Predicting Crop Yields with Machine Learning: A Comparative Study Using Environmental and Soil-Based Datasets

1st Gregor Pfister
Technical University of Applied Sciences Augsburg
Matrikelnr.: 2209779
gregor.pfister1@tha.de

***Abstract*—Crop yield prediction is a critical task in modern agriculture, with implications for food security, resource optimization, and sustainable farming practices. This study evaluates the predictive power of three distinct datasets: *Agriculture Crop Yield*, a synthetic dataset focused on environmental variables; *Crop Yield Prediction Dataset*, a real-world dataset emphasizing meteorological factors; and *Crop Yield Prediction*, which combines soil properties with environmental data. Using automated machine learning (AutoML) techniques through the *AutoGluon* framework, the datasets were analyzed for feature importance and prediction accuracy. Key findings revealed that rainfall was a dominant predictor in the synthetic dataset, while crop type and region were most influential in the real-world dataset. Soil properties such as nitrogen and potassium played a significant role in the soil-focused dataset. The study highlights the strengths and limitations of these datasets, emphasizing the need for diverse data sources to improve model robustness and applicability. Future work should explore the integration of complementary datasets, advanced modeling techniques, and temporal data to further enhance prediction capabilities.**

***Index Terms*—Crop yield prediction, machine learning, automated machine learning, synthetic data, environmental features, soil properties, AutoGluon, precision agriculture.**

I. Introduction

Predicting crop yields is a fundamental challenge in agricultural science and plays a critical role in ensuring food security and optimizing resource usage. As global populations rise and climate change impacts become more pronounced, accurate crop yield prediction models have become increasingly important for decision-making in agricultural practices. From informing planting schedules and irrigation planning to guiding policy development and market forecasting, yield prediction has far-reaching implications.

Modern advancements in data science and machine learning have introduced new opportunities to enhance traditional crop prediction methods. By leveraging environmental, soil, and management data, machine learning models can uncover complex patterns and interactions that were previously difficult to identify. However, challenges remain in achieving robust and generalizable predictions due to data variability, noise, and gaps in measuring critical features.

This paper investigates whether crop yields can be accurately predicted using datasets that focus on environmental features and soil properties. Through this work, we aim to:

1) Evaluate the predictive power of synthetic and real-world datasets to identify key yield determinants.
2) Highlight the relative importance of environmental (e.g., rainfall, temperature) and soil-based factors (e.g., nitrogen, potassium).
3) Explore the potential of machine learning, particularly ensemble models, to improve prediction accuracy and robustness.

The results of this study will provide insights into the utility of various data sources and modeling techniques for crop yield prediction, contributing to the development of data-driven, sustainable agricultural systems.

II. State of the Art

The application of Artificial Intelligence (AI) in agriculture has become a key driver of innovation, addressing challenges related to resource efficiency, crop monitoring, and yield prediction. Recent advancements have explored the integration of Internet of Things (IoT) devices with AI to create intelligent systems for crop management. IoT-enabled sensors and drones collect data on soil moisture, temperature, and crop health, which AI algorithms analyze to predict crop yields, detect anomalies, and optimize resource allocation [1]. Moreover, AI-powered image recognition systems identify early signs of pests and diseases, enabling timely interventions that reduce crop losses and the need for chemical treatments [1].

AI also enhances resource optimization by analyzing real-time and historical data to distribute water and fertilizers efficiently, ensuring sustainability and reducing environmental impact [1]. These technologies are central to modern precision farming practices, where AI-driven decision support systems offer personalized agronomic recommendations to improve yields [1].

Autonomous farming robots represent another significant advancement in precision agriculture. These robots integrate AI algorithms for decision-making and can autonomously navigate and perform farming activities such as weeding, spraying, and harvesting. Equipped with advanced sensors and actuators, these systems adapt to their environment, enabling both indoor and outdoor operations [2].

IoT-based systems also play a pivotal role in modern agriculture. Intelligent IoT systems integrate sensors for real-time data collection and machine learning models for decision-making. These systems enable precision irrigation and resource management at both farm and district levels. For instance, multi-agent IoT systems have demonstrated improved water-use efficiency in large-scale irrigation networks, showcasing their potential for sustainable farming practices [3].

Unmanned Aerial Vehicles (UAVs) combined with deep learning have emerged as transformative tools in precision farming. UAVs provide high-resolution imagery at a lower cost compared to traditional methods, while Convolutional Neural Networks (CNNs) enable accurate crop classification, segmentation, and yield estimation. This integration supports data-driven decision-making, enhancing productivity and sustainability in agriculture [6].

These advancements collectively illustrate the transformative potential of AI in agriculture, addressing critical issues such as food security, resource conservation, and sustainable farming practices. However, challenges such as data integration, accessibility for small-scale farmers, and ensuring the interpretability of AI models remain areas for further research.

III. Related Work

Recent advancements in artificial intelligence and machine learning have led to significant progress in crop yield prediction. This section discusses the current methodologies and their applications, focusing on deep learning techniques, remote sensing, IoT-based approaches, and relevant datasets.

A study leveraging MODIS data demonstrated the effectiveness of remote sensing for predicting crop yields, focusing on maize and soybean yields in the Central United States [4]. The researchers found that the Enhanced Vegetation Index (EVI2) outperformed the Normalized Difference Vegetation Index (NDVI) for yield prediction, particularly in regions with high crop leaf area. Additionally, the Normalized Difference Water Index (NDWI) was better suited for semi-arid regions due to its sensitivity to water stress and irrigation effects. Incorporating crop phenology metrics derived from MODIS significantly improved the accuracy of predictions, highlighting the importance of capturing temporal crop growth patterns.

Deep learning methods have emerged as state-of-the-art tools for agricultural tasks, including yield estimation. Convolutional Neural Networks (CNNs), a subset of deep learning, have been widely employed for their ability to process large-scale image data and extract complex patterns relevant to crop management. Semantic segmentation techniques based on CNNs allow for pixel-wise crop detection and yield estimation, while even improving on traditional challenges such as occlusion, even though other obstacles like variable lighting conditions still remain [5].

In addition to the above methods, several publicly available datasets and related projects have been explored to evaluate the potential of machine learning in crop yield prediction. The datasets used in this paper are:

- **Agriculture Crop Yield**, a comprehensive dataset containing crop yield data based on environmental conditions, has been analyzed for its predictive power. One example project explored crop productivity using exploratory data analysis and regression models to identify trends and assess feature importance [7], [8].
- **Crop Yield Prediction Dataset**, another publicly available dataset, focuses on the impact of environmental and meteorological factors on crop yield. This dataset has been utilized in various code projects to build prediction models [9], [10].
- **Crop yield Prediction**, which incorporates soil properties and weather data, has been used for experiments involving different ML approaches. A notable project compared the performance of various models on this dataset [11], [12].

These datasets form the foundation of the experiments conducted in this paper. Although there is already extensive research and existing solutions for crop yield prediction, further exploring these datasets and their potential remains valuable.

IV. Research Question

Crop yield prediction is a vital aspect of modern agriculture, with significant implications for food security, resource optimization, and sustainable farming. While advancements in artificial intelligence and machine learning have led to improved prediction models, challenges such as dataset variability, the inclusion of diverse environmental and soil factors, and the integration of synthetic and real-world data remain areas of active research.

In this study, we seek to address the following research question:

”To what extent can crop yields be accurately predicted using datasets that incorporate environmental and soil-based factors, and how do different machine learning models perform on synthetic versus real-world datasets?”

This research question guides the analysis of three datasets:

- **Agriculture Crop Yield**: A synthetic dataset focused on environmental variables and crop yields [7].
- **Crop Yield Prediction Dataset**: A real-world dataset emphasizing meteorological factors and crop productivity [9].
- **Crop yield Prediction**: A dataset combining soil properties and environmental conditions [11].

To explore this question, we:

- Conduct visual analyses to identify key patterns and trends within the datasets.
- Train and compare machine learning models using AutoML techniques.
- Evaluate the impact of dataset characteristics (e.g., synthetic vs. real-world) on prediction accuracy and feature importance.

The findings from this study aim to deepen our understanding of the factors influencing crop yield prediction and provide insights into the strengths and limitations of current machine learning approaches.

V. Concept

This study aims to investigate the potential of machine learning models to predict crop yields using three datasets that include environmental and soil-based factors. The process involves data exploration, visualization, and model evaluation, leveraging automated machine learning (AutoML) techniques to simplify and optimize the analysis.

To maintain clarity throughout the study, the datasets will be referred to as follows:

- **Dataset-1 (Agriculture Crop Yield)**: A synthetic dataset focused on environmental variables and crop yields [7].
- **Dataset-2 (Crop Yield Prediction Dataset)**: A real-world dataset emphasizing meteorological factors and crop productivity [9].
- **Dataset-3 (Crop Yield Prediction)**: A dataset combining soil properties and environmental conditions [11].

A. Data Analysis and Visualization

To gain insights into the datasets, the first step involves exploratory data analysis (EDA) and visualization:

- **Dataset-1**: This synthetic dataset will be analyzed to identify trends and correlations between environmental variables and crop yields.
- **Dataset-2**: Visualization techniques will be applied to explore the impact of meteorological factors on crop productivity.
- **Dataset-3**: Soil properties and environmental conditions will be assessed to uncover their influence on yields.

Key visualizations, such as scatter plots, heatmaps, and boxplots, will highlight feature relationships and the variability of crop yields across the datasets. These analyses aim to identify the most important factors contributing to yield prediction.

B. Machine Learning Models and AutoML

After understanding the datasets, machine learning models will be trained using the AutoML framework AutoGluon [13]. AutoGluon is selected for its ease of use and ability to automate key aspects of model selection and hyperparameter tuning, ensuring efficient and reproducible experimentation. The AutoML process will:

- Automatically select and optimize the best-performing models for each dataset.
- Evaluate models based on metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2.
- Compare the performance of individual models, such as Random Forest, Gradient Boosting, and Neural Networks, to ensemble methods.

Each dataset will be split into training and testing sets to validate model performance and prevent overfitting. Feature importance analysis will be conducted to assess which variables contribute most significantly to the prediction of crop yields.

C. Evaluation and Comparison

The evaluation will focus on:

- **Model Performance**: Comparing accuracy metrics across datasets to understand the differences between synthetic and real-world data.
- **Feature Importance**: Identifying the key predictors of crop yield in each dataset and comparing their consistency.
- **Dataset Characteristics**: Analyzing how the nature of each dataset (e.g., synthetic, environmental, or soil-focused) influences prediction accuracy and generalizability.

The results will provide a comprehensive evaluation of machine learning models for crop yield prediction and reveal insights into the strengths and limitations of each dataset. This will contribute to the broader understanding of how data characteristics impact model performance and agricultural decision-making.

VI. Realisation/Evaluation

A. Introduction

This section presents the practical implementation and evaluation of crop yield prediction using three datasets: **Dataset-1 (Agriculture Crop Yield)**, **Dataset-2 (Crop Yield Prediction Dataset)**, and **Dataset-3 (Crop Yield Prediction)**. The goal is to explore the potential of machine learning models in predicting crop yields and to evaluate the impact of different dataset characteristics on model performance.

The analysis is structured into three main parts for each dataset:

- **Exploratory Data Analysis (EDA):** Visualizations such as scatter plots, heatmaps, and distribution plots are used to identify trends, relationships, and key patterns within the data.
- **Machine Learning Model Training:** AutoML techniques, specifically the *AutoGluon* framework [13], are applied to train and evaluate machine learning models. Metrics such as Accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 are used to assess model performance.
- **Dataset Comparison:** Insights and results from each dataset are compared to evaluate the differences in prediction accuracy, feature importance, and the influence of dataset characteristics.

By analyzing these datasets in detail, this study seeks to uncover the key factors influencing crop yield prediction and assess the strengths and limitations of both synthetic and real-world datasets. The results aim to provide a deeper understanding of how data quality and feature composition affect machine learning models' ability to generalize and deliver actionable insights in agricultural contexts.

B. Dataset-1: Agriculture Crop Yield

1) Exploratory Data Analysis (EDA): The **Agriculture Crop Yield** dataset, referred to as **Dataset-1**, consists of synthetic data that includes environmental factors such as rainfall, temperature, soil type, and irrigation practices. Initial exploration revealed no missing values, allowing for a comprehensive analysis.

Key visualizations:

- **Feature Correlation Heatmap:** Figure 1 shows a heatmap of feature correlations. Rainfall (*Rainfall_mm*) exhibits the strongest correlation with crop yield (*Yield_tons_per_hectare*). Fertilizer usage and irrigation also demonstrate positive correlations with yield.
- **Scatter Plots:**
  - Figure 2 presents the relationship between rainfall and yield, showing a clear positive trend. Regions with higher rainfall generally achieve higher crop productivity.
  - Figure 3 shows the scatter plot of temperature vs. yield. While it seems like there is a slight positive correlation between higher temperature and yield this graph alone does not provide significant insights.
- **Pairplot:** Figure 4 displays the relationships between key features (*Rainfall_mm*, *Fertilizer_Used*, and *Irrigation_Used*), highlighting their combined effects on yield.

2) Machine Learning Model Training and Results: Machine learning models were trained using the *AutoGluon* framework [13], leveraging features such as *Rainfall_mm*, *Fertilizer_Used*, and *Irrigation_Used*. The performance metrics of the trained models are summarized in Table I.

TABLE I
PERFORMANCE METRICS OF AUTOGLUON MODELS ON DATASET-1.

Metric	Value
Root Mean Squared Error (RMSE)	0.501 tons/ha
Mean Squared Error (MSE)	0.251 tons^2/ha
Mean Absolute Error (MAE)	0.400 tons/ha
R^2 Score	0.913
Accuracy	91.41%

a) Interpretation of Results:

- Accuracy (91.41%): The model correctly predicted crop yields for 91.41% of the test cases. This high accuracy indicates that the model is effective in generalizing across the dataset.
- R^2 Score (0.913): The coefficient of determination, R^2, suggests that 91.3% of the variability in crop yield can be explained by the selected features. This demonstrates a strong relationship between the predictors and the target variable.
- Root Mean Squared Error (RMSE: 0.501 tons/ha): RMSE provides a measure of the average error in yield predictions. With an error of 0.501 tons/ha, the model achieves a high degree of precision, making it suitable for applications requiring detailed yield forecasting.
- Mean Absolute Error (MAE: 0.400 tons/ha): MAE measures the average magnitude of errors in predictions. The relatively low value of 0.400 tons/ha further confirms that the model's predictions are close to the actual values.
- Mean Squared Error (MSE: 0.251 tons^2/ha): MSE penalizes larger errors more heavily than smaller ones, ensuring that outliers are not ignored. The low MSE indicates that the model handles variability in the dataset effectively.

b) Key Insights: The model's strong performance metrics demonstrate that the selected features (*Rainfall_mm*, *Fertilizer_Used*, and *Irrigation_Used*) are effective predictors of crop yield in Dataset-1. High accuracy and R^2 values highlight the model's ability to generalize well, while the low RMSE and MAE indicate precise predictions. These results suggest that synthetic datasets like Dataset-1 can be valuable for understanding yield determinants, although real-world validation is necessary for broader applicability.

C. Dataset-2: Crop Yield Prediction Dataset

1) Exploratory Data Analysis (EDA): The **Crop Yield Prediction Dataset**, referred to as **Dataset-2**, contains real-world data emphasizing meteorological factors such as rainfall, temperature, and pesticide usage. This dataset provides insights into how these environmental factors influence crop productivity across various regions and crop types.

Key visualizations:

- **Overall Feature Correlation Heatmap:** Figure 5 shows the correlation matrix for selected features. No significant correlation between rainfall (*average_rain_fall_mm_per_year*), pesticide usage (*pesticides_tonnes*), or other features and yield (*hg/ha_yield*) can be observed from the heatmap alone, highlighting the complexity of interactions within the dataset.
- **Scatter Plots by Factors:**
  - Figure 6 illustrates the relationship between rainfall and crop yield. The plot suggests the existence of a sweet spot for rainfall, where an optimal amount leads to the best yields. Excessive rainfall appears to have a negative impact, potentially reducing crop productivity.
  - Figure 7 shows the relationship between temperature and yield. The data suggests that optimal temperatures for high yields vary by crop type. For instance, potatoes tend to achieve the highest yields at temperatures between 5 to 10°C, while soybeans grow best at around 25°C.
  - Figure 8 demonstrates the relationship between pesticide usage and crop yield. However, the dataset contains limited data for high pesticide usage, making it difficult to draw clear conclusions about its influence on crop yield.

2) Machine Learning Model Training and Results: Two models were trained using the *AutoGluon* framework [13]: one using all features and another with a subset excluding *Year* and *Unnamed: 0*. Performance metrics for both models are shown in Table II.

a) Interpretation of Results:

- Accuracy: The models achieved high accuracies of 92.14% (selected features) and 95.34% (full dataset), indicating strong predictive capabilities.

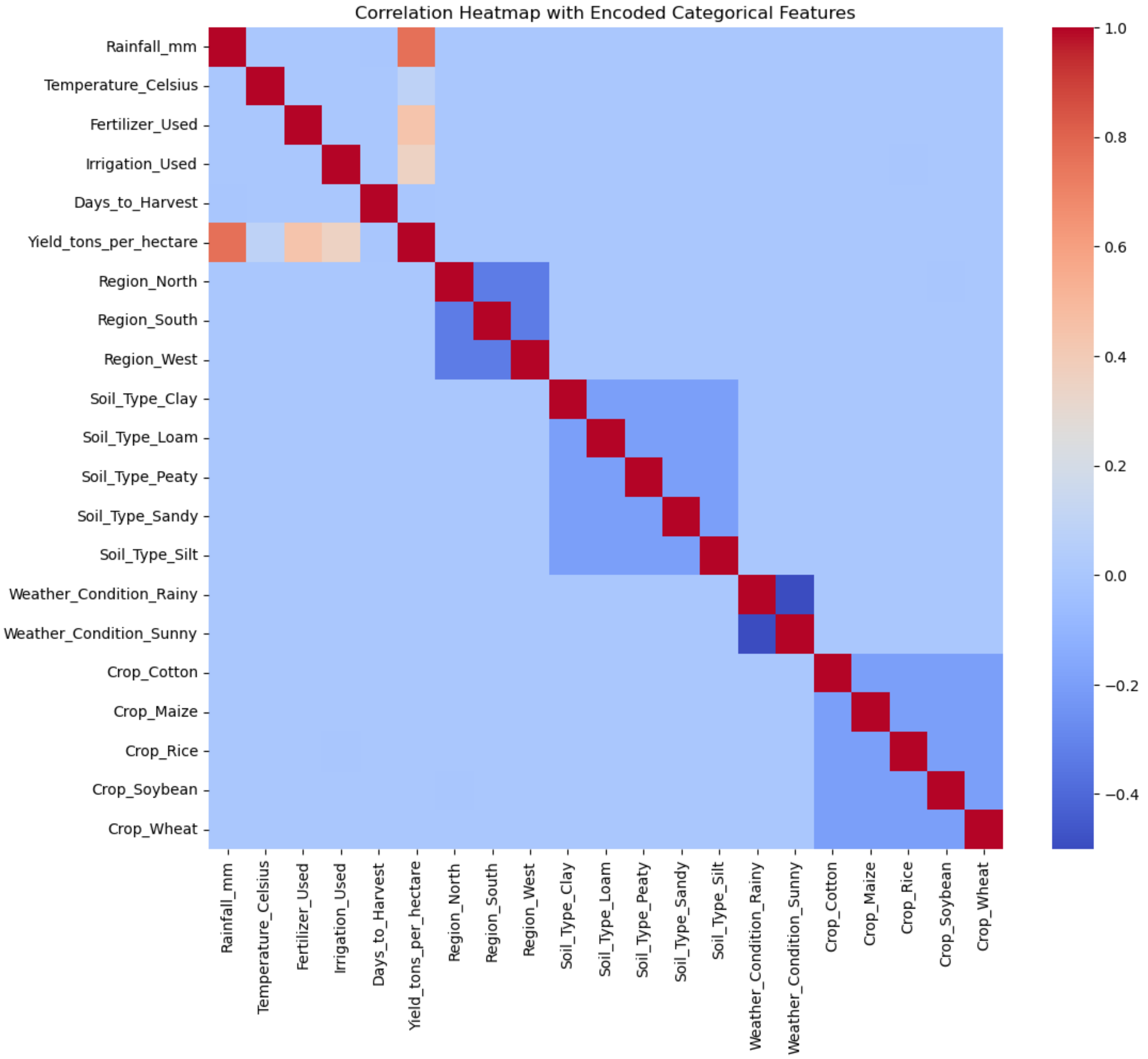

Fig. 1. Feature correlation heatmap for Dataset-1.

TABLE II
PERFORMANCE METRICS OF AUTOGLUON MODELS ON DATASET-2.

Metric	Selected Features Model	Full Dataset Model
MAE	6060.62	3591.36
MSE	1.82×10^8	6.95×10^7
RMS	13,505.67	8338.15
R^2 Score	0.975	0.990
Accuracy (%)	92.14	95.34

- R^2 Score: High R^2 values (0.975 and 0.990) suggest the models explain most of the variance in the yield predictions, with the full dataset model performing slightly better.
- Error Metrics: The lower MAE and RMSE of the full dataset model indicate it makes more precise predictions compared to the selected features model.

b) Feature Importance: The feature importance analysis (Figure 9) reveals the following insights:

- *Item* (crop type) is by far the most influential feature, with an importance value of 100,011, indicating that the type of crop being analyzed significantly affects yield predictions.
- *Area* (geographic region) is the second most critical feature, with an importance of 42,134, highlighting the influence of regional environmental and agricultural conditions on yield.
- *Year* also contributes to predictions, suggesting temporal trends or changes over time, such as shifts in climate or farming practices.
- Environmental factors such as *pesticides_tonnes*, *average_rain_fall_mm_per_year*, and *avg_temp* play moderately significant roles.

Figure 9 visualizes the relative importance of these features in predicting crop yield.

3) Insights: The following insights were derived from Dataset-2:

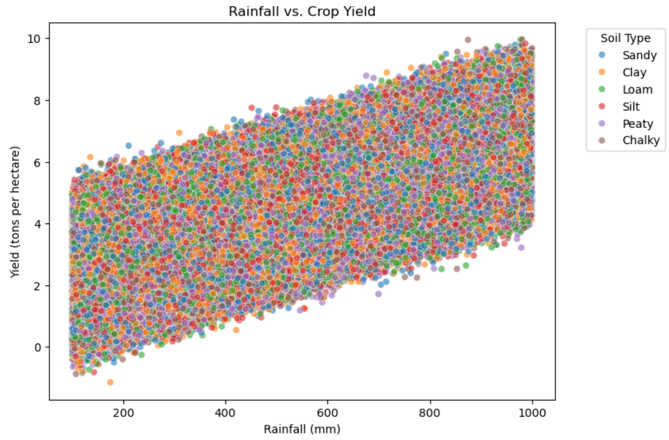

Fig. 2. Scatter plot of rainfall vs. yield in Dataset-1.

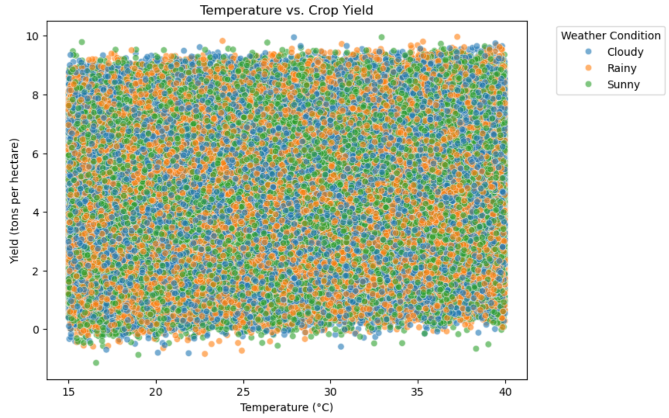

Fig. 3. Scatter plot of temperature vs. yield in Dataset-1.

- **Crop Type and Region are Key Drivers:** The analysis highlights *Item* (crop type) and *Area* (geographic region) as the most influential factors, suggesting that intrinsic crop characteristics and regional conditions dominate yield prediction.
- **Moderate Influence of Environmental Factors:** While environmental variables like *rainfall*, *temperature*, and *pesticides_tonnes* play a role, their importance is secondary to crop- and region-specific attributes.
- **Temporal Trends Matter:** The importance of *Year* suggests that temporal changes, such as climate variability or evolving agricultural practices, contribute to yield variability.

D. Dataset-3: Crop Yield Prediction

1) Exploratory Data Analysis (EDA): The **Crop Yield Prediction** dataset, referred to as **Dataset-3**, focuses on soil properties and environmental factors, such as rainfall, temperature, and fertilizer application, to predict crop yield. After data cleaning and preprocessing, 99 entries remained, ensuring a complete dataset for analysis.

Key visualizations:

- **Feature Correlation Heatmap:** Figure 10 illustrates the correlations between features. According to the heatmap, all features show strong positive correlations with crop yield (*Yield_Q_per_acre*) and with each other, except for *Temperature*, which exhibits a strong negative correlation with both crop yield and the other features.
- **Distribution of Yield:** Figure 11 shows the distribution of the target variable, *Yield_Q_per_acre*. The data is concentrated between 6 and 12 Q/acre, with peaks around 7 and 11 Q/acre.
- **Scatter Plots:**
  - Figure 12 illustrates the relationship between *Rainfall_mm* and yield, color-coded by fertilizer usage. The plot suggests that higher rainfall is associated with better yields. However, the items with the

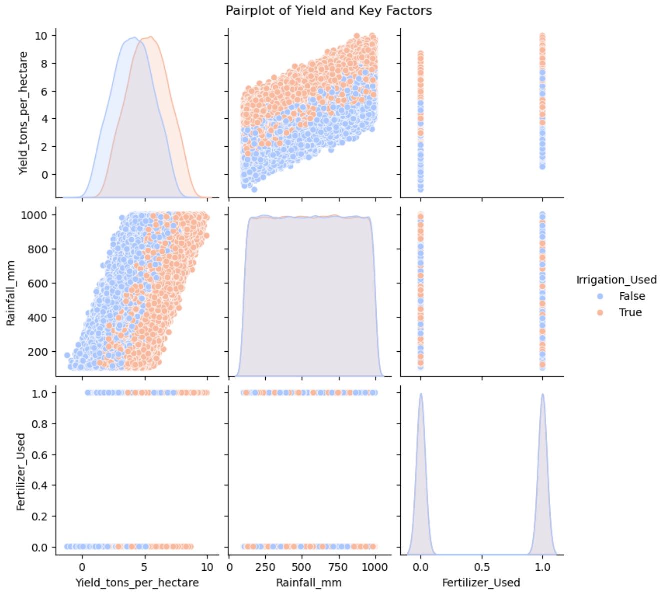

Fig. 4. Pairplot of key features in Dataset-1.

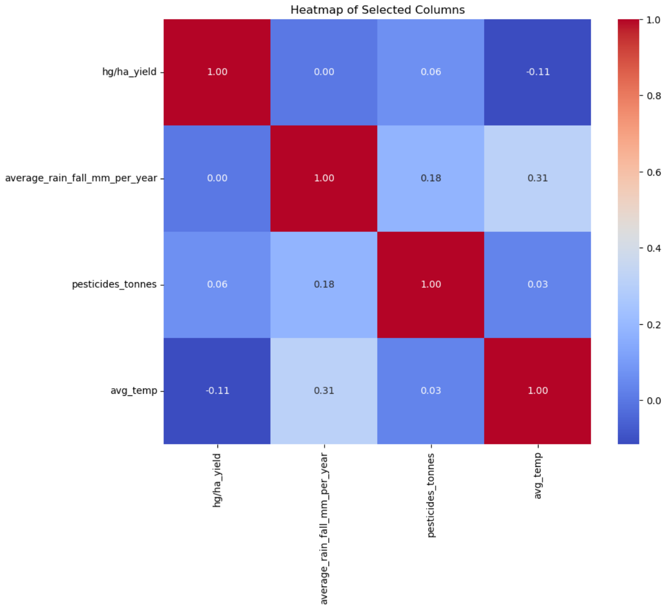

Fig. 5. Feature correlation heatmap for Dataset-2.

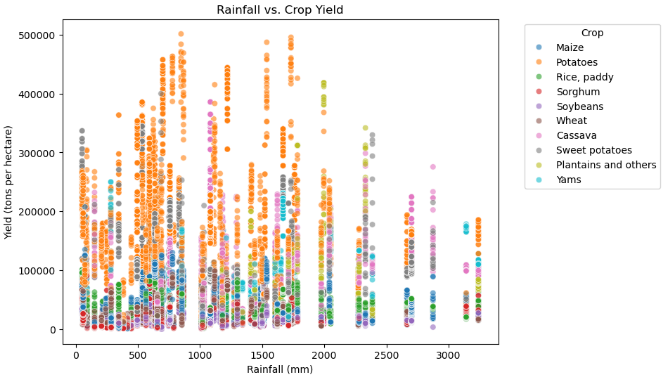

Fig. 6. Scatter plot of rainfall vs. yield in Dataset-2.

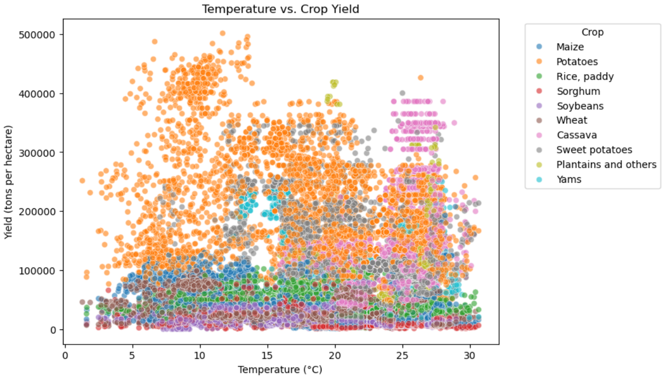

Fig. 7. Scatter plot of temperature vs. yield in Dataset-2.

highest rainfall also have the highest fertilizer usage, making it difficult to determine which factor is the primary driver of increased yield.

  - Figure 13 shows the impact of fertilizer usage on yield. While higher fertilizer levels generally correspond to increased yields, there is noticeable variability.

- **Boxplots and Line Graphs:** Figure 14 and Figure 15 summarize the average yield across different levels of *Rainfall_mm*, *Fertilizer*, *Nitrogen*, *Phosphorus*, and *Potassium*, showing distinct patterns for some features.

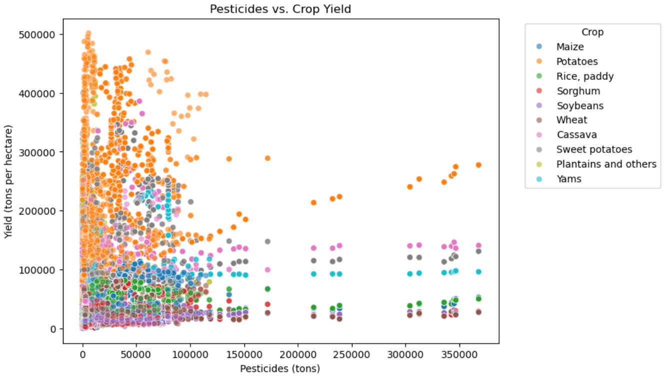

Fig. 8. Scatter plot of pesticide usage vs. yield in Dataset-2.

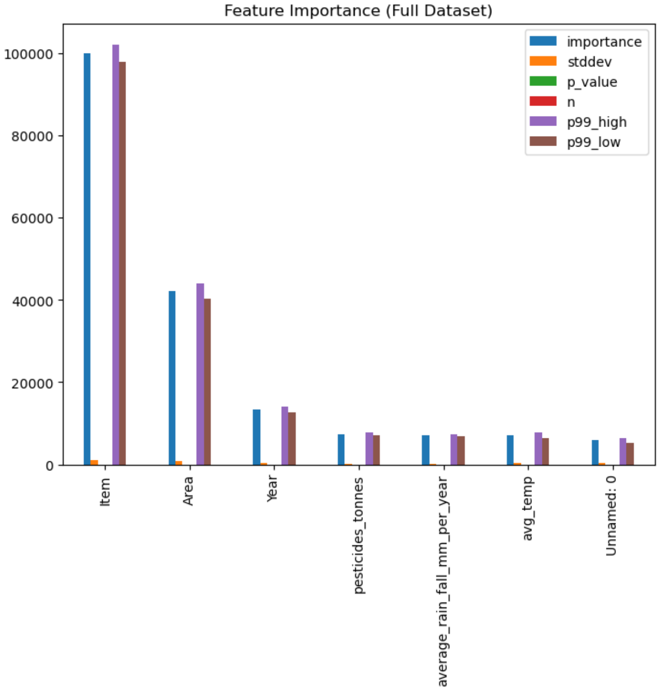

Fig. 9. Feature importance for Dataset-2.

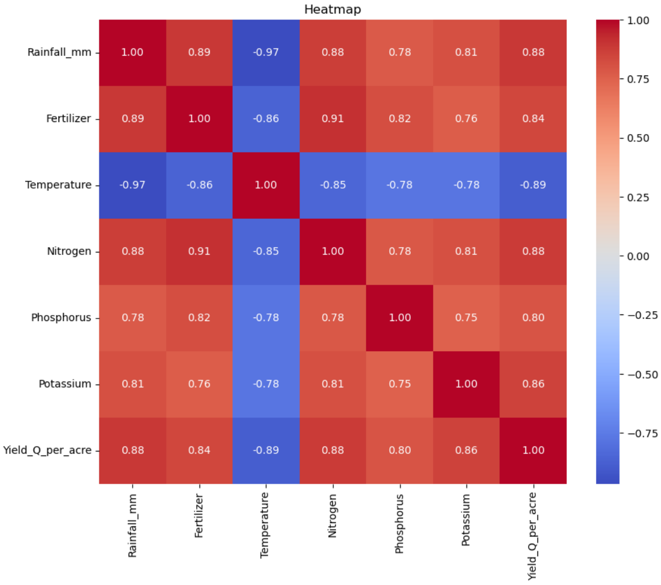

Fig. 10. Feature correlation heatmap for Dataset-3.

2) Machine Learning Model Training and Results: A machine learning model was trained using the *AutoGluon* framework [13], with the cleaned dataset split into 80% training and 20% testing. The performance metrics for the best model are shown in Table III.

a) Feature Importance: Feature importance analysis (Figure 16) reveals that soil properties, particularly *Potassium* and *Nitrogen*, are the most significant predictors of yield. Environmental factors such as rainfall, temperature, and fertilizer showed negligible or negative contributions.

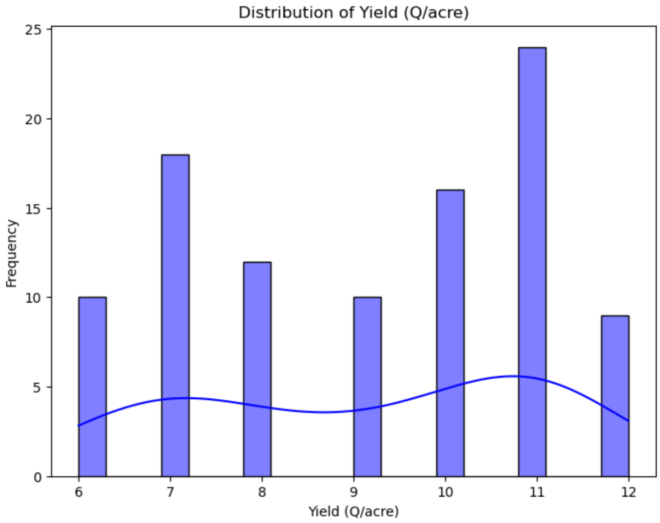

Fig. 11. Distribution of yield (Q/acre) in Dataset-3.

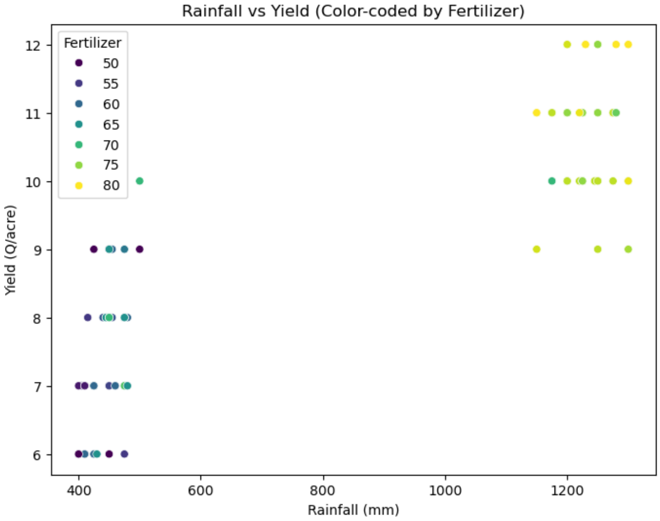

Fig. 12. Scatter plot of rainfall vs. yield, color-coded by fertilizer usage.

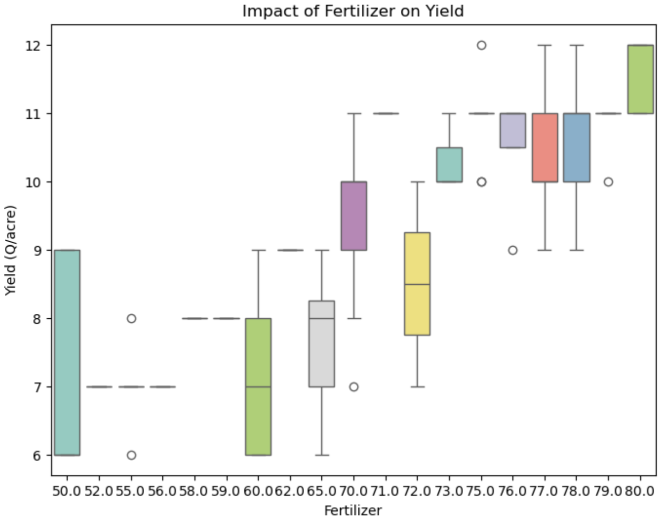

Fig. 13. Scatter plot of fertilizer usage vs. yield.

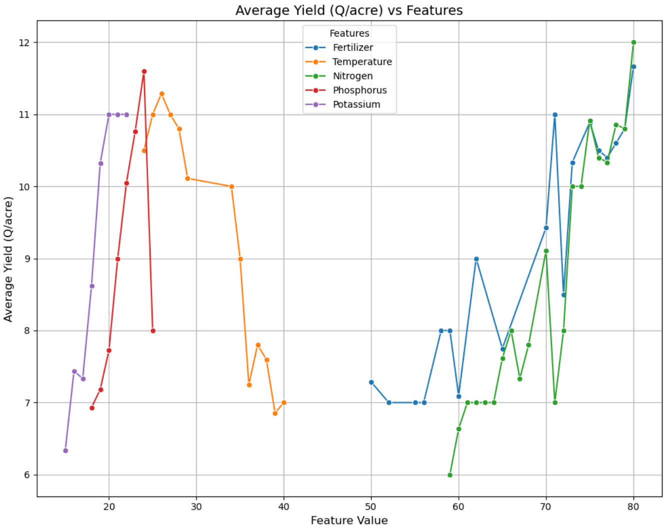

Fig. 14. Average yield (Q/acre) vs. soil and environmental features.

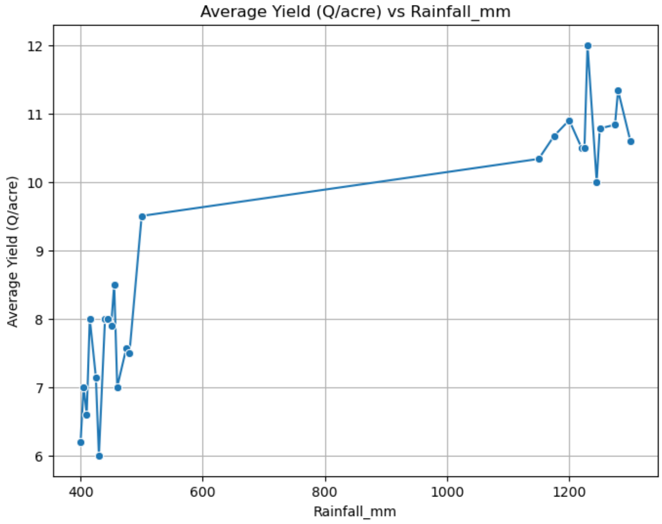

Fig. 15. Average yield (Q/acre) vs. soil and environmental features.

3) Insights: From the analysis of Dataset-3, the following conclusions were drawn:

- **Soil Properties Drive Yield Predictions:** *Potassium* and *Nitrogen* emerged as the dominant factors in predicting crop yield, consistent with the dataset's soil-focused nature.
- **Environmental Factors Play a Limited Role:** Unlike in Dataset-1 and Dataset-2, environmental factors such as rainfall and temperature were not significant predictors.

TABLE III
PERFORMANCE METRICS OF AUTOGLUON MODELS ON DATASET-3.

Metric	**Value**
Mean Absolute Error (MAE)	0.550
Mean Squared Error (MSE)	0.650
Root Mean Squared Error (RMSE)	0.806
R^2 Score	0.857
Accuracy (%)	94.27

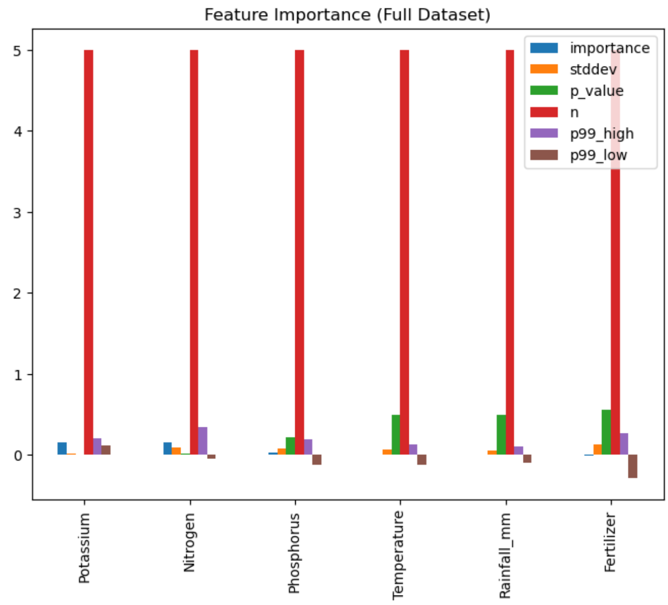

Fig. 16. Feature importance for Dataset-3.

- **High Model Accuracy:** The model achieved a strong accuracy of 94.27%, demonstrating that the selected features effectively capture the key determinants of yield in this dataset.

E. Comparison of Datasets

The three datasets analyzed in this study—**Agriculture Crop Yield**, **Crop Yield Prediction Dataset**, and **Crop Yield Prediction**—each provided unique insights into crop yield prediction. While Dataset-1 (*synthetic data*) emphasized environmental variables, Dataset-2 (*real-world data*) focused on meteorological factors, and Dataset-3 (*soil and environmental features*) combined properties of the soil and management practices. This section synthesizes the results of the exploratory data analysis (EDA) and machine learning experiments to compare their effectiveness in addressing the research question.

1) Exploratory Data Analysis (EDA): From the EDA, rainfall (*Rainfall_mm*) exhibited a strong positive correlation with yield in Dataset-1 (correlation coefficient of 0.76), emphasizing its critical role in the synthetic dataset. However, in Dataset-2, rainfall (*average_rainfall_mm_per_year*) showed only a moderate correlation, and in Dataset-3, its impact was negligible compared to soil-related features such as *Nitrogen* and *Potassium*. These differences highlight the varying contributions of environmental factors across datasets and reflect their distinct focuses.

2) Model Performance: The machine learning models trained on the three datasets exhibited notable differences in performance (Table IV). Dataset-2 achieved the highest R^2 score (0.990) and accuracy (95.3%), highlighting its strong predictive capability despite the challenges of working with real-world data. Dataset-1, with its synthetic nature and controlled feature relationships, achieved an R^2 score of 0.913 and an accuracy of 91.41%, demonstrating high performance but potentially limited applicability to practical scenarios. Dataset-3, with a focus on soil properties, achieved an R^2 score of 0.857 and an accuracy of 94.27%, emphasizing the importance of soil-related factors in yield prediction and offering a complementary perspective to the environmental focus of Datasets-1 and 2.

TABLE IV
COMPARISON OF MODEL PERFORMANCE METRICS ACROSS DATASETS

Metric	**Dataset-1**	**Dataset-2**	**Dataset-3**
R^2 Score	0.913	0.990	0.857
Accuracy (%)	91.41	95.3	94.27
RMSE	0.501	8.338	0.806
MAE	0.400	6.060	0.550

3) Insights and Implications: The analysis of these datasets highlights the trade-offs between synthetic and real-world data, as well as the impact of feature selection on prediction performance. Dataset-1's synthetic nature allowed for competitive accuracy (91.41%) and a high R^2 score (0.913) but may lack real-world applicability. Dataset-2's real-world focus ensures practical relevance and achieved the highest performance metrics (R^2 of 0.990 and accuracy of 95.3%). Dataset-3, with its emphasis on soil properties, provides critical insights into the importance of integrating environmental and soil-based features for holistic yield prediction. These findings suggest that combining datasets with complementary focuses could further enhance model performance and generalizability.

F. Reproducibility and Code Availability

All experiments, including data preprocessing, exploratory data analysis, and model training, were conducted using Jupyter Notebooks. The complete set of notebooks used to generate the results presented in this paper is publicly available at the following GitHub repository: https://github.com/Sidonitor/data-science

VII. CONCLUSION AND OUTLOOK

This study investigated the potential to predict crop yields using three datasets with varying focuses: **Agriculture Crop Yield** (synthetic, environmental variables), **Crop Yield Prediction Dataset** (real-world, meteorological factors), and **Crop Yield Prediction** (soil and environmental features). The results demonstrated that crop yields can be predicted with high accuracy using machine learning models, particularly those built with the *AutoGluon* framework.

Key findings include:

- Rainfall (*Rainfall_mm*) emerged as a dominant factor in yield prediction for Dataset-1, reflecting its synthetic design. However, in Dataset-2, crop type (*Item*) and geographic region (*Area*) were the most critical features, while environmental factors like rainfall played a secondary role.
- Dataset-3 emphasized the importance of soil properties, such as *Nitrogen* and *Potassium*, in determining yield, highlighting a complementary perspective to the environmental focus of Datasets-1 and 2.

- Dataset-2 achieved the highest model performance (R^2 of 0.990 and accuracy of 95.3%), underscoring its relevance for practical applications. Dataset-1, while slightly lower in performance (R^2 of 0.913 and accuracy of 91.41%), demonstrated the value of synthetic data for controlled studies. Dataset-3, with an accuracy of 94.27%, emphasized the importance of soil properties in yield prediction.

Despite these promising results, challenges remain. Dataset-specific limitations, such as the synthetic nature of Dataset-1 and the relatively small size of Dataset-3, indicate the need for further experimentation with more diverse and comprehensive datasets. Additionally, feature engineering and advanced preprocessing techniques could enhance model performance and interpretability.

A. Outlook

Future work should focus on:

- Combining datasets with complementary focuses (e.g., environmental and soil properties) to create more holistic predictive models.
- Exploring other machine learning frameworks and advanced techniques, such as deep learning or transfer learning, for yield prediction.
- Investigating the scalability of the proposed models to larger and more heterogeneous datasets, including global and multi-crop datasets.
- Incorporating temporal data, such as seasonal or multi-year trends, to improve predictions and account for climate variability.

By addressing these areas, future research can further enhance the accuracy, robustness, and practical applicability of crop yield prediction models, contributing to more sustainable and data-driven agricultural practices.

Acknowledgment

I would like to express my sincere gratitude to my friends and colleagues from university for their invaluable feedback and support throughout this research. Their input during discussions and their assistance with proofreading have greatly contributed to the quality of this work.

Additionally, I acknowledge the assistance of ChatGPT-4o in refining the language, grammar, and structure of this paper while ensuring the integrity and scientific rigor of the content. The tool was instrumental in enhancing the clarity and coherence of the manuscript.

References

[1] M. Hassan, K. Malhotra, and M. Firdaus, “Application of Artificial Intelligence in IoT Security for Crop Yield Prediction”, RRST, vol. 2, no. 1, pp. 136–157, Oct. 2022.

[2] I. Beloev, D. Kinaneva,G. Georgiev, G. Hristov, P. Zahariev, ”Artificial intelligence-driven autonomous robot for precision agriculture”, *Acta Technologica Agriculturae*, 2021 Mar 12;24(1), 48-54.

[3] A.-F. Jiménez, P.-F. Cárdenas, and F. Jiménez, ”Intelligent IoT-multiagent precision irrigation approach for improving water use efficiency in irrigation systems at farm and district scales”, *Computers and Electronics in Agriculture*, vol. 192, Art. no. 106635, 2022. [Online]. Available: https://doi.org/10.1016/j.compag.2021.106635. [Accessed: Jan. 25, 2025].

[4] D. K. Bolton and M. A. Friedl, ”Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics”, *Agricultural and Forest Meteorology*, vol. 173, pp. 74–84, 2013. [Online]. Available: doi: https://doi.org/10.1016/j.agrformet.2013.01.007. [Accessed: Jan. 25, 2025].

[5] M. Prabhakar, P. Raja, O. E. Apolo, and M. Pérez-Ruiz, ”Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques—A review”, *Frontiers in Plant Science*, vol. 12, p. 684328, Jun. 2021. [Online]. Available: doi: https://doi.org/10.3389/fpls.2021.684328. [Accessed: Jan. 25, 2025].

[6] A. Bouguettaya, H. Zarzour, A. Kechida, and others, ”Deep learning techniques to classify agricultural crops through UAV imagery: a review”, *Neural Computing and Applications*, vol. 34, pp. 9511–9536, 2022. [Online]. Available: doi: https://doi.org/10.1007/s00521-022-07104-9. [Accessed: Jan. 25, 2025].

[7] S. Oti-Attakorah, ”Agriculture Crop Yield Dataset”, *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield?resource=download. [Accessed: Jan. 25, 2025].

[8] S. Stanislavovich, ”Analyzing Crop Productivity”, *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/code/sergeistanislavovich/analyzing-crop-productivity. [Accessed: Jan. 25, 2025].

[9] R. Patel, ”Crop Yield Prediction Dataset”, *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset. [Accessed: Jan. 25, 2025].

[10] H. Mandhana, ”Crop Yield Prediction Model”, *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/code/hemanshumandhana/crop-yield-prediction-model. [Accessed: Jan. 25, 2025].

[11] Y. Minh, ”Crop Yield Prediction Dataset,” *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/yaminh/crop-yield-prediction. [Accessed: Jan. 25, 2025].

[12] D. Batra, ”Yield prediction”, *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/code/yaminh/yield-prediction.

[13] ”Fast and Accurate ML in 3 Lines of Code”, *AutoGluon*, [Online]. Available: https://auto.gluon.ai/stable/index.html. [Accessed: Jan. 25, 2025].