collocate each observation with soil, terrain, and weather variables to form a dense spatio-temporal sample set (Filippi et al., 2019).

Large-area studies, such as county-level maize analyses in the US Midwest or regional work in Northeast China, construct samples by merging official yield statistics with gridded or station-based climate data, soil maps, and multi-source satellite products (Figure 3) (Kang et al., 2020; Li et al., 2022). In Ghana, plot-level samples from hundreds of maize field trials are georeferenced and linked to 0-30 cm soil properties, climate variables during the planting season, and management practices, enabling model training across wide environmental and agronomic ranges (Asamoah et al., 2024).

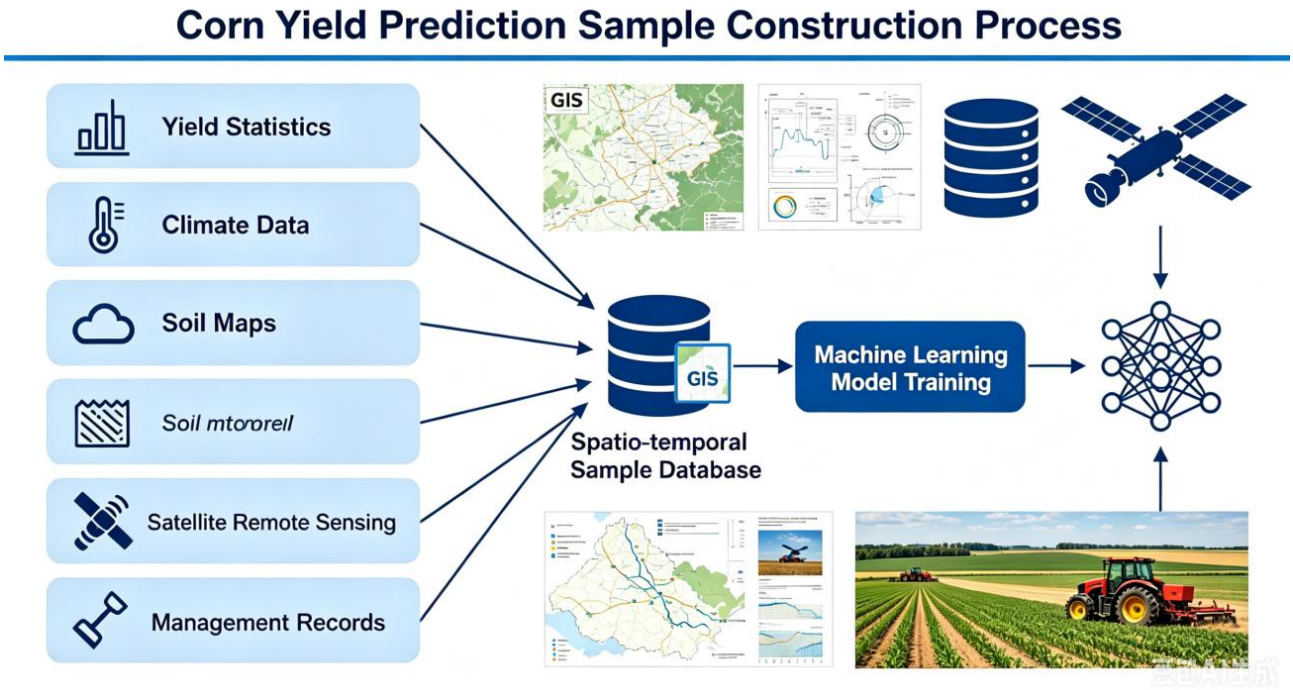

Figure 3 Workflow for integrating multi-source environmental and agricultural datasets into maize yield prediction samples

7.2 Comparative Analysis Of Multi-Model Prediction Results

Comparative studies consistently show that model performance depends strongly on algorithm choice and input richness. At the plot scale, combining vegetation indices, climate, soil, and fertilizer data, Random Forest and Adaptive Boosting clearly outperform linear regression, SVM, GPR, and KNN, with R^2 often above 0.85 and lowest RMSE values (Meng et al., 2021). In a Hungarian field using detailed spatio-temporal soil and micro-relief measurements, XGBoost surpassed neural and kernel methods, reaching test accuracies above 95%, while lattice-based smoothing further improved predictive AUC (Nyéki et al., 2021).

At regional scales, ensemble or tree-based machine learning models generally outperform both traditional regression and deep learning architectures. In the US Midwest, XGBoost provided the most accurate and stable county-level maize forecasts when hundreds of environmental features were used, while LSTM and CNN did not show clear advantages (Kang et al., 2020). Across Northeast China, an ensemble of several ML methods improved yield prediction over individual linear and ML models when integrating environmental and multi-sensor satellite data, explaining more than 70% of maize yield variability (Li et al., 2022).

7.3 Result validation and agricultural application analysis

Robust validation is essential to ensure that multi-model predictions have practical value. Studies highlight that naïve random data splits can substantially overestimate predictive skill, especially when the goal is true forecasting rather than interpolation within a season (Morales and Villalobos, 2023). More rigorous schemes, such as nested k-fold cross-validation across years and fields, or leave-one-field/leave-one-year-out designs, better