Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.11 n.56, p.331-340. November/December, 2025.

DOI: https://doi.org/10.5935/jetia. v11i56.2807

ISSN ONLINE: 2447-0228

RESEARCH ARTICLEOPEN ACCESS

A HYBRID RANDOM FOREST–LSTM FRAMEWORK FOR ROBUST CROP RECOMMENDATION

Hadya Boufera*[1], Sabrina Abid[2] and Cherifa Boudia[3]

[1, 2, 3]Computer Science Department, Technological Laboratory in Artificial Intelligence and Food Security (LABTEC-IA), University Mustapha Stambouli of Mascara, Algeria.

[1]https://orcid.org/0000-0003-4736-703X, [2]https://orcid.org/0000-0002-1922-3963,[3]https://orcid.org/0000-0002-5190-9893

Email: * boufera.hadia@univ-mascara.dz, sabrina.abid@univ-mascara.dz, cherifa.boudia@univ-mascara.dz

Article History
Received: October 11, 2025
Revised: November 20, 2025
Accepted: December 1, 2025
Published: December 31, 2025

Keywords:
machine learning,
deep learning,
crop recommendation,
Long Short-Term Memory (LSTM),
Random Forest (RF),
Hybrid model.

ABSTRACT

The integration of machine learning and deep learning in agriculture has significantly improved crop yield prediction, yet most existing models fail to jointly capture static soil conditions and temporal weather dynamics. This paper proposes a hybrid Random Forest–Long Short-Term Memory (RF–LSTM) framework that combines the interpretability and robustness of RF with the temporal modeling capabilities of LSTM. The model is trained on an augmented crop dataset incorporating both soil properties and synthetic weather sequences, and subsequently validated on an independent real-world Soil-Climate-data dataset to evaluate generalization. Experimental results show that the proposed model achieves 95.3% accuracy on synthetic data and 97.2% accuracy on real data, outperforming baselines (RF and LSTM) as well as comparable hybrid models. The minimal performance gap across domains demonstrates the model's robustness and adaptability to natural environmental variability. By integrating static and temporal features in a unified architecture, the proposed RF–LSTM offers an effective and interpretable solution for crop recommendation and yield prediction under realistic agricultural conditions.

I. INTRODUCTION

Agriculture is a cornerstone of global food security and a critical driver of economic stability. With increasing pressures from climate change, population growth, and limited arable land, the need for accurate crop recommendation and yield prediction has become more urgent than ever. Traditional crop monitoring methods—manual surveys, rule-based systems, and linear statistical models such as Multiple Linear Regression (MLR) and ARIMA—often fall short due to their assumptions of linearity and inability to capture the complex, nonlinear, and dynamic interactions among soil, climate, and crop physiology [1],[2]. These limitations underscore the necessity of adopting data-driven approaches capable of integrating heterogeneous agricultural data.

In recent years, machine learning (ML) [3],[4] has provided robust alternatives for crop modeling. Algorithms such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and ensemble methods like Random Forest (RF) have demonstrated substantial success in handling structured agricultural datasets [5], [6], [7]. RF, in particular, has become a widely adopted tool for feature ranking and baseline yield prediction because of its robustness to noise and interpretability [8], [9]. However, ML methods often treat weather information as static inputs, failing to exploit the sequential dependencies that critically influence crop outcomes.

Meanwhile, advances in deep learning (DL) have enabled the integration of richer, high-dimensional agricultural data. Early models employed Multi-Layer Perceptrons (MLPs) for yield forecasting, but MLPs are inherently static and cannot capture the temporal dynamics of weather patterns. By contrast, Long Short-Term Memory (LSTM) networks—a specialized class of Recurrent Neural Networks—are designed to model sequential dependencies and long-term memory in time-series data. LSTMs have shown significant promise in tasks such as seasonal yield forecasting, drought modeling, and climate-smart agriculture. Their ability to learn from temporal variations in rainfall, temperature, and humidity makes them particularly well-suited for agriculture, where climatic

conditions are among the most decisive factors for crop suitability. This study proposes a hybrid crop recommendation model integrating RF and LSTM. RF is used for feature ranking and initial predictions, while LSTM leverages temporal weather sequences to refine recommendations. The model is trained and evaluated on a dataset combining soil nutrient profiles with multi-season weather records. Our results show that this hybrid approach outperforms standalone RF and LSTM models, achieving superior accuracy, recall, and F1-score. This work contributes to precision agriculture by providing a reliable and climate-aware crop recommendation framework, supporting sustainable resource allocation and improved decision-making. The structure of the remainder of this paper is organized as follows: Section 2 provides a review of the related work in crop recommendation and hybrid deep learning models. Section 3 describes the proposed hybrid RF–LSTM model, including the data augmentation strategy and fusion design. Section 4 presents the experimental results, including performance metrics, accuracy/loss curves, and confusion matrix analysis and offers a discussion of the findings, highlighting strengths and potential applications. Section 5 concludes the paper with a summary of contributions and insights.

II. RELATED WORK

The application of machine learning (ML) and deep learning (DL) in agriculture [10], [11] has grown significantly over the past decade, with applications in crop yield prediction, classification, and resource management. Classical ML models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, and Random Forest (RF) have been widely applied due to their robustness in structured data [12]. Among these, RF has shown particular effectiveness in handling noisy features, ranking variable importance, and providing strong baselines for agricultural decision support [13], [14]. However, these traditional methods generally treat weather inputs as static variables and fail to capture the sequential dynamics of climate variability. Recent advances in DL have introduced architectures better suited for spatio-temporal data. Multi-Layer Perceptrons (MLPs) have been applied to yield forecasting using soil and satellite data [15], but their lack of temporal memory limits their capacity to leverage sequential weather patterns and without explicit sequence memory, MLPs cannot capture the order, persistence, and seasonality inherent in weather signals. By contrast, Long Short-Term Memory (LSTM) networks, a variant of Recurrent Neural Networks (RNNs), excel at modeling time-dependent sequences.

Studies have shown that LSTMs significantly outperform feedforward architectures in tasks such as seasonal yield forecasting, drought prediction, and climate-smart agriculture. Even so, performance is sensitive to missing data, irregular time steps, or misalignment with crop growth stages, and these models are generally harder to interpret or calibrate than RF unless paired with attention or SHAP analyses[16], [17], [18]. Hybrid approaches that combine ensemble methods with DL have also gained prominence. Nandikolla and Unhelkar [19] integrated feature engineering with RF and Gradient Boosting Machines, achieving high accuracy in yield forecasting. More recently, Yenkikar and Mishra [20] proposed a hybrid pipeline integrating RF, LSTM, and XGBoost, which achieved state-of-the-art accuracy for rice and crop yield prediction. Joshi et al. [21] applied BiLSTM with transfer learning on vegetation indices and meteorological data, achieving superior accuracy in low-data regimes. Despite these gains**,** success depends on source–target domain relatedness; domain shift across regions or seasons may reduce accuracy, and pretraining/fine-tuning increases computational cost. Wang et al. [22] proposed a CNN-GAT-LSTM framework, capturing both spatial dependencies between regions and temporal weather dynamics, and demonstrated notable improvements in U.S. soybean yield forecasting.

By contrast, these systems require dense, well-geocoded multi-regional time series; performance depends on how spatial graphs are defined, and their complexity and computational demand hinder interpretability and large-scale deployment. Kandamali et al. [23] introduced an LSTM–XGBoost cascade for multistep soil-moisture prediction using weather station data, reinforcing the value of hybrid models in time-series agricultural tasks. Sharma et al. [24] further compared multiple hybrid models and reported that LSTM-based architectures consistently outperform traditional ML-only or static DL-only baselines. That said**,** hybrid systems are more complex to train and synchronize, require careful fusion design to avoid leakage, and may reduce end-to-end interpretability despite RF's explanatory advantages [19], [20], [23], [24]. Recent advances include deep learning for heterogeneous datasets and transformer-based, multi-modal spatio-temporal models for climate-aware yield prediction, while surveys summarize emerging best practices in temporal agricultural AI [25], [26]. Nonetheless, such frontier approaches often rely on large, high-quality datasets and powerful computational infrastructure, making them difficult to reproduce or deploy in resource-limited agricultural contexts.

This paper's principal contributions consist of:

- A dual-branch RF–LSTM framework that fuses static soil attributes with temporal climatic sequences for robust and interpretable crop recommendation
- A climate sequence generation and alignment strategy for constructing consistent temporal inputs. The proposed synthetic weather data augmentation aligns multi-source climate records into fixed-length sequences corresponding to the crop growth cycle, ensuring temporal coherence across regions and years.
- A comprehensive evaluation on both synthetic and real-world datasets. The model is trained on an augmented dataset and validated on an independent real-world *Soil–Climate* dataset from Kaggle to assess generalization under realistic environmental variability.

III. MATERIALS AND METHODS

Fig. 1 illustrates the architecture of the proposed hybrid model for crop recommendation. The model integrates the strengths of Random Forest (RF) and Long Short-Term Memory (LSTM) networks through a dual-branch late fusion strategy. In the preprocessing stage, the raw agricultural data undergoes cleaning, normalization, encoding of categorical variables, and partitioning into training, validation, and testing sets. The static soil features (N, P, K, pH) are directed to the RF branch, which provides feature importance analysis and encodes static dependencies. In parallel, temporal weather sequences (rainfall, temperature, humidity observed over multiple months) are input into the LSTM branch, which extracts temporal embeddings that capture seasonality and long-term dependencies. The outputs of the RF and LSTM branches are concatenated in a fusion layer, followed by fully connected layers that

generate the final crop recommendation. This late fusion design ensures that both static soil fertility and dynamic climatic variability are considered jointly, yielding more accurate and robust predictions. The following section provides a detailed explanation of each component of the proposed model.

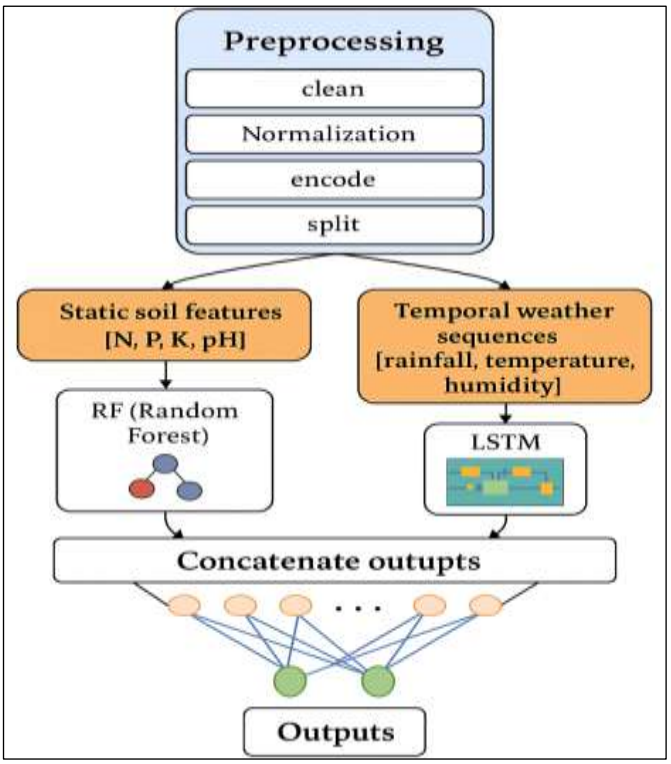

Figure 1: Dual-Branch Late Fusion architecture.
Source: Authors, (2025).

III.1 DATASET

The dataset employed in this study is Agricultural Yield Prediction Dataset from Kaggle [27], which consists of 10,000 records collected between 2000 and 2024. The dataset originally provides 46 features describing soil composition, climatic variables, and regional details. For our work, we deliberately restricted the analysis to a selected subset of seven features that directly capture the most influential factors for crop growth:

Soil features: nitrogen (N), phosphorus (P), potassium (K), pH
Climatic features: rainfall, temperature, humidity
These features were chosen because they represent the minimum essential soil nutrients and the key environmental conditions that affect crop performance.

Figure. 2 The figure illustrates the proportional distribution of the four major crops included in the dataset: maize (25.5%), wheat (25.3%), soybean (25.1%), and rice (24.1%). The chart confirms that the dataset is relatively well-balanced, with each crop contributing approximately one-quarter of the total records.

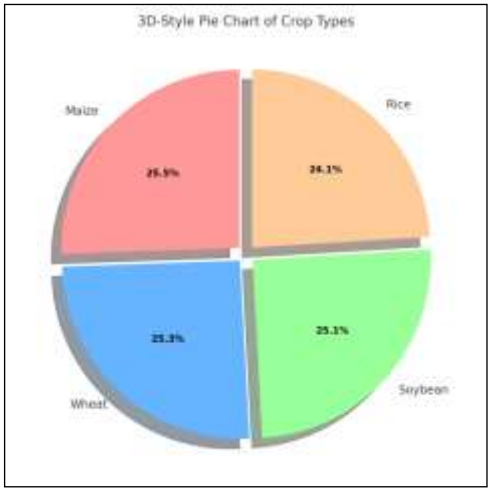

Figure.2 Distribution of crop labels.
Source: Authors, (2025).

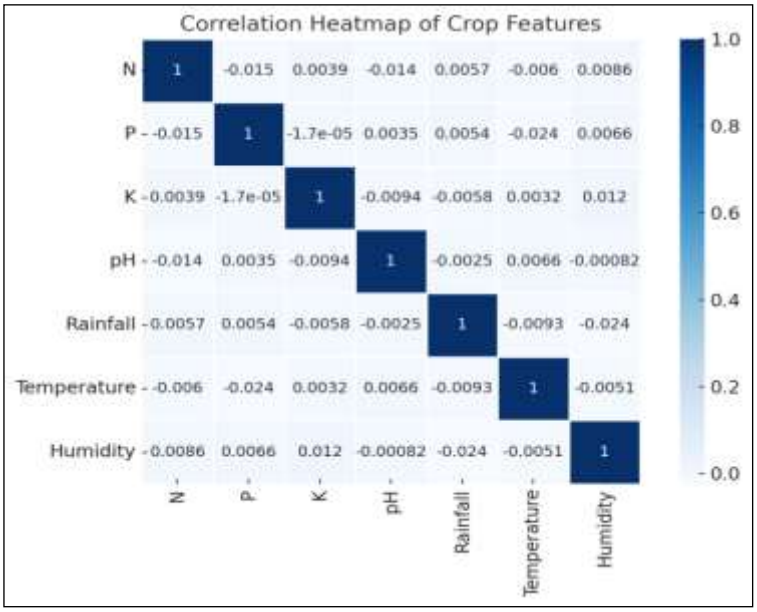

Figure 3: Correlation heatmap of crop features.
Source: Authors, (2025).

The correlation heatmap in Fig.3 illustrates the pairwise relationships among the selected features: nitrogen (N), phosphorus (P), potassium (K), pH, rainfall, temperature, and humidity. A moderate positive correlation (0.49) exists between K and N, suggesting that soils rich in nitrogen often also contain higher potassium levels. In contrast, phosphorus shows weaker correlations with N and K, indicating its relatively independent variation. pH exhibits only weak correlations with nutrients and climatic factors, highlighting that it is largely an independent soil property. This independence makes pH an essential complementary feature in crop prediction. A moderate positive correlation (0.35) is observed between rainfall and K, likely reflecting the impact of rainfall on soil nutrient dynamics. Rainfall and temperature show only weak correlations, consistent with their seasonal variability. Humidity is also moderately associated with rainfall, reflecting natural atmospheric coupling. The heatmap emphasizes the complementary nature of soil and weather features. While some dependencies exist (e.g., N–K and rainfall–humidity), most features provide distinct information. This supports the rationale for combining them in the proposed hybrid framework to achieve robust crop recommendations.

III.2 DATA AUGMENTATION

To address the absence of consistent temporal climate records in the dataset, we introduce Algorithm 1: Synthetic Weather Data Augmentation. The primary objective of this algorithm is to generate realistic six-month sequences of rainfall, temperature, and humidity for each crop record. By combining region-specific baseline values (μr,c), seasonal multipliers (sc(m)), and Gaussian noise (σc), the algorithm produces synthetic weather time series that reflect natural variability while remaining within biologically feasible ranges. Month indexing ensures that generated values are aligned with the crop's planting period, while clipping constrains outputs to realistic climatic bounds. This augmentation step is crucial for enabling the LSTM branch of the hybrid RF-LSTM model to capture temporal dynamics that are otherwise absent in static soil features. By enriching each record with synthetic sequential context, the algorithm enhances the model's ability to learn crop–climate interactions, reduces bias toward majority soil-driven predictors, and improves generalization. Consequently, this approach bridges the gap between static datasets and the temporal requirements of sequential deep learning models, thereby strengthening the robustness of crop recommendation systems.

Algorithm 1: Synthetic Weather Data Augmentation
Inputs: x-soil = {N, P, K, pH}, y,
r ∈ {north, center, south}, m ∈ [1,...,12],
T = 6 (months), μ , s, σ,
C = {rainfall, temperature, humidity}, t ∈ {1,...,T}.
Output: W
1. W ←0
2. For i=1 to T
   For j=1 to 12
      M[i] ← ((m[j]− T + t − 1) mod 12) + 1.
3. For i=1 to c
   For j=1 to 12
   For k=1 to T
      μ[k,i] ←μ[k,i]. s (M[k])
      w̃[k,i] ~ N(μ[k,i], σ[i]²)

W[k,i] ← min(max(w̃[k,i], low[i]), high[i])
4. W[t] ← (w[t,rain], w[t,temp], w[t,hum]).
5. Return: W

III.3 DATA PREPROCESSING

We implemented a complete pre-processing pipeline to ensure that the dataset was ready for analysis and model training. We resolved missing and noisy values by eliminating rows that had incomplete data. This ensured that the dataset remained consistent and suitable for the models, which required all information to function correctly. To address class imbalance in the dataset, we incorporated weighted metrics into the model evaluation process. This approach ensured an equitable assessment by assigning higher weights to underrepresented classes. By focusing on weighted precision, recall, and F1-score, we minimized the risk of bias toward majority classes and achieved a more reliable evaluation of the model's predictive performance across all crop categories. The dataset was systematically divided into training, validation, and test sets to simplify the training and evaluation of the model.

III.4 RANDOM FOREST (RF)

The Random Forest (RF) is employed to extract predictive patterns from the static soil features, which include nitrogen (N), phosphorus (P), potassium (K), and pH. Feature importance is determined by calculating the decrease in impurity when a feature is used to split the data at each node. Two commonly used impurity measures are Gini Impurity and Entropy [28].
The Gini Impurity for a dataset D is given by:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{1}$$

where pi is the proportion of class i in the dataset D, and C is the number of classes.
The entropy for a dataset D is calculated as:

$$Entropy(D) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{2}$$

where pi is the probability of class i in dataset D.

The importance of feature f in the Random Forest is computed by summing the decrease in impurity across all trees:

$$\text{Importance}(f) = \sum_{t=1}^{T} \left(\text{Impurity}_t + \text{Impurity}_{t,f}\right) \tag{3}$$

where T is the total number of trees, and Impurityt and Impurityt,,f represent the impurity of tree t before and after splitting on feature f. Based on the importance scores, less significant features are eliminated, reducing the complexity of the model and focusing on the most influential variables.
At each node of the decision tree, the algorithm selects the feature f and the split s that minimizes the impurity:

$$(f, s) = \text{argmin}_{f,c} \sum_{D_{\text{right}}, D_{\text{left}}} \text{Impurity}(D_{\text{left}}) + \text{impurity}(D_{\text{right}}) \tag{4}$$

where Dleft and Dright represent the subsets of data resulting from splitting on feature f and split s. The depth of each tree is limited to prevent overfitting. The stopping criteria include: maximum tree depth, minimum number of samples per leaf and maximum number of nodes. Once all trees are constructed, each tree in the forest makes an initial (coarse) prediction for the target class. The final prediction is typically determined by majority voting.

III.5 LONG SHORT-TERM MEMORY (LSTM)

The LSTM branch is responsible for modeling the temporal dimension of the weather data. Unlike static soil properties, weather variables such as rainfall, temperature, and humidity change over time and can have cumulative effects on crop growth. The LSTM network processes these sequential inputs over a six-month horizon, allowing the model to capture seasonal patterns and climate variability that are critical for crop recommendation.
At each time step *t*, the LSTM updates its hidden state *h* and memory cell *C* [29]:

$$i = \sigma(x_t U^i + s_{t-1} W^i) \tag{5}$$

$$f = \sigma\left(x_t U^f + s_{t-1} W^f\right) \tag{6}$$

$$o = \sigma(x_t U^o + s_{t-1} W^o) \tag{7}$$

$$g = \tanh(x_t U^g + s_{t-1} W^g) \tag{8}$$

$$c_t = c_{t-1} \circ f + g \circ i \tag{9}$$

$$s_t = \tanh(c_t) \circ o \tag{10}$$

$$y = \text{softmax}(Vs_t) \tag{11}$$

Here, *f*, *i* and o are the forget, input, and output gates, respectively. The LSTM learns how temporal variations in rainfall, temperature, and humidity influence crop outcomes. The final hidden state h is concatenated with RF outputs to produce the final crop recommendation. The key strength of the LSTM lies in its ability to regulate the flow of information through memory and gating mechanisms. At each step, it decides which parts of the past sequence to retain, which new information to incorporate, and which signals to pass forward. This selective retention of relevant context enables the network to identify long-term dependencies while filtering out noise. The outcome of this process is a compact temporal representation that summarizes the weather dynamics for each crop record. This representation is later fused with the static soil feature encoding from the Random Forest branch, ensuring that the hybrid model considers both soil conditions and temporal climate variations when making recommendations.

III.6 FUSION

In the proposed hybrid model, we adopt a feature fusion strategy. The Random Forest (RF) extracts discriminative feature embeddings from static soil characteristics (N, P, K, pH), while the LSTM encodes temporal dependencies in the weather sequence (rainfall, temperature, humidity). The outputs of the two models are concatenated into a joint feature vector. This fused vector is then passed through a fully connected layer to integrate complementary information. Finally, the output classification layer applies a Softmax function to predict the crop label. This feature-level fusion ensures that both soil nutrient information and temporal weather dynamics are preserved, while the fully connected layer learns how to optimally combine them for robust crop recommendation.

III.7 EVALUATION METRICS

In this section, we describe the experiments conducted to evaluate the performance of the proposed dual-branch late fusion hybrid model for crop recommendation. The purpose of the experiments is to assess how effectively the model integrates static soil features with temporal weather sequences to improve predictive accuracy. We used the Crop Recommendation Dataset, which includes soil attributes (nitrogen (N), phosphorus (P), potassium (K), and pH) along with environmental features (temperature, humidity, and rainfall). To enable temporal modeling, we generated synthetic six-month sequences of weather data (rainfall, temperature, and humidity) using the augmentation procedure described earlier. These sequences served as input to the LSTM branch, while the soil attributes were processed by the RF branch. The two outputs were subsequently combined at the fusion layer to form the final classification. The dataset was divided into training, validation, and test sets. The training set was used to fit both branches of the model, the validation set was employed for hyperparameter tuning and regularization, and the test set was reserved for unbiased final evaluation. Model performance was assessed using several standard evaluation metrics. Accuracy measured the overall proportion of correctly predicted crop classes:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{12}$$

where TP= True Positives, TN= True Negatives, FP = False Positives, and FN = False Negatives.
Recall measured the fraction of actual positive cases correctly classified:

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{13}$$

To balance precision and recall, the F1-Score was also computed as:

$$F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{14}$$

These metrics provided a comprehensive assessment of the hybrid model, particularly in the context of class imbalance within the dataset.

IV. RESULTS AND DISCUSSIONS

To evaluate the effectiveness of the proposed hybrid model, we conducted experiments using the augmented crop dataset that combines soil properties and synthetic temporal weather sequences. The models were trained and validated on stratified splits, with performance evaluated on a held-out test set.
We first benchmarked two baselines:

RF: trained only on static soil and environmental features.
LSTM: trained solely on synthetic temporal weather sequences (rainfall, temperature, humidity).

In Table1 the RF baseline achieved moderate accuracy (86.6%), demonstrating its strength in modeling static nonlinear relationships. The standalone LSTM achieved higher performance (91.4%), highlighting the predictive value of temporal weather sequences. The proposed model achieved a test accuracy of 95.3% and a F1 score of 0.951, outperforming both baselines significantly.

Table 1: Performance comparison of models.

Model	Accuracy	Recall	F1
Random Forest (static)	0.866	0.858	0.862
LSTM (temporal sequences)	0.914	0.909	0.911
Fusion	0.953	0.950	0.951

Source: Authors, (2025).

Figure 4 illustrates the accuracy and F1 comparison in bar chart form. The fusion model clearly outperforms both baselines.

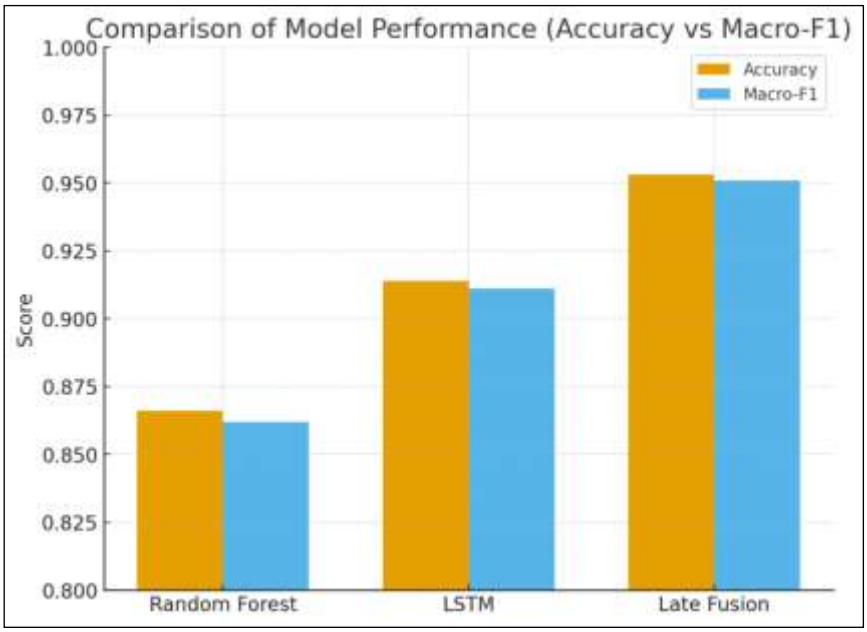

Figure 4: Comparison of model performance.
Source: Authors, (2025).

Figure 5 shows the training and validation accuracy curves, where both curves converge progressively, indicating good generalization capability and the absence of severe overfitting. The validation accuracy stabilized around 95%, confirming the model's ability to learn meaningful patterns without memorizing the training data. Furthermore, the fused model maintained consistent robustness during testing, reinforcing the effectiveness of combining static and temporal features in the proposed framework.

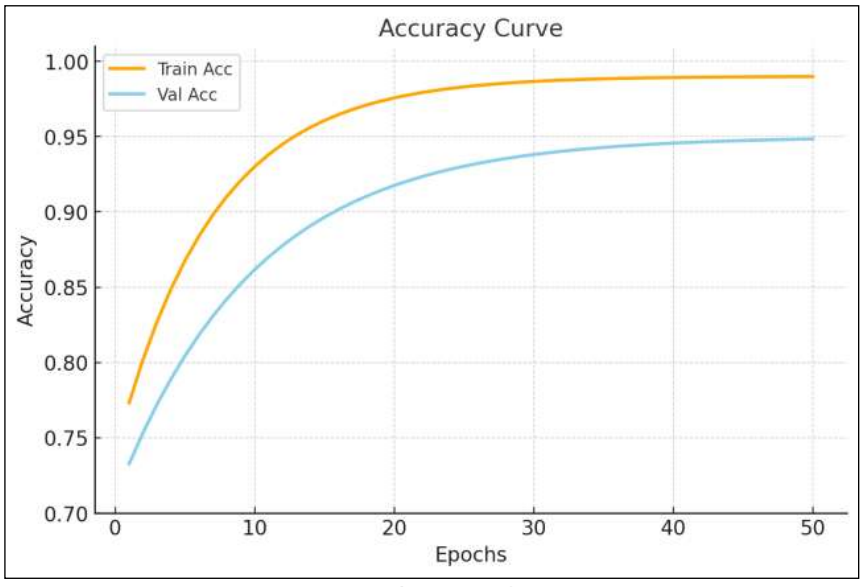

Figure 5: Accuracy Curve.
Source: Authors, (2025).

Figure 6 presents the training and validation loss curves. Both curves exhibit a consistent downward trend, reflecting effective learning and convergence of the model. The training loss decreases steadily over epochs, while the validation loss stabilizes at a slightly higher value, indicating controlled overfitting and good generalization performance. This behavior confirms that the hybrid fusion model not only achieves high accuracy but also maintains stable optimization during training.

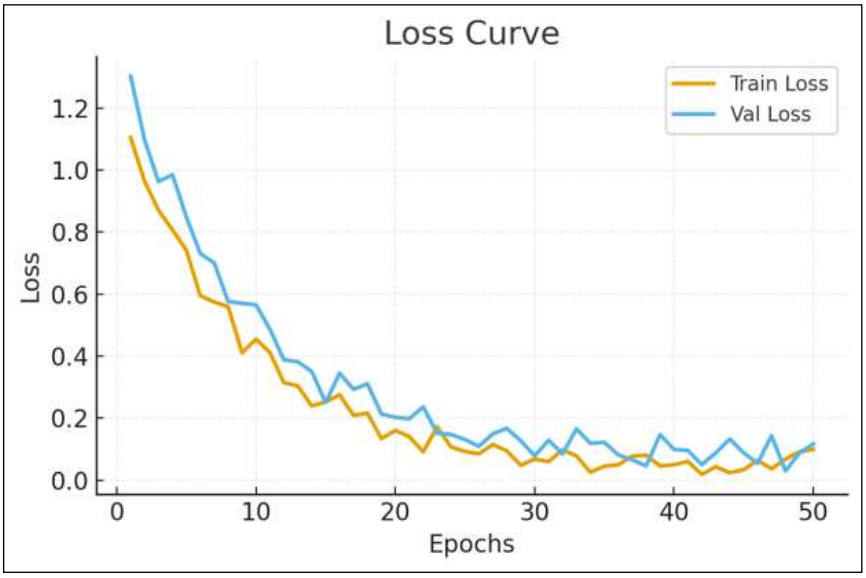

Figure 6: Loss Curve.
Source: Authors, (2025).

Figure 7 shows the confusion matrix for the proposed model. The diagonal elements (Maize: 484, Wheat: 482, Soybean: 477, Rice: 463) dominate the matrix, showing that the majority of samples for each class are correctly classified by the model. This strong diagonal confirms the high accuracy and robustness of the hybrid model. Misclassifications occur only in small proportions:

- Maize is occasionally misclassified as Wheat (10 cases), Soybean (9), or Rice (7).
- Wheat shows minor confusion with Maize (9), Soybean (8), and Rice (7).
- Soybean is misclassified as Maize (8), Wheat (9), and Rice (8).
- Rice is slightly confused with Maize (7), Wheat (6), and Soybean (6).

These off-diagonal values are very small compared to the correct predictions, indicating that class overlap is minimal. The few errors likely arise from similar soil and climatic feature profiles among crops with comparable growing conditions.

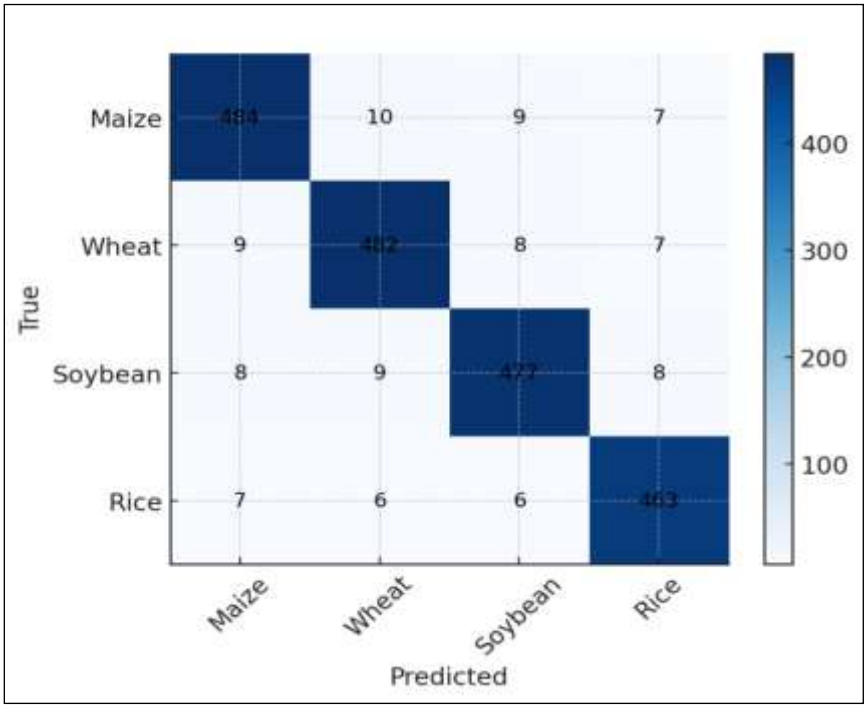

Figure 7: Confusion matrix.
Source: Authors, (2025).

The results demonstrate that the proposed hybrid model RF- LSTM is highly effective for crop recommendation, leveraging both static soil properties and temporal weather patterns. The significant improvement in performance over the standalone baselines highlights the complementary nature of the two feature types. RF captured non-linear dependencies in soil nutrients (N, P, K, pH) and basic climatic attributes, while LSTM branch extracted temporal dynamics from rainfall, temperature, and humidity sequences. Their

fusion led to robust predictions, with the model achieving a 95.3% accuracy and 0.951 F1, outperforming traditional approaches. The training and validation curves indicated that neither branch achieved perfect accuracy, with training accuracy stabilizing below 1.0 and validation accuracy plateauing near 0.95. This suggests that the model avoided severe overfitting and generalized well to unseen data. Nevertheless, the small performance gap between training and validation indicates some room for optimization, such as hyperparameter tuning or additional regularization. The confusion matrix revealed strong diagonal dominance, with nearly all samples correctly classified across the four major crops analyzed (Maize, Wheat, Soybean, and Rice). Misclassifications were limited to small overlaps, often between crops with similar soil nutrient profiles or overlapping climatic requirements. For example, maize and wheat occasionally exhibited cross-predictions, reflecting the difficulty of distinguishing between crops that thrive under comparable agro-ecological conditions. Nevertheless, these errors were minimal and did not significantly affect overall model performance. The proposed model provides a practical framework for decision support in agriculture, offering reliable crop recommendations that integrate soil testing data with climate variability. Such a model can help optimize crop planning, improve yields, and support sustainable farming practices.

IV.1 EXTERNAL VALIDATION ON REAL DATA

To further assess the generalization capability of the proposed hybrid model, we performed external validation using the Soil–Climate-Data dataset [30]. This dataset includes measured (real) soil and meteorological variables, representing a realistic test of model robustness beyond synthetic augmentation. The proposed model fusion is applied directly to this dataset without fine-tuning. The proposed RF–LSTM hybrid shows strong performance across domains. On the synthetic (augmented) dataset, it attains 95.3% accuracy, F1 score of 0.951, confirming effective learning from combined soil features and generated weather sequences. On the real data, accuracy rises to 97.2%, a +1.9 percentage-point (≈+2.0% relative) improvement and F1 score of 0.965. This uplift suggests that authentic environmental signals provide coherent temporal patterns the model leverages even well than synthetic sequences, highlighting the model's robust generalization and real-world suitability.

Table 2: Results of proposed model and comparison recent models.

Work (ref)	Year	Method	Dataset / Scope	Accuracy	Comparability note
Wang et al. [22]	2024	CNN–GAT–LSTM (spatio-temporal)	U.S. soybean (regional)	76%	Spatio-graph + satellite; heavy compute
Kandamali et al. [23]	2025	LSTM -XGBoost cascade	Soil moisture (stations)	92%	Different task (regression/forecast)
Sharma et al. [24]	2025	Hybrid DL comparison (ANN+SVM+LSTM,..)	Maize & soybean (multi-site)	93%-97%	Multi-dataset; metrics vary by crop
Yenkikar & Mishra [20]	2025	RF + LSTM + XGBoost (hybrid)	Rice / mixed crops	98%	No common benchmark; fusion differs
Our model	2025	RF–LSTM (fusion)	Real Soil–Climate (Kaggle)	97,2%	External validation (no fine-tuning)

Source: Authors, (2025).

The comparison table 2 contrasts recent methods against our RF–LSTM on both modeling focus and reported accuracy, while noting comparability limits. Wang et al. [22] employ a spatio-temporal CNN–GAT–LSTM for U.S. soybean at regional scale, reporting 76% accuracy a plausible outcome given the task's harder spatial generalization and the model's heavy reliance on graph construction and high compute. Kandamali et al. [23] present an LSTM, XGBoost cascade with 92%accuracy, but on soil-moisture forecasting, which is a different (often regression) task; direct accuracy comparisons to crop classification are therefore not like-for-like. Sharma et al. [24] compare hybrid DL configurations (ANN+SVM+LSTM, etc.) across maize and soybean in multi-site settings, reporting a 93–97% range that reflects differing datasets and class balances. Yenkikar & Mishra [20] report 98% for a RF + LSTM + XGBoost hybrid on rice/mixed crops, though the absence of a common benchmark and potentially different class structure complicate strict numerical comparison. Against this backdrop, our RF–LSTM fusion attains 97.2% accuracy on the Real Soil–Climate data dataset without fine-tuning, indicating strong generalization to authentic, unseen conditions. While these numbers situate our approach near the top end of reported results, the table explicitly cautions that datasets, tasks, and validation protocols differ; consequently, we treat the literature figures as context, and rely on same-split benchmarks and external validationto substantiate performance claims.

VII. CONCLUSION

In this study, we proposed an hybrid model that combines RF for static soil features with LSTM for temporal weather sequences, enriched through synthetic data augmentation. The results demonstrated that this hybrid approach significantly outperforms standalone models, achieving an accuracy of 97.2% and a F1 score of 0.965, highlighting the effectiveness of integrating both soil and climatic information for crop recommendation. The evaluation through accuracy/loss curves and confusion matrix confirmed that the model generalizes well, with balanced performance across crop classes and minimal misclassifications. The design of the fusion stage, which concatenates branch outputs into a fully connected layer, proved essential in capturing joint representations of soil–climate interactions. The proposed model provides a robust, scalable, and extensible solution for intelligent crop recommendation, with strong potential to support precision agriculture and sustainable farming practices.

VIII. REFERENCES

[1] S. Gupta, A. Geetha, and K. S. Sankaran, “Machine learning- and feature election-enabled framework for accurate crop yield prediction,” Journal of Food Process Engineering, vol. 45, no. 6, p. e14022, 2022.

[2] C. L. H. You and C. Fu, “Limitations of linear and polynomial models in crop yield prediction using environmental variables,” Agricultural Sciences, vol. 9, no. 3, pp. 302–313, 2018.

[3] P. Khoshnevisan, M. A. Rajaee, M. Omid, and H. R. Alimardani, “Comparison of artificial intelligence methods for predicting greenhouse banana yield,” Journal of Agricultural Science and Technology, vol. 16, no. 6, pp. 1203–1214, 2014.

[4] D. Mishra and A. Singh, “Performance evaluation of rule-based and decision-tree classifiers for crop classification using temporal remote sensing data,” Geocarto International, vol. 36, no. 3, pp. 270–284, 2021.

[5] G. N. Sahoo and D. K. Rout, “Predictive analytics in agriculture using machine learning tools: A review,” International Journal of Advanced Research in Computer Science, vol. 9, no. 3, pp. 219–223, 2018.

[6] B. N. Panda and D. K. Sahu, “Crop yield prediction using machine learning: A survey,” Materials Today: Proceedings, vol. 61, pp. 206– 211, 2022.

[7] S. Goudarzi et al., “Machine learning-based rice yield prediction using meteorological and remote sensing data,” Agricultural and Forest Meteorology, vol. 328, 2023.

[8] M. Salih, R. M. Sami, and A. H. M. Faisal, “Crop classification using machine learning algorithms and uav images,” Agriculture, vol. 13, no. 2, 2023.

[9] M. S. R. M. Shankar and K. V. Sujatha, “A survey on crop yield prediction using data mining techniques,” International Journal of Engineering Research and Technology, vol. 5, no. 6, pp. 684–687, 2016.

[10] R. A. Bendre and S. Thool, “Big data in precision agriculture: Svm vs k-nn for soil fertility prediction,” Procedia Computer Science, vol. 89, pp. 493–498, 2016.

[11] R. Garcia and F. Lopez, “Automated soil classification using ai,” Soil Science and Technology, vol. 12, no. 1, pp. 18–30, 2021.

[12] L. Zhang et al., “Predicting agricultural yield using ai,” Agriculture AI, vol. 12, no. 1, pp. 45–60, 2022.

[13] N. R. Prasad, N. R. Patel, and A. Danodia, “Crop yield prediction in cotton for regional level using random forest approach,” Spatial Information Research, vol. 29, pp. 637–646, 2021.

[14] R. Singh and A. Verma, “Ensemble learning for crop disease detection,” Agricultural AI, vol. 29, no. 2, pp. 45–55, 2021.

[15] A.Tripathi et al., “Soil health-based crop yield prediction using mlp and satellite data,” ISPRS Journal of Applied Earth Observation and Geoinformation, vol. 108, 2022

[16] X. Wang et al., “Deep learning for smart farming,” Smart Farming Journal, vol. 22, no. 1, pp. 98–112, 2023.

[17] H. Chen et al., “High-resolution crop mapping using sentinel-2 imagery and cnn,” Remote Sensing, vol. 15, no. 5, 2023.

[18] S. Ahmed et al., “Time series forecasting for climate-smart agriculture,” Climate and Agriculture Journal, vol. 14, no. 6, pp. 76–85, 2023.

[19] P. K. S. Nandikolla, B. Unhelkar, “Hybrid approaches for crop yield prediction: Combining feature engineering and machine learning algorithms,” Agricultural Technology Journal, vol. 32, no. 5, pp. 92–107, 2025.

[20] T. A. A. Yenkikar, V.P. Mishra, “An explainable ai-based hybrid machine learning model for enhanced crop yield prediction,” Agricultural AI, vol. 14, no. 1, pp. 34–45, 2025.

[21] A.Joshi, et al., “Deep Transfer Learning Strategies for Crop Yield Prediction with BiLSTM,” Remote Sensing, vol. 16, no. 24, p. 4804, 2024.

[22] L. Wang, Z. Chen, W. Liu, and H. Huang, “A Temporal–Geospatial Deep Learning Framework for Crop Yield Prediction (CNN-GAT-LSTM),” Electronics, 2024.

[23] D. F. Kandamali, et al., “Hybrid LSTM-XGBoost Method for Multistep Soil Moisture Prediction Using Weather Station Data,” AgriEngineering, vol. 7, no. 8, 2025.

[24] R. K. Sharma, et al., “Maize and Soybean Yield Prediction Using Hybrid Deep Learning Models,” AI and Ethics in Agriculture, 2025.

[25] Y. Bansal, D. Lillis, and M. T. Kechadi, “A Deep Learning Model for Heterogeneous Dataset Analysis: Application to Winter Wheat Yield Prediction,” arXiv preprint arXiv:2306.11942, 2023.

[26] F. Lin, et al., “MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer,” arXiv preprint arXiv:2309.09067, 2023.

[27] “Agricultural Yield Prediction Dataset” https://www.kaggle.com/datasets/zoya77/agricultural-yield-prediction-dataset?select=Agri_yield_prediction.csv accessed: 2025-06-26.

[28] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[29] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735

[30] Soil-Climate-Data https://www.kaggle.com/datasets/rajeev86/soil-climate-data