CROP YIELD PREDICTION USING ENSEMBLE MODEL

1st Hemant Kumar Singh
Msc in Computing(Data Analytics)
Dublin City University, Dublin
hemant.singh5@mail.dcu.ie

2nd Sailesh Krishnan
Msc in Computing(Artificial Intelligence)
Dublin City University, Dublin
sailesh.krishnan2@mail.dcu.ie

***Abstract*—Crop yield prediction is vital to agriculture, influencing global food security, economic stability, and environmental sustainability. In the context of machine learning, Traditional prediction methods have relied on machine learning (ML) and deep learning (DL) algorithms such as Random Forests, LASSO, XGBoost, Linear Regression and Convolutional Neural Networks (CNNs). Despite their effectiveness, an emerging trend of ensemble models demonstrates improved prediction performance by amalgamating multiple algorithms. This research presents a novel ensemble model, the *L-Hist Ensemble*, which integrates the Light Gradient Boosting Machine (LGBM) Regressor and the Histogram-based Gradient Boosting Regressor, harmonised through a Linear Regression meta-model. Results show that the *L-Hist Ensemble* model outperforms traditional approaches such as Random Forest and Optimised Weighted Ensemble, exhibiting a substantial prediction improvement of approximately 28%. Furthermore, the model delivers a Train RMSE of 13.838, Test RMSE of 13.732, and Test R^2 score of 0.8479, along with a balanced Train and Test RRMSE, showcasing its prediction accuracy and stability. Despite being marginally outperformed by the complex CNN-RNN Framework, the *L-Hist Ensemble* model holds advantages in computational efficiency and ease of interpretation, making it a promising alternative for real-world agricultural applications. This work contributes to evolving efficient, accurate, and interpretable crop yield prediction systems.**

***Index Terms*—Crop Yield Prediction, Machine Learning, Ensemble Models, Light Gradient Boosting Machine, Histogram-based Gradient Boosting, Blending, Linear Regression, Bayesian Optimization, Deep Learning, CNN-RNN, Precision Agriculture.**

I. Introduction

Agriculture forms the backbone of many economies worldwide, directly influencing food security, economic stability, and environmental sustainability [29][30]. As an integral part of agricultural management, the prediction of crop yields plays a pivotal role in planning and policy formulation. With modern technologies and computational methods, machine learning (ML) and deep learning (DL) have emerged as powerful tools in refining crop yield prediction models, catalysing substantial advancements in precision agriculture.

Much research has been dedicated to deploying machine learning and deep learning models to forecast crop yields. These endeavours encompass diverse algorithms, ranging from Random Forests [3][4] to Convolutional Neural Networks (CNNs) [19], Deep Neural Networks (DNNs), and their hybrid forms [17]. Integrating multi-source data, including climatic, remote sensing, and soil data, alongside intricate model variables has been harnessed to offer promising results. Empirical research works [17][20][21][22] suggest employing ensemble models that amalgamate multiple algorithms has surfaced as a significant development in crop yield prediction, demonstrating a considerable improvement in prediction performance.

This research presents a novel ensemble approach for crop yield prediction – the *L-Hist Ensemble* model. Building upon the strengths of the Light Gradient Boosting Machine (LGBM) Regressor and the Histogram-based Gradient Boosting Regressor, this ensemble model leverages the blending method. In this approach, multiple base models are trained using unique learning algorithms, and their predictions are subsequently amalgamated using a meta-model, the Linear Regression model in this case. This work aims to explore the efficacy of this ensemble model, which combines the merits of these algorithms while balancing model complexity and interpretability.

Experimental results showcase the superior predictive ability of the *L-Hist Ensemble* model, which outperforms traditional machine learning approaches like Random Forest and Optimised Weighted Ensemble. Moreover, the *L-Hist Ensemble* model records a high-performance metric, highlighting its ability to generalise effectively to unseen data. While marginally lagging behind complex deep learning frameworks like the CNN-RNN Framework [14] regarding raw performance metrics, the *L-Hist Ensemble* model presents clear advantages, including computational efficiency and ease of interpretation. These attributes underscore its potential for practical implementation in real-world agricultural contexts, where operational efficiency is paramount.

This research work underscores the continued evolution and significance of modern computational models in the agricultural realm, aiming to contribute to the development of efficient, accurate, and easily interpretable crop yield prediction systems.

We aim to find out these particular things in our study:

1. Can our method that combine multiple machine learning models enhance the predictability of agricultural yield outcomes, and if so, to what extent?

2. How our method performs against well-established methods like Random forest, Optimised weighted ensemble and CNN RNN Framework.

This paper unfolds as follows: The "Related Work" section sheds light on prior work in machine learning (ML) and deep learning (DL) for crop yield prediction. Next, the "Data Gathering" section provides a detailed outline of the dataset used for this study. In contrast, the "Data Preprocessing" section elucidates the steps to make the raw data suitable for our ML and DL models. The "Model Selection" section thoroughly reviews the other prediction models used for comparison in this study. The construction of the proposed *L-Hist Ensemble* model is then discussed in the "Model Architecture" section. The measures for assessing the models' performance are further detailed in the "Performance Metrics" section. The "Experimentation" section takes you through the journey of our model development, explaining various attempts and their results. Our findings, including the *L-Hist Ensemble* model's performance against traditional ML and DL models, are evaluated in the "Results and Discussion" section. Finally, the paper concludes with the "Conclusion" section, where the overall findings are summarised, and the superiority of the *L-Hist Ensemble* model is reinforced, providing insights into potential future work in crop yield prediction.

II. Related Work

The successful prediction of crop yields has been a significant area of research and carries profound implications for agricultural planning, especially in economies heavily reliant on agriculture. This section examines existing literature and draws on various studies to understand crop yield prediction's current state and future directions using ensemble models.

Studies have consistently emphasised the role of machine learning algorithms in crop yield prediction. Notably, [2] conducted an extensive review, highlighting the importance of integrating various aspects, such as remote sensing and disease recognition, to enhance prediction accuracy. Complementing this, [1] presented the usefulness of coupling crop modelling and machine learning for improving corn yield predictions. They underlined that machine learning models require more than just weather information and should include additional hydrological inputs for increased accuracy.

Various research, including those by [3] and [4], explored the effectiveness of different machine learning algorithms. Both studies concluded that the choice of the machine learning algorithm, such as Random Forest or Decision Tree, significantly influences the prediction accuracy. Furthermore, a hybrid approach combining machine learning and deep learning techniques proposed by [5] showed promising results, suggesting a potential direction for future studies.

Other research endeavours have also utilised machine learning techniques in crop yield prediction. Studies by [6] and [7] showcased how these techniques can facilitate accurate predictions, helping farmers make informed decisions. In addition, [8] suggested the efficiency of the Kalman filter algorithm in yield prediction systems.

Integrating AI techniques into yield prediction has also proven beneficial.[9] used Bayesian Model Averaging (BMA) and multiple deep neural networks for a probabilistic estimate of soybean crop yield. At the same time, [10] demonstrated the effectiveness of Random Forests (RF) in predicting crop yield responses globally and regionally.

Simultaneously, satellite imagery and remote sensing have emerged as a potent tool for crop yield prediction. Studies by [11][12] and [15] highlighted how UAV-based multimodal data fusion and MODIS-NDVI data can be employed effectively within this domain. Furthermore, deep learning frameworks, like the one proposed by [14], have significantly improved prediction accuracy.

Ensemble models have been increasingly used to predict crop yields in recent years. [17] used novel CNN-DNN machine learning ensemble models to predict corn yields, highlighting the superiority of homogenous ensembles over heterogeneous ones. A study by [18] on 'Chok Anan' mangoes yield under different irrigation regimes using RF models, and the application of CNNs based on UAV data [19] pointed to the potential of machine learning models in agricultural engineering.

Further extending the potential of ensemble models, [20] proposed an ensemble of CNN-LSTMs to predict yearly soybean yields and daily strawberry yields and prices. [22] proposed a framework to forecast corn yields using ensemble models, which were found to be the most precise. In contrast to the machine learning approaches, [21] used the Agricultural Production Systems Simulator (APSIM) to predict and explain crop yields and soil dynamics, underscoring the importance of initial conditions and early season measurements.

In summary, these studies form a solid foundation for our research on crop yield prediction using ensemble models. The key to successful crop yield prediction lies in integrating machine learning algorithms, deep learning techniques, and many data sources, including climatic, soil, and plant conditions data. This provides a more comprehensive understanding of the factors influencing crop yield and improves the accuracy of predictions, providing valuable insights for agricultural planning and decision-making.

III. Data Gathering

We have compiled an extensive dataset of county-level corn production statistics ranging from 1981 to 2018. The data, sourced from the USDA National Agricultural Statistics Service [21], comprises 10,672 individual records of annual corn production across 293 distinct counties. This dataset reflects critical environmental and managerial influences on crop yields. However, genotype data could not be included due to the lack of public accessibility.

Our dataset includes four primary categories, totalling 431 variables, inclusive of the target variable:

A. Planting Progress (Planting Date)

This category includes 52 variables, providing detailed data on the weekly progression of corn planting across each state [23].

B. Weather

This category comprises 312 attributes related to six unique weather factors, tracked weekly. The data is sourced from NASA POWER [28] and the Iowa Environmental Mesonet [27]. The six weather factors include:

- Daily rainfall (mm/day)
- Daily sunlight (J/m²/day)
- Snow water equivalent (mm)
- Daily lowest air temperature (°C)
- Daily highest air temperature (°C)
- Daily vapour pressure (kPa)

C. Soil Details

To provide an in-depth understanding of soil conditions, we have included data on soil density, clay percentage, cation exchange capacity at pH 7, coarse fragments, total nitrogen, organic carbon density, organic carbon stock, pH in H2O, sand, silt, and soil organic carbon content. These variables are measured at various depths: 0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm. This data is sourced from the Web Soil Survey [22].

D. Corn Yield

This variable refers to the annual corn yield, measured in bushels per acre, sourced from USDA-NASS [21].

IV. Data Preprocessing

Our raw dataset underwent numerous preprocessing steps to optimise it for training the predictive model, enhancing performance by refining the data, removing irrelevant information, and filling gaps. The initial 431 features were reduced to 393 by eliminating redundant variables from the 'Planting Progress' category that did not contribute to our corn yield prediction. Features unrelated to the prediction were dropped. Missing values, a common dataset issue, were imputed using the mean of the respective variables. The cleaned dataset was then divided into feature set (X) and target variable (corn yield, y) and standardised to a mean of 0 and a standard deviation of 1 using the StandardScaler function from the sklearn library. To prevent overfitting and ensure model validation, we divided the dataset into a base dataset for initial training and a meta-dataset for training and testing the meta-model, employing a 70-30 split. The meta-dataset was divided into training and testing subsets in an 80-20 ratio. Overall, these preprocessing steps helped transform the raw data into a machine-readable format, improved model performance and enabled reliable model evaluation through practical training and testing.

V. Model Selection

To ascertain the effectiveness of our proposed *L-Hist Ensemble* model for crop yield prediction, we must examine its performance with other existing models. For this purpose, we have selected three established models, inspired by prominent studies in the domain, to serve as our benchmarks: Random Forest [24], Optimised Weighted Ensemble [23], and CNN-RNN Framework [14]. The subsequent sections present an overview of each model and the specific techniques we utilised for their deployment.

A. Random Forest

Building upon the foundational work of Breiman [24], the first model employed was the Random Forest Regressor. This model is renowned for its capacity to manage a large number of features without overfitting. Carefully choosing the hyperparameters is essential to the model's success. We used BayesSearchCV, a Bayesian optimisation technique, to achieve this. This method combines prior assessments to develop a probabilistic model for discovering optimal parameters, making it preferable to conventional approaches like grid and random search. The factors we took into account were the number of trees, the depth of the tree, and various circumstances involving splitting and leaf nodes. We established the ranges for these based on accepted procedures and computational limitations. After several Bayesian search and cross-validation cycles, the optimum parameters were chosen for our model's training. The most remarkable results came from building the final model with particular values for these parameters, such as 100 trees and a maximum tree depth of 39.

B. Optimized Weighted Ensemble

Next, we introduced an Optimised Weighted Ensemble model to enhance the predictive performance. This model was inspired by the work of [23], who proposed an optimisation problem that minimises the mean squared error (MSE) of predictions. The optimisation problem used out-of-bag predictions from each base learner obtained through k-fold cross-validation.

The optimisation problem can be mathematically expressed as below:

$$\begin{aligned} \text{Minimize:} \quad & \frac{1}{n}\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{k} W_j \cdot \bar{Y}_{ij}\right) - Y_i\right)^2 \\ \text{Subject to:} \quad & \sum_{j=1}^{k} W_j = 1 \\ & W_j \geq 0, \quad \text{for } j = 1, \ldots, k \end{aligned}$$

where W_j is the weights corresponding to base model j ($j = 1, \ldots, k$), n is the total number of instances, Y_i is the actual value of observation i, and $\bar{Y}_{ij}$ is the prediction of observation i by base model j.

The optimisation problem is a non-linear convex problem, and the weights of base learners are optimised using the 'SLSQP' method. Weights were constrained to sum to 1, and each weight was restricted to a range [0, 1]. Once optimal weights were acquired, the base learners were trained on the entire training set. The final predictions were computed using the optimised weights as a weighted average of the base learner predictions.

C. CNN-RNN Framework

The final model employed was a hybrid CNN-RNN Framework, following the work of [14]. The model was chosen for its ability to handle time-series data effectively. This model's CNN and RNN components allowed it to learn and capture spatial and temporal dependencies in the data.

After data preprocessing, the model was split into three branches: weather, soil, and planting time. Different input layers represented each branch. The model was then trained using the RMSprop optimiser, with the mean squared error (MSE) serving as the loss function. An Early Stopping mechanism was used to prevent overfitting and monitor the training process. Although, as suggested by the original work, a high limit for the training process was used, the actual training ceased when there was no improvement in validation loss for a tangible period.

By comparing the performance of our proposed *L-Hist Ensemble* model against these three established models, we have provided a comprehensive evaluation of its predictive capabilities in crop yield prediction. The following sections will discuss these comparative analyses' implementation details and results.

VI. Model Architecture

The architecture employed for the proposed *L-Hist Ensemble* model is an ensemble of regressor models, specifically the Light Gradient Boosting Machine (LGBM) Regressor and the Histogram-based Gradient Boosting Regressor. The ensemble design is predicated on the blending method, an efficient ensemble learning technique. This method involves training multiple base models, each employing a unique learning algorithm, and subsequently amalgamating their predictions using another model, termed the meta-model.

A. Base Models

The ensemble's foundation comprises two regressor models renowned for accurately modelling intricate, non-linear relationships in large datasets.

1) Light Gradient Boosting Machine (LGBM) Regressor: This model uses a gradient-boosting framework that employs tree-based learning algorithms. The algorithm is designed for distributed and efficient modelling with notable advantages such as faster training speed, higher efficiency, lower memory usage, superior accuracy, and support for parallel and GPU learning.

2) Histogram-based Gradient Boosting (HistGradientBoosting) Regressor: This model is a variant of the Gradient Boosting Machines, a robust ensemble machine learning technique that constructs a model stage-wise. In this model, continuous input variables are binned into discrete bins, accelerating computation and reducing memory footprint while preserving the model's ability to capture complex relationships in the data.

Hyperparameters were calibrated using a Bayesian optimisation technique to optimise the performance of these base models. We employed the BayesSearchCV function from the Scikit-Optimize library for this calibration. Bayesian Optimisation is an efficient technique for global Optimisation of black-box functions. It is particularly well-suited for the Optimisation of high-cost functions, situations where the balance between exploration and exploitation is essential, and hyperparameter tuning, known to outperform other methods like grid or random search.

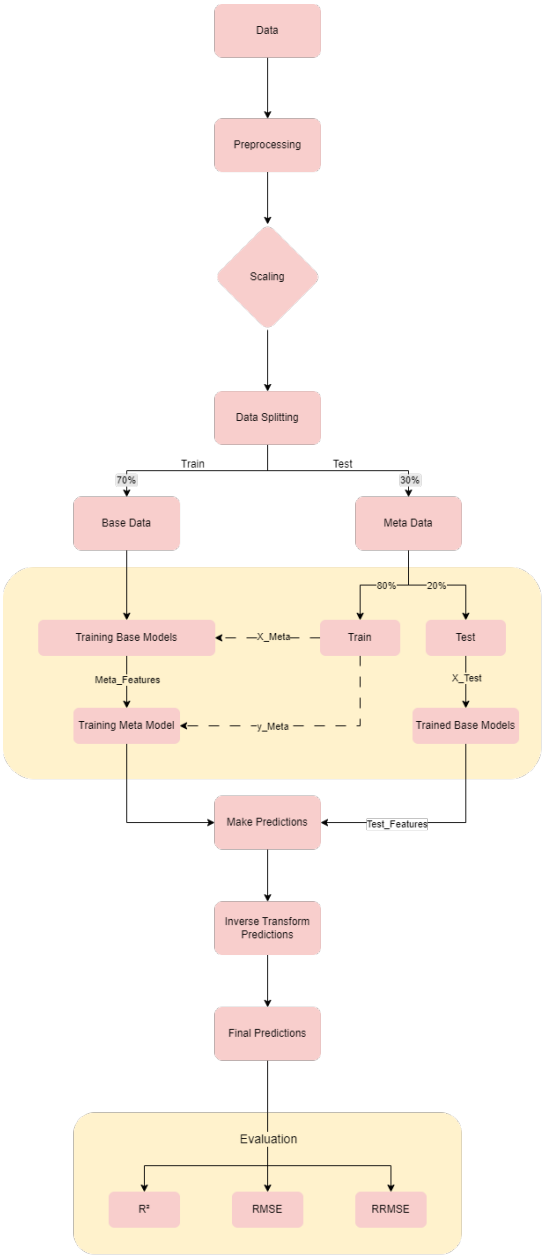

Fig. 1. *L-Hist Ensemble* Architecture

B. Meta Model

A Linear Regression model was utilised as the meta-model to harmonise the predictions of the base models. We chose Linear Regression, a simple yet powerful algorithm, as the meta-model due to its ease of interpretation, computational

efficiency, and ability to prevent overfitting in the ensemble model. The meta-model is trained to learn the optimal way of amalgamating the predictions from the base models to generate the final prediction.

This ensemble model's architecture offers a high degree of flexibility. It can be readily extended by integrating more base models or replacing the existing ones with alternative models, depending on the specific requirements of the task. Moreover, the meta-model can also be swapped with a more complex model if the complexity of the task demands it.

VII. Performance Metrics

The model's effectiveness is evaluated using the following metrics:

1) **Coefficient of Determination (R^2)** : R^2 indicates the model's predictive capability [31]. The higher the R^2 (closer to 1), the better the model predicts the unseen data. It is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

2) **Root-Mean-Square Error (RMSE)** : It measures the average magnitude of prediction errors [31]. Lower RMSE values suggest better model performance. It is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{2}$$

3) **Relative Root-Mean-Square Error (RRMSE)** : It provides a relative measure of error expressed as a percentage, allowing for direct comparisons between models [31]. It is given by:

$$RRMSE = \frac{RMSE}{\max(y_i) - \min(y_i)} \times 100\% \tag{3}$$

These metrics comprehensively evaluate the model's accuracy and predictive capability.

VIII. Experimentation

Our research progressed through a series of experimental stages, each paving the way to our proposed *L-Hist Ensemble* Model. The experimentation began with an attempt to optimise performance using a Bayesian search with a continuous variable applied in conjunction with a random forest. However, this approach led to overfitting, hence limiting its effectiveness.

Subsequently, we ventured into combining the Random Forest Regressor with the Gradient Boosting Regressor, aiming to improve the predictive accuracy by enhancing the R-squared, RMSE, and RRMSE values. Regrettably, not only were the results suboptimal, but the computational cost was significantly high, marking this approach as inefficient.

Shifting gears, we experimented with LightGBM, paired with Bayesian continuous variable. This showed promising results on the training set but faltered on the validation set, underscoring the model generalisation's challenge. Nonetheless, this experiment's insights facilitated the development of our proposed *L-Hist Ensemble* Model, leveraging LightGBM and HistGBM as base models.

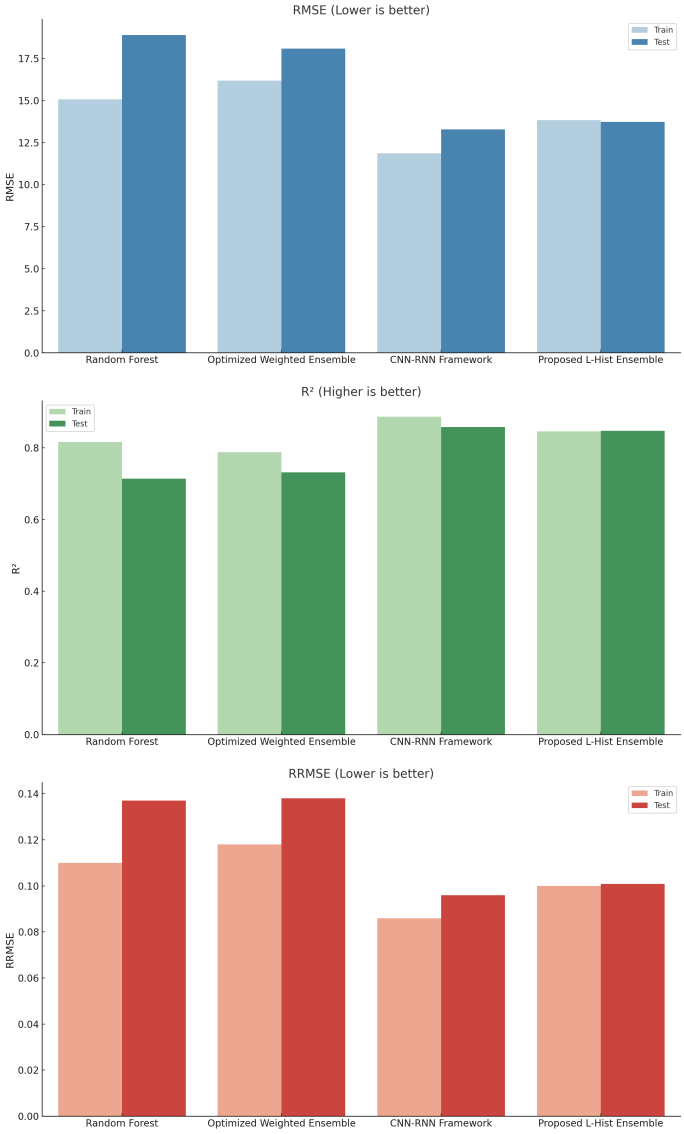

Fig. 2. Evaluation Metrics Results

The next phase of our research involved investigating two ensemble learning approaches: stacking and blending. The stacking approach, despite the application of regularisation methods such as using Ridge regression (L2 regularisation) and Lasso regression (L1 regularisation), and even Elastic Net regression (combining L1 and L2) as the final estimator, needed to be more balanced. Furthermore, adjusting the train/test ratio from 0.8/0.2 to 0.7/0.3 did not improve.

In contrast, the blending approach demonstrated better generalisation after implementing strategic changes and modifying the learning rate range, altering the n-estimator range in the base models, introducing Linear Regression as the meta-model, and adjusting the test size from 0.2 to 0.3 improved performance. The observations from this phase ultimately

informed the design and tuning of our final *L-Hist Ensemble* Model.

IX. Results and Discussion

Our experimental results, obtained from a comprehensive evaluation of various machine learning models for crop yield prediction, show that our Proposed *L-Hist Ensemble* model outperforms traditional machine learning approaches like Random Forest and Optimised Weighted Ensemble. Further, it is competitively aligned with more complex deep learning frameworks like CNN-RNN, with certain added advantages.

The Proposed *L-Hist Ensemble* model achieves a lower test RMSE of 13.732 compared to Random Forest's 19.016 and the Optimised Weighted Ensemble's 19.086. This reflects a substantial improvement of approximately 28% in predicting crop yields, thus demonstrating the superior predictive ability of our model. The Proposed *L-Hist Ensemble* model also records a high test R^2 score of 0.847, which indicates the model's ability to explain the variations in the yield data. One

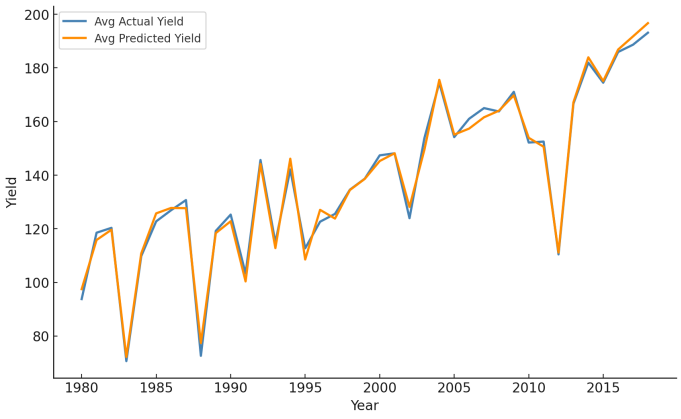

Fig. 3. Average Actual Vs Average Predicted Yield

key differentiating aspect of the *L-Hist Ensemble* model is its generalisation ability. This is highlighted by the similar performance metrics in the training and testing phases, with a training RMSE of 13.838 and a test RMSE of 13.732. The corresponding R^2 scores are also quite close, with 0.846 on the training set and 0.847 on the test set. It suggests the model fits the training data and can generalise effectively to unseen data.

Compared to the CNN-RNN Framework, the Proposed *L-Hist Ensemble* model is slightly lagging in performance. The CNN-RNN Framework achieves a test RMSE of 13.276 and an R^2 score of 0.858. However, it is noteworthy that the CNN-RNN Framework presents a more marked performance discrepancy between the training and testing phases. Its train RMSE is 11.853, demonstrating that the model might be overfitting to some extent on the training data and failing to generalise as effectively on the test data.

While the Proposed *L-Hist Ensemble* model trails slightly in raw performance metrics, it offers distinct advantages over

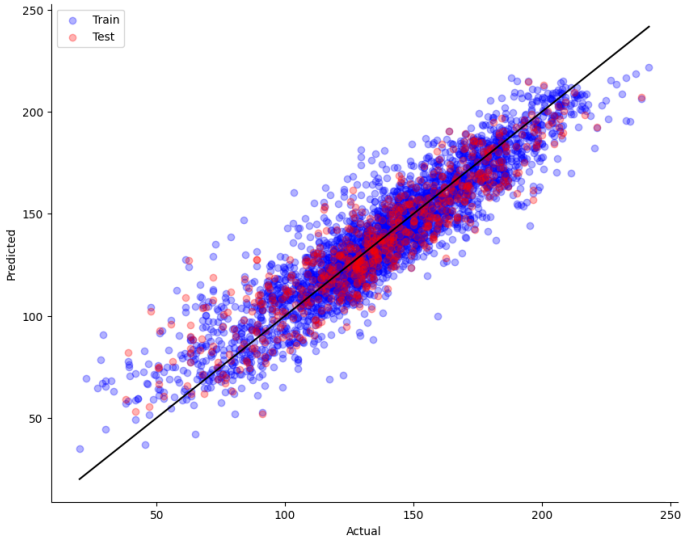

Fig. 4. Scatter Plot: True target values VS Predicted target values

the CNN-RNN Framework. Firstly, it is computationally less demanding, ensuring quicker model training and prediction times. This aspect is crucial in practical scenarios where real-time predictions are necessary and computational resources are limited. Furthermore, the simpler architecture of the *L-Hist Ensemble* model is more interpretable and easier to troubleshoot, which can be a decisive factor in production environments.

In summary, the Proposed *L-Hist Ensemble* model presents a promising alternative for crop yield prediction. It achieves high performance while maintaining good generalisation ability and is more computationally efficient than complex deep-learning models. These qualities make it ideal for real-world applications where performance and operational efficiency are paramount.

X. Conclusion

This study undertook the challenging task of predicting crop yields, an essential aspect of agriculture. It capitalised on Machine Learning (ML) and Deep Learning (DL) methodologies, culminating in the development of the *L-Hist Ensemble* model. The underpinnings of this model comprised two proficient regressor models - Light Gradient Boosting Machine (LGBM) Regressor and Histogram-based Gradient Boosting Regressor. These models, harmonised via a Linear Regression meta-model, demonstrated a compelling performance that superseded conventional ML models like Random Forest and Optimised Weighted Ensemble.

The experimental findings attest to the efficacy of the proposed *L-Hist Ensemble* model, as evinced by a substantial improvement of approximately 28% in predicting crop yields when juxtaposed with traditional ML approaches. It exhibited a high degree of generalisation, striking a delicate balance between training and testing performances, thus precluding overfitting.

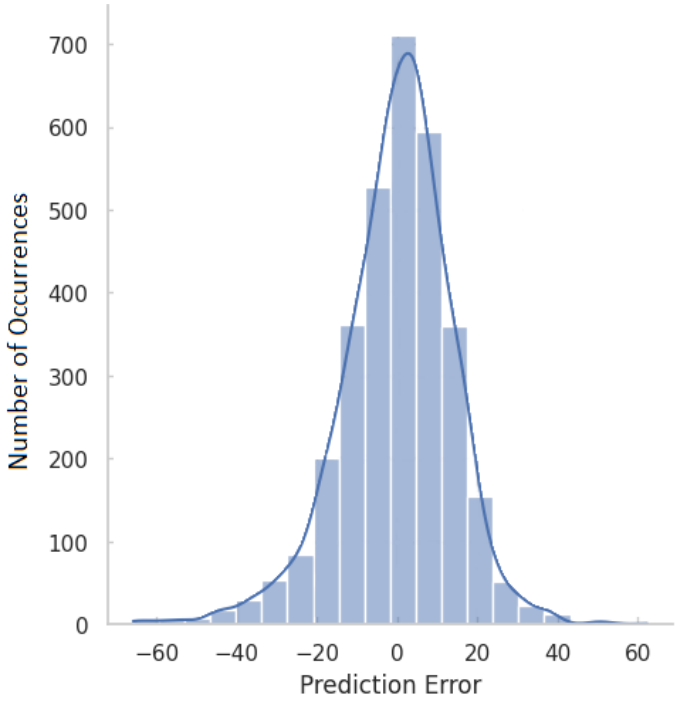

Fig. 5. Error Distribution in Training Data

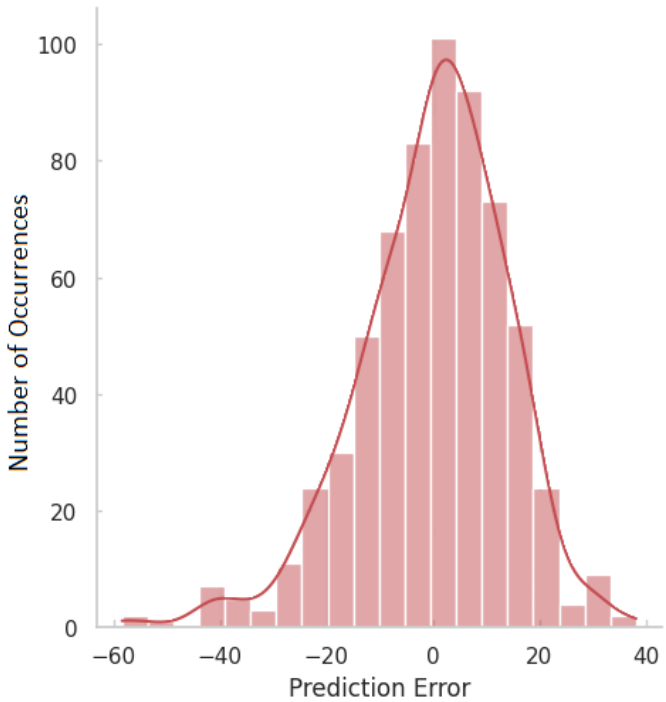

Fig. 6. Error Distribution in Testing Data

Although the sophisticated CNN-RNN Framework exhibited marginally superior performance metrics, the *L-Hist Ensemble* model bears its own merits. It offers simpler interpretability and requires less computational resources and shorter training times. This positions it as a favourable alternative in resource-constrained environments.

The study has added a shred of new evidence to the discussion of the potential of ensemble methods in agricultural yield prediction. Future research avenues could explore different ensemble strategies and the integration of other ML and DL models as base learners. Also, the utilisation of multiple sources of information, like information from simulation models like APSIM and data from satellite resources, can be considered for future research for their influence on the model's performance and Optimisation. There is also scope to delve deeper into feature engineering and selection to boost model performance. We looked into how ML and DL approaches may enhance agricultural practices and policy in our modest attempt. This groundwork will contribute to ensuring a more sustainable and safe food supply in the future.

REFERENCES

[1] Mohsen Shahhosseini, Guiping Hu, Sotirios V. Archontoulis, Isaiah Huber. "Coupling Machine Learning and Crop Modeling Improves Crop Yield Prediction in the US Corn Belt." 2020.

[2] Mamunur Rashid, Bifta Sama Bari, Yusri Yusup, Mohamad Anuar Kamaruddin, Nuzhat Khan, Nuzhat Khan. "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction." 2021. *IEEE Access*, doi:10.1109/access.2021.3075159.

[3] Kavita Jhajharia, P. Mathur, Sanchit Jain, Sukriti Nijhawan. "Crop Yield Prediction using Machine Learning and Deep Learning Techniques." 2023. *Procedia Computer Science*, doi:10.1016/j.procs.2023.01.023.

[4] S. Ramani, Dr. K. Merriliance. "AGRICULTURE EXPENDITURE VISUALISATION AND CROP YIELD PREDICTION USING MACHINE LEARNING." 2022.

[5] Sonal Agarwal, Sandhya Tarar. "A HYBRID APPROACH FOR CROP YIELD PREDICTION USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS." 2021, doi:10.1088/1742-6596/1714/1/012012.

[6] Ashwini I. Patil, Ramesh Medar, Vinod Desai. "Crop Yield Prediction Using Machine Learning Techniques." 2020, doi:10.32628/ijsrset20736.

[7] Guna Sekhar Sajja, Subhesh Saurabh Jha, Hicham Mhamdi, Mohd Naved, Samrat Ray, Khongdet Phasinam. "An Investigation on Crop Yield Prediction Using Machine Learning." 2021, doi:10.1109/icirca51532.2021.9544815.

[8] R. Rao. "CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS." 2022, doi:10.26562/ijirae.2022.v0906.08.

[9] Peyman Abbaszadeh, K. Gavahi, Atieh Alipour, Proloy Deb, Hamid Moradkhani. "Bayesian Multi-modeling of Deep Neural Nets for Probabilistic Crop Yield Prediction." 2022, doi:10.1016/j.agrformet.2021.108773.

[10] Jig Han Jeong, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis Timlin, Kyo Moon Shim, James S. Gerber, Vangimalla R. Reddy, Soo-Hyung Kim. "Random Forests for Global and Regional Crop Yield Predictions." 2016, doi:10.1371/journal.pone.0156571.

[11] Maitiniyazi Maimaitijiang, Vasit Sagan, Paheding Sidike, Sean Hartling, Flavio Esposito, Felix B. Fritschi. "Soybean yield prediction from UAV using multimodal data fusion and deep learning." 2020, doi:10.1016/j.rse.2019.111599.

[12] Jichong Han, Zhao Zhang, Juan Cao, Yuchuan Luo, Liangliang Zhang, Ziyue Li, Jing Zhang. "Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China." 2020, doi:10.3390/rs12020236.

[13] K. Gavahi, Peyman Abbaszadeh, Hamid Moradkhani. "DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting." 2021, doi:10.1016/j.eswa.2021.115511.

[14] Saeed Khaki, Lizhi Wang, Sotirios V. Archontoulis. "A CNN-RNN Framework for Crop Yield Prediction." 2020, doi:10.3389/fpls.2019.01750.

[15] Manasah S. Mkhabela, Paul R. Bullock, S. Raj, Shusen Wang, Y. Yang.

"Crop yield forecasting on the Canadian Prairies using MODIS NDVI data." 2011, doi:10.1016/j.agrformet.2010.11.012.
[16] Saeed Khaki, Lizhi Wang. "Crop Yield Prediction Using Deep Neural Networks." 2019, doi:10.3389/fpls.2019.00621.
[17] Mohsen Shahhosseini, Guiping Hu, Saeed Khaki, Sotirios V. Archontoulis. "Corn Yield Prediction with Ensemble CNN-DNN." 2021.
[18] Shinji Fukuda, Wolfram Spreer, Eriko Yasunaga, Kozue Yuge, Vicha Sardsud, Joachim Müller. "Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes." 2013, doi:10.1016/j.agwat.2012.07.003.
[19] Petteri Nevavuori, Nathaniel Narra, Tarmo Lipping. "Crop yield prediction with deep convolutional neural networks." 2019, doi:10.1016/j.compag.2019.104859.
[20] Mohamed Sadok Gastli, Lobna Nassar, Fakhri Karray. "Satellite Images and Deep Learning Tools for Crop Yield Prediction and Price Forecasting." 2021, doi:10.1109/ijcnn52387.2021.9534388.
[21] USDA-NASS. "USDA - National Agricultural Statistics Service." 2023, https://www.nass.usda.gov/.
[22] Web Soil Survey. (2023). Natural Resources Conservation Service, United States Department of Agriculture. https://websoilsurvey.nrcs.usda.gov/.
[23] Shahhosseini, Mohsen, Guiping Hu, and Hieu Pham. "Optimising ensemble weights for machine learning models: A case study for housing price prediction." Smart Service Systems, Operations Management, and Analytics: Proceedings of the 2019 INFORMS International Conference on Service Science. Springer International Publishing, 2020.
[24] Breiman, L. "Random forests." Mach. Learn. 45 (2001): 5–32, doi: 10.1023/A:1010933404324.
[25] Archontoulis, Sotirios V., et al. "Predicting crop yields and soil-plant nitrogen dynamics in the US Corn Belt." Crop Science 60.2 (2020): 721-738.
[26] Shahhosseini, Mohsen, Guiping Hu, and Sotirios V. Archontoulis. "Forecasting corn yield with machine learning ensembles." Frontiers in Plant Science 11 (2020): 1120.
[27] Iowa Environmental Mesonet. (2023). Iowa State University, Department of Agronomy. https://mesonet.agron.iastate.edu/.
[28] NASA's Prediction of Worldwide Energy Resource. (2023). NASA Langley Research Center. https://power.larc.nasa.gov/.
[29] World Bank Organization https://www.worldbank.org/en/topic/agriculture/overview
[30] Newman, Carol, Saurabh Singhal, and Finn Tarp. "Introduction to understanding agricultural development and change: Learning from Vietnam." Food Policy 94 (2020): 101930.
[31] Sajid, Saiara Samira, et al. "Integrating Crop Simulation and Machine Learning Models to Improve Crop Yield Prediction." 2022 17th Annual System of Systems Engineering Conference (SOSE). IEEE, 2022.