

effectively exploit engineered features (Taşan et al., 2022). Gradient-boosting families (CatBoost, LightGBM, XGBoost) have also shown excellent performance for general crop yield prediction and for eggplant yield based on genotype-related variables, where CatBoost provided accurate and robust forecasts, highlighting the suitability of tree-based boosting for eggplant yield modeling under varying environmental and management conditions (Islam et al., 2023; Mahesh and Soundrapandiyan, 2024).

5 Evaluation and Optimization of Yield Prediction Performance

5.1 Comparison of regression and machine learning algorithms

Crop yield prediction studies consistently show that machine learning algorithms often outperform simple regression when relationships between climate, management, and yield are nonlinear and complex. Comparative evaluations across linear regression, decision trees, random forests, support vector machines, and neural networks report that ensemble methods such as Random Forest and Gradient Boosting generally achieve higher accuracy and better generalization than traditional linear models, especially when diverse environmental and management variables are included (Kurmi and Singh, 2025). However, linear models remain competitive when relationships are close to linear, offering advantages in interpretability and lower computational cost (Nazir et al., 2025).

Broader multi-crop comparisons confirm that advanced tree-based models and k-nearest neighbors often provide lower error and higher correlation with observed yields than multiple linear regression, particularly when many climatic and soil predictors are used. Recent work further extends comparisons to deep learning (e.g., LSTM and Bi-LSTM), showing that optimized recurrent networks can substantially reduce prediction error relative to support vector regression and time-series models such as ARIMA and VAR, demonstrating the value of capturing temporal dependencies in climate and yield series (Kumar et al., 2023).

5.2 Accuracy assessment using evaluation indicators

Evaluation of yield prediction models relies on multiple complementary indicators to capture both error magnitude and explanatory power. Common metrics include root mean squared error (RMSE), mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2), which together provide a comprehensive view of prediction bias, dispersion, and goodness-of-fit (Kurmi and Singh, 2025; Nazir et al., 2025). Studies comparing regression and machine learning approaches typically rank models by minimizing RMSE and MAE while maximizing R^2 , revealing clear performance hierarchies among algorithms under different data conditions (Pant et al., 2025).

Large-scale forecasting frameworks and ensemble systems also employ normalized RMSE (NRMSE) and additional agreement indices to compare machine-learning baselines against operational forecasting systems or process-based crop models, emphasizing reproducibility and robustness across crops, regions, and seasons (Paudel et al., 2020; Singh et al., 2025). In practice, these metrics are often computed under cross-validation or using independent test years, allowing rigorous assessment of generalization and facilitating fair comparison of alternative algorithms for integrating fertilization and climate variables in yield prediction (Sowmya and Prasad, 2024).

5.3 Optimization of model parameters and prediction stability

Model performance and stability depend strongly on appropriate hyperparameter tuning and feature selection. Grid-search and other systematic optimization methods applied to tree-based ensembles such as Random Forest and Gradient Boosting have been shown to significantly improve RMSE, MAE, and R^2 compared with default configurations, with tuned ensembles delivering more robust rice yield predictions under variable climatic conditions (Hoque et al., 2024; Sowmya and Prasad, 2024). Similarly, combining multiple tuned base learners in stacked or adaptive ensembles can further reduce prediction error relative to any single model, demonstrating the benefits of leveraging diverse algorithmic strengths (Sánchez et al., 2014).

Advanced frameworks integrate hybrid feature selection and metaheuristic optimization to enhance both accuracy and efficiency. For example, coupling clustering and correlation-based filters with feature selection methods, followed by hyperparameter optimization of support vector regression via an improved Crayfish Optimization