

Green Intelligence-Driven Explainable and Energy-Efficient AI Models for Climate-Resilient Smart Agriculture

Sanjay SINGLA

Post Doc Research Scholar, Multimedia University, Cyberjaya, Malaysia.
Department of Computer Science & Engineering, University Institute of Engineering,
Chandigarh University, Gharuan, Punjab, India

dr.ssinglacs@gmail.com

Chinnaiyan SENTHILPARI

Centre for Advanced Devices and Systems, Faculty of Artificial Intelligence and Engineering
(FAIE), Multimedia University, Cyberjaya, Malaysia

c.senthilpari@mmu.edu.my (corresponding author)

Chu-Liang LEE

Centre for Advanced Devices and Systems, Faculty of Artificial Intelligence and Engineering
(FAIE), Multimedia University, Cyberjaya, Malaysia

cllee@mmu.edu.my

Abstract

Crop-yield prediction is crucial for agricultural planning and food-security management, but there is a risk that models validated only in the "known" years and regions may offer only partial evidence of generalization in the "unknown" ones. In this study, we present a heterogeneous expert-fusion framework with residual refinement for crop-yield prediction in the presence of distribution shift, namely, GreenzakaAI+. The framework is composed of complementary machine-learning and deep-learning experts by using sample-adaptive gating and then models the residual error after fusion. GreenzakaAI+ is compared to 10 independent baselines that vary across bagging, boosting and deep-learning architectures and all of both temporal and spatial targets are not included in any development and selection decisions within a leakage-controlled protocol. With a feature representation of 68 dimensions and a training set of 362,633 observations, GreenzakaAI+ is able to achieve a validation of 92.17%, surpassing the best independent baseline, Wide & Deep, by 2.99 percentage points, along with a validation of 86.82% in the temporal aspect and 95.34% in the spatial aspect, which equates to an unseen-average of 91.08%. Stable performance estimates are provided by Bootstrap confidence intervals, and significant differences from evaluated baselines are provided by paired statistical comparisons. The proposed framework outperforms the independent baselines on the unseen-generalization score on the temporal split, although it doesn't perform as well as GreenzakaAI+. The results show the effectiveness of heterogeneous expert fusion with residual refinement in achieving a good balance in cross-distribution evaluation.

Keywords: Crop Yield Prediction; Expert Fusion; Residual Refinement; Agricultural ai ; machine learning, deep learning

1. Introduction

Crop yield forecasting as a base of agricultural planning, food security management, resource management, and decision making under climate change. Yield formation is the result of multi-interactions (in most cases multi-linear) of climatic conditions, soil characteristics, crop genotype, and management, which render conventional statistical estimation of yield unreliable in various agricultural settings. Therefore, yield prediction has gained importance in the context of machine learning (ML) and deep learning (DL), remote sensing and multimodal environmental data for crop and area-level predictions.

In this transition, predictive modelling has slowly evolved to ensemble and hybrid architectures, rather than one-for-one modelling for individual learners. Tree-based models like Random Forest and gradient-boosting algorithms are competitive in structured agricultural data sets, and deep architectures offer more capacity to learn complex spatial, temporal and multimodal relationships. Yet good validation in a traditional random or fixed validation does not guarantee good prediction for new years or space. This is especially crucial for agricultural use where climatic and environmental conditions are naturally highly nonstationary, as in this case.

Another weakness is that the development and generalization evaluations of the models are usually performed in fairly comparable data distributions that may conceal failures in the face of distribution shifts. Furthermore, typical approaches to expert fusion are heterogeneous and have the drawback of assuming that there are no systematic errors left after fusion, so they cannot adequately model the systematic errors that occur after fusion. This restriction raises a crucial methodological issue: whether combinations of heterogeneous predictive models can yield higher accuracy than conventional models and be robust even if the training and evaluation distributions are different?

In order to solve this problem, the present study introduces a heterogeneous expert-fusion framework with residual refinement for crop-yield prediction called GreenzakaAI+. To address this question, the present study proposes a heterogeneous expert-fusion framework with residual refinement for crop-yield prediction, named GreenzakaAI+. The framework achieves integration of complementary ML and DL experts via sample-adaptive fusion, whereby the contribution of those experts can vary depending on the sample, and then learns the residual error remaining after aggregation. 10 independent how-to models are tested using the same protocol consisting of traditional validation, temporally unseen and spatially unseen data, all with leakage control. Temporal and spatial targets are not included in any model development and selection decisions, but are only used in the post-selection evaluation.

There are three key contributions from this study. Firstly, GreenzakaAI+ integrates multiple experts using sample-adaptive weights instead of fixed-weighted averaging. Second, it also accounts for residual refinement to explicitly model systematic error due to expert aggregation. Third, it introduces a leakage-controlled evaluation framework which critically compares the conventional overall predictive accuracy, the temporal generalization capabilities, the spatial generalization capabilities and the robustness of statistical errors of the proposed model against independent, test baselines. The design has the advantage that the study can quantify the impact of distribution shift as well as see if the improvements are still meaningful.

This paper is organized as follows: in the rest of the paper, we describe the remaining details. The related works and the research gap are discussed in Section 2. The proposed methodology and leakage-controlled evaluation protocol is detailed in Section 3. The experimental results and discussion are presented in Section 4. The paper ends in Section 5, which provides a discussion of future research.

A. Related Work

2.1 Machine Learning and Ensemble-Based Yield Prediction

For tabular data, such as in agriculture, gradient-boosting and ensemble methods are competitive. For tabular data, gradient-boosting like the boosting-based regressors are able to model non-linear yield relationships [1]; stacked meta-models are ensembles of different regressors which leverage complementary error structure across learners [2]; and inverse-error weighting augments the boosting fusion by favoring the members with lower validation errors [3]. Such advances are always reported on one-test, whereby distribution-shift robustness has remained un-tested. This incentivises the use of five independent boosting/bagging baselines (Random Forest, Extra Trees, XGBoost, LightGBM and CatBoost) assessed here in terms of explicit shift and not just in-domain accuracy.

2.2 Deep and Spatio-Temporal Learning

In deep architectures, the spatial and temporal structure is explicitly modeled and the tabular ensembles are extended: convolutional–graph-attention–recurrent fusion models local, inter-regional, and historical dependencies [4]; climate-conditioned deep learning models with climate-change, fertilizer, and crop-area effects are more competitive than statistical and machine-learning results for Australian crops [5]. Both studies do not differentiate between temporal and spatial generalization, so any benefits can not be credited to one dimension versus the other. This encourages us to consider five neural baselines (MLP, Residual MLP, Wide & Deep, Deep Cross Network, Feature Transformer) with the same temporal/spatial protocol.

2.3 Heterogeneous Expert Fusion and Adaptive Weighting

Static averaging is outperformed by all heterogeneous-expert fusion methods reported: a fuzzy-optimized ensemble of several learners tuned with a GA, where the learners are LIME/SVR/XGBoost, and the post hoc learner interpretation is done with LIME/SHAP [6]; inverse-error reweighting of the prediction of several LSTM/SVR/XGBoost learners [3]; using multimodal remote-sensing–meteorological fusion [14]; and learned sample-adaptive gating over several information sources [15]. None of these frameworks, however, accounts for the residual error remaining after fusion, thus leaving any systematic error common to all experts uncorrected. This opening encourages the use of two-stage designs, the first of which is to use sample-adaptive expert gating, and the second of which is to explicitly correct the residuals on the fused output as tested here.

2.4 Distribution Shift and Generalization

Evaluations in agricultural ML are being under increasing examination of evaluation protocol. The Leave-One-Year-Out validation is used in CY-Bench, since using random splits could result in correlated observations from the same year ending up in both partitions, causing leakage and possibly "cold start" estimates [7]. Spatial transfer of drought models in both European and Asian settings demonstrates the need for retraining or local recalibration [10] and spatial analysis of yield gaps for winter wheat confirms domain-dependent performance [11]. The drought–yield causal relevance varies with phenological stage [12] and with the type of drought (meteorological, soil, and vegetation) [13]. Large-scale out-of-distribution benchmarking of deep spatio-temporal models also indicates that the accuracy of in-domain deep models is not predictive of out-of-domain model rankings when the region is held out, and that the degradation is significantly larger than when the year ahead is held out. All of these results confirm that the traditional

validation is not a reliable measure of deployment performance. This encourages to perform identical conventional, temporal, and spatial evaluation of each model in this study in a single pipeline of leakage-controlled preprocessing before training.

2.5 Complementary Agricultural AI Directions

Relative model performance is dependent not only on the crop but also on the region and the resource context, as demonstrated by examples of interpretable feature-importance analysis for legume yield [8], energy-aware model design for smart agriculture [9] and comparative multi-model benchmarking for potato [16, 17] and soybean [18] and even landscape studies using remote-sensing [19]. The directions are intended to facilitate wide-spread multi-paradigmatic baselining without being offered as demonstrated properties of the proposed model.

2.6 Research Gap and Motivation

Three conclusions can be drawn from the literature reviewed. Conventional validation results in better performance than simpler baselines for ensemble or deep-learning methods [1]–[5]. Single-model prediction is enhanced by the introduction of heterogeneous fusion such as adaptive fusion that are based on modelling and learned-gating weighting [3, 6, 14, 15]. It is demonstrated through distribution-shift studies, that usual validation performance is not indicative of temporal or spatial generalization [7], [10]–[13] and [20]. What is missing is the combined treatment of these findings—currently, there is no assessment of the temporal or spatial generalization in existing fusion frameworks and no fusion with residual-correction in existing generalization studies. This is important because in the future year and/or unseen region, which is when the yield forecast is needed to be useful, a model chosen based on in-domain validation might under-perform. GreenzakaAI+ aims to directly explore this gap by being a heterogeneous expert-fusion framework that is designed and evaluated as a robust cross-distribution crop-yield prediction system, compared to independent machine-learning and deep-learning baselines, under conventional system validation, temporal generalization and spatial generalization within a single leakage-controlled pipeline. The design does not assume that dominance over all baselines in all splits will be observed but rather whether a design that results in more balanced generalization when evaluated across a range of evaluation regimes provides a more effective solution than any single evaluated baseline.

3. Methodology

3.1 Overall Framework

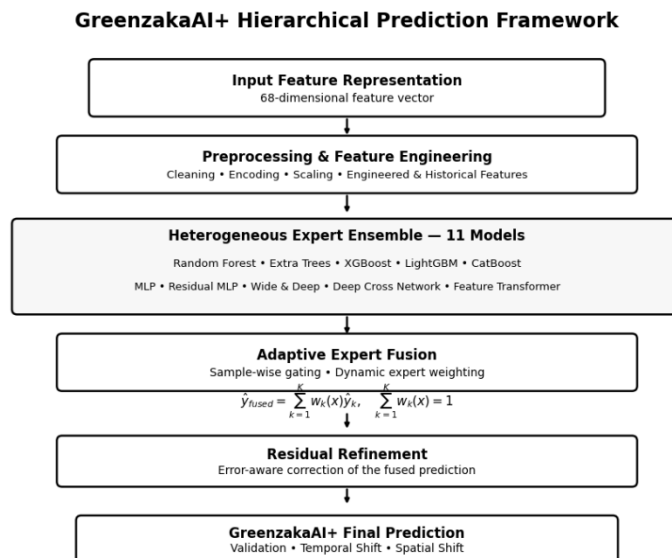


Fig.1. Overall Framework

GreenzakaAI+ is a heterogeneous expert-fusion framework with residual refinement for crop-yield prediction with distribution shift. The pipeline consists of five steps: (i) leakage-controlled data preparation, (ii) independent benchmarking of ten predictors that are heterogeneous in nature, (iii) sample-adaptive expert fusion, (iv) residual-error correction, and (v) evaluation under conventional, temporal, and spatial regimes. The final model has 68 predictors. All optimization and model-selection decisions are made without taking the temporal or spatial targets into account, and only used for post-selection evaluation. This figure 1 shows the whole pipeline from feature representation to expert fusion, residual correction, and cross-distribution evaluation.

3.2 Data Preparation and Leakage Control

Suppose $(N = 362,633)$ is the number of training observations and $(X \in \mathbb{R}^{N \times 68})$ is the predictor matrix and $(y \in \mathbb{R}^N)$ is the target vector. All the preprocessing statistics (imputation values, standardization parameters, categorical encodings) are fitted based on the training partition and applied as is to the other partitions. Therefore, there is no temporal or spatial-target information to affect the model fitting, hyperparameter selection, and optimization of fusion.

3.3 Independent Baseline Benchmark

The protocol consists of 10 predictors: Random Forest, Extra Trees, XGBoost, LightGBM, CatBoost, MLP, Residual MLP, Wide & Deep, Deep Cross Network, and Feature Transformer. All models are represented in the same form, and the same validation, temporal, and spatial partitions are fixed, allowing the comparisons to be made at the architecture level, with the same experimental setup.

3.4 GreenzakaAI+: Adaptive Fusion and Residual Correction

GreenzakaAI+ is a heterogeneous expert fusion and residual refinement. Let $(\hat{y}_i, 1, \dots, \hat{y}_i, K)$ be the predictions of K heterogeneous experts for observation (x_i) . For sample-specific scores, a $(g_k(x_i))$ is used to normalize the scores using

$$w_{i,k} = \frac{\exp(g_k(x_i))}{\sum_{j=1}^K \exp(g_j(x_i))}, \quad \sum_{k=1}^K w_{i,k} = 1 \quad (1)$$

The fused prediction is

$$\hat{y}_i^F = \sum_{k=1}^K w_{i,k} \hat{y}_{i,k} \quad (2)$$

The expert pool consists of four models: MLP, LightGBM, Wide & Deep and Feature Transformer, as well as the previously developed GreenzakaNet component which is not used as a baseline, but rather as a developmental expert.

The remaining prediction error is equal to

$$r_i = y_i - \hat{y}_i^F \quad (3)$$

A residual learner estimates this error as

$$\hat{r}_i = f_r(x_i, \hat{y}_i^F) \quad (4)$$

and the final prediction is

$$\hat{y}_i^G = \hat{y}_i^F + \alpha \hat{r}_i \quad (5)$$

A correction strength was chosen based upon the internal holdout described in Section 3.5. For the evaluated residual learners, the best internal holdout performance was found to be by a medium-capacity Extra Trees regressor, which was retained with $(\alpha = 1)$. This is a two-stage design, which decouples the heterogeneous predictive fusion from the systematic residual correction.

3.5 Development Protocol and Evaluation Metrics

The internal split consisted of 65,246 development observations and 16,311 holdout observations, which were used for fusion and development of the residual model. This holdout was only used for expert-fusion optimization, residual-model selection, and correction-strength tuning. Temporal and spatial targets were not employed in the development phase and were only examined after the modelling is done. The performance of the system is evaluated by using Mean Absolute Error *MAE*, Root Mean Square Error *RMSE* and (R^2) :

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (8)$$

The unseen-generalization score is reported as

$$R_{\text{unseen}}^2 = \frac{R_{\text{temporal}}^2 + R_{\text{spatial}}^2}{2} \quad (9)$$

The individual temporal and spatial results are separately reported while.

3.6 Statistical Validation and Reproducibility

95% confidence intervals are provided using the bootstrap method; paired Wilcoxon error comparisons; paired effect-size estimation; target-range analysis; and error-quantile diagnostics. Fixed partitions, leakage controlled pre-processing are used for all models. Base prediction matrices are pre-computed and reused later when performing fusion and statistical analysis, without retraining and enabling comparisons. The above is the workflow that ensues:

Input Features → Independent Experts → Adaptive Fusion → Residual Correction → GreenzakaAI+

4. Results and Discussion

4.1 Overall Predictive Performance and Benchmark Comparison:

In terms of overall performance when compared with the best baseline, in the 11 models evaluated, the proposed framework, GreenzakaAI+, achieved an R^2 of 92.17% compared to 89.18% for the best baseline, which represents a clear advantage, albeit not consistent, in the proposed framework when evaluated under the conventional framework. In the aspect of temporal shift, GreenzakaAI+ achieved an R^2 value of 86.82%, while in the aspect of spatial shift, it achieved an R^2 value of 95.34%, and the average of the unseen values was 91.08%. Specifically, it showed the best temporal performance of all the models evaluated (87.21%) and the best spatial performance of all the independent baselines (95.08%), which means that no single architecture is clearly superior in all the evaluation regimes. The R^2 performance for all of the evaluation regimes is presented in Table 1 and the comparison of the models is illustrated in Fig. 2 below.

Table 1. Overall Performance Comparison of All Evaluated Models

Rank	Model	Validation R^2 (%)	Temporal R^2 (%)	Spatial R^2 (%)	Unseen Average R^2 (%)
1	GreenzakaAI+	92.1671	86.8213	95.3425	91.0819
2	MLP	88.6248	87.2136	94.6633	90.9384
3	Random Forest	88.2930	86.5784	94.6569	90.6176
4	Wide & Deep	89.1840	86.0851	94.8001	90.4426
5	Feature Transformer	89.0270	86.4329	94.3500	90.3914
6	LightGBM	88.7041	85.5291	95.0789	90.3040
7	Extra Trees	88.0703	86.1963	94.1590	90.1776
8	XGBoost	88.6227	85.3399	94.9338	90.1369
9	CatBoost	88.6947	84.6361	94.5764	89.6062
10	Deep Cross Network	88.9167	85.4576	93.2153	89.3365
11	Residual MLP	88.9701	84.0367	93.9335	88.9851

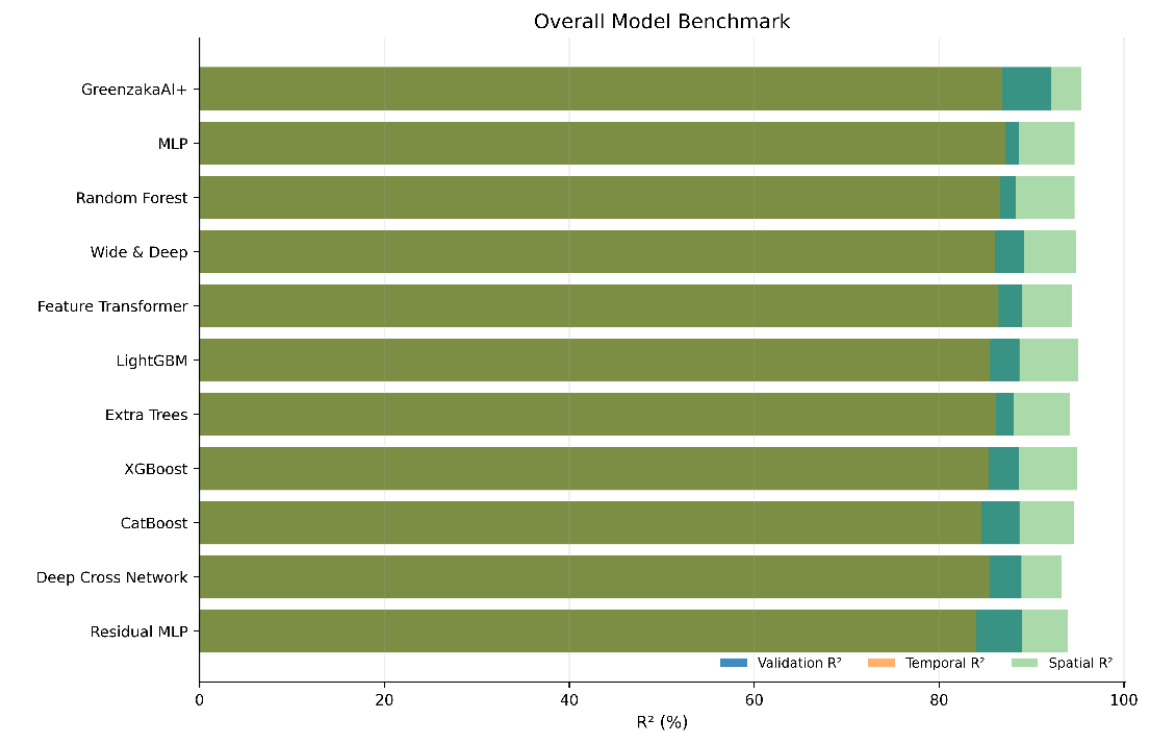


Fig. 2. Overall model benchmark

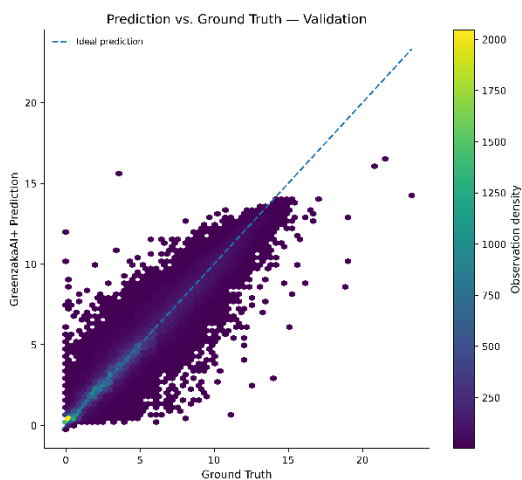


Fig. 3. Prediction vs ground truth

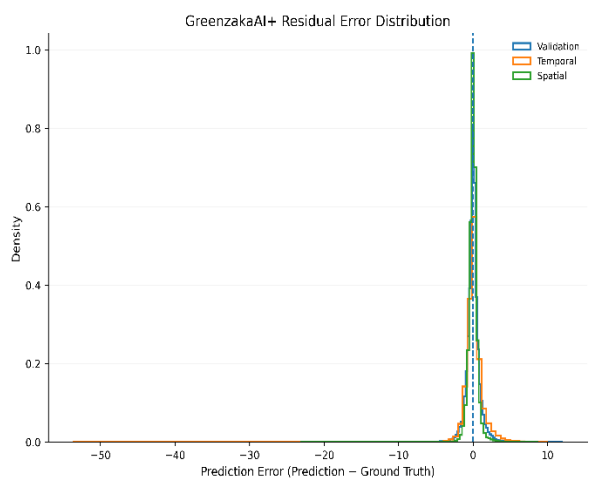


Fig. 4. Residual error distribution

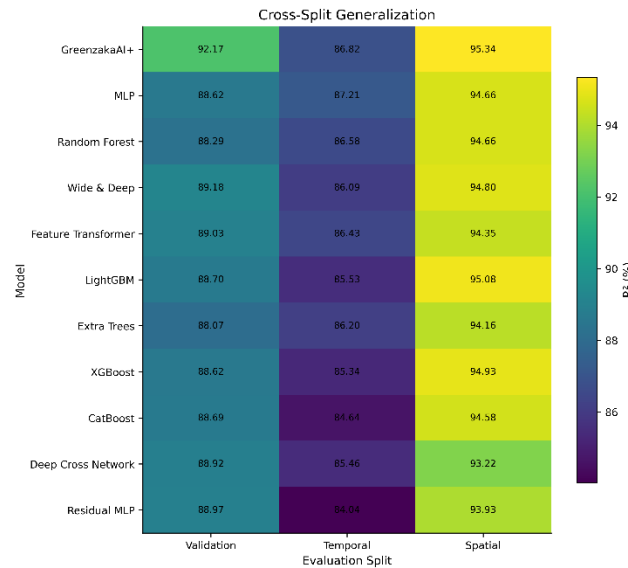


Fig. 5. Cross-split R^2 heatmap

Finally, the validation MAE (0.5979) and RMSE (0.9035) of the evaluated models were the lowest, highlighting the benefit of residual refinement in addition to heterogeneous fusion. The prediction behavior against ground truth is shown in Fig. 3 and the distribution of the prediction residual and the characteristics of the prediction errors are shown in Fig. 4. The cross-split performance of all the models evaluated is also shown in Fig. 5, where the differences in R^2 between validation, temporal and spatial evaluation regimes are highlighted. These results, taken together, suggest that the proposed framework provides a significant improvement over the traditional predictive accuracy (which, in turn, did not involve exposing the proposed test to temporal information or spatial information during the development process).

4.2 Generalization, Error Behavior, and Statistical Robustness

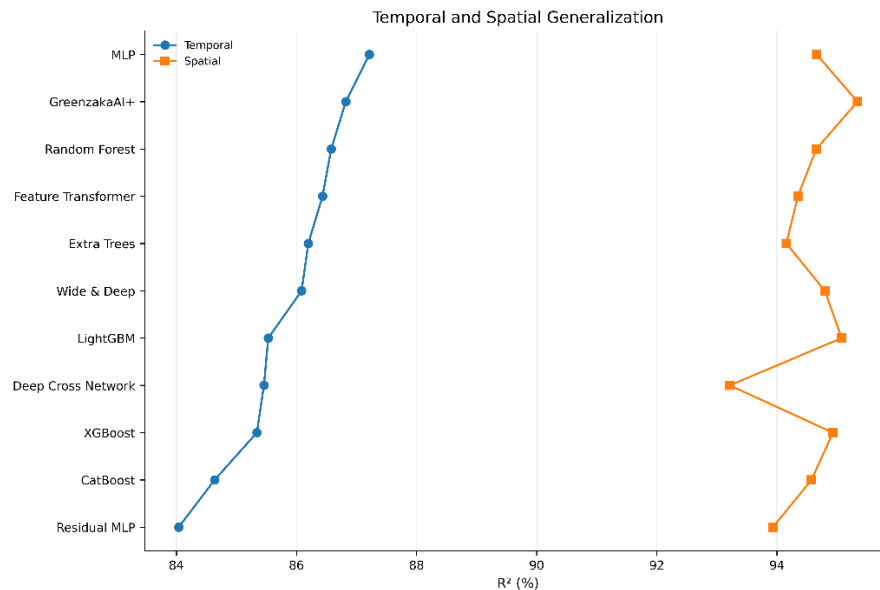


Fig. 6. Temporal and spatial generalization



Fig. 7. Target-range error analysis

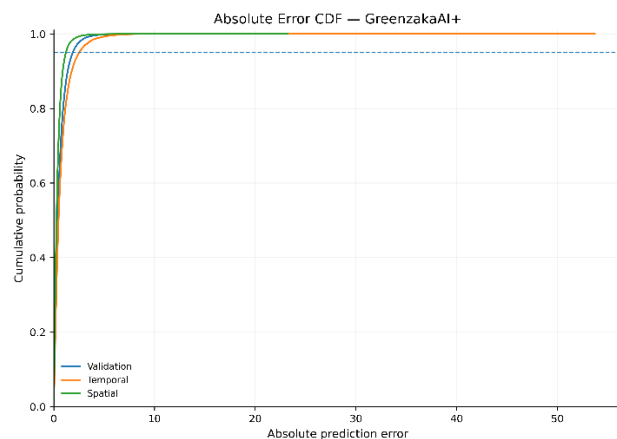


Fig. 8. Error quantile CDF

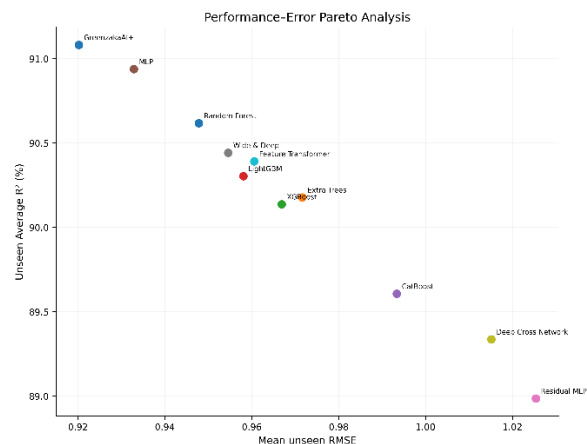


Fig. 9. Performance error Pareto analysis

The results of cross-distribution evaluation show that the predictive performance is very different across temporal and spatial regimes, indicating that relying on traditional validation approaches is not sufficient to describe the actual deployment behaviour in Fig. 6. GreenzakaAI+ scored a R^2 of 86.82% in temporal, 95.34% in spatial, and an unseen-average score of 91.08%. While MLP slightly outperformed GreenzakaAI+ on the temporal split, the proposed framework still achieved the best unseen performance across all independently evaluated baseline models, in which its improvement seemed to be due to its balanced cross-distribution robustness instead of any single axis. A full comparison with the best baseline for each regime is given in Table 2.

Table 2. GreenzakaAI+ versus Best Independent Baselines

Evaluation Regime	GreenzakaAI+ R^2 (%)	Best Independent Baseline	Baseline R^2 (%)	ΔR^2 (pp)
Validation	92.1671	Wide & Deep	89.1840	+2.9830
Temporal	86.8213	MLP	87.2136	-0.3923
Spatial	95.3425	LightGBM	95.0789	+0.2636
Unseen Average	91.0819	MLP	90.9384	+0.1434

This picture is further confirmed by error analysis, which found GreenzakaAI+'s median absolute error to be 0.383, 95th percentile error as 1.884, and 99th percentile error as 3.192. The model's main generalization problem seems to be temporal, as evidenced by the higher tail uncertainty for the temporal prediction (4.546 at the 99th percentile), while the spatial prediction's uncertainty was comparatively stable throughout the distribution of errors. The evaluation of the target-range behavior is shown in Fig. 7, the cumulative error distribution is shown in Fig. 8 and the performance error tradeoff for the different models tested is further illustrated in Fig. 9.

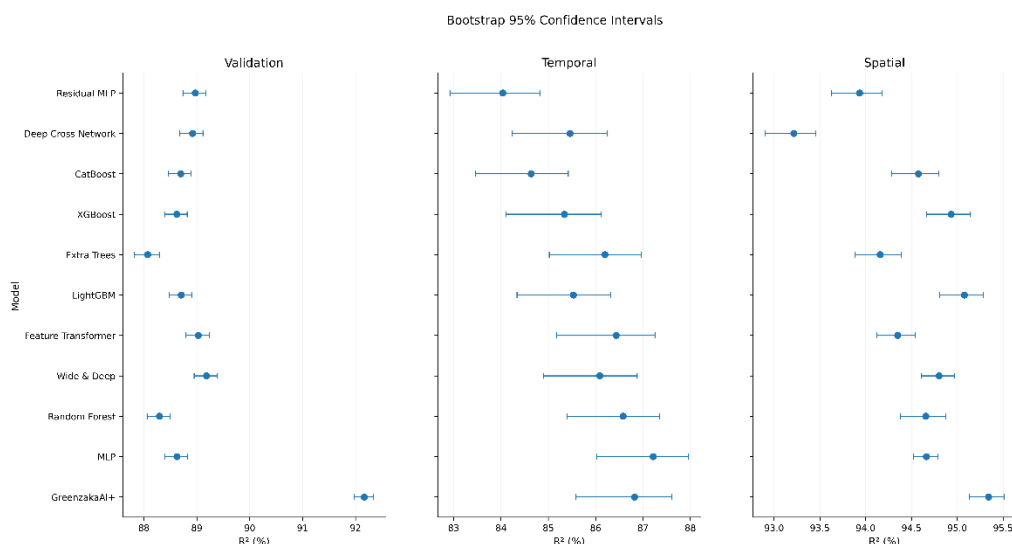


Fig. 10. Bootstrap 95% confidence intervals

The reported estimates were stable across bootstrapping as demonstrated by 95% confidence intervals of 91.98–92.34% for validation R², 85.59–87.61% for temporal R², and 95.14–95.51% for spatial R² in Fig. 10. Comparisons of the statistical errors of the observed differences to the evaluated baselines also showed that there were no statistically significant error differences were found between the evaluated baselines when considering the paired statistical comparisons, and so the observed performance advantage should be considered in light of the confidence intervals reported.

4.3 Development Analysis and Contribution of GreenzakaAI+

The proposed framework is progressively improved, as seen from Table 3 and Fig. 11, greenzakaNet gets an unexpected R² of 90.64%, while reaching 90.95% with GreenzakaAI_v2, and finally 91.08% with the final configuration of GreenzakaAI+. This shift, from GreenzakaAI_v2 to GreenzakaAI+ led to a 2.62% increase in validation R² while leaving comparatively small increases in temporal and spatial R² (+0.16% and +0.11%, respectively) - showing that most of the benefit of residual correction seems to reside in conventional validation, and that there were small improvements under distribution shift.

Table 3. Development and Ablation Analysis of GreenzakaAI+

Development Stage	Validation R² (%)	Temporal R² (%)	Spatial R² (%)	Unseen Average R² (%)
GreenzakaNet	89.6042	86.3123	94.9752	90.6437
GreenzakaAI_v2	89.5466	86.6660	95.2350	90.9505
GreenzakaAI+	92.1671	86.8213	95.3425	91.0819
GreenzakaNet → GreenzakaAI+ (Δ pp)	+2.5629	+0.5090	+0.3673	+0.4381

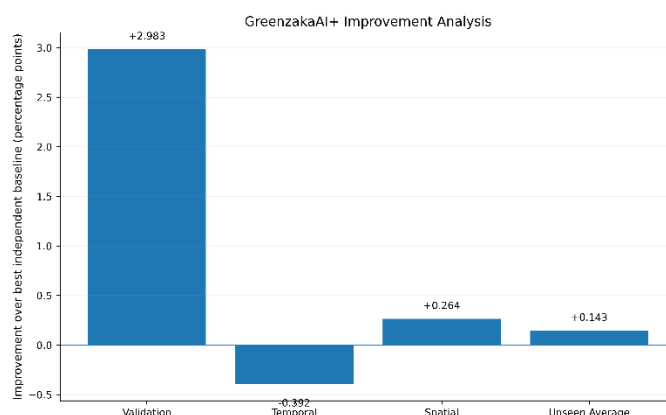


Fig. 11. GreenzakaAI+ improvement analysis

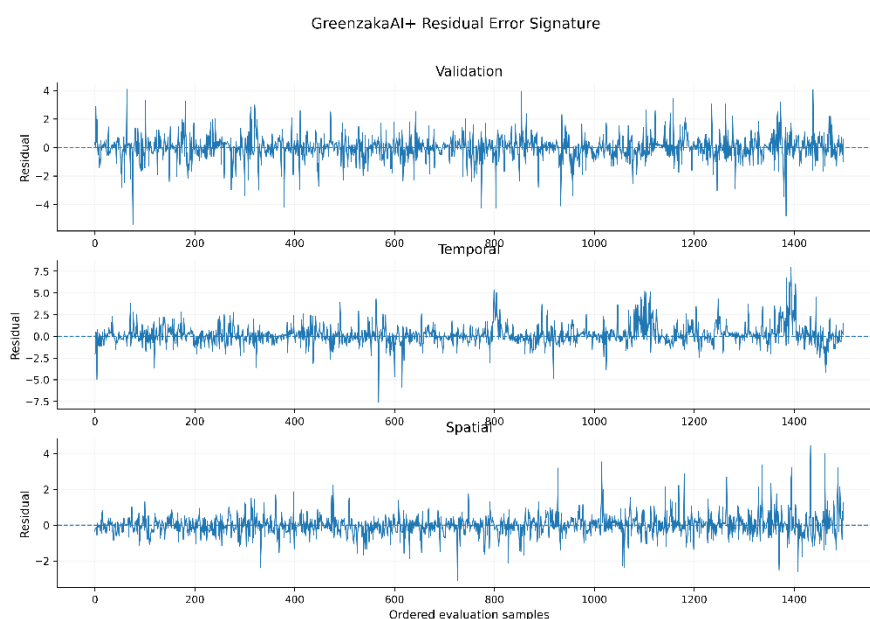


Fig. 12. Residual error signature

The result of the analysis of the residual-error signature is provided in Fig. 12. Overall these analyses suggest that the main strength of GreenzakaAI+ is not that it is always the best on every individual evaluation split, but that it combines the two key principles of strong conventional accuracy in a way that is compatible with the main strength of GreenzakaAI: heterogeneous expert fusion to boost accuracy and simultaneously residual refinement to maintain competitive robustness in both temporal and spatial distribution shift. The installed factor is quoted as the overall comparison with the strongest independent baselines, in Table 2.

5. CONCLUSION

In this work, we proposed a heterogeneous expert-fusion framework for crop-yield prediction under distribution shift: GreenzakaAI+. It integrates complementary machine-learning and deep-learning predictors, using adaptive fusion, and models the residual prediction error by residual correction. A leakage-controlled evaluation protocol was used in the validation, in the temporal unseen configuration and in the spatially unseen configuration, with the temporal and spatial targets being separate from the model development and selection process. GreenzakaAI+ was validated at 92.17%, 86.82% and 95.34%, with an unseen-average of 91.08%. The model was significantly better than the best validation baseline, and performed competitively under temporal and spatial distribution shifts. The results were further ensured by bootstrap confidence intervals, paired statistical comparisons, residual analysis and evaluation during the development stage.

However, although these results were obtained, temporal generalization is somewhat difficult in comparison to the MLP that performed slightly better on the temporal . Further research is thus warranted in the following areas of external-dataset validation, stronger temporal adaptation, and greater assessment of computational-efficiency. The results suggest heterogeneous expert fusion followed by residual refinement is a good method to enhance the robustness of crop yield prediction under varying data distributions.

References

- [1] P. Mahesh and R. Soundrapandiyan, "Yield prediction for crops by gradient-based algorithms," *PLoS One*, vol. 19, no. 8, Art. no. e0291928, 2024, doi: 10.1371/journal.pone.0291928.
- [2] V. Ramesh and P. Kumaresan, "Stacked ensemble model for accurate crop yield prediction using machine learning techniques," *Environmental Research Communications*, vol. 7, no. 3, Art. no. 035006, 2025, doi: 10.1088/2515-7620/adb9c0.
- [3] A. Ravikumar, V. Periyasamy, K. M. K. Devi, R. G. Krishnamoorthy, and O. K. Mariavalavan, "A hybrid SERWI ensemble model for crop yield prediction using an inverse RMSE weighting strategy," *Scientific Reports*, vol. 15, Art. no. 45085, 2025, doi: 10.1038/s41598-025-31987-y.
- [4] L. Wang, Z. Chen, W. Liu, and H. Huang, "A temporal–geospatial deep learning framework for crop yield prediction," *Electronics*, vol. 13, no. 21, Art. no. 4273, 2024, doi: 10.3390/electronics13214273.
- [5] H. Demirhan, "A deep learning framework for prediction of crop yield in Australia under the impact of climate change," *Information Processing in Agriculture*, vol. 12, no. 1, pp. 125–138, 2025, doi: 10.1016/j.inpa.2024.04.004.
- [6] A. Ikram, S. Ikram, E.-S. M. El-Kenawy, A. Hussain, A. H. Alharbi, and M. M. Eid, "A fuzzy-optimized hybrid ensemble model for yield prediction in maize-soybean intercropping system," *Frontiers in Plant Science*, vol. 16, Art. no. 1567679, 2025, doi: 10.3389/fpls.2025.1567679.
- [7] M. Kallenberg et al., "CY-Bench: a comprehensive benchmark dataset for sub-national crop yield forecasting," *Earth System Science Data*, vol. 18, no. 6, pp. 3997–4018, 2026, doi: 10.5194/essd-18-3997-2026.
- [8] T. Petropoulos, L. Benos, R. Berruto, G. Miserendino, V. Marinoudi, P. Busato, C. Zisis, and D. Bochtis, "Interpretable machine learning for legume yield prediction using satellite remote sensing data," *Applied Sciences*, vol. 15, no. 13, Art. no. 7074, 2025, doi: 10.3390/app15137074.

- [9] E. Elbasi, Y. I. Alzoubi, A. E. Topcu, and M. Nadeem, "Green AI for smart agriculture: Energy-efficient predictive models for crop yield and resource management," *IEEE Access*, vol. 13, pp. 204924–204953, 2025, doi: 10.1109/ACCESS.2025.3637027.
- [10] H. Sahu, P. K. Garg, S. Vijay, and A. Dasgupta, "Predictive drivers and transferability of multi-scale machine learning based crop yield prediction under drought across European and Asian climates," *Agricultural Water Management*, vol. 327, Art. no. 110254, 2026, doi: 10.1016/j.agwat.2026.110254.
- [11] S. R. Mousavi, V. A. Jahandideh Mahjenabadi, B. Khoshru, and M. Rezaei, "Spatial prediction of winter wheat yield gap: agro-climatic model and machine learning approaches," *Frontiers in Plant Science*, vol. 14, Art. no. 1309171, 2024, doi: 10.3389/fpls.2023.1309171.
- [12] S. Yeşilköy, Ö. Baydaroğlu, and I. Demir, "Linking drought indicators and crop yields through causality and information transfer: a phenology-based analysis," *Scientific Reports*, vol. 16, Art. no. 2373, 2026, doi: 10.1038/s41598-025-32185-6.
- [13] S. Chang, E. Isaev, H. Chen, B. F. Wu, N. N. Yan, Z. H. Ma, and J. H. Meng, "Exploring the linkages between different types of drought and their impacts on crop production in Kyrgyzstan," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 4566–4580, 2024, doi: 10.1109/JSTARS.2024.3359429.
- [14] A. D. Yewle, L. Mirzayeva, and O. Karakuş, "Multi-modal data fusion and deep ensemble learning for accurate crop yield prediction," *Remote Sensing Applications: Society and Environment*, vol. 38, Art. no. 101613, 2025, doi: 10.1016/j.rsase.2025.101613.
- [15] F. Mena, D. Pathak, H. Najjar, C. Sanchez, P. Helber, B. Bischke, P. Habelitz, M. Miranda, J. Siddamsetty, M. Nuske, M. Charfuelan, D. Arenas, M. Vollmer, and A. Dengel, "Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction," *Remote Sensing of Environment*, vol. 318, Art. no. 114547, 2025, doi: 10.1016/j.rse.2024.114547.
- [16] E.-S. M. El-Kenawy, A. A. Alhussan, N. Khodadadi, S. Mirjalili, and M. M. Eid, "Predicting potato crop yield with machine learning and deep learning for sustainable agriculture," *Potato Research*, vol. 68, pp. 759–792, 2025, doi: 10.1007/s11540-024-09753-w.
- [17] A. Alsaber, A. Satpathi, M. Alsabah, and P. Setiya, "Optimizing potato yield predictions in Uttar Pradesh, India: a comparative analysis of machine learning models," *Scientific Reports*, vol. 15, Art. no. 26897, 2025, doi: 10.1038/s41598-025-12719-8.
- [18] L. B. Santos, D. Gentry, A. Tryforos, L. Fultz, J. Beasley, and T. Gentimis, "Soybean yield prediction using machine learning algorithms under a cover crop management system," *Smart Agricultural Technology*, vol. 8, Art. no. 100442, 2024, doi: 10.1016/j.atech.2024.100442.
- [19] F. Meng, Z. Ren, P. Zhang, C. Wang, S. Hong, R. Geng, W. Hong, X. Wang, B. Huang, B. Zhang, and Y. Bai, "Estimation of the relationship between urban landscape pattern and crop yield by remote sensing data and field measurement," *Remote Sensing*, vol. 17, no. 22, Art. no. 3667, 2025, doi: 10.3390/rs17223667.
- [20] A. Chakravarty, "Out-of-distribution generalization in climate-aware crop yield prediction with Earth observation data," in *Proc. ICCV 2025 Workshop on Sustainability with Earth Observation and AI*, 2025, arXiv:2510.07350.