

Enhancing Agricultural Forecasting: A Hybrid LightGBM- BiLSTM Ensemble for Crop Yield Prediction

Sagana C¹, Manjula Devi R², Thangatamilan M³, Adithyan M⁴, Ajay M⁵ and Chinnathambi C⁶

¹Kongu Engineering College, Perundurai, Tamil Nadu, India

Email: sagana.c@gmail.com

²⁻⁶KPR Institute of Engineering and Technology, Coimbatore, India

Email: rmanjuladevi.gem@gmail.com, thangatamil.m@gmail.com, adithyanm.22cse@kongu.edu, ajaym.22cse@kongu.edu, chinnathambic.22cse@kongu.edu

Abstract— When it comes to farmers in Tamil Nadu, it is necessary to predict crop yield correctly since the result of each harvest determines their own livelihood and food provision of the whole area. In an attempt to introduce some form of certainty into this process of high uncertainty, we carried out thorough research to determine the most trustworthy and effective AI-based prediction model of crop yield. We assessed different machine learning models, such as Random Forest, Stacking Ensemble, and hybrid models that stacked gradient boosting and deep learning models. The optimized random forest model had achieved R-squared value of 0.914 which is around 91 percent. More sophisticated approaches, including Stacking Ensembles and hybrid ones, which employed LightGBM to extract features and neural networks to learn about the non-linear data. To achieve a higher level of accuracy, we created a new hybrid system that has LightGBM and a Bidirectional LSTM (BiLSTM) network. The LightGBM element effectively learns the interaction of complex features in agricultural data whereas the BiLSTM element learns the temporal and contextual relationships among the features. An impressive model was obtained with the suggested LightGBM + BiLSTM hybrid ensemble which achieved R-Squared of 0.9214. This study proves that machine learning and deep learning methods can be systematically experimented and integrated to create strong predictive models that can be used to make data-driven decisions about sustainable agriculture in Tamil Nadu and other regions.

Index Terms— Predictive Models, Machine Learning, Deep Learning, Stacking Ensemble, Crop Yield Prediction.

I. INTRODUCTION

Within the agricultural scene of the Tamil Nadu region, the prosperity of the farmer is often determined by one, important question which may only be answered once the season is done: What will my harvest be? The solution to this query reverberates throughout the complete farming ecosystem as it not only affects the economic wellbeing of individual farming families, it also affects the food security and economic policies of the region. It has been experience, intuition and historical record keeping, a fraught process, over decades this prophecy has been a blend of the three. The age of data though is a chance to apply this uncertainty to calculated insight.

It is the enormous volumes of agricultural data that have been gathered over the years that contain concealed patterns and intricate relationships that cannot be discerned through conventional analysis. This is where machine learning (ML) and deep learning (DL) can change the game, as they give us the means to create a model that can learn the past to make predictions that are as accurate as possible in the future.

Our personal research process commenced with the investigation of diverse models of great power, including Random Forests to complex gradient boosting models. Although some of them worked well, we found that there was no single architecture that was well-suited to cope with the entire complexity of the data. This prompted the main hypothesis of our work that a real breakthrough in performance would be reached not through the discovery of a more successful single model, but through the creation of a more successful team. This paper presents a new hybrid architecture, which exactly does so. We have been able to structure a collaboration between two expert artificial intelligence scientists.

The initial one is LightGBM, the latest generation of gradient boosting models that are fast and have the outstanding capability of identifying complex, non-linear forms in structured data such as farm records. The second one is a Bi-directional Long Short-Term Memory (BiLSTM) network, which is a deep learning model with strong context and sequence capabilities in information, where the features are viewed as an interrelated narrative.

By combining the forecasts of these two different yet complementary models we have come up with a forecasting system that proves to be more accurate and robust than all those of the single models we tested. The following sections will explain the design, implementation and evaluation of this hybrid LightGBM + BiLSTM model. We will display our approach, both in data preprocessing and the details of the ensemble architecture, and give a clear explanation of its better performance. It is a useful blueprint of what to make next generation agricultural implementations that would bring a new dimension of clarity and confidence to the agricultural societies of Tamil Nadu and beyond.

II. RELATED WORKS

According to [1], climate change has had a severely negative impact on crop yields in India over the last twenty years. The authors propose an interactive web-based prediction model using Random Forest to predict crop yield with respect to crop-specific weather conditions such as temperature, humidity, rainfall, and soil moisture [1]. Similarly, [2] implemented several ML methods — Random Forest, Support Vector Machine, Gradient Descent, LSTM, and Lasso — for crop yield estimation. They report Random Forest outperforming others with metrics around $R^2 = 0.963$, $RMSE = 0.035$, and $MAE = 0.0251$, and propose practical applications for crop selection. [3] shows that in recent years agriculture forecasting has increasingly used AI techniques. Common features include temperature, soil type, and vegetation indices, while frequently applied algorithms are Linear Regression, Random Forest, Gradient Boosting, CNN and LSTM. Hybrid ML-based models have also been explored. In their study of Swiss farms, [4] combined farm data with eleven climate indices (ECA&D) and compared Random Forest and boosting methods. Random Forest achieved strong performance (reported as $R^2 = 0.9589$, $MSE = 908.80$), showing ML's potential for practical yield prediction under climate influences. [5] proposed an Improved Optimizer Function (IOF) embedded in LSTM to enhance crop yield prediction for Andhra Pradesh (India). Using rainfall, wind, temperature, solar radiation and crop stats, they addressed underfitting/overfitting and reported $RMSE = 2.19$, outperforming CNN, RNN and vanilla LSTM.

A hybrid approach combining Random Forest, Decision Tree, ANN and deep models (RNN, LSTM) was presented by [6].

They emphasize the influence of soil nutrients (N, P, K), crop rotation, temperature and precipitation on yield and on selecting the best crop for a given land. [7] discuss how land type, resource availability and weather changes affect crop yield. Their findings show temperature and rainfall have the greatest impact, and they encourage the use of advanced ML and DL methods for improved accuracy. According to [8], genotype, environment and genotype-by-environment interactions are central to yield. They applied a Deep Neural Network (DNN) to maize yield prediction and reported higher accuracy than lasso, shallow networks and regression trees.

Large-scale studies such as [9] use environmental and soil factors (State, District, Crop, season, space, production) on hundreds of thousands of records (250k observations) and evaluate Random Forest, Gradient Boost and Decision Tree ensembles to reduce errors and improve predictions. [10] used historical climatic and groundwater data for yield-related analyses. They applied KNN for ground-water classification and future value prediction to assess water availability for crop production.

III. PROPOSED METHODOLOGY

The method of constructing the prediction model was a planned, step-by-step process. This process began with a solid basis of information, carefully ready it, searched a vault of mighty implements, and lastly contrived a tailor made, problem that is to be solved in the most accurate way. This section describes each of those vital steps.

A. Dataset Description

There was a historical beginning of our research. data bank of agricultural records in all of Tamil. Nadu, spanning from 1997 to 2022. To ensure our model we could learn to do good things by examples. curated version of this data, preprocessed (gt 200).csv. which specializes in crops with a greater level of historical information. This data gave us the Five significant inputs into our model are district, crop, season, area. (in hectares) and year—which we promised our single target: the final production in kilos per hectare.

B. Data Preprocessing

Raw data, such as a raw material, must be cleaned, then it can be put to use such that it is powerful. Our data preparation was an important process in data preprocessing. The refining process began by focusing on the most crucial records of cultivation. Our data were reduced to those records where the production values were above 200. This helped to remove potential data entry noise and made the model more stable, useful for agriculture and reliable. One of our problems were the immense difference in the yields between different crops. Crops like sugarcane can be used to produce very large yield compared to rice or wheat. This massive scale difference may defeat a model. To solve this, we applied a logarithmic transformation (np.log1p) to reduce the range of yield values such that the model can learn subtle details more effectively. The prediction of the model was converted back to original scale for our final results.

C. System Architecture and Workflow

Our hybrid system architecture as shown in figure 1 is a shift from the single-modeled system of the past to an advanced two-expert interaction system. On top of the full data-preprocessing pipeline described earlier in the sections above, our methodology proposes a new ensemble architecture that balances the mutually complementary strengths of both the gradient boosting and deep learning paradigm.

D. Models Evaluated

Our search to the best prediction model was an extensive and competitive analysis. It was not a case of selecting a model but a set of experiments when we tested a variety of strong architectures with their own distinct learning mechanisms on the data. This intensive method enabled us to compare best practices of high performance and get to know which combination would give the optimal results.

Random Forest

The first high-performance base line was the Random Forest. Consider this model to be a giant committee of individual decision-makers (decision trees). All the Trees are presented with a random, slightly different, subset of the farm data and trained to make their own predictions. The model uses an average of the predictions made by all the trees in the forest to obtain a final answer by simply voting democratically, and it was very successful with a strong R-squared of 0.91.

Stacking Ensemble

The architecture is more advanced, and works similarly to a group of highly specialized experts (Random Forest, LightGBM) controlled by an older expert (Ridge CV). Our first step was to train various models on the data which we refer to as our specialists. At that point a final meta-model was introduced to examine the predictions of the specialists and combine their results in the best intelligent manner helping to achieve R2 of 0.9119.

LightGBM + MLP:

This is a more sophisticated hybrid model, which operates in two different phases. A strong LightGBM model was trained in the first stage, but it was no part of its work to make the last prediction. We instead used it as a smart feature engineer to process the raw data and get the complicated patterns it found. The second step involved feeding this enriched and high-level information to a Multi-Layer Perceptron (MLP), a form of neural network. This approach, which applies one expert model to train smarter data to a different expert model, was remarkably good with a leading R2 of 0.9088.

- Among all these sophisticated architectures, a special hybrid model which had a direct cooperation between LightGBM and a Bi-directional LSTM, was shown to perform better than all other methods. Combining the

performance of our best-performing machine learning model with a sequence and context-aware deep learning model enabled a stunning improvement in performance, giving a R-squared of 0.9214.

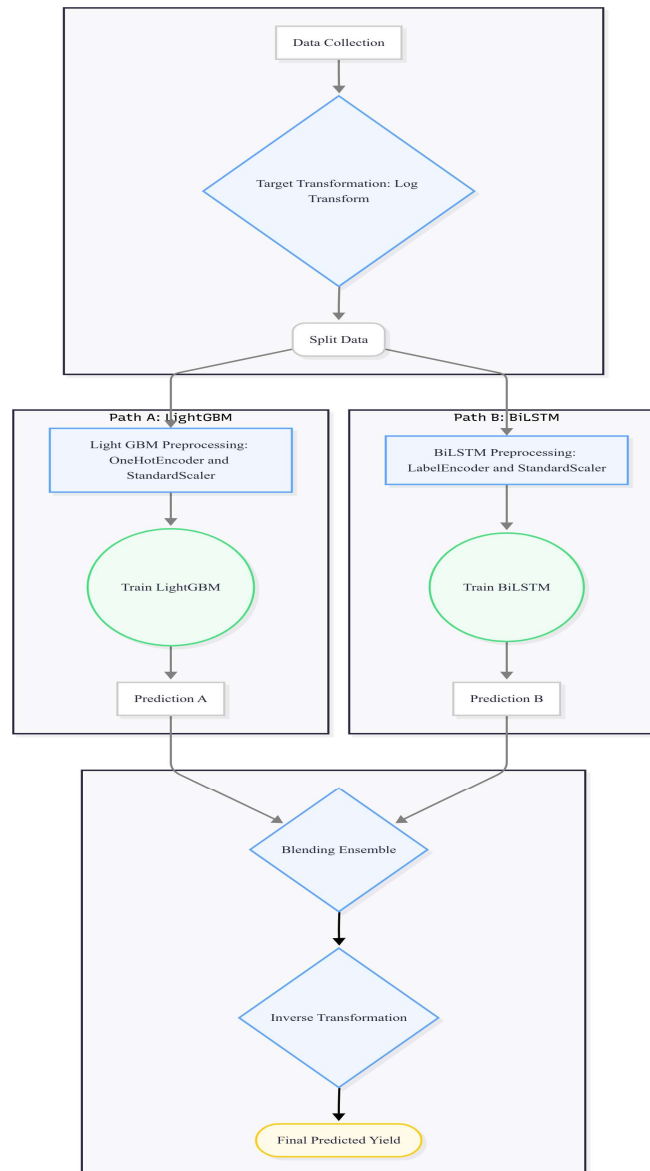


Figure 1: Workflow of the Model

E. Optimization & Training

Our last model settled on was a hybrid arrangement of LightGBM and Bi-LSTM, and at this point, we were finished and moved to the most crucial stage, the training of our team of professionals to work at their fullest. It was not a straightforward press play operation, but a well-thought-out procedure that was aimed at unleashing the full capabilities of both the machine learning and deep learning elements. Our training approach centered on one of the techniques, referred to as blending, in which we train our two experts separately to become masters of their field and thereafter blend their inputs together.

- You can train the Pattern-Finder (LightGBM): LightGBM is our first state-of-the-art machine learning model. Its training is a beautiful and strong art of refinement. It operates by creating a cascade of decision trees, except there is a very clever twist: the new tree is created in particular to eliminate the errors of the

previous tree. This progressive, error correction algorithm enables LightGBM to quickly pick up the non-linear, complex trends in the farm data and it is an extremely fast and accurate pattern- finder.

- **Training the Context Expert (Bi-LSTM):** Our second expert, the Bi-directional LSTM, is a deep learning model and its training is a more complicated procedure for example, teaching a student, a new complicated language.
 - **Learning the Vocabulary (Embeddings):** The Bi-LSTM needed to learn the rich meaning of our categories before it could begin to understand the data.
 - **The AdamW Optimizer:** In order to support this complex learning process, we employed the AdamW optimizer, which is a state-of-the-art algorithm that acts as a coach to the neural network.
 - **Preventing Memorization (Dropout):** The most frequent threat of the deep learning models is that they can merely memorize the training data as opposed to learning the structure. In order to avoid it, we adopted a method known as Dropout that randomly switches sections of the network during training.

IV. RESULTS & EVALUATION

The last and most important thing we did after designing and training our models was to test them on unexplored test data. It is the part where the quantitative findings of our experiments are presented, and the implications of these findings are discussed, giving a clear picture why our suggested hybrid model is the best way to go.

A. Evaluation Metrics

Performance assessment of a machine learning model is an important process of establishing its reliability where the results may have a direct effect on the agricultural planning and livelihoods of farmers. Thus, the testing of the suggested hybrid LightGBM–BiLSTM model was conducted on the basis of a comprehensive set of statistical indicators that gauge the accuracy, consistency, and strength together. Such measures are Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Root Mean Squared Logarithmic Error (RMSLE) which helps us to identify the optimal performance of the model.

Coefficient of Determination (R^2) measures the ratio of the change in the actual crop yields that can be attributed to model forecasts which is shown in equation 1. The metric is expressed as:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (1)$$

Root Mean Squared Error (RMSE) is often regarded as the standard measure of the scale of the prediction errors is shown in equation (2). It quantifies the distance between the average of the values that were predicted and the actual observed yield as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

Mean Absolute Error (MAE) is an easy to understand and interpret value given by equation (3). It is a mean difference between the predicted and actual yield, and it is defined as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

Root Mean Squared Logarithmic Error (RMSLE) shown in equation 4 will be used to overcome this problem by evaluating the difference between the forecasted and the actual values in the relative sense as opposed to the absolute sense. It is given by:

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(1 + \hat{y}_i) - \log(1 + y_i))^2} \quad (4)$$

B. Experimental Results

A detailed comparison of predictive performance of different machine learning models and hybrid deep learning models is given in table I. A quantitative evaluation will use four important statistical measures, namely Coefficient of Determination (R Squared), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Root Mean Squared Logarithmic Error (RMSLE), that combine to evaluate the consistency and accuracy of the models. All the findings as shown in table I point to the fact that a gradient boosting (LightGBM) and bidirectional recurrent neural network (BiLSTM) can be used to create a stronger and more generalized

prediction model, increasing the capabilities of the model to adapt to different agroclimatic conditions, and provide effective yield predictions.

TABLE I. OVERALL PERFORMANCE METRICS OF PREDICTION MODELS

Model	R2	RMSE	MAE	RMSLE
Random Forest	0.9144	668.21	111.32	0.3861
RF + LightGBM (Stacking Ensemble)	0.9119	678.15	110.70	0.1942
LightGBM + MLP	0.9088	689.86	110.33	0.2109
LightGBM + BiLSTM (Proposed Hybrid Model)	0.9214	640.30	102.57	0.1972

C. Discussion

As Figure 2 shows, the graphical representation of the real and predicted rice production in different districts of Tamil Nadu. The proximity between the actual and the predicted numbers shows the fact that the model can easily generalize across various districts, seasons, and the cultivation areas, which were all key predictors during the training period. The observed minor deviation in some districts may indicate that the model is affected by the variation in the data distribution especially in the areas where the quantity of training data can be relatively small.

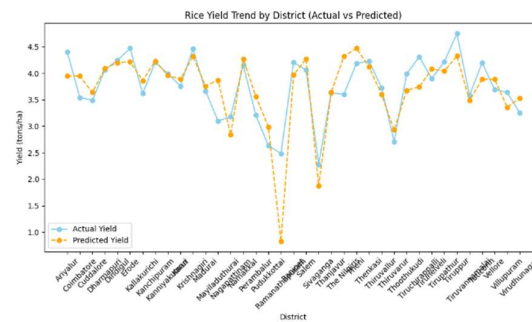


Figure 2. Rice Yield Trend by District (Actual vs. Predicted)

Figure 3 undertakes the comparative analysis of the actual and predicted sugarcane yield in various districts. The graph shows the validity of the suggested machine learning model in the ability to record a regional variation in the yield of agro- climatic and soil characteristics. The fact that the actual and predicted values are near each other is evidence of the fact that the model has a high level of predictive accuracy and can generalize; this proves its reliability in predicting the real world situations of agriculture

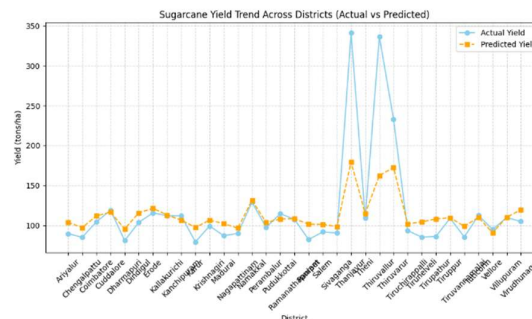


Figure 3. Sugarcane Yield Trend by District (Actual vs. Predicted)

Figure 4 shows the actual and predicted coconut yield of different districts. The graph shows that the model can follow the actual yield trends quite well, which shows that there is a great correlation between actual and estimated values. This consistency indicates the high capacity of the prediction model to perform well in the condition of yield changes depending on the soil fertility, climatic variables, and district-specific cultivation processes. Minor differences in the actual and predicted yields in some areas can be due to local effects like variations in rainfall, presence of some pests or unequal use of inputs which cannot always be captured in historical records.

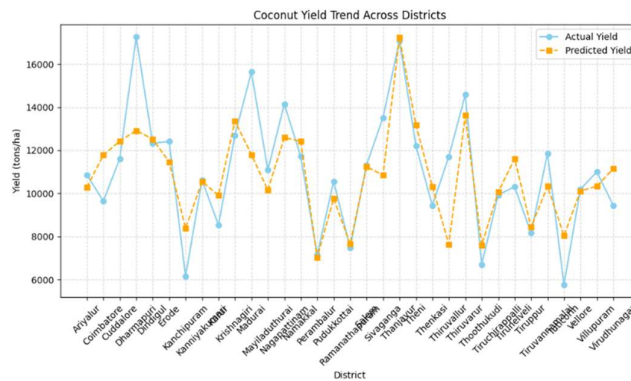


Figure 4. Coconut Yield Trend by District (Actual vs. Predicted)

Figure 5 illustrates actual and predicted yield of groundnut across all the districts and gives an insight into how the model can predict yield variation at a regional level with a good level of accuracy. The two trends are nearly similar which means that the developed model is capable of establishing the existing relationship between the agro-climatic variables and yield outcomes. This uniformity demonstrates the accuracy and dependability of the model in forecasting yield performance of the groundnut.

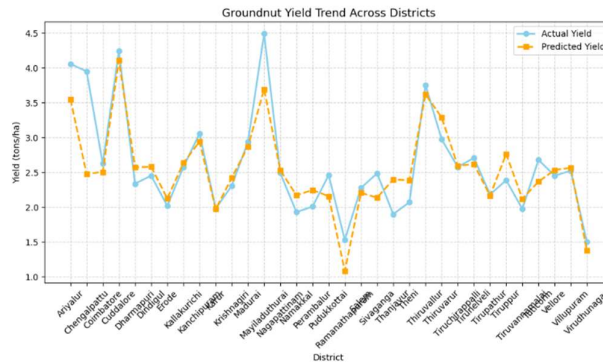


Figure 5. Groundnut Yield Trend by District (Actual vs. Predicted)

Figure 6 shows the relative yields of cotton in different districts with the actual and modeled values. The fact that the suggested model is good in terms of predicting the spatial variation. The low differences in the actual and predicted yields in most of the districts indicates the strong predictive power and generalization of the model. The LightGBM model and its gradient boosting mechanism is highly effective in identifying the principal predictive trends and nonlinear relationships in the structured farm data. The Bi-directional LSTM, in its turn, contributes by understanding the contextual information of the relations between the features, and viewing them as a part of each other, instead of individual entities.

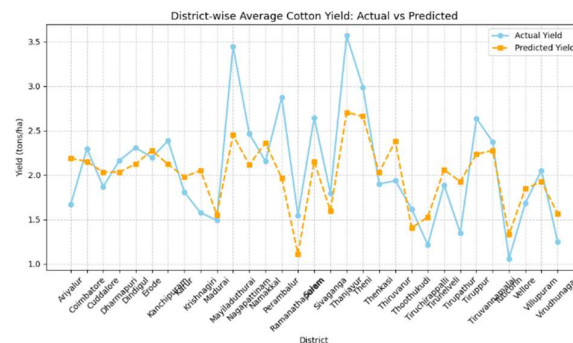


Figure 6. Cotton Yield Trend by District (Actual vs. Predicted)

V. CONCLUSION AND FUTURE WORK

This process was started with a basic problem of farmers throughout Tamil Nadu, and it is the uncertainty of the crop. This uncertainty, as we have shown in this paper, can be greatly mitigated by the clever use of contemporary artificial intelligence. Our journey took us through a wide variety of machine learning and deep learning models and at a stage of intensive screening, we found that the improvement in the performance cannot be achieved with one model, but rather with a cooperation. A hybrid model that combines the merits of the Light GBM pattern-finder and Bi-directional LSTM context expert, our last suggested model which is a testimony to this strategy with R2 value of 0.9214.

This model has demonstrated the capacity to explain the enormous proportion of the variation in yield using a high extent of accuracy. The results are not only statistically significant but also represent a practical step towards a reliable, data-driven decision-support tool to our agricultural communities. Although our present model is able to respond to the question, what will my harvest be? Our plan is to respond to an even more powerful question to the farmer in the future, which crop should I plant to get the best possible harvest? The second step of this study will be to develop our prediction engine into a full-fledged prescriptive Crop Recommendation Engine.

This system of the future would be a shift from forecasting to advising. To do this, the model will be expanded so as to learn with a far richer array of inputs which will include:

- Soil Characteristics: Essential information such as the type of soil (e.g., clay, loam, sandy), the pH level, and the content of such major nutrients as Nitrogen (N), Phosphorus (P), and Potassium (K).
- Climatic conditions: Climatic forecasts, Local weather information and short-term weather forecasts of the season ahead.

REFERENCES

- [1] M. Champaneri, D. Chachpara, C. Chandvidkar, and M. Rathod, "Crop yield prediction using machine learning," *International Journal of Science and Research (IJSR)*, vol. 9, no. 4, pp. 645–650, 2020.
- [2] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, "Crop yield prediction using machine learning and deep learning techniques," in *International Conference on Machine Learning and Data Engineering*, vol. 218. Elsevier, 2023, pp. 406–417.
- [3] S. M. Shawon, F. B. Ema, A. K. Mahi, F. L. Niha, and H. Zubair, "Crop yield prediction using machine learning: An extensive and systematic literature review," *Smart Agricultural Technology*, vol. 10, p. 100718, 2025.
- [4] N. B. Wibisono and S. Saiful, "Crop yield prediction using random forest algorithm and XGBoost machine learning model," *International Journal of Research and Innovation in Social Science (IJRISS)*, vol. 9, no. 3, pp. 1983–1990, 2025.
- [5] U. Bhimavarapu, G. Battineni, and N. Chintalapudi, "Improved optimization algorithm in LSTM to predict crop yield," *Computers*, vol. 12, no. 1, p. 10, 2023.
- [6] S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," *Journal of Physics: Conference Series*, vol. 1714, no. 1, p. 012012, 2021.
- [7] N. Bali and A. Singla, "Emerging trends in machine learning to predict crop yield and study its influential factors: A survey," *Archives of Computational Methods in Engineering*, vol. 29, pp. 95–112, 2022.
- [8] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers in Plant Science*, vol. 10, p. 621, 2019.
- [9] N. Banupriya, D. Tejasvi, and P. Vaishnavi, "Crop yield prediction based on Indian agriculture using machine learning," *International Journal of Modern Agriculture*, vol. 10, no. 3, pp. 73–80, 2021. [Online]. Available: https://www.researchgate.net/publication/357622040_Crop_Yield_Prediction_BASED_ON_INDIAN_AGRICULTURE_USING_MACHINE_LEARNING
- [10] V. L. Jothi, A. Neelambigai, S. Nithish Sabari, and K. Santhosh, "Crop yield prediction using KNN model," in *International Conference on Recent Trends in Information, Communication and Computer Technology (RTICCT)*, vol. 8, no. 12. IJERT, 2020, pp. 4–6.