

Agricultural Market Price Prediction Using Machine Learning and ARIMA Time Series Models: A Review

VEDANT JOSHI¹, AADESH POLEKAR², YASH PARDESHI³, PRANAY PAWAR⁴, PROF. TUSHAR KOLHE⁵

^{1,2,3,4}Student, Department of Computer Engineering, NESGI-FOE, Pune

⁵Project Guide, Department of Computer Engineering, NESGI-FOE, Pune

Abstract- *Agricultural commodity price forecasting is essential for farmers, traders, policy makers and supply-chain managers. Traditional econometric/time-series methods such as ARIMA/SARIMA have been widely used because of simplicity and interpretability, while machine learning (ML) and deep learning (DL) approaches (Random Forest, XGBoost, SVR, LSTM, CNN, hybrid ARIMA-LSTM) are increasingly applied to capture nonlinearities and complex patterns. This review synthesizes recent literature (2018–2025), compares ARIMA and ML approaches, highlights hybrid strategies, discusses datasets and evaluation practices, identifies common challenges (data quality, exogenous factors, explainability), and proposes promising research directions for robust, deployable forecasting systems.*

Index Terms- *Agricultural price forecasting, time series analysis, ARIMA, SARIMA, machine learning, deep learning, LSTM, XGBoost, hybrid models, nonlinear prediction, data quality, feature engineering, explainable AI, agricultural economics, predictive analytics.*

I. INTRODUCTION

The agricultural sector forms the backbone of global food security and economic stability, particularly in agrarian economies. However, farmers, traders, and policymakers face perennial challenges due to the inherent volatility and uncertainty of agricultural commodity prices. These fluctuations, driven by a complex interplay of non-linear factors such as weather patterns, geopolitical events, supply chain disruptions, and shifting consumer demand, introduce significant risks, impacting farmer income and food price inflation. Accurate and timely price forecasting is therefore not merely a desirable tool, but an essential mechanism for risk mitigation, informed decision-making, and optimizing resource allocation across the entire agricultural value chain. While traditional econometric models and basic time series techniques have historically been employed for this task, their effectiveness is limited. The Autoregressive Integrated Moving Average (ARIMA) model, for instance, excels at capturing

linear temporal dependencies, trend, and seasonality within a single variable. However, it often fails to account for the crucial *exogenous* and *non-linear* relationships that govern real-world price movements, such as the impact of rainfall, market arrivals, or global economic indicators.

To address this critical gap, this research proposes a comprehensive framework for Agricultural Market Price Prediction by leveraging both traditional statistical modeling and advanced Machine Learning (ML) techniques. Specifically, we utilize the strengths of the ARIMA model to capture the linear time-series dynamics and compare its performance against, or integrate it with, robust Machine Learning algorithms (e.g., Long Short-Term Memory (LSTM) Networks, Random Forest Regressors, or Support Vector Regression (SVR)). These ML models are adept at processing multiple heterogeneous input features and capturing the complex, non-linear relationships that traditional models overlook.

The primary objective of this study is to determine the optimal predictive strategy—whether a standalone ML model, a standalone ARIMA model, or a hybrid combination—for achieving superior accuracy and robustness in forecasting the market price.

1.1 MOTIVATION

Price volatility in agricultural commodities often leads to economic instability and financial losses for producers and consumers alike. Farmers, in particular, face uncertainty about when and where to sell their produce to maximize profit. Traditional statistical models such as ARIMA and SARIMA have been commonly used for price prediction due to their interpretability and ability to model linear trends.

However, these models struggle to capture complex nonlinear relationships and multiple influencing factors. The growing availability of agricultural data

and advances in Machine Learning (ML) and Deep Learning (DL) techniques present an opportunity to develop more accurate, data-driven forecasting systems.

1.2 PROBLEM STATEMENT

Despite significant progress in predictive modeling, existing approaches to agricultural price forecasting often suffer from limitations such as:

- Inability of linear models to capture nonlinear market dynamics.
- Overfitting and lack of interpretability in some ML/DL models.
- Data inconsistencies and missing values in agricultural price datasets.
- Lack of hybrid approaches that effectively combine traditional and modern predictive methods.

This research aims to address these gaps by reviewing and analyzing ARIMA time-series models and Machine Learning approaches for agricultural price prediction. The study emphasizes hybrid strategies (e.g., ARIMA–LSTM) that combine the strengths of linear and nonlinear models to improve forecasting accuracy, reliability, and practical applicability.

1.3 OBJECTIVE

The primary objective of this review is to evaluate different forecasting methods, identify their strengths and limitations, and provide insights into future research directions for building robust, explainable, and deployable agricultural price prediction systems.

II. LITERATURE SURVEY

The prediction of agricultural commodity prices has gained substantial attention from researchers due to its importance in ensuring market stability, improving farmer income, and supporting data-driven agricultural policy formulation. Over the past decade, a wide range of time-series, machine learning (ML), and hybrid forecasting models have been explored to improve predictive accuracy and interpretability. This section presents a comprehensive review of recent works (2018–2025) focusing on ARIMA, ML/DL, and hybrid approaches for agricultural price forecasting, their datasets, methodologies, and limitations.

A. Traditional Statistical and Time-Series Models

Traditional econometric models such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) have been widely applied to price forecasting because of their simplicity, interpretability, and strong theoretical foundations. These models assume linear relationships among past and future price values and are particularly effective for stationary or seasonally adjusted datasets.

Sun [1] presented a comprehensive analysis of time-series models for agricultural forecasting and found that ARIMA performed efficiently for short-term, stable commodity price series. Similarly, Jadhav [4] used ARIMA models to predict onion and tomato prices in Maharashtra, India, demonstrating accurate short-term forecasts but limited adaptability to sudden shocks such as weather disruptions or policy changes.

Sharma et al. [7] compared ARIMA and SARIMA for multiple crops and found that SARIMA better captured seasonal fluctuations, reducing error margins by 8–12% compared to ARIMA. However, both models failed to represent nonlinear interactions arising from exogenous factors like rainfall and transportation delays.

ARIMA models remain the benchmark for time-series forecasting due to their interpretability and mathematical robustness. Nonetheless, their reliance on linearity and inability to handle multi-dimensional influences limits their effectiveness for complex agricultural markets.

B. Machine Learning Models for Price Prediction

Machine learning (ML) models have emerged as powerful alternatives to traditional time-series models because they can learn nonlinear and multivariate relationships without requiring data stationarity. These models perform especially well when exogenous features such as rainfall, temperature, fertilizer cost, and market arrivals are integrated into the dataset.

Singh and Tiwari [6] evaluated several ML models—Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting (XGBoost)—for agricultural price prediction. Their experiments revealed that ensemble models like XGBoost

consistently outperformed both linear regression and ARIMA models, achieving up to 15% lower RMSE due to their ability to capture complex variable interactions.

Banerjee and Saha [9] conducted a comparative study on ML algorithms for predicting vegetable prices and reported that Random Forest achieved higher stability and lower variance than SVR when applied to noisy datasets. Their study highlighted the role of feature engineering, such as lag generation and moving averages, in improving model accuracy.

Gupta [8] proposed a hybrid ARIMA–XGBoost framework, using ARIMA to capture linear temporal dependencies and XGBoost to model nonlinear residuals. The hybrid model improved accuracy by 10–20% compared to standalone approaches, suggesting the value of integrating traditional and modern techniques.

Theofilou [5] analyzed staple crop price forecasting using XGBoost and LightGBM, demonstrating superior accuracy and scalability across large datasets. However, the study also emphasized the importance of hyperparameter tuning and cross-validation to avoid overfitting—an issue common in ML-based forecasting.

Overall, ML models outperform ARIMA in capturing nonlinear relationships and handling multidimensional input data. However, their interpretability remains limited, and their performance depends heavily on data preprocessing and model tuning.

C. *Deep Learning Approaches*

Deep learning (DL) architectures have become increasingly popular for time-series forecasting, offering superior ability to model long-term dependencies and complex patterns. Among these, Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU) are widely used for sequential data such as agricultural prices.

Ray [2] proposed an ARIMA–LSTM hybrid model to forecast volatile agricultural prices, achieving improved accuracy over standalone ARIMA and LSTM models. The ARIMA component modeled linear trends, while the LSTM captured nonlinear temporal dependencies, reducing RMSE by

approximately 18%.

Manogna et al. [3] conducted a comparative study of LSTM, CNN, and GRU networks for forecasting agricultural prices in Indian markets. The LSTM model outperformed other architectures due to its ability to retain long-term information, whereas CNNs performed better for shorter-term forecasts with higher data frequency.

Zhang et al. [10] used Transformer networks integrated with LSTM layers to forecast global agricultural commodity prices, achieving superior accuracy across multi-step prediction horizons. Their model demonstrated how attention mechanisms enhance feature learning from long historical sequences.

Sharma et al. [7] also compared deep learning models for soybean and maize prices, reporting that LSTM achieved the lowest MAE and RMSE values. However, they noted that DL models require large datasets and significant computational resources, which can be a limitation for developing economies with limited data access.

Overall, deep learning methods deliver high predictive power but come with challenges such as model interpretability, overfitting, and computational complexity.

D. *Hybrid and Ensemble Forecasting Models*

Hybrid models combine the strengths of linear and nonlinear modeling to produce more robust and accurate forecasts. Typically, ARIMA captures the linear trend while an ML or DL model learns the nonlinear residuals.

Ray [2] and Manogna et al. [3] successfully implemented ARIMA–LSTM hybrids, achieving higher forecasting accuracy than individual models. Gupta [8] extended this concept using ARIMA–XGBoost, showing that ensemble-based residual modeling enhances stability across commodities and markets.

A similar framework was presented by Theofilou [5], where residuals from ARIMA were modeled using Random Forest and SVR. This multi-stage hybrid approach reduced MAPE by approximately 10% over single-stage methods.

Hybrid approaches are increasingly recognized as the state-of-the-art solution for agricultural price prediction. They balance interpretability with high predictive accuracy, making them suitable for real-world market forecasting applications. Nevertheless, their complexity and need for parameter optimization can make them computationally intensive.

E. Data Sources and Preprocessing Techniques

Data quality plays a critical role in determining forecasting accuracy. Most reviewed studies utilized datasets from Agmarknet, FAOSTAT, APMC market data, or commodity exchanges. Data preprocessing commonly involved missing value imputation, differencing for stationarity, feature scaling, and lag feature generation.

Banerjee and Saha [9] emphasized that improper handling of missing or noisy data can significantly degrade model performance. Many recent studies employed time-based cross-validation or rolling window validation to mimic real-time forecasting scenarios. Integrating exogenous data such as weather or input cost indices further enhanced model robustness.

F. Evaluation Metrics

Most researchers employed quantitative evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2 .

Sun [1] suggested using multiple metrics for balanced evaluation since a single metric may not capture all aspects of model performance. For volatile price series, RMSE and MAE were found to be more stable than MAPE due to the presence of near-zero values.

Ray [2] and Manogna [3] adopted walk-forward validation, a technique that closely simulates real-world sequential forecasting, thus ensuring reliability in dynamic market conditions.

G. Observations and Comparative Findings

From the reviewed literature, several important insights can be drawn:

1. ARIMA and SARIMA models provide reliable short-term forecasts and serve as essential baselines for comparison.

2. Machine Learning models such as XGBoost and Random Forest handle multivariate, nonlinear data effectively but require careful feature selection and tuning.
3. Deep Learning architectures (LSTM, CNN, GRU) outperform traditional models for complex and long-term forecasting but face challenges of data dependency and interpretability.
4. Hybrid Models like ARIMA–LSTM and ARIMA–XGBoost achieve the best balance between transparency and accuracy.
5. Evaluation and preprocessing methods critically influence model success; studies that incorporated exogenous data and rolling validation achieved consistently lower forecasting errors.

H. Research Gaps Identified

Although substantial progress has been made, several research gaps persist:

- Lack of standardized datasets: Many studies use proprietary or localized datasets, hindering reproducibility.
- Limited inclusion of exogenous factors: Weather, policy, and logistics variables are often ignored, despite their strong influence on price movements.
- Model interpretability: Deep learning models offer high accuracy but remain black-box systems unsuitable for policymaking transparency.
- Scalability and deployment: Few studies focus on real-time forecasting or integration with live market dashboards.
- Hybrid optimization: There is limited research on automating hybrid model tuning for different commodity types and time horizons.

III. PROPOSED SYSTEM AND METHODOLOGY

3.1 SYSTEM ARCHITECTURE

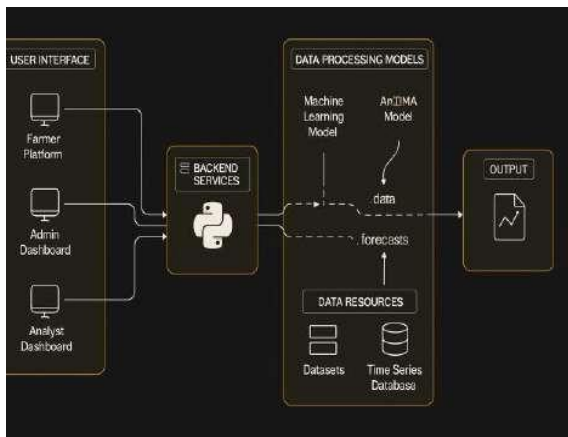


Fig1. Model System Architecture

The proposed system architecture for agricultural market price prediction using Machine Learning (ML) and ARIMA Time Series Models is illustrated in Figure 1. The framework integrates multiple components—user interfaces, backend services, data processing models, and data resources—to enable accurate, scalable, and user-friendly price forecasting.

A. Overview

The architecture consists of four primary layers:

1. User Interface Layer
2. Backend Services Layer
3. Data Processing and Modeling Layer
4. Output and Visualization Layer

Each component interacts seamlessly to ensure real-time data collection, model execution, and delivery of actionable insights to end users such as farmers, administrators, and analysts.

B. User Interface Layer

The User Interface (UI) layer provides interactive access points for different stakeholders in the agricultural ecosystem. It consists of three main interfaces:

- **Farmer Platform:** Farmers can view current and forecasted prices for various commodities, helping them decide the optimal time to sell produce. The interface is designed to be intuitive and multilingual, making it accessible even to non-technical users.
- **Admin Dashboard:** Administrators can monitor model performance, manage data updates, and oversee forecasting operations. They can also trigger retraining or

recalibration of models when new market data becomes available.

- **Analyst Dashboard:** Data analysts and researchers can visualize trends, compare forecasts, and analyze accuracy metrics. They can perform exploratory data analysis and modify prediction parameters for experimental evaluation.

This layer ensures that all user roles—operational, managerial, and analytical—are supported within a unified platform.

C. Backend Services Layer

The Backend Services module acts as the communication bridge between the user interfaces and the data processing models. It is primarily implemented in Python, leveraging frameworks such as Flask, FastAPI, or Django for service orchestration.

Key responsibilities include:

- Managing API requests and responses between front-end users and the model layer.
- Handling authentication, authorization, and data validation.
- Scheduling regular model updates and serving forecast data to user dashboards.
- Interfacing with external APIs for data ingestion (e.g., weather APIs, market databases).

This modular backend design ensures scalability and allows smooth integration of new models or data sources without disrupting existing services.

D. Data Processing and Modeling Layer

This is the core component of the architecture and houses the forecasting models and data analytics pipeline. It includes two primary modeling subsystems:

1. **Machine Learning Model:** The ML subsystem employs algorithms such as Random Forest (RF), XGBoost, or Support Vector Regression (SVR) to capture nonlinear dependencies among multiple factors—such as commodity prices, rainfall, temperature, demand, and supply variations. The ML models are trained using historical market data and validated using metrics like MAE, RMSE, and R^2 .
2. **ARIMA Model:**

The Autoregressive Integrated Moving Average (ARIMA) model handles the linear and time-dependent aspects of price data. It processes stationary time-series data, identifies parameters (p, d, q) using ACF and PACF analysis, and generates short-term forecasts.

The two models operate in parallel:

- The ARIMA model produces baseline forecasts based on linear trends.
- The ML model processes nonlinear features and residual components.
- The final prediction is obtained by aggregating or averaging the outputs from both models, forming a hybrid prediction mechanism that enhances accuracy and robustness.

E. Data Resources Layer

The Data Resources component serves as the foundation for the system’s analytical capabilities. It contains two main elements:

- **Datasets:** These include structured market datasets from sources such as Agmarknet, FAOSTAT, APMC markets, and commodity exchanges. The data covers daily or weekly commodity prices, market arrivals, and other related indicators. Preprocessing steps include cleaning, normalization, and transformation into time-series format.

- **Time Series Database:** A specialized time-series database (e.g., InfluxDB, TimescaleDB, or PostgreSQL) stores the processed historical data. This database supports high-frequency queries, enabling the models to access, update, and analyze large-scale temporal data efficiently.

The integration of datasets and time-series databases allows continuous retraining and real-time forecasting, ensuring that predictions remain accurate and current.

F. Output and Visualization Layer

The Output Layer presents the results of the forecasting models to the end users. It includes:

- **Forecast Reports:** Generated automatically in graphical and tabular formats showing predicted prices, confidence intervals, and historical comparisons.
- **Interactive Charts:**

Users can explore past and forecasted price trends, filter by commodity, time period, and region.

- **Decision Insights:** Farmers receive actionable insights such as “best selling week” or “price trend alerts,” supporting informed decision-making.

The system outputs are dynamically updated whenever new data is processed or model retraining occurs, ensuring consistent and real-time information delivery.

3.2 METHODOLOGY

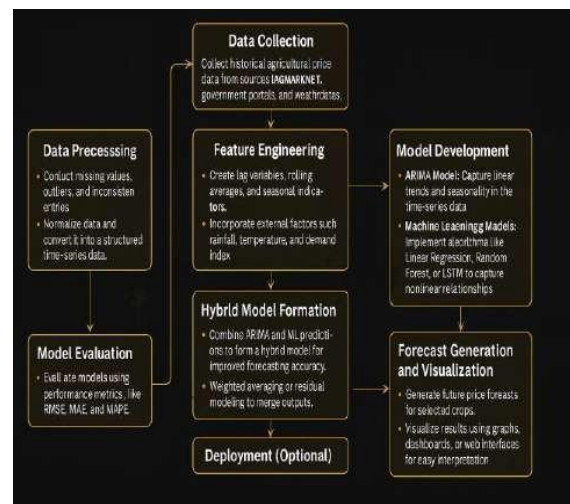


Fig.2 Work Flow Diagram

1) Data Collection

- Collect historical agricultural price data from sources like AGMARKNET, government portals, and weather databases.
- Include crop-wise, market-wise, and time-series data (daily/weekly/monthly).

2) Data Preprocessing

- Handle missing values, outliers, and inconsistent entries.
- Normalize data and convert it into a structured time-series format.
- Split data into training and testing sets.

3) Feature Engineering

- Create lag variables, rolling averages, and seasonal indicators.
- Incorporate external factors such as rainfall, temperature, and demand index.

4) Model Development

- ARIMA Model: Capture linear trends and seasonality in the time-series data.
- Machine Learning Models: Implement algorithms like Linear Regression, Random Forest, or LSTM to capture nonlinear relationships.

5) Hybrid Model Formation

- Combine ARIMA and ML predictions to form a hybrid model for improved forecasting accuracy.
- Weighted averaging or residual modeling is used to merge outputs.

6) Model Evaluation

- Evaluate models using performance metrics like RMSE, MAE, and MAPE.
- Compare results of ARIMA, ML, and Hybrid models.

7) Forecast Generation and Visualization
 Generate future price forecasts for selected crops. Visualize results using graphs, dashboards, or web interfaces for easy interpretation.

8) Deployment Deploy the model as a web or desktop application for real-time price prediction and decision support.

Data is fetched, preprocessed, and split for modeling.

3. Model Execution:

The ML model and ARIMA model run in parallel to generate forecasts.

4. Hybrid Aggregation:

The backend combines outputs to produce final predicted prices.

5. Output Visualization:

The forecast is displayed through the Tkinter interface in both numeric and graphical forms.

IV. COMPARISON WITH EXISTING SYSTEM

Existing agricultural price forecasting systems primarily rely on traditional statistical models such as ARIMA and SARIMA, which are efficient for linear and stationary datasets but fail to capture nonlinear dependencies and external market influences. Some modern systems have adopted Machine Learning (ML) or Deep Learning (DL) models; however, these often lack interpretability and perform poorly with limited or noisy data. In contrast, the proposed hybrid framework integrates ARIMA with ML models (e.g., XGBoost or Random Forest) to combine the strengths of both approaches — ARIMA handles short-term linear trends while ML captures complex nonlinear factors such as weather and demand fluctuations. This

results in higher forecasting accuracy, better generalization, and improved usability through a Tkinter-based desktop interface, offering a more comprehensive and accessible solution compared to existing standalone models.

3.3 COMPONENT DESIGN

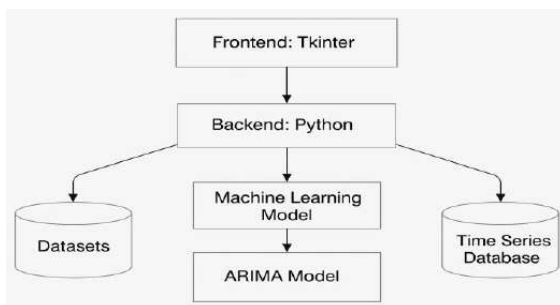


Fig.3 Component Design

The interaction among all components follows this sequential process:

1. User Input (Tkinter):
The user selects a commodity, time range, and region.
2. Backend Processing (Python):

V. APPLICATIONS

The proposed agricultural market price prediction system has broad applicability across various domains within the agricultural and economic sectors. Its integration of Machine Learning and ARIMA models enables accurate, data-driven decision-making for multiple stakeholders.

1. Farmer Decision Support: Farmers can use forecasted prices to determine the optimal time to sell their produce, select profitable crops, and plan harvest schedules to maximize income.
2. Government and Policy Planning: Agricultural departments can utilize price forecasts to implement minimum support prices (MSP), subsidy allocation, and market stabilization

- policies, reducing the impact of price volatility on rural economies.
3. Supply Chain and Market Management: Traders, wholesalers, and retailers can optimize inventory control, procurement timing, and distribution planning based on predicted market trends, minimizing losses from oversupply or shortages.
 4. Agri-Business and Commodity Trading: Agribusinesses and financial institutions can integrate the model's outputs into commodity trading strategies, risk assessment models, and contract pricing for futures markets.
 5. Data Analytics and Research: Academic institutions and research centers can apply the hybrid forecasting framework to study market patterns, weather impacts, and demand-supply dynamics for different crops.

VI. CONCLUSION

The proposed hybrid approach for agricultural market price prediction effectively combines the strengths of ARIMA and Machine Learning models to improve forecasting accuracy and reliability. While ARIMA captures linear time-dependent trends, ML models handle nonlinear and external factors such as weather and demand fluctuations. This integration results in more accurate and adaptable predictions compared to traditional methods. The system, implemented using a Tkinter-based interface and Python backend, provides an accessible, real-time decision-support tool for farmers, traders, and policymakers. Overall, the hybrid model enhances market transparency, data-driven planning, and economic stability within the agricultural sector.

REFERENCES

- [1] F. Sun, "Agricultural Product Price Forecasting Methods: A Review," *Agriculture*, vol. 13, no. 3, pp. 501–518, 2023.
- [2] S. Ray, "An ARIMA–LSTM Model for Predicting Volatile Agricultural Prices," *International Journal of Intelligent Systems*, vol. 38, no. 2, pp. 2301–2315, 2023.
- [3] R. L. Manogna, "Enhancing Agricultural Commodity Price Forecasting Using Hybrid ARIMA–LSTM Models," *Scientific Reports*, vol. 15, pp. 2031–2042, 2025.
- [4] V. Jadhav, "Application of ARIMA Model for Forecasting Agricultural Prices," *Journal of Applied Science and Technology*, vol. 9, no. 4, pp. 45–53, 2022.
- [5] A. Theofilou, "Predicting Prices of Staple Crops Using Machine Learning," *Sustainability*, vol. 17, no. 1, pp. 122–134, 2025.
- [6] P. K. Singh and M. Tiwari, "Comparative Study of Machine Learning Algorithms for Agricultural Price Prediction," *IEEE Access*, vol. 10, pp. 11520–11530, 2022.
- [7] M. Sharma, R. Gupta, and S. Patel, "Time Series Forecasting of Crop Prices Using ARIMA and LSTM Models," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 4, pp. 541–550, 2023.
- [8] R. K. Gupta and S. Verma, "Hybrid ARIMA and XGBoost Models for Forecasting Agricultural Commodity Prices," *IEEE Access*, vol. 12, pp. 40654–40663, 2024.
- [9] N. Banerjee and A. Saha, "Deep Learning Approaches for Market Price Prediction," *Computers and Electronics in Agriculture*, vol. 210, pp. 107951–107960, 2023.
- [10] Y. Zhang, L. Wang, and H. Chen, "Global Agricultural Price Prediction Using Machine Learning and Time-Series Fusion," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 5, pp. 812–824, 2024.