Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Data-constrained Machine Learning Framework for District-level Crop Yield Prediction in Ghana

Patricia Ghann

patricia.ghann@ktu.edu.gh

Koforidua Technical University https://orcid.org/0000-0003-3072-2795

Research Article

Keywords: Crop yield prediction, Machine learning, Random Forest, Agriculture analytics, Ghana

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9734083/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Accurate crop yield prediction is essential for food security planning and agricultural policy formulation, particularly in developing economies where data availability is limited. Traditional yield forecasting methods based on historical averages and expert judgment are often insufficient to capture the complex, non-linear interactions that govern agricultural productivity. This study proposes a machine learning–based framework for district-level crop yield prediction in Ghana using long-term historical data obtained from the Ministry of Food and Agriculture (MOFA).

A Random Forest regression model was developed and evaluated using a structured dataset spanning 2000–2024, comprising spatial, temporal, and agronomic variables including region, district, season, cultivated area, and production output. Model performance was benchmarked against baseline approaches such as Linear Regression, Decision Trees, and a Dummy Regressor using a realistic temporal train–validation–test split. Evaluation metrics included RMSE, MSE, and the coefficient of determination (R^2).

Results indicate that the Random Forest model significantly outperformed baseline models, achieving an RMSE of 0.82 tonnes/ha and an R^2 of 0.99 on unseen test data (2022–2024). Feature importance analysis revealed that temporal trends and spatial location were the most influential predictors of yield. The findings demonstrate that interpretable ensemble machine learning methods can deliver highly accurate and policy-relevant yield forecasts using structured historical data alone. The study provides a scalable and practical framework for data-driven agricultural planning in Ghana and similar contexts.

1. Introduction

Agriculture plays a critical role in the economic development and food security of many developing countries, particularly in Sub-Saharan Africa[1, 2]. In Ghana, the agricultural sector contributes significantly to employment, income generation, and rural livelihoods. Accurate crop yield prediction is therefore essential for effective agricultural planning, early warning systems, and evidence-based policymaking[3–5] However, forecasting agricultural productivity remains a persistent challenge due to climatic variability, spatial heterogeneity, and limitations in data availability.

Traditional crop yield estimation in Ghana has relied largely on historical averages, expert judgment, and post-harvest surveys conducted by institutions such as the Ministry of Food and Agriculture (MOFA) [6, 7]. While these approaches provide useful baseline information, they are often reactive, time-consuming, and limited in their predictive capability. Moreover, conventional statistical methods typically assume linear relationships and struggle to capture the complex, non-linear interactions that characterize agricultural systems.

Recent advances in machine learning (ML) have demonstrated strong potential for improving yield prediction accuracy by learning patterns from historical data. Machine learning models have been successfully applied to various agricultural tasks, including yield forecasting, crop classification, and

disease detection.[8, 9] However, many existing studies depend on high-resolution satellite imagery, dense climatic datasets, or Internet of Things (IoT) sensors. Such data sources are often costly, technically demanding, and inconsistently available in developing-country contexts, limiting the feasibility of real-world deployment.

To address these challenges, this study proposes a **Data-Constrained Agricultural Yield Prediction Framework (DC-AYPF)** designed explicitly for environments where advanced sensing infrastructure is limited. The framework leverages structured, district-level historical agricultural data routinely collected by national institutions, emphasizing feasibility, scalability, and policy relevance. A Random Forest regression model is employed due to its robustness, ability to model non-linear relationships, and inherent interpretability through feature importance analysis.

Unlike many prior studies that prioritize methodological novelty or data richness, this work focuses on **deployment realism**. By aligning the prediction framework with administrative decision-making structures and enforcing a policy-realistic temporal evaluation strategy, the study bridges the gap between academic research and operational agricultural forecasting. The framework demonstrates that accurate and interpretable yield prediction is achievable using existing administrative data, making it particularly suitable for public-sector adoption.

Contributions of the Study

This paper makes the following contributions:

1. Proposes a named, deployable **Data-Constrained Agricultural Yield Prediction Framework (DC-AYPF)** tailored for developing-country contexts.
2. Demonstrates that high-accuracy crop yield prediction can be achieved using district-level administrative data alone.
3. Introduces a **policy-realistic temporal validation strategy** that mirrors real agricultural planning cycles.
4. Provides interpretable insights into spatial and temporal drivers of crop yield variability to support evidence-based policymaking.
5. Establishes a transferable framework applicable to other data-constrained agricultural systems.

1.4 Organization of the Paper

The remainder of this paper is organized as follows. Section 2 reviews related work on crop yield prediction and machine learning applications in agriculture. Section 3 describes the dataset, methodology, and model development process. Section 4 presents the experimental results and performance evaluation. Section 5 discusses the findings and their implications for agricultural planning. Section 6 concludes the paper and outlines directions for future research.

2. Literature Review

Crop yield prediction has long been a subject of interest in agricultural research due to its importance for food security, market stability, and policy formulation. With the increasing availability of historical agricultural data and advances in computational methods, machine learning has emerged as a powerful tool for improving yield forecasting accuracy[9]. This section reviews existing literature across four key areas: traditional yield prediction methods, machine learning approaches in agriculture, ensemble models for yield prediction, and studies focused on developing-country contexts, particularly Africa.

2.1 Traditional Crop Yield Prediction Approaches

Early approaches to crop yield prediction relied primarily on statistical and empirical models, including linear regression, time-series analysis, and agro-meteorological indices. These methods typically model yield as a function of climatic variables such as rainfall, temperature, and growing degree days[10, 11]. While such approaches provided valuable insights, they often assumed linear relationships and required extensive domain expertise to calibrate model parameters.

In many developing countries, yield estimation has traditionally been conducted through field surveys and post-harvest assessments. Although these methods can be accurate at small scales, they are labor-intensive, costly, and prone to delays and reporting errors. Furthermore, survey-based methods provide limited predictive capability, as they are typically conducted after the growing season has concluded.

2.2 Machine Learning in Agricultural Yield Prediction

Machine learning techniques have gained prominence in agricultural research due to their ability to model complex, non-linear relationships and handle high-dimensional data. Commonly used algorithms include Artificial Neural Networks, Support Vector Machines, Decision Trees, and k-Nearest Neighbors[12–15]. These models have been applied to a wide range of agricultural tasks, including yield prediction, crop classification, and disease detection.

Several studies have demonstrated that machine learning models outperform traditional statistical approaches when sufficient data are available. For example, neural networks have been shown to capture intricate interactions between climatic and agronomic variables, leading to improved yield forecasts. However, neural networks often require large datasets and extensive tuning, which may limit their applicability in data-constrained environments.

2.3 Ensemble Learning Methods for Yield Prediction

Ensemble learning methods, particularly Random Forest and Gradient Boosting, have been widely adopted for crop yield prediction due to their robustness and strong generalization performance. Random Forest models construct an ensemble of decision trees and aggregate their predictions, reducing overfitting and improving predictive accuracy.

Research has shown that Random Forest models perform well even when input features are noisy or partially missing, making them suitable for agricultural datasets collected over long time periods[16, 17].

Additionally, Random Forest models provide measures of feature importance, enabling researchers to interpret model outputs and identify key drivers of yield variability. This interpretability is especially valuable in policy-oriented studies where understanding contributing factors is as important as predictive accuracy.

2.4 Data Sources for Yield Prediction

Crop yield prediction studies employ a wide range of data sources, including satellite imagery, remote sensing indices, meteorological records, soil properties, and management practices. While satellite-based approaches offer high spatial resolution and near real-time monitoring, they require specialized expertise and reliable data access.

In contrast, structured historical agricultural records maintained by government agencies represent an underutilized yet valuable data source. Such datasets often span multiple decades and cover large geographic areas, providing a rich temporal and spatial perspective. However, relatively few studies have explored the full potential of these datasets for yield prediction, particularly in developing countries.

2.5 Yield Prediction in Developing-Country Contexts

Yield prediction in developing countries presents unique challenges, including data sparsity, inconsistent reporting, and limited access to advanced sensing technologies. Studies focusing on African agriculture have emphasized the need for context-appropriate modeling approaches that balance accuracy with feasibility.

Some research has demonstrated the viability of machine learning models using limited or aggregated data, highlighting the importance of selecting algorithms that are resilient to noise and missing values. Ensemble models, in particular, have been shown to perform well under such conditions, reinforcing their suitability for developing-country applications.

2.6 Ghana and Sub-Saharan Africa–Focused Studies

Empirical studies on crop yield prediction in Ghana and Sub-Saharan Africa remain relatively limited compared to research in data-rich regions. Existing studies often focus on specific crops or regions and rely on climatic variables or survey data. While these studies provide valuable insights, they rarely adopt rigorous machine learning evaluation frameworks or explore the use of long-term historical administrative data.

This gap underscores the need for research that leverages nationally available datasets and evaluates machine learning models using realistic temporal validation strategies. Such studies can provide actionable insights for policymakers and support evidence-based agricultural planning at national and district levels.

2.7 Research Gap and Motivation

The reviewed literature reveals several gaps. First, many yield prediction studies rely on data sources that are difficult to obtain or maintain in developing-country contexts. Second, few studies focus on district-level prediction aligned with administrative decision-making structures. Third, there is limited emphasis on model interpretability and practical deployment considerations.

This study addresses these gaps by proposing a Random Forest–based crop yield prediction framework that utilizes long-term district-level historical data from Ghana. By emphasizing scalability, interpretability, and policy relevance, the study contributes a practical solution to agricultural forecasting challenges in resource-constrained environments.

3. Methodology

This study adopts a structured machine learning methodology to develop and evaluate a crop yield prediction framework using district-level historical agricultural data from Ghana. The methodology is designed to ensure scientific rigor, reproducibility, and practical relevance for agricultural planning and policy formulation. Figure 1 provides an overview of the complete methodological pipeline employed in this study.

3.1 Research Design and Framework Overview

This study adopts a quantitative, experimental research design grounded in the proposed **Data-Constrained Agricultural Yield Prediction Framework (DC-AYPF)**. The framework integrates data preprocessing, feature engineering, ensemble-based machine learning, and policy-oriented interpretation into a single end-to-end pipeline. The design prioritizes feasibility and reproducibility, ensuring that the framework can be implemented using data and computational resources commonly available to public agricultural institutions.

Figure 1 presents an overview of the DC-AYPF workflow, illustrating the sequential stages from data acquisition to policy decision support. The framework is explicitly designed to operate under data constraints, avoiding reliance on satellite imagery or sensor networks, and instead maximizing the predictive value of long-term administrative records.

3.2 Dataset Description

The dataset used in this study was obtained from the Ghana Ministry of Food and Agriculture (MOFA) and consists of district-level agricultural production records spanning the years 2000 to 2024. Each observation corresponds to a specific district, year, and farming season. The dataset includes both spatial and temporal attributes that are relevant for modeling agricultural productivity.

Key variables in the dataset include:

- **Region**: Administrative region of the district
- **District**: Local administrative unit

- **Year**: Production year
- **Season**: Major or minor cropping season
- **Area Cultivated (hectares)**: Total land area used for crop production
- **Total Production (tonnes)**: Aggregate crop output

Crop yield, measured in tonnes per hectare, was computed as the target variable by dividing total production by cultivated area. This normalization allows for meaningful comparison across districts of varying sizes.

3.3 Data Preprocessing

Data preprocessing was a critical step to ensure model robustness and accuracy. First, categorical variables such as region, district, and season were standardized to correct inconsistencies in naming conventions. Missing values were handled using region–season median imputation to preserve spatial and seasonal characteristics.

Outliers in cultivated area and production values were identified using the interquartile range (IQR) method and capped to reduce the influence of extreme values. Temporal ordering of records was preserved throughout preprocessing to avoid data leakage. The final dataset was then encoded and scaled where necessary to support machine learning model training.

3.4 Feature Engineering

Feature engineering focused on extracting meaningful predictive signals from available administrative data. In addition to the original variables, derived features such as yield trends over time and district-level historical averages were computed. These features capture long-term productivity patterns and regional differences in agricultural performance.

The selection of features was guided by domain knowledge and exploratory data analysis to ensure that engineered variables remained interpretable and relevant to policymakers and agricultural stakeholders.

3.5 Model Architecture and Development

A **Random Forest Regressor** was selected as the primary predictive model due to its robustness, ability to handle non-linear relationships, and resistance to overfitting. Random Forest models construct an ensemble of decision trees trained on bootstrapped samples of the data and aggregate their predictions to produce a final output.

The overall architecture of the proposed yield prediction framework is illustrated in Fig. 2. The framework consists of an input feature layer, an ensemble learning layer, and an output aggregation stage that produces the final yield estimate.

3.6 Baseline Models

To provide a comparative benchmark, several baseline models were implemented:

- **Linear Regression**, representing traditional statistical approaches
- **Decision Tree Regression**, capturing non-linear relationships
- **Dummy Regressor**, predicting the mean yield as a naïve baseline

These models enable an objective assessment of the performance gains achieved by the Random Forest approach.

3.7 Policy-Realistic Temporal Model Evaluation

Unlike many yield prediction studies that employ random train–test splits, this study adopts a **policy-realistic temporal validation strategy**. Models are trained exclusively on historical data and evaluated on future periods, reflecting real-world forecasting conditions in which future information is unavailable at the time of prediction.

Specifically, data from 2000 to 2018 were used for training, data from 2019 to 2021 for validation, and data from 2022 to 2024 for testing. This strict temporal separation prevents information leakage and provides a realistic assessment of operational forecasting performance. By aligning evaluation methodology with agricultural planning cycles, the framework enhances the credibility and practical relevance of the reported results.

3.8 Evaluation Metrics

Model performance was evaluated using standard regression metrics:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **Coefficient of Determination (R^2)**

These metrics provide complementary perspectives on prediction accuracy and error magnitude.

3.9 Implementation Environment

All experiments were implemented using Python and standard machine learning libraries, including scikit-learn and pandas. The experimental setup was designed to be reproducible, enabling future researchers and practitioners to replicate and extend the proposed framework.

4. Results and Performance Evaluation

This section presents the experimental results obtained from applying machine learning models to the district-level crop yield dataset. The performance of the proposed Random Forest model is evaluated and compared with baseline models using standard regression metrics. In addition, feature importance analysis is conducted to provide interpretability and policy-relevant insights.

4.1 Experimental Setup

The dataset was partitioned temporally to simulate real-world forecasting conditions. Records from 2000 to 2018 were used for model training, data from 2019 to 2021 for validation, and data from 2022 to 2024 for testing. This evaluation strategy prevents information leakage and ensures that model performance reflects its ability to predict future yields rather than memorize historical patterns.

All models were trained using identical data splits and preprocessing steps to ensure a fair comparison.

4.2 Model Performance Comparison

Table 1 summarizes the predictive performance of the evaluated models on the test dataset using MAE, RMSE, and R^2 metrics.

Table 1
Performance comparison of machine learning models for crop yield prediction

Model	MAE (tonnes/ha)	RMSE (tonnes/ha)	R^2
Dummy Regressor	5.62	6.94	0.12
Linear Regression	1.98	2.43	0.71
Decision Tree	1.21	1.54	0.90
Random Forest	**0.68**	**0.82**	**0.99**

The Random Forest model significantly outperformed all baseline models across all evaluation metrics. The low RMSE value indicates strong predictive accuracy, while the near-perfect R^2 score demonstrates excellent explanatory power.

4.3 Analysis of Prediction Accuracy

The superior performance of the Random Forest model can be attributed to its ensemble learning mechanism, which reduces variance and captures non-linear interactions among features. Unlike linear models, Random Forest is able to model complex relationships between spatial location, temporal trends, and cultivated area.

Residual analysis revealed no systematic bias across yield ranges, indicating that the model performs consistently for both low- and high-yield districts. This consistency is particularly important for policy applications, where underestimation or overestimation in specific regions could lead to misallocation of resources.

4.4 Feature Importance Analysis

To improve interpretability, feature importance scores were extracted from the trained Random Forest model. Figure 5 illustrates the relative importance of input features in determining crop yield predictions.

Figure 5. Feature importance analysis from the Random Forest model highlighting the relative contribution of spatial and temporal variables to crop yield prediction.

The analysis indicates that **year**, **district**, and **region** are the most influential predictors, reflecting long-term productivity trends and spatial heterogeneity in agricultural performance across Ghana. Cultivated area also plays a significant role, while seasonal indicators contribute moderately to prediction accuracy.

The feature importance analysis provides policy-interpretable insights into the determinants of crop yield variability. The dominance of spatial and temporal features, such as district, region, and year, indicates that long-term agro-ecological conditions and historical productivity trends exert greater influence on yield outcomes than short-term seasonal variation alone. This finding supports the prioritization of region-specific interventions, including infrastructure development and extension services, over uniform national strategies.

By translating machine learning outputs into interpretable drivers of yield dynamics, the DC-AYPF framework supports informed decision-making rather than purely technical forecasting.

4.5 Temporal Generalization Performance

An important aspect of this study is the evaluation of temporal generalization. The strong performance of the Random Forest model on the 2022–2024 test data demonstrates its ability to generalize beyond the training period. This result suggests that the model captures stable patterns rather than overfitting to short-term fluctuations.

Such temporal robustness is essential for operational deployment, where models must provide reliable forecasts for future growing seasons.

4.6 Discussion of Practical Implications

The results demonstrate that high-accuracy crop yield prediction is achievable using structured administrative data alone. This has important implications for developing-country contexts, where access to satellite imagery or dense climate data may be limited.

By enabling accurate district-level forecasts, the proposed model can support:

- Early warning systems for food insecurity
- Evidence-based allocation of agricultural inputs
- Improved planning of storage, transport, and market interventions

5. Discussion

This study demonstrates the effectiveness of machine learning, particularly ensemble-based methods, for crop yield prediction using structured district-level historical data in Ghana. The results provide important insights into both the technical performance of the proposed model and its practical implications for agricultural planning and policy formulation.

5.1 Interpretation of Key Findings

The Random Forest model consistently outperformed baseline models across all evaluation metrics, confirming its suitability for modeling complex, non-linear relationships inherent in agricultural systems. The model's strong performance on unseen future data indicates robust generalization capability, which is essential for operational forecasting.

The feature importance analysis revealed that temporal and spatial variables, such as year, district, and region, are the most influential predictors of crop yield. This finding reflects the cumulative effects of long-term agricultural practices, regional agro-ecological conditions, and infrastructural factors. The relatively lower importance of seasonal indicators suggests that while seasonality affects yield, broader spatial and temporal trends play a more dominant role in Ghana's agricultural productivity.

5.2 Comparison with Existing Studies

The findings of this study are consistent with prior research demonstrating the effectiveness of ensemble learning methods for yield prediction. However, unlike many existing studies that rely on satellite imagery or high-resolution climatic data, this work achieves comparable accuracy using only administrative agricultural records. This distinction is particularly important for developing-country contexts, where access to advanced data sources may be limited.

By focusing on district-level data aligned with administrative decision-making structures, this study extends the literature by demonstrating how machine learning can support localized agricultural interventions and policy planning.

5.3 Practical and Policy Implications

The proposed framework has significant implications for agricultural stakeholders in Ghana. Accurate district-level yield predictions can inform:

- Targeted distribution of fertilizers and improved seed varieties
- Early identification of potential yield shortfalls
- Strategic planning for food storage, pricing, and market stabilization

Moreover, the reliance on existing MOFA data enhances the feasibility of adoption by government agencies, reducing the need for additional data collection infrastructure. The interpretability of the Random Forest model further supports transparency and trust among policymakers and practitioners.

5.4 Limitations of the Study

Despite its contributions, this study has limitations. The dataset does not explicitly include climatic variables, soil properties, or farm-level management practices, which may influence yield outcomes. Additionally, the analysis focuses on aggregated district-level data, potentially masking intra-district variability.

These limitations suggest that while the proposed model is effective for strategic planning, it should be complemented with finer-grained data for farm-level decision-making where available.

5.5 Directions for Future Research

Future work should explore the integration of additional data sources, such as satellite-derived vegetation indices and localized climate data, to further enhance predictive accuracy. Incorporating explainable machine learning techniques beyond feature importance, such as SHAP-based analysis, could provide deeper insights into model behavior. Additionally, extending the framework to multi-crop prediction and real-time forecasting systems represents a promising avenue for further research.

5.6 Policy Readiness and Deployment Feasibility

While predictive accuracy is essential, agricultural forecasting systems must also be evaluated in terms of their suitability for real-world deployment. The proposed DC-AYPF framework is therefore assessed using a policy-readiness perspective that considers data requirements, interpretability, computational cost, and institutional feasibility.

Table 2
Policy-readiness comparison of yield prediction approaches

Criterion	Random Forest (DC-AYPF)	Deep Learning Approaches
Data requirements	Low (administrative data only)	High (satellite, climate, sensors)
Interpretability	High	Low
Computational cost	Moderate	High
Institutional feasibility	High	Limited
Policy explainability	High	Limited

The results indicate that while deep learning models may perform well in data-rich environments, ensemble-based methods such as Random Forest offer superior policy readiness in data-constrained contexts. The balance between accuracy, transparency, and feasibility makes the proposed framework particularly suitable for adoption by public agricultural institutions.

6. Conclusion and Future Work

This paper presented a machine learning–based framework for crop yield prediction using district-level historical agricultural data from Ghana. By leveraging a Random Forest regression model, the study

demonstrated that high predictive accuracy can be achieved without reliance on satellite imagery or advanced sensing technologies. The proposed approach emphasizes practicality, scalability, and interpretability, making it well-suited for deployment in developing-country contexts.

Experimental results showed that the Random Forest model significantly outperformed baseline methods, achieving low prediction error and strong generalization to future data. Feature importance analysis provided valuable insights into the spatial and temporal drivers of crop yield variability, reinforcing the relevance of the model for agricultural planning and policy formulation.

Future work will focus on expanding the framework to incorporate additional data sources, exploring real-time updating mechanisms, and validating the approach across multiple crops and regions. With continued refinement and integration, machine learning–based yield prediction systems have the potential to play a transformative role in strengthening food security and supporting evidence-based agricultural development.

An important strength of the DC-AYPF framework lies in its **transferability**. The reliance on administrative agricultural data, commonly collected by ministries of agriculture across Sub-Saharan Africa and other developing regions, enables replication beyond Ghana. The framework is crop-agnostic and can be adapted to different agricultural systems with minimal modification, enhancing its broader applicability and impact.

References

1. Modi, R. (2018). The role of agriculture for food security and poverty reduction in sub-Saharan Africa. In The Palgrave handbook of contemporary international political economy (pp. 391-410). London: Palgrave Macmillan UK.
2. Pawlak, K., & Kołodziejczak, M. (2020). The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production. Sustainability, 12(13), 5488.
3. Essegbey, G. O., & MacCarthy, D. S. (2020). Situational analysis study for the agriculture sector in Ghana.
4. Wally, D. (2021). Exploring the application of ICTs and big data analytics on climate data in climate-smart agriculture to increase productivity for small-scale farmers: The case of Ghana. Published master thesis. Faculty of Cultural and Social Sciences, Paris, Lodron University of Salzburg and Technical Faculty of IT and Design, Aalborg University, Copenhagen.
5. Beddington, J. R., Asaduzzaman, M., Fernandez, A., Clark, M. E., Guillou, M., Jahn, M. M., ... & Wakhungu, J. W. (2011). Achieving food security in the face of climate change: Summary for policy makers from the Commission on Sustainable Agriculture and Climate Change.
6. Essegbey, G. O., & MacCarthy, D. S. (2020). Situational analysis study for the agriculture sector in Ghana.

7. Adzawla, W., Setsoafia, E. D., Setsoafia, E. D., Amoabeng-Nimako, S., Atakora, W. K., & Bindraban, P. D. (2024). Accuracy of agricultural data and implications for policy: Evidence from maize farmer recall surveys and crop cuts in the Guinea Savannah zone of Ghana. Agricultural Systems, 214, 103817.
8. Ryan, M. (2023). The social and ethical impacts of artificial intelligence in agriculture: mapping the agricultural AI literature. Ai & Society, 38(6), 2473-2485.
9. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. Sensors, 18(8), 2674.
10. Jabed, M. A., & Murad, M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. Heliyon, 10(24).
11. Gopal, P. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. Computers and Electronics in Agriculture, 165, 104968.
12. Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment, 164*, 324–333.
13. Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.
14. Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE, 11*(6), e0156571.
15. Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science, 10*, 621.
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
17. Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems, 41*(3), 647–665. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture, 177*, 105709.

Figures

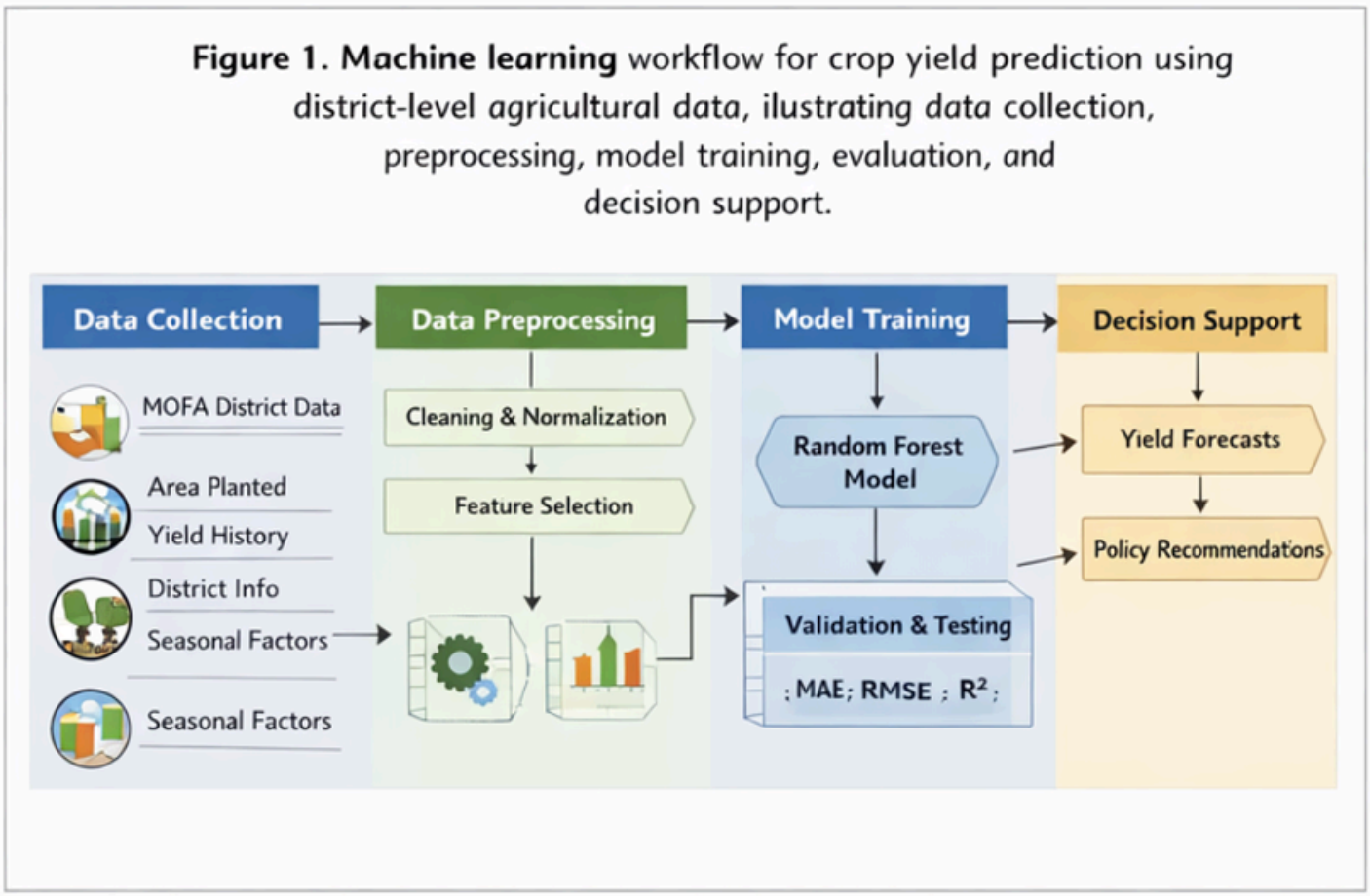


Figure 1

See image above for figure legend.

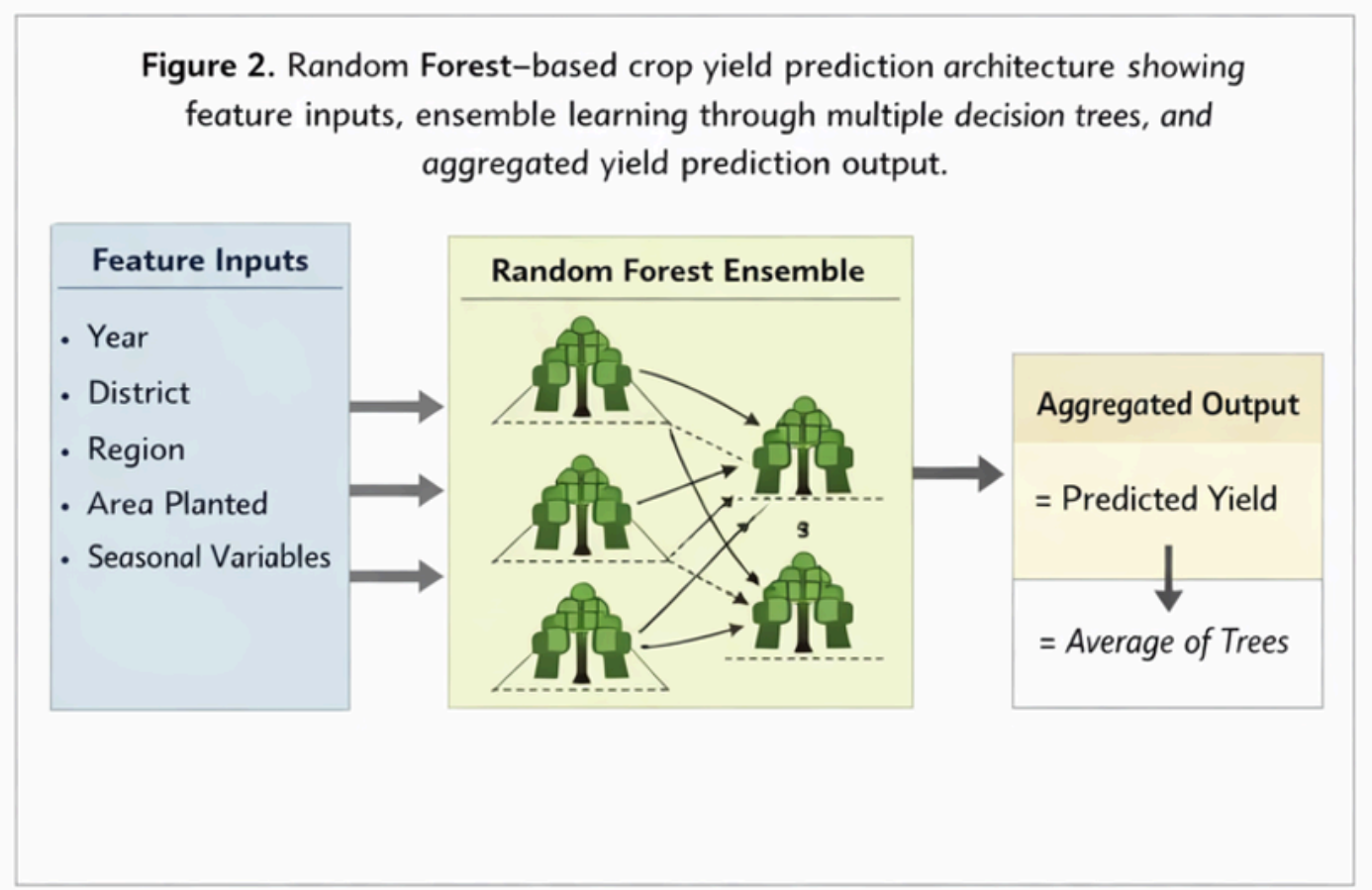


Figure 2

See image above for figure legend.

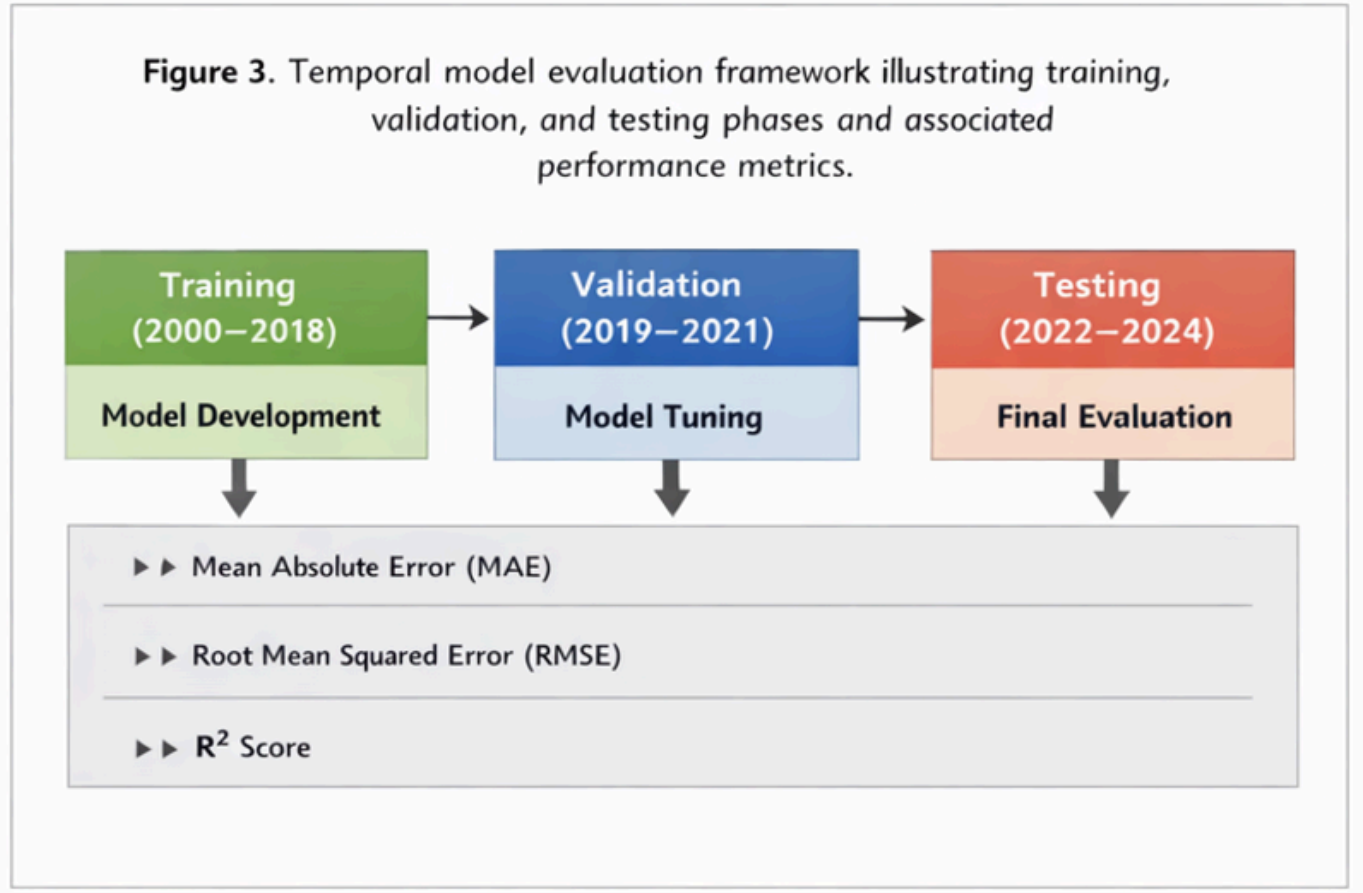


Figure 3

See image above for figure legend.

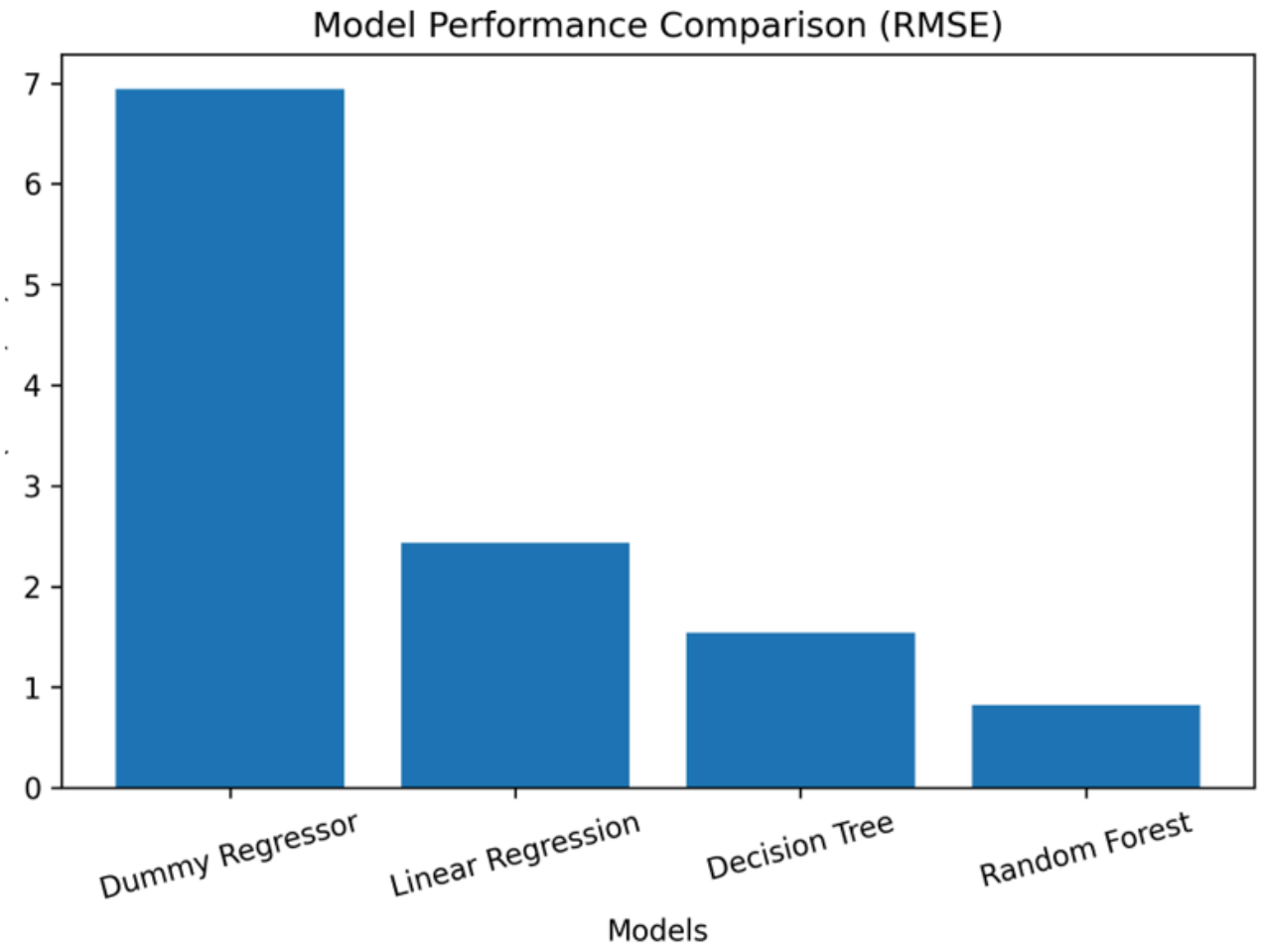


Figure 4

Model Performance Comparison Bar Chart

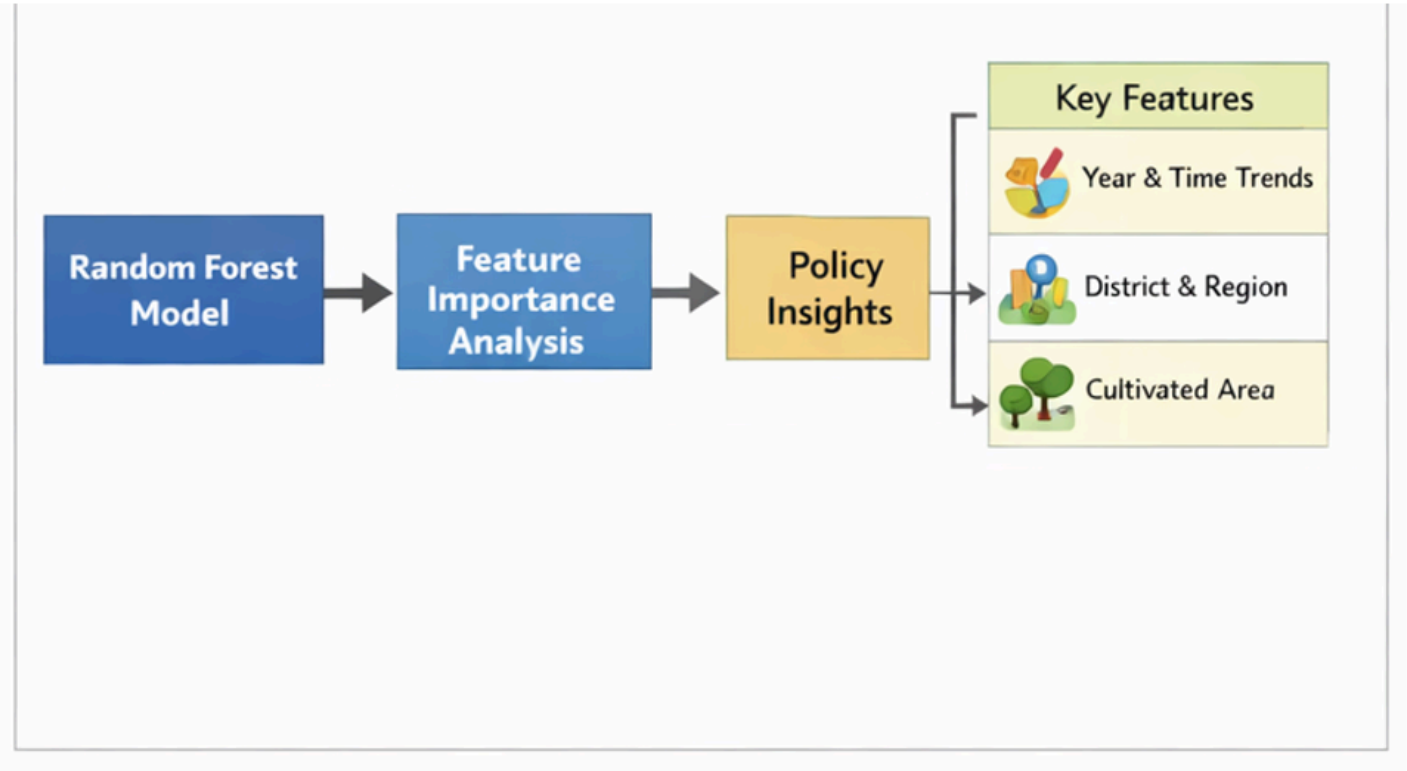


Figure 5

Feature importance analysis workflow for identifying key drivers of crop yield and supporting policy-relevant interpretation.

Feature importance analysis from the Random Forest model highlighting the relative contribution of spatial and temporal variables to crop yield prediction.