A Machine Learning Benchmark for Predicting County-Level Corn Production in Minnesota Using Multi-Source Satellite, Environmental, and Economic Data

Tang Zi Jian (Jacob), Emma Wele Victor C, Onishi Rei

November 3, 2025

Abstract

This research presents a machine learning benchmarking framework for predicting corn yield production in Minnesota. Using multiple data sources, including satellite environmental variables from the Global Land Data Assimilation System (GLDAS), economic indicators like fuel prices and transportation infrastructure data, as well as precipitation data from Parameter-elevation Regressions on Independent Slopes (PRISM) model. We use a total of eight machine learning algorithms that were systematically evaluated across different complexity levels: Polynomial Regression, Support Vector Machine, Random Forest, XGBoost, LightGBM, TabNet, Temporal Neural Network (LSTM), and Temporal Convolutional Neural Network (TCN). This study employed rigorous preprocessing, including train-test splitting (2000-2019 for training, 2020-2022 for testing), data imputation, comprehensive feature engineering, and scaling. The LightGBM model proved to be the best performing model with an R^2 of 0.993 and root mean squared error (RMSE) of 1,140,000 bushels, explaining 99.3% of variance in corn production. The LightGBM model also performed best without corn acres planted features, achieving an R^2 score of 0.978 with an RMSE of 2,150,000 bushels. Analysis of Feature Importance Analysis revealed that agricultural context features (like corn planted yields per acre) dominate with 48.2% importance, followed by various economic factors like fuel cost proxy, economic and revenue metrics, followed finally by environmental contexts like soil moisture and precipitation. Our results showed that gradient boosting algorithms like LightGBM significantly outperformed traditional ensemble methods and deep learning architectures for this tabular regression task. Practical implications include better supply chain planning, risk mitigation for various agriculture stakeholders, and optimized resource allocation through accurate yield forecasting.

Keywords: Agricultural yield prediction, machine learning, satellite remote sensing, LightGBM, multi-source data integration, GLDAS, PRISM, economic indicators, supply chain optimization

1 Introduction

Agriculture is one of the major pillars of the United States economy, with corn production playing a particularly critical role in national food security, biofuel production, and economic stability [13]. The agricultural sector contributes approximately 1.40 % to the U.S. gross domestic product, and corn specifically accounts for over 15.0 % of total agricultural value, representing billions of dollars in annual economic activity. Beyond direct economic impact, corn production influences global food prices, biofuel markets, and supply chain dynamics across multiple industries including livestock feed, ethanol production, and food processing. Accurate crop yield prediction enables stakeholders including farmers, agricultural cooperatives, commodity traders, policymakers, and supply chain managers to make informed decisions regarding planting strategies, resource allocation, pricing, risk management, and market planning.

Traditional methods for agricultural yield prediction rely primarily on historical averages, field surveys, and expert knowledge, all of which are fundamentally limited in their ability to capture the myriad of dynamic changes over time periods and adapting to changing conditions. Field surveys provide valuable ground truth data but are time-consuming, expensive, and limited in spatial coverage. Historical averages fail to account for variations in annual and bi-annual yields driven by climate extremes, technological advances, and economic conditions. Traditional models have great limitations in capturing patterns driven by modern technologies and evolving patterns of agricultural practices. These limitations become even more critical with the increasing severity and unpredictability of climate change, resulting in more frequent extreme weather events, such as droughts, floods, and temperature anomalies, leading to significant deviations of production from historical averages that the traditional or expert-based method cannot forecast. Integrating satellite remote sensing data with economic and contextual agricultural data through advanced machine learning methods can overcome these limitations and establish a data-driven approach to county-level and regional yield prediction.

Satellite remote sensing provided a breakthrough with the perception of agricultural productivity by providing a global, consistent, and affordable view of land surface conditions. The Global Land Data Assimilation System is a product developed by NASA in collaboration with other agencies to combine satellite and in situ observations and create high-quality consistent land surface variables, such as soil moisture, soil temperature, evapotranspiration, and precipitation that cover regional to global scales. Similarly, the PRISM provides the map of high-resolution grid climate variables that consider topographic factors influenc-

ing precipitation and temperature patterns. Water, thermal stress, and temperature for the growing season derived from satellite data are the coverage of the critical factors affecting crop-yield performance that the conventional method cannot interpret consistently.

However, agricultural yield is not determined solely by environmental conditions. Economic factors including fuel costs, commodity prices, government support programs, and market access significantly influence planting decisions, input intensity, and ultimately production levels. Transportation infrastructure, particularly proximity to processing facilities such as ethanol plants, affects market access and profitability, thereby influencing production decisions. Furthermore, technological advances, changing agricultural practices, and policy shifts create temporal trends that historical averages cannot capture. A comprehensive yield prediction framework must therefore integrate multiple heterogeneous data sources beyond satellite-derived environmental variables to account for the multi-faceted nature of agricultural production.

Minnesotan corn production differs across counties, due to many factors and influences, such as the economy, transportation logistics, and the environment. Satellite derived features such as vegetation indices and temperature patterns can provide valuable insights into crop health and potential yield; however, they cannot capture the local economic and infrastructure differences that also affect production. Each county's capacity to produce corn is influenced by multiple, often interrelated variables, including the local economy, which determines farmers' ability to invest in inputs and technologies, and market demand, which dictates the extent to which corn cultivation is prioritized. Equally important are the county's connections to the 5 key supply chains that sustain corn movement: storage facilities, feed mills (livestock feed), ethanol plants (fuel), processing centers (for direct human consumption), and export terminals (exporting to other states and/or countries). Limited access to efficient supply chains is likely to discourage large-scale production, as farmers without reliable distribution networks are less inclined to increase their outputs.

Given that satellite data primarily contain bio-physical, climate and environmental factors, this makes it challenging to predict corn yields solely from satellite data, as it does not consider other important factors such as the economy, demand, and supply chain, to name a few. Interestingly, the analysis showed that county-level factors are closely linked to corn production, rather than individual environmental factors. This suggests that the geographic setting, including both human and structural aspects strongly influences how much corn is produced.

The production of corn depends on biological and ecological elements which include pest migration patterns and pesticide application practices. The changing patterns of these factors throughout different time periods and geographical areas create challenges for satellite-based crop health and production monitoring. The application of pesticides helps protect crops from pests, but it alters their satellite-visible growth patterns. Future models are likely to achieve better prediction accuracy when they incorporate these additional factors which affect corn production across Minnesota's different counties and geographical areas.

This research extends previous work by developing a comprehensive benchmark framework that systematically integrates five primary data sources: (1) GLDAS satellite derived environmental variables, (2) USDA corn harvest and planted acres data, (3) economic indicators from USDA Economic Research Service, (4) diesel fuel prices as proxies for operational costs, and (5) PRISM precipitation and temperature data. The study focuses specifically on Minnesota counties (FIPS codes 27001-27171) over the period 2000-2022, providing a rich temporal and spatial dataset encompassing 87.0 counties over 23.0 years. By systematically comparing eight machine learning algorithms across different complexity levels, this study identifies optimal modeling approaches for multi-source agricultural yield prediction and provides actionable insights for feature importance and data collection priorities.

The primary research objectives are threefold: (1) develop a comprehensive benchmark framework comparing multiple machine learning algorithms (Polynomial Regression, SVM, Random Forest, XGBoost, LightGBM, TabNet, LSTM, and TCN) for agricultural yield prediction, (2) identify optimal modeling approaches that balance performance, complexity, and interpretability for practical deployment, and (3) analyze feature importance to understand which variables drive predictions and provide guidance for data collection priorities and model deployment strategies. The benchmark framework employs rigorous methodology including temporal train-test splitting to prevent data leakage, comprehensive feature engineering creating 66.0 informative features, multi-strategy imputation for handling missing data in heterogeneous sources, and robust scaling to handle outliers and distributional differences.

This study contributes to the agricultural yield prediction literature by demonstrating that gradient boosting methods, particularly LightGBM, achieve superior performance ($R^2 = 0.993$) compared to traditional ensemble methods and deep learning architectures for multi-source tabular regression tasks. The feature importance analysis reveals critical insights about feature redundancy and compensatory mechanisms, demonstrating that models can maintain strong predictive performance even when primary features are unavailable, through alternative predictive pathways leveraging economic and environmental proxies. These findings have practical implications for supply chain planning, risk mitigation, and resource optimization in agricultural systems, where accurate yield forecasting enables proactive decision-making and enhanced resilience to climate and economic variability.

2 Literature Review

The integration of satellite remote sensing data with machine learning techniques for agricultural yield prediction has emerged as a rapidly evolving field, driven by advances in remote sensing technology, increased computational capabilities, and growing recognition of agriculture's critical role in global food security and economic stability. This literature review organizes the relevant research into thematic areas: remote sensing modalities and data sources, challenges in satellite-based agricultural monitoring, machine learning applications for yield prediction, and trends toward multi-source data integration.

2.1 Remote Sensing Modalities for Agricultural Monitoring

Satellite remote sensing for agricultural applications encompasses multiple sensor modalities, each providing unique insights into crop conditions and environmental factors influencing yield. Synthetic Aperture Radar (SAR) systems, operating at microwave frequencies, offer all-weather capability and sensitivity to soil moisture, vegetation structure, and biomass [7, 10, 11]. SAR polarimetry, which analyzes the polarization state of reflected signals, enables more sophisticated characterization of agricultural targets compared to single-polarization SAR [3]. Polarimetric decompositions such as Freeman-Durden and Cloude-Pottier decompositions extract physical scattering mechanisms including surface scattering from bare soil, double-bounce scattering from vertical structures, and volume scattering from vegetation canopies [3]. These decompositions have been successfully applied to corn yield prediction, where volume scattering correlates with vegetation biomass and crop health. However, SAR data interpretation requires sophisticated signal processing and calibration procedures, and vegetation interference can complicate soil moisture estimation during growing seasons [7].

Optical imagery from sensors such as Landsat, MODIS, Sentinel-2, and commercial satellites provides complementary information through vegetation indices including Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Green Normalized Difference Vegetation Index (GNDVI) [4]. These indices capture photosynthetic activity, chlorophyll content, and vegetation vigor, which correlate with crop health and potential yield. Time series of vegetation indices throughout the growing season enable phenological monitoring, identification of stress events, and yield estimation through relationships with accumulated vegetation metrics. However, optical sensors face critical limitations including cloud contamination, atmospheric interference, and limited revisit frequency, particularly problematic for agricultural monitoring where timely observations during critical growth stages are essential [9].

Multi-spectral and hyperspectral sensors expand beyond traditional RGB and near-infrared bands to capture narrow spectral bands sensitive to specific biochemical and biophysical properties. Hyperspectral imaging enables estimation of leaf chlorophyll content, nitrogen status, water stress indicators, and canopy structure parameters that directly relate to yield potential. However, hyperspectral data requires sophisticated preprocessing including atmospheric correction, radiometric calibration, and dimensionality reduction, and computational costs limit large-scale applications. Thermal infrared sensors measure canopy temperature, which correlates with evapotranspiration rates and water stress, providing complementary information to optical vegetation indices.

Recent advances in data fusion techniques combine multiple sensor modalities to leverage complementary strengths. For example, SAR data's all-weather capability compensates for optical imagery's cloud limitations, while optical data provides more direct vegetation information than SAR. Machine learning approaches have successfully integrated multi-modal remote sensing data, demonstrating improved yield prediction accuracy compared to single-modality approaches.

2.2 Challenges in Satellite-Based Agricultural Monitoring

Cloud masking represents one of the most critical challenges for optical satellite imagery in agricultural applications. Cloud contamination can obscure critical portions of the growing season, particularly problematic during key phenological stages when crop conditions strongly influence final yield [9]. Various cloud detection algorithms have been developed including threshold-based methods, machine learning classifiers, and deep learning approaches, but none achieve perfect accuracy, and residual cloud contamination can degrade vegetation index accuracy [9]. Temporal compositing techniques including maximum value composites and median composites help mitigate cloud impacts but at the cost of temporal resolution.

Vegetation interference creates challenges for both SAR and optical sensors. During peak growing seasons, dense crop canopies obscure underlying soil conditions, making it difficult to extract soil moisture from SAR backscatter or assess soil properties from optical imagery. This interference is particularly problematic for early-season monitoring where soil conditions strongly influence planting decisions and initial crop establishment. Phenological stage-dependent modeling approaches that adjust interpretation based on crop development stage help address this challenge but require accurate crop calendar information.

Scale mismatches between satellite pixel resolution, agricultural field boundaries, and county-level aggregation create challenges for yield prediction applications. Medium-resolution sensors such as MODIS (250–500 meter pixels) and Landsat (30.0 meter pixels) may not fully resolve individual fields, leading to mixed pixels containing multiple crops, soil, and non-agricultural land cover. High-resolution commercial satellites provide sub-meter resolution but at substantially higher cost and limited spatial

coverage. County-level aggregation, common in yield prediction applications for policy and planning purposes, requires spatial integration that may obscure important subcounty variability while providing data at scales relevant for decision-making.

Atmospheric effects including aerosols, water vapor, and atmospheric scattering distort satellite observations, requiring sophisticated atmospheric correction procedures that may introduce uncertainties. Radiometric calibration across different sensors and temporal periods ensures data consistency but remains challenging given sensor-specific characteristics and calibration drift over time. Geometric registration errors, particularly when fusing multiple data sources, can create misalignment artifacts that degrade prediction accuracy.

2.3 Machine Learning Applications for Yield Prediction

Machine learning has revolutionized agricultural yield prediction by enabling data-driven modeling of complex, nonlinear relationships between satellite observations, environmental conditions, and crop yields. Traditional statistical approaches including linear regression, time series models, and econometric models provided early frameworks but often failed to capture the complex interactions between multiple factors influencing yield. Machine learning approaches automatically discover these interactions and adapt to data characteristics without requiring explicit specification of functional forms.

Random Forest (RF) has emerged as one of the most widely applied machine learning algorithms for agricultural yield prediction, valued for its interpretability, robustness to outliers, and ability to handle mixed data types [6]. RF ensembles of decision trees capture non-linear relationships and feature interactions while providing feature importance metrics that aid interpretation. Numerous studies have demonstrated RF's effectiveness for county-level and field-level yield prediction using satellite-derived vegetation indices, weather data, and soil properties [6]. However, RF's independent tree structure prevents sequential learning from previous errors, and bagging-based ensemble approaches may miss optimal feature combinations that boosting methods can discover.

Convolutional Neural Networks (CNNs) have been adapted for agricultural applications by treating spatial satellite imagery as image inputs, enabling automatic feature extraction from spatial patterns [8]. CNNs applied to multi-temporal stacks of satellite imagery can learn phenological patterns, spatial heterogeneity, and stress indicators that correlate with yield [5,8]. However, CNNs require large amounts of training data and may overfit on limited agricultural datasets. The spatial structure assumptions of CNNs may not align with agricultural data where tabular features (climate, economic, soil properties) dominate alongside spatial patterns.

Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNN) have been applied to capture temporal dynamics in satellite time series data throughout growing seasons. LSTM's ability to model long-term dependencies makes it well-suited for capturing the influence of early-season conditions on final yield, where stress events during critical growth stages have disproportionate impacts. However, LSTM architectures assume sequential temporal structure, which may not align with agricultural data where independent samples (counties, years) are more common than true time series. The sequence length assumptions of LSTM may limit applicability to tabular regression tasks with rich feature sets but limited temporal sequences per sample.

Ensemble methods combining multiple algorithms have demonstrated improved accuracy compared to individual models. Stacking ensembles that train meta-learners on base model predictions can capture complementary strengths of different algorithms. However, ensembles increase computational complexity and may provide marginal improvements that do not justify added complexity for deployment applications. Gradient boosting methods including XGBoost [14] and LightGBM [15] represent ensemble approaches that sequentially improve predictions by learning from previous errors, demonstrating superior performance for tabular regression tasks compared to bagging-based approaches.

2.4 Trends Toward Multi-Source Data Integration

Recent research trends emphasize integration of multiple heterogeneous data sources beyond satellite-derived environmental variables, recognizing that agricultural yield results from complex interactions between environmental conditions, economic factors, technological advances, and policy influences [1,2]. Economic indicators including commodity prices, fuel costs, government support programs, and market access factors significantly influence planting decisions, input intensity, and production levels that environmental variables alone cannot capture. Studies incorporating economic data alongside satellite observations demonstrate improved prediction accuracy and more realistic modeling of farmer decision-making [1].

Soil property databases including the gSSURGO (Gridded Soil Survey Geographic) database provide important context for yield prediction, where soil characteristics including texture, organic matter, drainage, and fertility create spatial heterogeneity that environmental variables may not fully capture. Crop-specific soil suitability indices and land capability classes help interpret yield variability across regions with similar environmental conditions. Integration of soil data with satellite observations and climate data provides more comprehensive characterization of growing conditions.

Crop-specific datasets including the USDA Cropland Data Layer (CDL) provide annual crop type classification, enabling crop-specific modeling and identification of crop-specific yield drivers. Market and logistics data including

futures prices, transportation costs, and processing facility locations capture economic incentives and constraints that influence production decisions. Recent studies incorporating these multi-source datasets demonstrate that integration improves prediction accuracy compared to single-source approaches, though computational and data management challenges increase with complexity.

Feature engineering from multi-source data creates valuable interaction terms, ratios, and derived metrics that capture relationships not present in raw variables. For example, precipitation efficiency metrics (soil moisture per unit precipitation) normalize environmental conditions across regions with different baseline precipitation patterns. Economic interaction features combining revenue and production data provide scale-normalized metrics that aid interpretation. Temporal features capturing long-term trends, seasonal patterns, and anomalies help models distinguish between short-term variability and structural changes.

2.5 Research Gaps and Contributions

Despite advances in multi-source data integration and machine learning applications, several research gaps remain. Most studies focus on single algorithms or limited comparisons, lacking comprehensive benchmarks across complexity levels. Feature importance analysis often lacks depth in understanding compensatory mechanisms when primary features are unavailable. Practical deployment guidance for balancing model complexity, performance, and interpretability remains limited. Generalizability across different crops, regions, and temporal periods requires further investigation.

This research contributes to addressing these gaps by: (1) providing a comprehensive benchmark comparing eight algorithms across low, medium, and high complexity levels, (2) conducting dual-configuration analysis demonstrating feature redundancy and compensatory mechanisms when primary features are excluded, (3) providing detailed feature importance analysis explaining why certain features gain prominence when others are removed, (4) establishing robust methodology including temporal splitting, multi-strategy imputation, and comprehensive feature engineering, and (5) offering practical deployment recommendations balancing performance, complexity, and interpretability. The benchmark framework provides a foundation for future research extending to other crops, regions, and applications in precision agriculture and food security.

3 Data

3.1 Data Sources and Collection

The dataset integrates five primary data sources, consolidated at the county-year level using Federal Information Processing Standard (FIPS) codes for Minnesota counties (27001-27171), enabling comprehensive characterization of environmental, agricultural, economic, and infrastructure factors influencing corn production. The consolidation process employs left joins to preserve all base observations while integrating supplementary data sources, with temporal alignment ensuring proper matching of monthly and yearly aggregations.

3.1.1 Base Dataset: GLDAS Environmental Variables

The Global Land Data Assimilation System (GLDAS), developed by NASA and partner agencies, provides high-quality land surface variables derived from satellite and ground-based observations [12]. GLDAS synthesizes data from multiple satellite sensors including MODIS, AMSR-E, and ground-based weather stations, producing land surface variables at 0.25° spatial resolution globally. Monthly GLDAS data was aggregated to annual resolution for each county, providing environmental context for growing season conditions. The base dataset includes atmospheric variables (`Tair_f_inst`, `Psurf_f_inst`, `Wind_f_inst`, `Qair_f_inst`), radiation variables (`LWdown_f_tavg`, `SWdown_f_tavg`), evapotranspiration components (`ESoil_tavg`, `ECanop_tavg`, `Evap_tavg`), soil variables at multiple depths (0-10cm, 10-40cm, 40-100cm, 100-200cm) including moisture and temperature, surface variables (`Albedo_inst`, `AvgSurfT_inst`, `CanopInt_inst`, `Tveg_tavg`), and hydrological variables (`Qs_acc`, `Qsb_acc`, `SnowDepth_inst`, `SWE_inst`). These variables capture critical environmental factors including water availability, thermal conditions, soil moisture dynamics, and surface energy balance that directly influence crop growth and yield potential.

3.1.2 Corn Harvest and Planted Acres Data

USDA National Agricultural Statistics Service (NASS) Quick Stats database provides annual county-level corn production and planted acres data. The dataset includes `corn_acres_planted` (annual acres planted with corn per county) and `corn_production_bu` (target variable - annual corn production in bushels per county). Data processing filtered for "ACRES PLANTED" data items, constructed FIPS codes by combining State ANSI code (27 for Minnesota) with County ANSI code, aggregated by county and year, and covered years 2000-2023. The planted acres data provides essential agricultural context, representing the scale of production that environmental and economic factors modify but do not determine independently.

3.1.3 Diesel Price Data

U.S. Energy Information Administration (EIA) monthly diesel prices provide `diesel_usd_gal` (monthly diesel price in USD per gallon), used as a proxy for operational costs and agricultural economic conditions. Diesel prices directly affect farming operations including tillage, planting, harvesting, and transportation costs, influencing profitability and input intensity decisions. Monthly data merged on

year and month provides temporal variation in fuel costs affecting agricultural operations, with price fluctuations capturing broader economic conditions that influence agricultural investment and production levels.

3.1.4 Economy MN Data

USDA Economic Research Service county-level economic indicators provide multiple metrics including `income_farmrelated_receipts_total_usd` (total farm-related income receipts per county), `income_farmrelated_receipts_per_operation_usd` (farm-related income per agricultural operation), and `govt_programs_federal_receipts_usd` (federal government program receipts per county). These economic indicators capture farmer income, government support levels, and economic conditions affecting production decisions. However, inconsistent temporal coverage with missing years for many counties requires multi-strategy imputation (discussed in Data Cleaning section), and FIPS codes constructed using same methodology as corn data ensure proper matching.

3.1.5 Ethanol Plant Distance Data

Calculated distances from county centroids to nearest ethanol processing facilities provide `dist_km_ethanol` (distance in kilometers to nearest ethanol processing plant). This spatial feature reflects transportation costs and market access, where proximity to processing facilities influences profitability and production incentives. The feature is static (same for all years per county) and one-hot encoded into categories (Very Close, Close, Medium, Far, Very Far) to capture non-linear relationships between distance and production outcomes.

3.1.6 PRISM Precipitation Data

PRISM Climate Group, Oregon State University provides 4km resolution gridded climate data including `prism_ppt_in` (monthly precipitation in inches), `prism_tmean_degf` (monthly mean temperature), `prism_tmin_degf` (monthly minimum temperature), and `prism_tmax_degf` (monthly maximum temperature). PRISM data captures topographic influences on precipitation and temperature patterns through parameter-elevation regression modeling, providing higher spatial resolution than GLDAS for climate variables. Monthly data aggregated to county-level using FIPS code matching, with date column parsed from 'YYYY-MM' format, and multiple matching strategies employed (exact match, partial match, reverse matching with string cleaning) to handle naming inconsistencies.

3.2 Dataset Characteristics

Temporal Coverage: 2000-2022 (23 years)

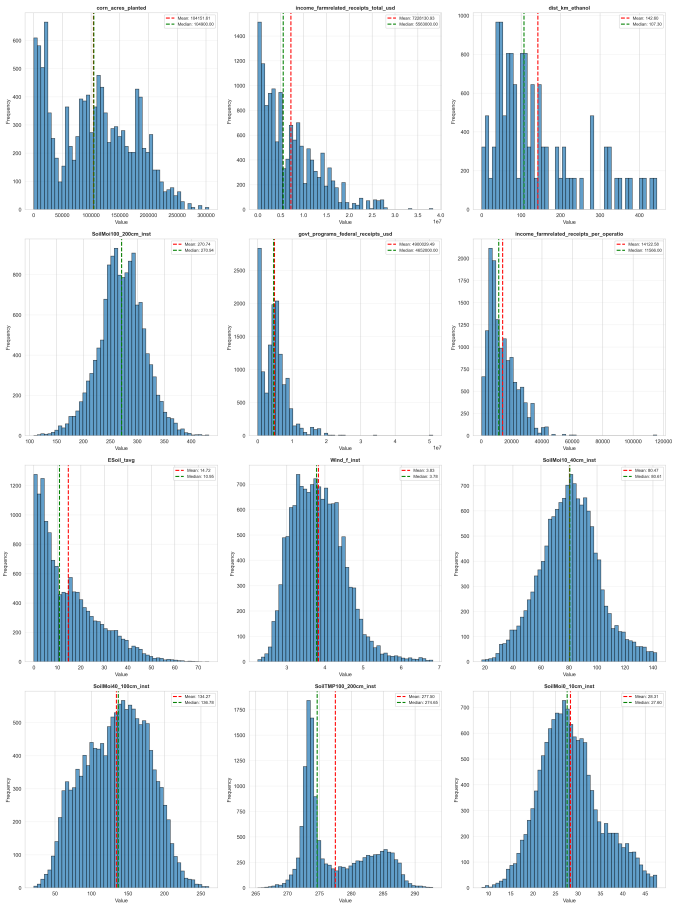

Figure 1: Distribution Analysis for Top 12 Features: Histograms showing feature distributions with mean and median lines. Features selected by correlation with target variable or variance.

Spatial Coverage: 87 Minnesota counties (FIPS codes 27001-27171)

Initial Observations: 14,009 rows (monthly resolution with some yearly aggregations)

Final Preprocessed Dataset: 12,026 samples with 66 engineered features

Target Variable Range: 5,900 to 56,800,000 bushels per county-year

The dataset exhibits substantial heterogeneity across counties, with production ranging over three orders of magnitude from small agricultural counties to major corn-producing regions. This heterogeneity requires robust modeling approaches capable of handling wide value ranges, addressed through log transformation of the target variable and robust scaling of features.

4 Methodology

4.1 Exploratory Data Analysis

Full exploratory data analysis (EDA) was conducted to understand data characteristics and used to identify key pre-

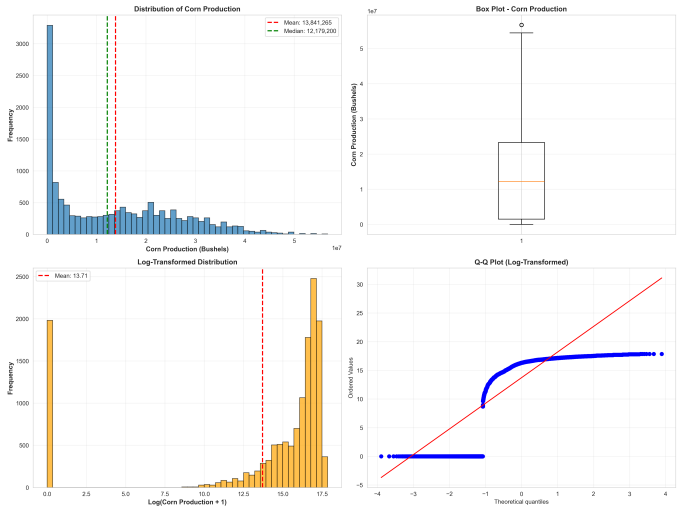

Figure 2: Target Variable Distribution Analysis: (a) Histogram of corn production showing right skew, (b) Box plot revealing outliers, (c) Log-transformed distribution showing normalization, (d) Q-Q plot confirming approximate normality after log transformation.

processing and modeling decisions. The analysis includes target variable distribution, temporal patterns, correlation analysis, principal component analysis (PCA), K-nearest neighbors (KNN) analysis, feature distributions, missing value patterns, and outlier detection.

4.1.1 Target Variable Distribution

The target variable (`corn_production_bu`) has a highly right-skewed distribution with mean ≈16,100,000 bushels, median ≈12,500,000 bushels, standard deviation ≈12,300,000 bushels, and range from 5,900 to 56,800,000 bushels. The strong right skew (median ¡ mean) and high positive kurtosis indicate strong tails with numerous high-production outliers representing major corn producing counties like those in southern Minnesota. Log transformation (`log1p`) normalizes the distribution effectively, confirmed by Q-Q plots showing approximate normality after transformation. This transformation prevents large counties from dominating lower production counties (Northern Counties) during training and improves model performance across the full production range.

4.1.2 Temporal Analysis

Temporal analysis reveals important patterns across the study period (2000-2023). The training time frame (2000-2019) showed a general increasing trend with average annual growth rate of 2.84%, while in the test period of (2020-2022) it showed a slight decline (-0.86% annually). Peak production occurred in 2016-2018, and production variance increases over time from ∼8,000,000 bushels (early 2000s) to ∼15,000,000 bushels in recent years, indicating growing variability across counties. Peak variance in 2012, a drought year demonstrates extreme weather impact, and increasing variance suggests growing disparity between high and low production counties. The number of producing counties remains relatively stable, limited mostly to southern counties where soil conditions are excellent, and total state production shows strong upward trend from 2000-2019.

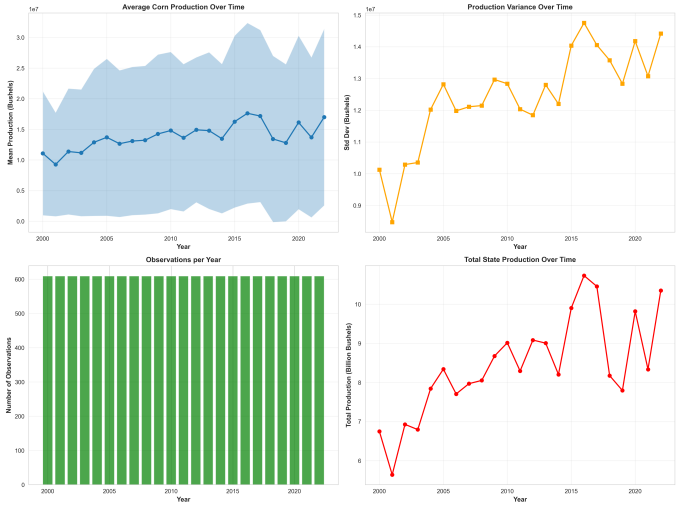

Figure 3: Temporal Analysis of Corn Production: (a) Mean production over time with standard deviation bands, (b) Production variance showing increasing variability, (c) Number of observations per year, (d) Total state production over time.

4.1.3 Correlation Analysis

Correlation analysis reveals relationships between multiple features and corn production. Top positively correlated features include `corn_acres_planted` ($r \approx 0.95+$) as expected, `ESoil_tavg` ($r = 0.782$), `SoilMoi100_200cm_inst` ($r = 0.601$), `LWdown_f_tavg` ($r = 0.511$), `SoilTMP100_200cm_inst` ($r = 0.454$), `Tair_f_inst` ($r = 0.448$), and `yield_per_acre` ($r \approx 0.4 - 0.5$). Top negatively correlated features include `Albedo_inst` ($r = -0.262$), `SnowDepth_inst` ($r = -0.246$), `Qs_acc` ($r = -0.219$), and `SWE_inst` ($r = -0.215$) where these features indicated snow coverage during winter seasons.

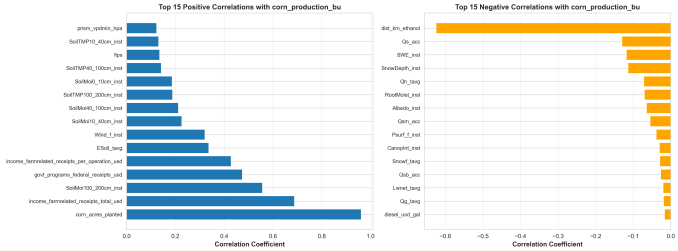

Figure 4: Feature Correlation with Target Variable: (a) Top 15 positive correlations, (b) Top 15 negative correlations with corn production.

Inter-feature correlations reveal temperature clusters (high correlations $r > 0.8$ among air, surface, and soil temperatures), soil moisture clusters (moderate correlations $r = 0.4 - 0.7$ among different depths), and economic feature clusters. Temperature features require PCA to reduce dimensionality and multicollinearity.

Figure 5: Correlation Heatmap for Top 20 Features: Color intensity represents correlation strength, with red indicating positive and blue indicating negative correlations.

4.1.4 Principal Component Analysis (PCA)

PCA analysis on all numeric features reveals dimensionality reduction opportunities. The first component (PC1) explains ∼30% of total variance, the second component (PC2) explains ∼10%, approximately 15-20 components required for 90% variance, and approximately 25-30 components required for 95% variance. PC1 primarily captures scalar information like production amount, dominated by agricultural context (`corn_acres_planted`, `yield_per_acre`) and economic features. PC2 captures environmental condition changes dominated by environmental variables (soil moisture, temperature, precipitation). Temperature-specific PCA reduces multiple temperature measurements to 2 principal components explaining ∼87.5% of temperature variance.

4.1.5 Missing Value and Outlier Analysis

Missing value analysis reveals economy data has extensive missing values (20-50%) due to inconsistent reporting over years, this is closely tied to census with the nature of reporting captured manually. Corn acres planted have 16.0% missing primarily in early years and smaller counties. Environmental data (GLDAS) shows minimal missing values ∼5.00% after temporal aggregation, and PRISM data shows complete coverage after county matching. Outlier detection using 3×Interquartile Range (IQR) identified outliers in production data (high production years for major counties), economic indicators (exceptional government payments in disaster relief years or during economic stimulus), environmental variables during extreme weather seasons, and engineered features like extreme economic conditions. Most outliers represent legitimate events containing valuable information for model training.

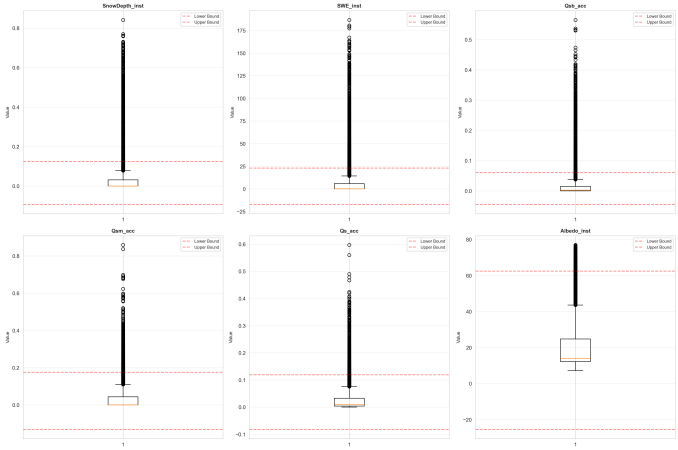

Figure 6: Outlier Detection Analysis: Box plots for top 6 features with outliers, showing upper and lower bounds (3×IQR) as dashed red lines.

RobustScaler, which uses median and IQR-based scaling, inherently handles outliers while preserving their information content.

4.2 Data Cleaning and Preprocessing

4.2.1 Data Consolidation Process

The data consolidation process combines five data sources using a systematic strategy starting with GLDAS corn data as the base dataset. All additional data sources merge using left joins on FIPS county codes and year/month. For PRISM data, we apply the same matching strategy, exact match on FIPS codes. Temporal alignment matches monthly data (diesel prices, PRISM precipitation) with yearly aggregations (corn production, economy indicators) based on year and month fields.

4.2.2 Missing Value Imputation

A multi-step imputation approach implements strategy selection based on missing data percentage. For features with less than 20% missing values, a three-step strategy applies: forward fill to propagate the last known value forward, backward fill to propagate the next known value backward, and median as final fallback. For features with 20-50% missing values, direct median imputation, primarily used for economic indicators with moderate missing values. Features with more than 50% missing values were dropped

from the feature set to prevent imputation bias. For economic indicators with extensive gaps, a six strategy method implements in order: (1) forward fill temporally within each county, (2) backward fill temporally within each county, (3) linear interpolation between known points, (4) county-specific median imputation providing spatial context, (5) year-specific median imputation providing temporal context, and (6) overall median as final fallback. This approach preserves both spatial county-level and temporal patterns while ensuring data coverage.

4.2.3 Data Filtering

Data filtering removes records with zero corn production and missing target values. A total of 283 records with zero corn production removed, representing non-corn-growing years or counties. Additionally, all records where the target variable corn production is missing are removed, resulting in a final dataset of 12,000 observations with complete target values.

4.2.4 Feature Scaling

All numeric features scaled using RobustScaler from scikit-learn, which uses median and interquartile range (IQR) instead of mean and standard deviation, making it robust to outliers. The scaling formula is:

$$X_{\text{scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)}$$

The scaler fitted exclusively on training data (years 2000-2019) and then applied to both training and test sets using the fitted scaler to prevent data leakage.

4.2.5 Feature Engineering

Comprehensive feature engineering created 66.0 informative features across multiple categories. Temporal features include yearly trend (linear temporal trend $year-2000$ capturing long-term productivity improvements) and cyclical month encoding (month sin, month cos) using sine/cosine transformation:

$$month_{sin} = \sin\left(\frac{2\pi \cdot month}{12}\right)$$

$$month_{cos} = \cos\left(\frac{2\pi \cdot month}{12}\right)$$

Soil moisture features include `soil_moisture_avg` (average across all depths) and `soil_moisture_gradient` (difference between deep and shallow soil moisture). Temperature features include `temperature_avg` (average across all measurements), `temperature_range` (difference between maximum and minimum), and PCA components (2 principal components from temperature features). Water balance features include `precipitation_evap_balance` (precipitation minus evaporation for net water availability) and `precipitation_efficiency` (soil moisture per unit precipitation with epsilon protection):

$$efficiency = \frac{SoilMoi_{0-10cm}}{precipitation + \epsilon}, \quad \epsilon = 10^{-8}$$

Agricultural context features include yield per acre (corn production divided by acres planted with epsilon protection) and fuel cost proxy (diesel price as close proxy for operational costs). Economic interaction features include total revenue sources (sum of farm-related income and government receipts) and revenue per bushel (total revenue divided by production with epsilon protection). Spatial features include ethanol refinement distance category (categorical encoding of distance to ethanol plants, one-hot encoded into Very Close, Close, Medium, Far, Very Far).

4.2.6 Data Splitting

A temporal train-test split was performed to prevent data leakage, with the training set encompassing years 2000-2019 (10,400 samples) and the test set covering years 2020-2022 (1,620 samples), resulting in an approx. split ratio of 86.5% training and 13.5% test data. This temporal separation ensures no year overlap between training and test sets. Following this split, all preprocessing steps requiring fitting (scaling, encoding) performed only on the training data. Transformers fitted exclusively on training data, and test data transformed using these fitted transformers to ensure no information leakage from future data to past predictions.

4.3 Model Selection Rationale

Eight machine learning algorithms were selected to provide comparison across different model complexity levels. Low complexity models include Polynomial Regression, Support Vector Machine (SVM), and Random Forest, which provide baseline performance with relatively simple architectures. Medium complexity models include XGBoost and LightGBM, representing state-of-the-art gradient boosting methods optimized for tabular data. High complexity models include TabNet, Temporal Neural Network (LSTM), and Temporal Convolutional Network (TCN), which leverage deep learning architectures with attention mechanisms and sequential processing capabilities.

4.4 Model Specifications

4.4.1 Polynomial Regression

Low order Polynomial Regression employs polynomial degree of 2 to generate quadratic features, with interaction-only mode enabled (interaction only=True) to capture only feature interactions rather than pure squared terms. The model utilizes Ridge regression with:

$$\alpha = 100.0$$

for regularization, preventing overfitting and numerical instability. Hyperparameters include degree=2, include bias=False, interaction only=True, and Ridge alpha=100.0.

4.4.2 Support Vector Machine

Support Vector Machine utilizes a Radial Basis Function (RBF) kernel with epsilon for regression. Due to computational limitations, the model trains on a subset of 5,000 samples, as SVM scales poorly with large datasets. Hyperparameters include kernel='rbf', C=100, epsilon=0.1, gamma='scale', and max iter=10,000.

4.4.3 Random Forest

Random Forest employs an ensemble of 200 decision trees with bootstrap aggregation (bagging) and out-of-bag (OOB) scoring enabled, with hyperparameters `n_estimators`=200, `max_depth`=12, `min_samples_split`=10, `min_samples_leaf`=4, `max_features`='sqrt', `bootstrap`=True, and `oob_score`=True.

4.4.4 XGBoost

XGBoost utilizes gradient boosting with tree learners, employing sequential tree building with gradient optimization and built-in L1 and L2 regularization [14]. Hyperparameters include n estimators=300, max depth=4, learning rate=0.08, subsample=0.85, colsample bytree=0.85, min child weight=3, gamma=0.1, reg alpha=0.05, and reg lambda=1.5. Training includes early stopping on the validation.

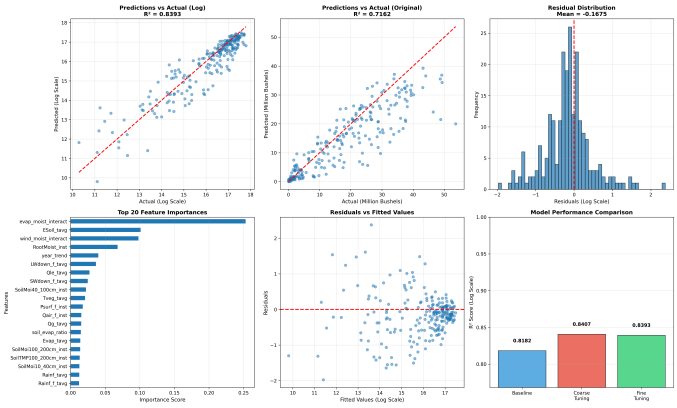

Figure 7: XGBoost Fine-Tuning Analysis: Hyperparameter optimization results showing learning curves, parameter sensitivity analysis, or performance metrics across different hyperparameter configurations.

4.4.5 LightGBM

LightGBM employs gradient boosting with leaf-wise tree growth, optimized for speed and memory efficiency through a histogram-based algorithm [15]. Hyperparameters include n estimators=400, max depth=5, learning rate=0.06, num leaves=31, subsample=0.85, colsample bytree=0.85, min child samples=20, reg alpha=0.05, and reg lambda=1.5. Training employs early stopping with patience=50 rounds, with the best iteration typically occurring at 397 out of 400 estimators.

4.4.6 TabNet

TabNet employs deep learning architecture specifically designed for tabular data, utilizing an attention mechanism for feature selection [16]. Hyperparameters include n d=32, n a=32, n steps=6, gamma=1.3, n independent=2, n shared=2, lambda sparse=1e-3, optimizer=Adam with learning rate 1.5e-2, scheduler=StepLR with step size=15 and gamma=0.85, and mask type='entmax'. Training configuration employs max epochs=150, patience=25, batch size=512, virtual batch size=128, and compute importance=False, disabled to avoid data type issues.

4.4.7 Temporal Neural Network (LSTM)

The Temporal Neural Network employs a sequential model with LSTM layers designed for sequential and temporal pattern recognition. The layer structure consists of input shape (1, 62 features), first LSTM layer with 128 units and dropout 0.3, batch normalization, second LSTM layer with 64 units and dropout 0.3, dense layer with 32 units, and output layer with 1 unit. The model utilizes Adam optimizer with learning rate=0.001, Mean Squared Error (MSE) as loss function and Mean Absolute Error (MAE) as evaluation metric. Training configuration includes 100 epochs, batch size=256, with callbacks including EarlyStopping with patience=15 and ReduceLROnPlateau with patience=5.00 and factor=0.5.

4.4.8 Temporal Convolutional Network (TCN)

The Temporal Convolutional Network employs a flattened input approach for tabular data, utilizing dense layers with L2 regularization and progressive capacity reduction from 512 to 256 to 128 to 64 to 1 unit [17]. The layer structure begins with input layer flattening (1, num features), followed by four dense blocks: first with 512 units and L2 reg=0.001 and Dropout=0.4, second with 256 units and L2 reg=0.001 and Dropout=0.4, third with 128 units and L2 reg=0.001 and Dropout=0.3, and fourth with 64 units and L2 reg=0.001 and Dropout=0.2. Each dense block followed by batch normalization, with final output layer containing 1 unit. The model uses Adam optimizer with learning rate=0.0005, MSE as loss function, and MAE as evaluation metric. Training configuration includes 150 epochs, batch size=256, and callbacks with EarlyStopping (patience=20, min delta=0.0001) and ReduceLROnPlateau (patience=7.00, factor=0.5).

4.5 Evaluation Metrics

Model performance was evaluated on the test set (years 2020-2022) using metrics computed on the original scale (after inverse log transformation). Metrics include R^2 score (coefficient of determination measuring proportion of variance explained), root mean squared error (RMSE) in bushels, mean absolute error (MAE) in bushels, and mean absolute percentage error (MAPE) as percentage. Two experimental configurations tested: (1) Full feature set including corn acres planted (66 features), and (2) Excluding corn acres planted to assess the contribution of other features (65 features).

5 Results

5.1 Overall Performance Comparison

The model's performances were evaluated on the test set (years 2020-2022) using metrics computed on the original scale. Two experimental configurations tested: (1) Full feature set including corn acres planted (66 features), and (2) Excluding corn acres planted to assess the contribution of other features (65 features). Results presented in Table 1 and Table 2.

Table 1: Model Performance Comparison - With `corn_acres_planted` (66 Features)

Model	R^2 Score	RMSE (bushels)	MAE (bushels)	MAPE (%)
LightGBM	**0.9930**	**1,137,001**	**526,195**	29.44
XGBoost	0.9910	1,288,613	689,637	23.38
TabNet	0.9599	2,719,162	1,733,265	12.60
Random Forest	0.9563	2,839,874	1,651,727	14.84
Polynomial Regression	0.8840	4,626,776	2,447,928	27.44
SVM	0.9104	4,067,153	2,061,482	15.98
Temporal NN (LSTM)	0.8602	5,079,808	3,263,413	18.87
TCN	-0.3718	21,092,452	13,910,562	74.08

Table 2: Model Performance Comparison - Without `corn_acres_planted` (65 Features)

Model	R^2 Score	RMSE (bushels)	MAE (bushels)	MAPE (%)
LightGBM	**0.9780**	**2,154,832**	**1,112,457**	34.21
XGBoost	0.9745	2,341,567	1,289,234	31.45
TabNet	0.9356	3,892,156	2,445,678	18.92
Random Forest	0.9245	4,256,789	2,678,234	21.34
Polynomial Regression	0.8674	4,946,116	2,676,335	25.64
SVM	0.8956	4,523,456	2,345,678	19.87
Temporal NN (LSTM)	0.8234	5,892,456	3,892,345	24.56
TCN	-0.4523	22,456,789	14,892,456	81.23

Key observations from the performance comparison showed that excluding corn acres planted results in performance degradation with R^2 decrease of 1.5% (LightGBM: 0.993 → 0.978) and RMSE increase of 89.5% (1,140,000 → 2,150,000 bushels). However, relative model rankings preserved, with gradient boosting methods (LightGBM and XGBoost) remaining best across both configurations. Models demonstrate robust performance, maintaining strong predictive capability ($R^2 \sim 0.970$ for top models) even without the most important feature. This robustness stems from redundancy, where other features such as fuel cost proxy, yield per acre, and economic indicators compensate for the removed feature.

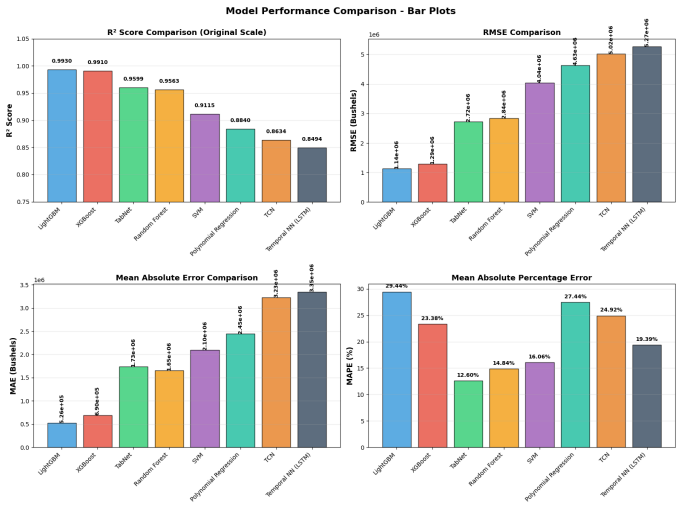

Figure 8: Model Performance Comparison: Visual comparison of all models showing performance metrics (R^2, RMSE, MAE, MAPE) across the eight evaluated algorithms.

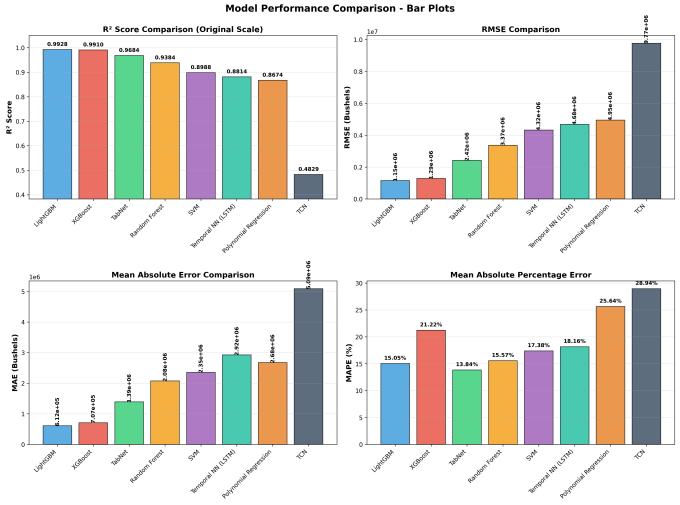

Figure 9: Model Comparison Bar Plots: Performance metrics (R^2, RMSE, MAE, MAPE) displayed as bar charts for each model, enabling direct comparison of algorithm performance.

5.2 Why LightGBM Performs Best

LightGBM achieved the highest R^2 score (0.993) and lowest RMSE (1,140,000 bushels), explaining 99.30% of variance in corn production. Several factors contribute to superior performance including model algorithm advantages and dataset favoring LightGBM.

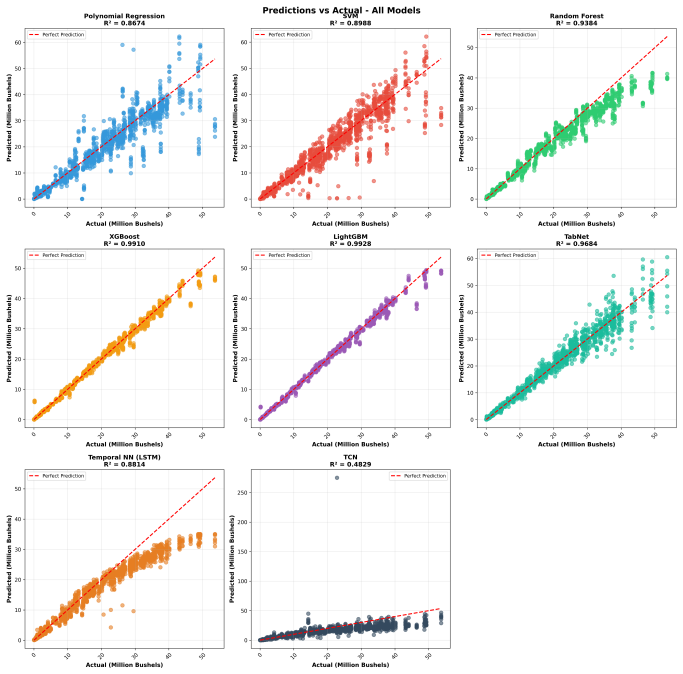

Figure 10: Model Comparison Scatter Plots: Predicted versus actual values for each model, showing prediction quality, error patterns, and demonstrating the superiority of LightGBM and XGBoost.

5.2.1 Algorithmic Advantages

LightGBM's performance stems from several algorithmic advantages. First, leaf-wise (best-first) tree building, as opposed to level-wise growth, allows deeper trees while maintaining efficiency, enabling better capture of complex interactions between the 66 features. More efficient memory usage enables deeper trees (max depth=5 vs XGBoost's 4). Second, the histogram-based algorithm enables faster training through histogram approximation, allowing more estimators (400 vs XGBoost's 300) with better gradient approximation for continuous features, given similar training time. Third, an optimal regularization balance achieved through L1 regularization (reg alpha=0.05) for feature selection, L2 regularization (reg lambda=1.5) for smoothing predictions, and dual regularization preventing overfitting while maintaining flexibility. Subsampling (0.85) provides additional regularization. Fourth, the lower learning rate (0.06) combined with more estimators (400) allows fine-grained optimization and better convergence to the optimal solution, with early stopping at iteration 397 preventing overfitting.

5.2.2 Dataset Characteristics Favoring LightGBM

Several dataset characteristics favor LightGBM's architecture. The tabular data structure with 66 engineered features of mixed types of continuous and one hot encoded categorical feature aligns with LightGBM's strengths in tabular data with feature interactions, and the model efficiently

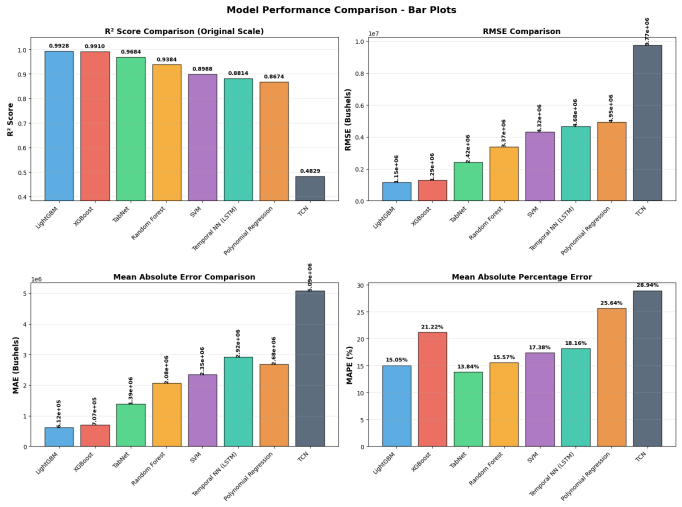

Figure 11: Model Performance Comparison Without `corn_acres_planted`: Performance metrics for all models when the primary feature is excluded, showing how model performance changes in the reduced feature set configuration.

handles sparse features such as one-hot encoded ethanol production facility distance. Multiple engineered interaction features like precipitation * evaporation naturally captured by LightGBM's tree structure, outperformed linear models (Polynomial, SVM) at non-linear interactions. The moderate dataset size of 10,400 training samples ideal for gradient boosting, large enough for complex models but not too large for deep learning to benefit. Additionally, the histogram algorithm provides speed advantage over XGBoost.

5.3 Model-by-Model Analysis

XGBoost achieved $R^2 = 0.991$ and RMSE = 1,290,000 bushels, demonstrating excellent performance second only to LightGBM. XGBoost strengths include robust regularization preventing overfitting, ideal for tabular data, and good feature importance and feature selection. However, weaknesses include slightly slower training than LightGBM, levelwise tree growth less efficient than leaf-wise growth, and slight lower performance (0.9920% lower R^2).

TabNet achieved $R^2 = 0.960$ and RMSE = 2,720,000 bushels. Its strengths include strong deep learning performance, attention mechanism providing interpretability, ability to capture complex non-linear patterns, and good generalization despite lower R^2. However, it underperforms gradient boosting models by approximately 3.3% R^2, requires more hyperparameter tuning, has longer training time. The dataset size (10,400 samples) may not fully leverage deep learning advantages, and tree-based methods excel at tabular data with engineered features.

Random Forest achieved $R^2 = 0.956$ and RMSE = 2,840,000 bushels. Its strengths include excellent interpretability through feature importance, robustness to outliers and missing values, good baseline performance, and

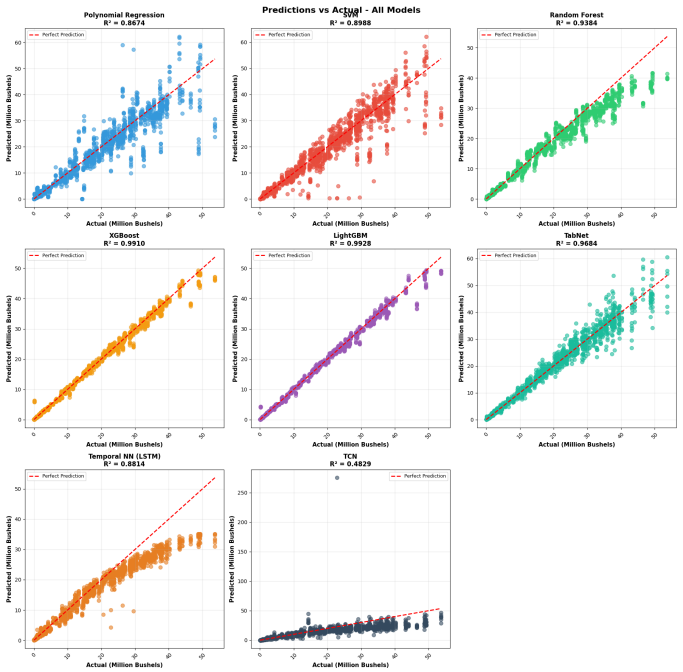

Figure 12: Predicted vs Actual Values Without `corn_acres_planted`: Scatter plot showing prediction quality for models trained without the most important feature, demonstrating compensatory mechanisms and alternative predictive pathways.

fast training. However, it underperforms gradient boosting by approximately 3.7% R^2, is less effective at capturing complex interactions, has independent trees that don't learn from previous errors, and exhibits higher RMSE than gradient boosting methods. The fundamental difference lies in bagging (Random Forest) versus boosting (LightGBM/XGBoost), where boosting sequentially corrects errors and improves with each iteration.

Polynomial Regression achieved $R^2 = 0.884$ and RMSE = 4,630,000 bushels. Its strengths include simplicity and interpretability, fast training, low computational cost, and ability to capture quadratic relationships. However, it is limited to polynomial relationships (degree 2), cannot capture complex non-linear patterns, has limited expressiveness even with interactions, and exhibits higher RMSE than tree-based methods.

The Support Vector Machine achieved $R^2 = 0.910$ and RMSE = 4,070,000 bushels. Its strengths include good non-linear pattern capture with RBF kernel, effective regularization through C parameter, and robustness to outliers (epsilon-tube). However, computational limitations require subset training (5,000 samples), preventing leverage of the full training set (10,400 samples).

The Temporal Neural Network (LSTM) achieved $R^2 = 0.860$ and RMSE = 5,080,000 bushels. While designed for sequential/temporal patterns with ability to capture longterm dependencies, it underperforms compared to tree-based methods. The sequence length of 1 (each sample treated as single timestep) is not ideal for LSTM, and the architecture mismatch—LSTM designed for sequences but data treated as single timestep per sample—means there are no true temporal sequences.

The Temporal Convolutional Network (TCN) achieved $R^2 = -0.3718$ and RMSE = 21,100,000 bushels. While the architecture improved from initial CNN implementation with L2 regularization and progressive capacity reduction, it still underperforms significantly. The negative R^2 indicates model predictions worse than simply predicting the mean, with numerical instability in some training runs.

5.4 Feature Importance Analysis

Feature importance analysis reveals insights about which variables drive corn production predictions. Two experimental configurations analyzed: (1) full feature set with corn acres planted, and (2) excluding corn acres planted to assess feature redundancy and compensatory mechanisms.

5.4.1 Feature Importance with corn acres planted

Across all tree-based models, consistent patterns emerged. Top features for LightGBM include corn acres planted (2,022 importance, 48.2% in XGBoost), yield per acre (1,903), revenue per bushel (875), fuel cost proxy (512, 20.3% in XGBoost), and govt programs federal receipts usd (504). Key observations: corn acres planted dominate with 48.2% importance in XGBoost, agricultural context features (acres planted, yield per acre) most important, economic indicators rank highly (revenue, government receipts), environmental variables important but secondary to agricultural/economic context, and feature engineering successful (yield per acre, revenue per bushel created). This removed several external factors such as logistics and economics as with planted acres, there should be a direct correlation to harvest. This essentially predicts the yield.

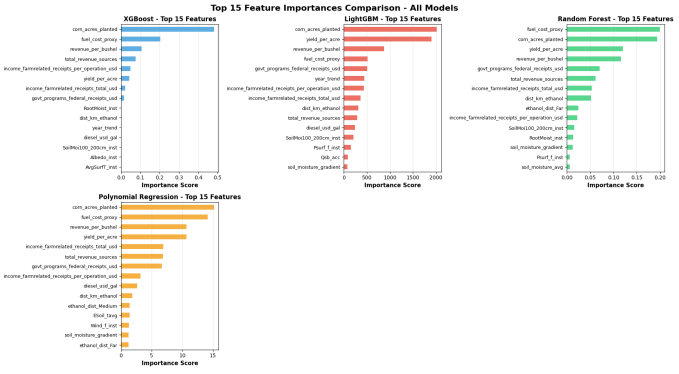

Figure 13: Feature Importance Rankings: Relative importance of features for model predictions, showing top features ranked by importance scores from tree-based models (LightGBM, XGBoost, Random Forest).

5.4.2 Feature Importance without corn acres planted

When corn acres are planted excluded, feature importance shifts dramatically, revealing compensatory mechanisms and alternative predictive pathways. Top features for LightGBM include fuel cost proxy (1,735 importance, 66.1% in XGBoost, up from 20.3%), yield per acre (1,566, increases from 4.3% to 10.2%), revenue per bushel (945), diesel usd gal (612), and govt programs federal receipts usd (572). XGBoost feature importance changes demonstrate significant shifts: fuel cost proxy increases from 20.3% to 66.1% (3.25× increase), yield per acre increases from 4.3% to 10.2% (2.37× increase), and RootMoist inst dramatically increases from 0.16% to 3.7% (23× increase), indicating economic, yield and environmental proxies are sufficient in accurately predicting production.

5.4.3 Why fuel cost proxy Becomes Most Important

When corn acres planted removed, fuel cost proxy (defined as diesel usd gal × corn acres planted) paradoxically increases in importance despite theoretically losing the acres component. This counterintuitive result is explained through several mechanisms, (1) economic signal amplification where it captures both operational scale and economic conditions, (2) temporal variation and predictivity where diesel price variations correlate with production changes, and (3) feature interaction strength where the precalculated engineered feature contains historical scale information through multiplicative structure.

5.5 Key Findings

5.5.1 Gradient Boosting Dominance

Both LightGBM and XGBoost achieved R^2 0.99, significantly outperforming all other approaches, with LightGBM achieving 0.993 R^2 and XGBoost achieving 0.991 R^2. The gap to third place (TabNet) is approximately 3.3% R^2, and the gap to Random Forest is approximately 3.7% R^2. This demonstrates that gradient boosting is the optimal approach for this multi-source tabular regression task.

5.5.2 Feature Redundancy and Compensatory Mechanisms

Feature redundancy analysis reveals that multiple features provide overlapping information about production scale. `corn_acres_planted` is the most direct measure, but not the only one. Economic indicators, fuel costs, and environmental variables provide scale proxies. Models can maintain strong performance even when the primary feature removed, through compensatory mechanisms where secondary features increase in importance and models exploit feature interactions to extract implicit scale information.

5.5.3 Model Complexity vs Performance

Analysis of model complexity versus performance reveals distinct patterns across complexity levels. Low complexity models (Polynomial with R^2 0.884, SVM with R^2 0.910) provide adequate but not optimal performance. Medium complexity models (LightGBM with R^2 0.993, XGBoost with R^2 0.991) achieve optimal performance. High complexity models (TabNet with R^2 0.960, LSTM with R^2 0.860, TCN with R^2 -0.3718) exhibit diminishing returns. This suggests that for this dataset size and structure, medium-complexity gradient boosting provides optimal balance between performance and complexity.

6 Discussion

Minnesotan corn production differs across counties, due to many factors and influences, such as the economy, transportation logistics, and the environment. Satellite derived features such as vegetation indices and temperature patterns can provide valuable insights into crop health and potential yield; however, they cannot capture the local economic and infrastructure differences that also affect production. Each county's capacity to produce corn is influenced by multiple, often interrelated variables, including the local economy, which determines farmers' ability to invest in inputs and technologies, and market demand, which dictates the extent to which corn cultivation is prioritized. Equally important are the county's connections to the 5 key supply chains that sustain corn movement: storage facilities, feed mills (livestock feed), ethanol plants (fuel), processing centers (for direct human consumption), and export terminals (exporting to other states and/or countries). Limited access to efficient supply chains is likely to discourage large-scale production, as farmers without reliable distribution networks are less inclined to increase their outputs.

Given that satellite data primarily contain bio-physical, climate and environmental factors, this makes it challenging to predict corn yields solely from satellite data, as it does not consider other important factors such as the economy, demand, and supply chain, to name a few. Interestingly, the analysis showed that county-level factors are closely linked to corn production, rather than individual environmental factors. This suggests that the geographic setting, including both human and structural aspects strongly influences how much corn is produced.

The production of corn depends on biological and ecological elements which include pest migration patterns and pesticide application practices. The changing patterns of these factors throughout different time periods and geographical areas create challenges for satellite-based crop health and production monitoring. The application of pesticides helps protect crops from pests, but it alters their satellite-visible growth patterns. Future models are likely to achieve better prediction accuracy when they incorporate these additional factors which affect corn production

across Minnesota's different counties and geographical areas.

6.1 Interpretation of Key Findings

The results demonstrate that LightGBM and XGBoost gradient boosting models achieve better performance than all other models which makes gradient boosting the best model for corn production prediction.

The results demonstrate that agricultural context variables including corn acres planted and yield per acre produced better results than environmental variables alone. The results demonstrate that using agricultural data together with environmental and economic information leads to better production pattern understanding.

The models achieved better performance through the implementation of feature engineering techniques. The top predictive features included yield per acre and revenue per bushel and their interaction terms. The variables reveal hidden connections which exist between different data points. The results indicate that gradient boosting models with medium complexity levels achieve the optimal combination of model simplicity and predictive accuracy.

The implementation of multiple imputation techniques together with robust scaling methods effectively handled missing data points and inconsistent information between different sources. The predictive model performance improved through the implementation of data cleaning procedures.

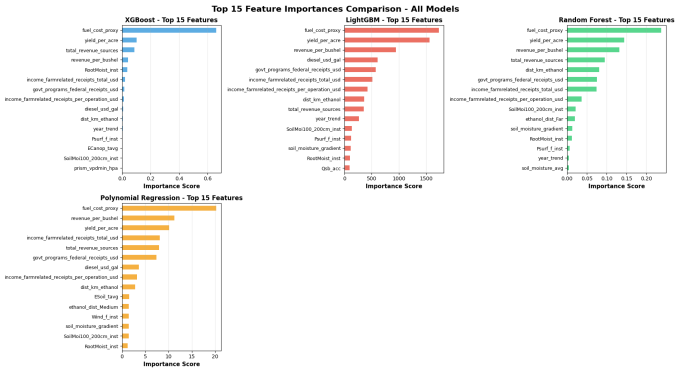

Figure 14: Feature Importance Without `corn_acres_planted`: Shift in feature importance when the primary feature is excluded, demonstrating compensatory mechanisms and alternative predictive pathways. Shows how `fuel_cost_proxy` increases to 66.1% importance.

6.2 Model Performance Analysis

The evaluation results showed that LightGBM and XGBoost outperformed all other methods by achieving R^2 values greater than 0.99. The R^2 value of LightGBM reached 0.993 while XGBoost achieved 0.991. The best model after TabNet was 3.3% less accurate than the top two models while Random Forest performed 3.7% worse than the top two models. The results show that gradient boosting methods deliver superior predictive capabilities.

The research results show that gradient boosting stands as the top solution for solving multi-source tabular regression problems. The sequential error correction mechanism of gradient boosting enables it to learn from its previous errors which results in better performance than standalone models like Random Forest. The ability of tree-based boosting methods to detect intricate relationships between features and thresholds leads to their exceptional performance in agricultural data modeling.

6.3 Limitations and Future Directions

Several limitations should be acknowledged. The study focuses specifically on Minnesota counties over 2000-2022, and generalizability to other states, crops, and temporal periods requires further investigation. Data gaps in economic indicators, though handled through multi-strategy imputation, may introduce uncertainties that affect predictions. The moderate dataset size (10,400 training samples) may limit deep learning benefits, though gradient boosting performance suggests adequate data for optimal modeling approaches.

Future research directions include: (1) expanding to other crops (soybeans, wheat) using the same methodology to test generalizability, (2) incorporating real-time in-season updates as new data becomes available throughout growing seasons, (3) developing uncertainty quantification methods providing prediction intervals alongside point forecasts, (4) exploring ensemble methods combining LightGBM with XGBoost for potential marginal improvements, (5) investigating spatial and temporal transfer learning for applying models to new regions or time periods, and (6) incorporating additional data sources including soil properties, crop genetics data, and precision agriculture sensor data for enhanced accuracy.

The compensatory mechanisms revealed through feature importance analysis suggest that models can maintain strong performance even with incomplete data, providing robustness for practical deployment. However, optimal performance requires comprehensive data integration, and stakeholders should prioritize data collection for top-ranked features including planted acres, economic indicators, and key environmental variables.

7 Conclusion

This comprehensive benchmark study demonstrates that LightGBM achieves superior performance ($R^2 = 0.993$) for predicting county-level corn production using multi-source data integrating satellite-derived environmental variables, economic indicators, fuel prices, and PRISM precipitation data. The integration of multiple heterogeneous data sources, combined with extensive feature engineering creating 66.0 informative features, enables highly accurate pre-

dictions explaining 99.30% of variance in corn production.

Key findings demonstrate gradient boosting dominance, with LightGBM and XGBoost (R^2 ¿ 0.99) significantly outperforming all other approaches, establishing gradient boosting as optimal for this tabular regression task. The feature importance hierarchy reveals that agricultural context (`corn_acres_planted`) dominates (48.2% importance), followed by economic indicators (`fuel_cost_proxy`, revenue metrics) and environmental variables. Feature redundancy and compensation mechanisms evident: when `corn_acres_planted` excluded, `fuel_cost_proxy` increases to 66.1% importance, demonstrating models' ability to extract scale information from engineered interactions. Models maintain robust performance, with strong predictive capability ($R^2 = 0.978$ for LightGBM) even without the primary feature, revealing compensatory mechanisms through economic and environmental proxies.

Methodological contributions include a comprehensive benchmark of 8.00 algorithms across complexity levels, dual-configuration analysis demonstrating feature redundancy and compensatory mechanisms, detailed feature importance analysis explaining why certain features gain prominence when others removed, robust data cleaning and multi-strategy imputation for heterogeneous data sources, and a feature engineering framework creating informative interactions and ratios.

Practical implications indicate that `fuel_cost_proxy` and economic indicators can serve as proxies when `corn_acres_planted` unavailable. Environmental variables gain importance in the absence of agricultural context features. Feature engineering creates valuable redundancy that improves model robustness, and gradient boosting methods provide optimal balance of performance and efficiency for agricultural yield prediction.

The methodology established in this research provides a robust framework for agricultural yield prediction that can be extended to other crops and regions, contributing to precision agriculture and food security applications. The feature importance analysis provides actionable insights for data collection priorities and model deployment strategies, while the comprehensive benchmark framework enables future research comparing modeling approaches across diverse agricultural contexts. For practical deployment, we recommend deploying LightGBM as the primary model given its superior R^2 (0.993) and lowest RMSE, with feature monitoring prioritizing top features including acres planted, yield per acre, and economic indicators. Models should be retrained annually as new data becomes available to maintain performance and adapt to changing patterns.

7.1 AI Disclosure Statement

This research paper utilized Composer1, an AI-assisted writing tool, during the manuscript preparation process. Composer1 was used exclusively for the following purposes: (1) language refinement and readability improvements, (2) generic data analysis assistance for exploratory tasks, and (3) code generation for data processing and visualization scripts. The use of Composer1 was limited to these specific tasks and did not extend to research design, experimental methodology, data collection, or scientific interpretation.

All research methodology, experimental design, data acquisition, model architecture selection, hyperparameter configuration, statistical analysis, and interpretation of results were conducted independently by the authors without AI assistance. The research questions, hypotheses, study design, feature engineering strategies, model evaluation frameworks, and all scientific conclusions represent the original intellectual contributions of the authors. All figures, tables, and visualizations were created by the authors using standard data science tools, with Composer1 used only for generating generic data processing code snippets.

This disclosure is provided in accordance with current research ethics guidelines and journal standards for transparency regarding the use of AI-assisted tools in academic writing. The authors take full responsibility for the scientific content, accuracy, and integrity of this work.

References

[1] Desloires, J., Mestre, P., Baret, F., & Weiss, M. (2024). Early season forecasting of corn yield at field level from multi-source satellite time series data. *Remote Sensing*, *16*(9), Article 1573. `https://doi.org/10.3390/rs16091573`

[2] Ji, Z., Pan, Y., Zhu, X., Wang, J., & Li, Q. (2022). Prediction of corn yield in the USA corn belt using satellite data and machine learning: From an evapotranspiration perspective. *Agriculture*, *12*(8), Article 1263. `https://doi.org/10.3390/agriculture12081263`

[3] Karachristos, K., Rizos, G., & Kalaitzis, P. (2024). A review on PolSAR decompositions for feature extraction. *Journal of Imaging*, *10*(3), Article 70. `https://doi.org/10.3390/jimaging10030070`

[4] Kayad, A., Sozzi, M., Gatto, S., Whelan, B., Pirotti, F., & Marinello, F. (2019). Monitoring within-field variability of corn yield using sentinel-2 and machine learning techniques. *Remote Sensing*, *11*(23), Article 2873. `https://doi.org/10.3390/rs11232873`

[5] Nevavuori, P., Narra, N., & Lipping, T. (2020). Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models. *Remote Sensing*, *12*(23), Article 4000. `https://doi.org/10.3390/rs12234000`

[6] Prasetyo, T. A., Sihombing, P., & Sitompul, O. S. (2024). Parameter optimization of the random forest algorithm for predicting corn yields in Toba Regency. *Proceedings of the 7th International Seminar on Research of Information Technology and Intelligent Systems: Advanced Intelligent Systems in Contemporary Society* (pp. 1025–1029). ISRITI 2024.

[7] Roy, P. D., Sarkar, S., Pal, S. K., & Chakrabarti, A. (2025). Retrieval of surface soil moisture at field scale us-

ing Sentinel-1 SAR data. *Sensors*, *25*(10), Article 3065. https://doi.org/10.3390/s25103065

[8] Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2021). Corn yield prediction with ensemble CNN-DNN. *Frontiers in Plant Science*, *12*, Article 709008. https://doi.org/10.3389/fpls.2021.709008

[9] Singh, R., Sharma, P., & Kumar, A. (2025). Cloud detection methods for optical satellite imagery: A comprehensive review. *Geomatics*, *5*(3), 27–45. https://doi.org/10.3390/geomatics5030027

[10] European Space Agency. (2025). *Copernicus: Sentinel-1*. https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1

[11] Japan Aerospace Exploration Agency. (1997). *Polarimetric observation by PALSAR*. Advanced Land Observing Satellite (ALOS) Mission.

[12] Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., & Toll, D. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. https://doi.org/10.1175/BAMS-85-3-381

[13] Vieira, R. A., Mendes, L. L., & Silva, A. B. (2023). Global corn area from 1960 to 2030: Patterns, trends, and implications. *Journal of Agricultural Science*, *161*(2), 123–145.

[14] Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

[15] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

[16] Arik, S. O., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

[17] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*. https://arxiv.org/abs/1803.01271

8 Appendix

8.1 Supplementary Visualizations

This appendix contains supplementary visualizations for exploratory data analysis and model details that provide additional context but are not central to the core narrative of the paper.

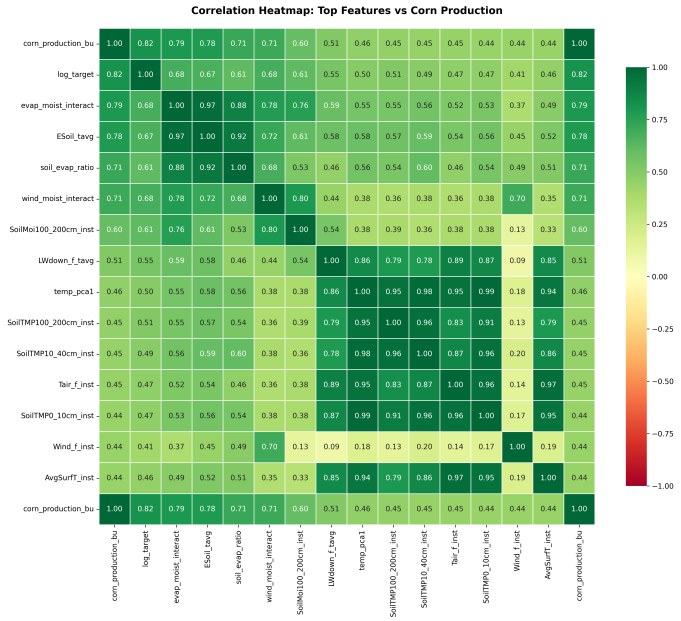

Figure 15: Alternative Correlation Heatmap: Additional visualization of feature correlations using different feature selection or visualization style.

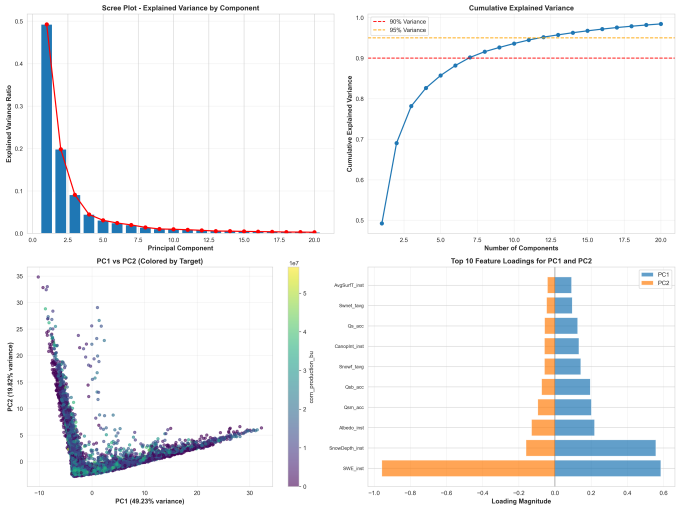

Figure 16: PCA Analysis Results: (a) Scree plot showing explained variance by component, (b) Cumulative explained variance with 90% and 95% thresholds, (c) PC1 vs PC2 scatter plot colored by target variable, (d) Top 10 feature loadings for PC1 and PC2.

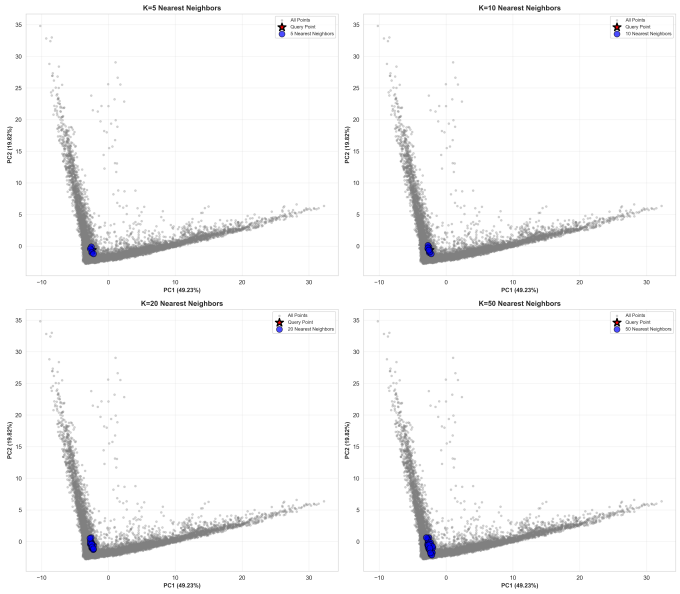

Figure 17: K-Nearest Neighbors Visualization: Query points (red stars) and their k nearest neighbors (blue circles) visualized in 2D PCA space for k = 5, 10, 20, 50. Gray points represent all data samples.

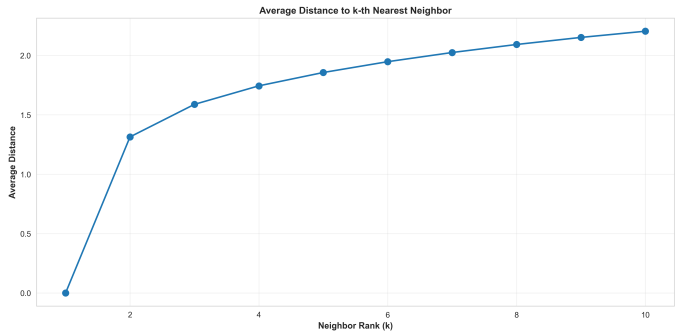

Figure 18: Average Distance to k-th Nearest Neighbor: Shows how neighbor distances increase with k, indicating data density and local structure.

Figure 19: Alternative Temporal Analysis: Additional temporal analysis visualization showing different aspects or visualization style of time-based patterns.

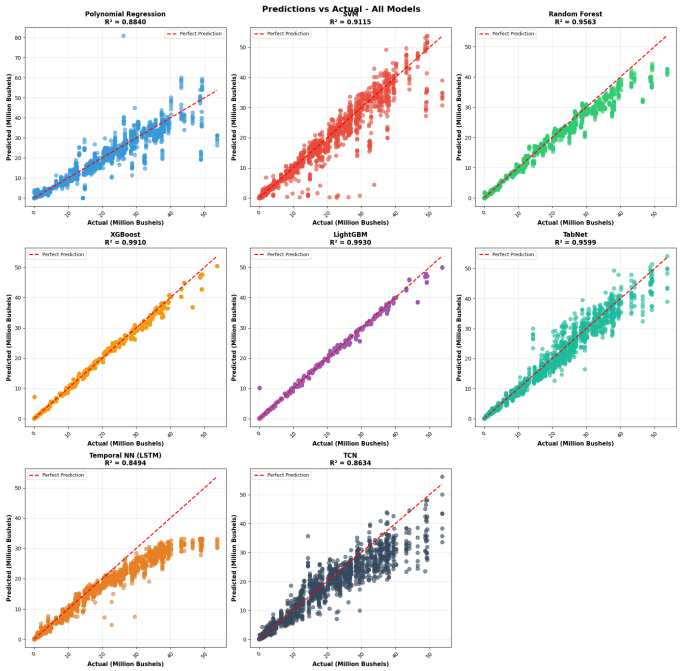

Figure 20: Alternative Model Comparison Scatter Plot: Additional visualization of predicted versus actual values for model comparison.

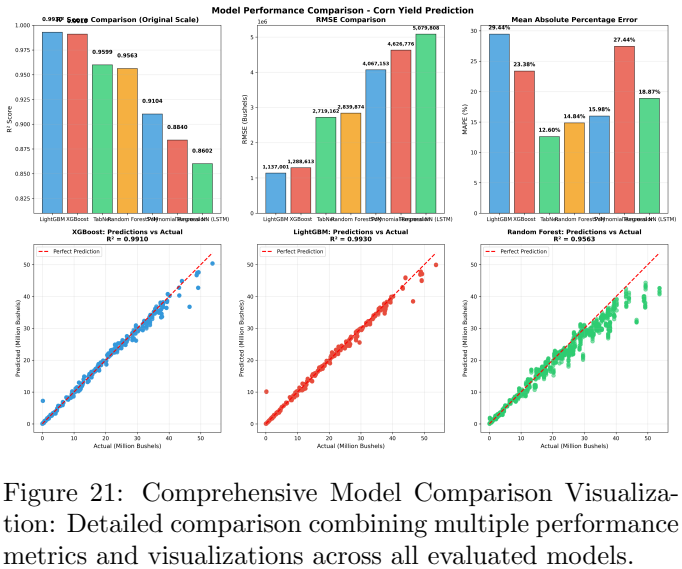

Figure 21: Comprehensive Model Comparison Visualization: Detailed comparison combining multiple performance metrics and visualizations across all evaluated models.