Sarikonda et al., 2025). To avoid bias and overfitting, workflows also emphasize correct partitioning schemes and prevention of information leakage, along with modular feature creation from weather, soil, remote sensing, and crop-model outputs (Paudel et al., 2020; Morales and Villalobos, 2023).

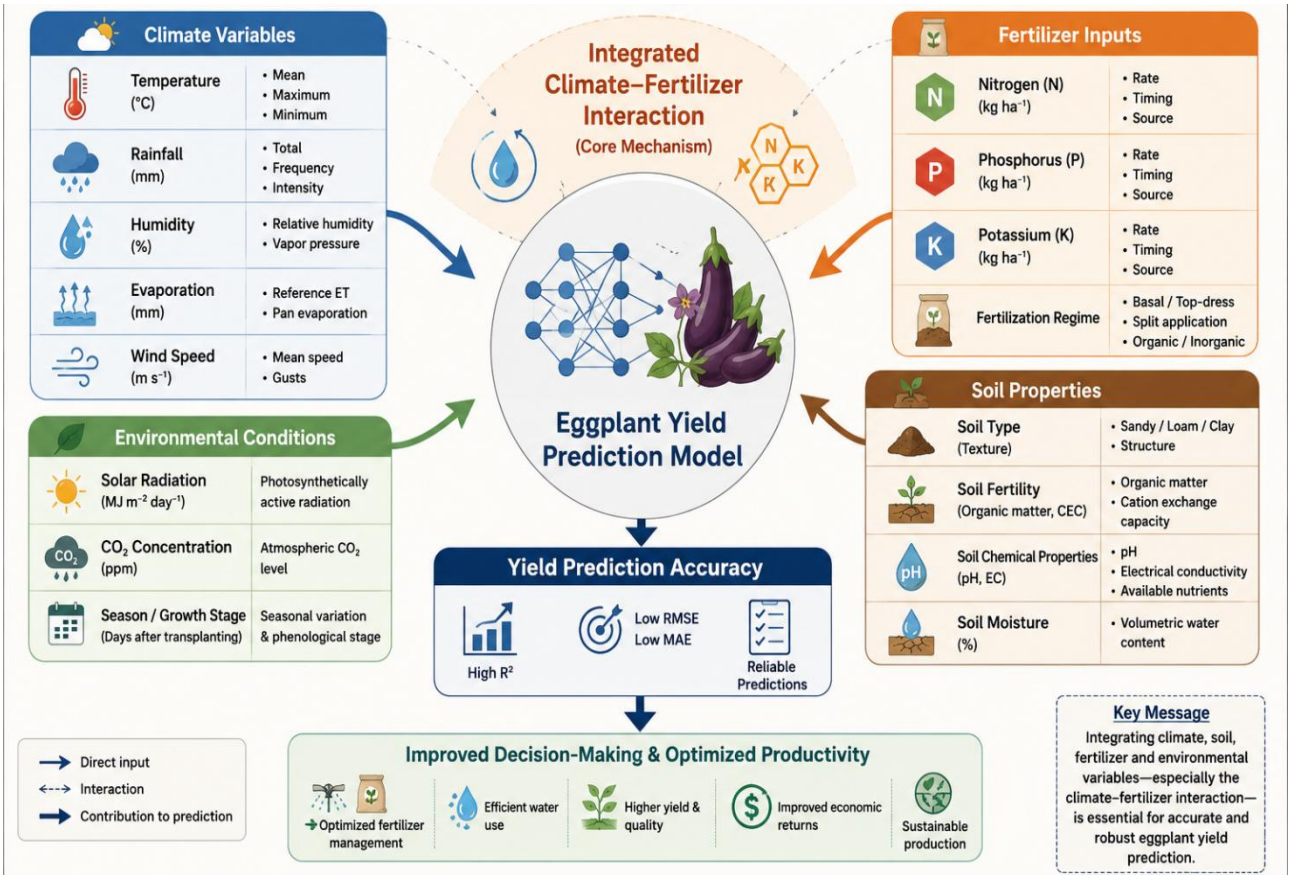

Figure 1 Conceptual framework of key input variables used in machine learning-based eggplant yield prediction models. Climate, soil, fertilizer, and environmental variables jointly influence prediction accuracy and crop productivity responses

Feature selection and extraction are key to reducing redundancy and improving generalization. Relief-based feature selection and linear discriminant analysis have been used to isolate the most discriminative predictors before training support vector machines, k-nearest neighbors, and random forests for yield classification or regression (Gupta et al., 2022). Hybrid approaches combine correlation-based filters, clustering, and recursive feature elimination to build reduced, information-rich datasets that, together with optimized support vector regressors, substantially improve prediction accuracy while lowering computational cost, illustrating the value of systematic feature engineering pipelines (Abdel-Salam et al., 2024).

4.3 Development of statistical and machine learning models

A wide range of statistical and ML algorithms has been applied to crop yield prediction, offering guidance for constructing eggplant-specific models. Linear regression, random forest, gradient boosting trees, and related methods are among the most widely used, with random forest and boosting-based techniques often achieving strong performance across diverse environments and crops (Mahesh and Soundrapandiyan, 2024; Shawon et al., 2024). Ensemble models that integrate multiple learners (e.g., Extra Trees, gradient boosting, or stacked approaches) have repeatedly reached very high R^2 and low error metrics, suggesting that ensemble strategies are promising for capturing complex fertilization-climate-yield relationships (Iniyan et al., 2023; Nossam et al., 2024).

For eggplant specifically, machine learning models using spectral vegetation indices, days after planting, and irrigation-related coefficients have successfully predicted yield; principal component analysis-based inputs combined with artificial neural networks achieved very high accuracy, indicating that nonlinear models can