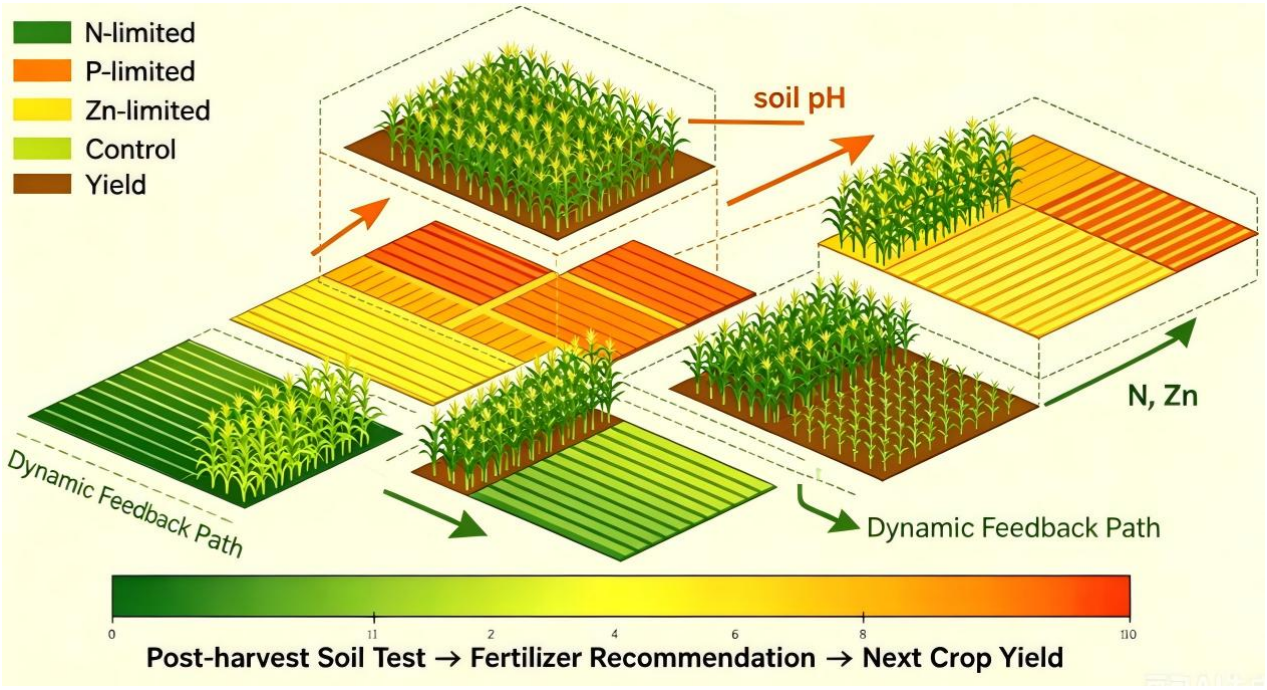

Figure 2 Spatial heterogeneity of soil nutrient limitations and their effects on maize yield

4.3 Feature selection and dimensionality reduction methods

High-dimensional soil-climate datasets require effective feature selection (FS) to avoid overfitting and reduce computational cost. Reviews of machine-learning yield models emphasize that optimal feature sets, obtained by FS, are essential because only a subset of soil, climate, and management variables truly drive prediction accuracy (Hara et al., 2021). In a dedicated framework for yield prediction, a Relief-based FS step was combined with linear discriminant analysis feature extraction, before applying machine-learning classifiers, which markedly improved accuracy over models using all raw variables (Gupta et al., 2022).

Comparative studies of dimensionality reduction for crop yield forecasting show that combining FS and feature extraction (FX) can outperform either alone. In rice yield models based on vegetation and temperature indices, a hybrid approach (FSX) integrating FS with principal component-type FX improved RMSE by up to 60% relative to using all features, and FSX-based models outperformed pure FS or FX in most regions (Pham et al., 2022). More recent works in crop yield prediction apply hybrid FS pipelines (e.g., correlation-based filters, ANOVA, ensemble FS) coupled with advanced learners such as XGBoost or optimized SVR, consistently reporting higher predictive accuracy and lower error once redundant and noisy predictors are removed.

5 Methods for Prediction Model Construction

5.1 Traditional statistical modeling methods

Traditional statistical methods for yield prediction are mainly based on linear or polynomial relationships between yield and a limited set of explanatory variables, often weather indices. Multiple linear regression and its variants have long been used as benchmarks when comparing newer machine learning approaches for maize and other crops, typically using growing-season temperature and precipitation plus a time trend to represent technological progress (Leng and Hall, 2020). Extensions such as quadratic, interaction, and polynomial regression have also been applied to maize and other cereals, and can achieve reasonable accuracy when relationships are approximately linear and the number of predictors is small (Shastry et al., 2017).

More recent work has introduced penalized regression techniques (LASSO, Elastic Net, ridge), which perform variable selection and effectively handle multicollinearity among many weather indices (Vashisth and Aravind, 2026). For maize in semi-arid New Delhi, Elastic Net outperformed stepwise multiple linear regression across vegetative, flowering, and grain-filling stages, with the lowest RMSE and normalized RMSE, highlighting the