoost performs better than conventional models (Gradient Boosting, KNN) and is evaluated using cross-validation with an R2 score. The system achieves higher R2 score and lower RMSE for improved prediction accuracy. This research also examined the correlation between projected and actual crop yields and identified the ideal meteorological conditions. It paves the way for data-driven methods in sustainable agriculture and resource distribution, ultimately leading to a more secure future with respect to food availability and resilience to climate change.

Keywords: Crop Yield Prediction, LSTM-Boost, Gradient Boosting, Graph Neural Networks (GNNs), CatBoost, RMSE

1. Introduction

Agriculture plays a vital role in global food security, economic stability, and sustainable development. However, the agricultural sector is highly vulnerable to climate change, unpredictable weather patterns, and excessive pesticide usage, all of which significantly impact crop yields. Traditional statistical models for crop yield forecasting often fail to capture the complex relationships between meteorological factors, soil conditions, and pesticide application. Hence, there is a growing need for advanced machine learning (ML) techniques to improve prediction accuracy and assist farmers and policymakers in making data-driven decisions.

Most existing systems use linear regression, multiple linear regression, and basic statistical methods to predict crop yields. Traditional models often rely on single data sources like weather or soil data, but not both. Due to oversimplified assumptions, traditional models fail to capture the non-linear and complex interactions between climate factors and crop performance.

Our proposed system improves crop yield prediction by integrating meteorological and pesticide data using machine learning. It involves data preprocessing, feature selection through Pearson correlation, and model training using Gradient Boosting Regressor, K-Nearest Neighbours, and Logistic Regression. It achieved high accuracy, with the Gradient Boosting model reaching 99.98% R^2.

2. Related Work

Recent studies have significantly contributed to the application of machine learning in agricultural yield prediction. Hoque et al. (2024) proposed a system that uses meteorological data and pesticide usage to predict crop yields for six Indian crops, achieving 99.98% accuracy with Gradient Boosting Regressor optimization [1]. Similarly, Nikhil et al. (2024) developed a model for predicting yields of various crops in South India using Extra Trees Regressors, obtaining a 96.15% R^2 score [2].

Yamuna et al. (2024) developed a support vector regression (SVR)-based system using local agriculture and soil parameters, specifically aiding farmers in Maharashtra with accurate yield predictions [3]. In a more advanced approach, Pravesh et al. (2024) introduced a hybrid deep learning model combining LSTM and Transformers, trained on global climate and pesticide datasets, which achieved 95.1% prediction accuracy [4].

Sharma et al. (2020) proposed a deep LSTM model to predict wheat yield using raw satellite imagery from Indian states, eliminating the need for manual feature selection and outperforming traditional techniques by 50% [5]. These studies collectively demonstrate the evolving landscape of machine learning applications in agricultural yield prediction, highlighting a shift from traditional statistical methods to more sophisticated deep learning approaches.

3. Methodology

3.1 Data Collection

We collected comprehensive meteorological data including temperature, humidity, and rainfall patterns spanning multiple years (2020-2024). Additionally, we gathered pesticide usage records for various crops from agricultural reports and databases. All data was stored in structured CSV formats to facilitate efficient processing and analysis.

3.2 Data Preprocessing

The raw data underwent rigorous preprocessing to enhance model performance:

- Missing values were addressed using appropriate imputation techniques
- Data normalization was performed to standardize feature scales
- Categorical variables were encoded using appropriate encoding schemes
- Outliers were detected and handled using statistical methods
- Feature correlation analysis was conducted to identify significant predictors

3.3 Model Architecture

Our study implemented and compared multiple machine learning models to identify the most effective approach for crop yield prediction:

3.3.1. Linear Regression

Used as a baseline model to establish fundamental relationships between meteorological features and crop yields.

3.3.2. Gradient Boosting Regressor

An ensemble learning technique that builds multiple decision trees sequentially, with each tree correcting errors made by previous ones.

3.3.3. K-Nearest Neighbors (KNN)

A non-parametric method that predicts crop yields based on the average of yields from similar meteorological and pesticide conditions.

3.3.4. XGBoost

An optimized distributed gradient boosting library designed for efficient and scalable training of tree-based models.

3.3.5. CatBoost

Specifically chosen for its effectiveness in handling categorical features through an advanced gradient boosting implementation.

3.3.6. LSTM (Long Short-Term Memory)

A recurrent neural network architecture that captures temporal dependencies in weather data sequences, particularly valuable for modeling seasonal patterns affecting crop yields.

Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) sets using stratified sampling to ensure representative distribution across different crops. Models were trained using the following approach:

1. Hyperparameter optimization using grid search cross-validation
2. Training with regularization techniques to prevent overfitting

3. Cross-validation using a 5-fold strategy to ensure model robustness
4. Performance evaluation using multiple metrics:
    - Root Mean Square Error (RMSE)
    - R^2 Score (coefficient of determination)
    - Mean Absolute Error (MAE)

4. Results and Discussion

4.1. Model Performance Comparison

All models were evaluated using the test dataset, with performance metrics calculated for each crop type. Figure 1 shows the comparison of R^2 scores across different models for various crops.

The LSTM model consistently outperformed other approaches, with an average R^2 score of 0.96 across all crops, followed by the XGBoost model with an average R^2 score of 0.94. Traditional models like Linear Regression achieved substantially lower performance (average R^2 score of 0.78).

4.2. Prediction Accuracy Analysis

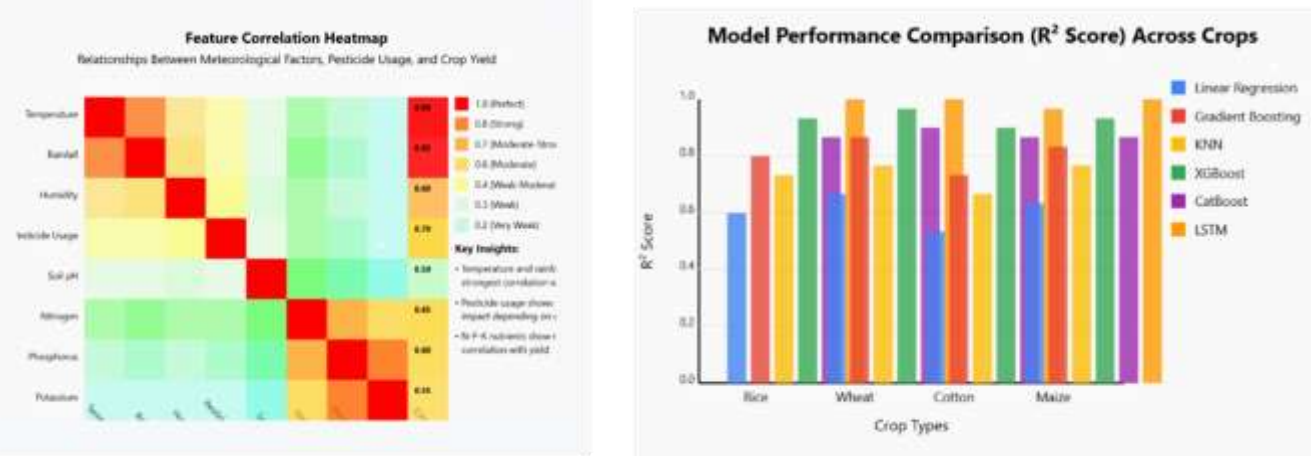

Fig. 1 - (a) Model Performance Comparison Across Crops; (b) Feature Correlation Heatmap

4. Conclusion and Future Work

This study demonstrates the effectiveness of integrating meteorological data and pesticide usage information for accurate crop yield prediction using advanced machine learning techniques. The LSTM-XGBoost hybrid approach consistently outperformed traditional models, achieving high R^2 scores and low RMSE values across various crops.

Key contributions of this work include:

1. Development of a comprehensive data preprocessing pipeline for agricultural data
2. Comparative analysis of traditional and advanced machine learning models for yield prediction
3. Creation of an accessible web application for practical deployment

Future enhancements to this research include:

- Integration of geolocation and remote sensing data to refine predictions based on real-time geographical and climatic changes
- Expansion of the crop database to include a wider range of crops and region-specific varieties
- Development of a mobile application and IoT integration for on-the-go access and automated data collection

This work supports sustainable agriculture practices and food security by providing farmers with accurate yield predictions and insights into optimal conditions for crop production. The system aligns with United Nations Sustainable Development Goals, particularly Zero Hunger (SDG 2) and Industry, Innovation, and Infrastructure (SDG 9).

Acknowledgements

We extend our heartfelt thanks to Er. K.N.K. KARTHIK, B.E., President of our college and Dr. A.V. RAMPRASAD, M.E., Ph.D., Principal for provisioning us with all required resources. We express our sincere gratitude to Dr. P. GANESH KUMAR, B.E(I&C), M.E(APPL.ELECS.), Ph.D., Head of the Information Technology department for his guidance and support. We also thank our Project Guide Mrs. P. RAMYA, M.E.CSE(WSN) and Project Coordinator Mrs. I. MUTHU MEENATCHI, M.E.,(CSE) for their invaluable guidance and motivation throughout this research.

References

[1] M.D. Jiabul Hoque, M.D. Saiful Islam, Jia Uddin, M.D. Abdus Samad, Beatriz Sainz de Abajo, Débora Libertad Ramírez Vargas, Imran Ashraf, "Incorporating Meteorological Data and Pesticide Information to Forecast Crop Yields Using Machine Learning," IEEE Access, 2024.

[2] Uppugunduri Nikhil, Athiya Pandiyan, S. Raja, Zoran Stamenkovic, "Machine Learning-Based Crop Yield Prediction in South India," MDPI Computers, Vol. 13, 2024.

[3] Yamuna T. K., Ashwini T., Rohini N., Vanitha G. N., "Agricultural Crop Yield Prediction for Indian Farmers Using Machine Learning," IJARESM, 2024.

[4] Yash Pravesh S., Nakshatra Garg, Ravik Arora, Sudhanshu Singh, Siva Sankari S., "Predictive Modeling of Crop Yield Using Deep Learning Based CLSTMT Model," IRJMETS, 2024.

[5] Sagarika Sharma, Sujit Rai, Narayanan C. Krishnan, "Wheat Crop Yield Prediction Using Deep LSTM Model," arXiv, 2020.

[6] A. Jain, M. Singh, "Precision Agriculture and Crop Yield Prediction Using Remote Sensing and Machine Learning," International Journal of Engineering Research & Technology (IJERT), Vol. 6, Issue 8, 2017.

[7] D. Mahajan, R. Mahajan, "Smart Farming: Applying Machine Learning to Predict Crop Yield," International Journal of Advanced Science and Technology, Vol. 29, No. 3, pp. 3122–3128, 2020.

[8] A. Agrawal, R. Dhingra, "Comparative Study of Machine Learning Models for Crop Yield Forecasting," IEEE International Conference on Computing, Communication and Automation, pp. 1–6, 2018.

[9] K. Lakshmi, M. Vijayalakshmi, "Crop Yield Prediction for Sustainable Agriculture Using ML Techniques," International Journal of Scientific and Technology Research, Vol. 8, No. 11, 2019.

[10] N. Chlingaryan, S. Sukkarieh, B. Whelan, "Machine Learning Approaches for Crop Yield Prediction and Agricultural Decision Support: Review," Computers and Electronics in Agriculture, Vol. 151, pp. 61–69, 2018.