Received: 9 September 2025 | Revised: 12 January 2026 | Accepted: 11 March 2026 | Published online: 13 April 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–27
DOI: 10.47852/bonviewJCCE62027619

BON VIEW PUBLISHING

Multi-Source Data Fusion and Machine Learning for Soybean Crop Price Forecasting in India

Vilas Damodhar Ghonge[1,*] and Yogesh Kulkarni[1]

[1] *Department of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, India*

Abstract: Soybean is a major oilseed crop in India, and its market prices have exhibited significant volatility in recent years. Such price fluctuations create serious challenges for small and medium-scale farmers. Accurate price forecasting is essential to support informed decision-making by farmers and agri-business stakeholders. This study focuses on forecasting soybean prices in the Indian market using data spanning January 2015 to June 2025. The dataset integrates multiple heterogeneous sources, including daily market data from Agmarknet, weather information from the India Meteorological Department, regional production, and trade statistics. Extensive exploratory data analysis is conducted to examine price distributions, temporal trends, regional variations, and inter-variable relationships. Several predictive approaches are evaluated, including traditional time-series models (Autoregressive Integrated Moving Average), ensemble machine learning models, and deep learning models. In addition, a hybrid Long Short-Term Memory–Gated Recurrent Unit framework (AgroNET) is proposed to effectively model complex temporal dependencies across heterogeneous data sources. Model performance is assessed using k-fold cross-validation and evaluated through root mean squared error, mean squared error, MAPE, and R^2 metrics in a Python-based implementation. The results demonstrate that deep learning-based models outperform conventional approaches, with AgroNET achieving the highest R^2 and lowest error values. To enhance model transparency, explainable artificial intelligence using the local interpretable model-agnostic explanation technique is incorporated to identify key factors influencing individual price predictions. Overall, the proposed framework offers an effective and interpretable solution for soybean price forecasting in India and supports future research on multi-source integration and real-time agricultural price prediction.

Keywords: soybean price prediction, time-series forecasting, agricultural market, Indian agriculture, machine learning

1. Introduction

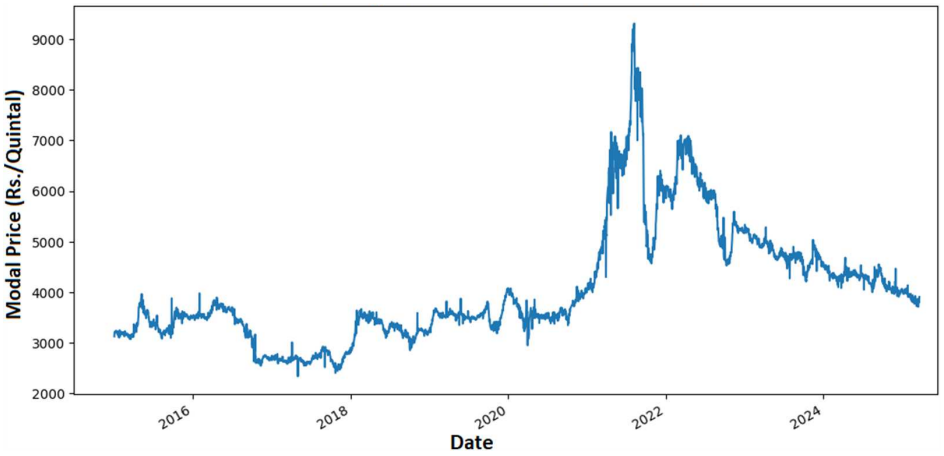
Soybean (*Glycine max*) is among the significant oilseed crops in India, which is mainly cultivated in the Kharif (monsoon) season. India is the fifth-largest soybean producer in the world, and it is estimated that the country produced 130.5 lakh tons of soybean in 2023–2024, which is approximately one-third of the total production of all oilseeds in the country [1]. The key states in soybean production are Maharashtra, Madhya Pradesh, and Rajasthan, which give the largest production. Soybeans are a very important source of edible oil and protein meal; they are consumed in the form of cooking oil and animal feeds. Considering this importance, the soybean market price has a direct influence on the income of farmers, the price of food consumed by consumers, the supply of an essential industrial raw material, and the balance of the trade [2].

The prices of soybeans in India have been fluctuating significantly in recent years. Until much of the 2010s, the price of mandi (wholesale market) soybeans has been increasing at a comparatively linear pace, but since the end of 2020, it has started to go up at a steep rate. This spike was driven by a mixture of both local and international forces. In 2020, over-rain and pest attacks destroyed soybean cultivation in major soybean growing states such as Madhya Pradesh and Maharashtra, decreasing domestic supply. At the same time, the supply in the world became stricter because of droughts in the main producers of soybeans (Argentina and Brazil) and the increased demand in the world. This ideal storm resulted in the skyrocketing of prices in soybeans by the end of 2021, with the local prices of soybeans in certain markets increasing by more than two times. The prices were as high as in history before stabilizing, which eventually saw a decrease in 2023 as supply returned to normal. As an example, the market price dropped to approximately 4300 per quintal in mid-2024, which is lower than in 2021 and even lower than the minimum support price (MSP) set by the government [1]. This volatility presents a challenge to farmers, who are unable to make any decision or plan on how to sow, store, market, and export their products. The average price trend of soybeans is shown in Figure 1.

Proper price forecasting can be a useful instrument to overcome this uncertainty [3]. Predictive assurance of crop price allows farmers to determine either what and when to plant or sell the

***Corresponding author:** Vilas Damodhar Ghonge, Department of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, India. Email: vilas.damodhar@mitwpu.edu.in

Figure 1
Soybean average price trend over time



produce, assists government agencies in planning the release of buffer stock and import/export policies, and aids industries to manage supply chains. Nevertheless, the agricultural commodity prices are notoriously hard to predict as they are affected by a high number of unpredictable variables, including weather shocks, pest outbreaks, and the global market forces [4]. Crop price forecasting has frequently been done using traditional econometric and time-series models such as the Autoregressive Integrated Moving Average (ARIMA). Although ARIMA models are useful in capturing short-term trends, they assume that there are linear relationships and cannot work well when price changes are catalyzed by nonlinear and intricate relationships. Linear models may also produce unreliable forecasts in situations where there are sudden shocks or regime changes in the volatile agricultural markets.

Methods such as machine learning (ML) [5–7] and deep learning (DL) [8, 9] have become increasingly popular in recent years, as they can predict prices with complex nonlinear models and no parametric assumptions. Recent studies indicate the usefulness of ML in the development of intelligent agriculture. In the study by Santos et al. [10], it was shown that the ensemble-based ML models can be used to predict the soybean yield in the management of cover crops by modeling the complex relationships among the weather, soil, and management factors. On the same note, Sowmya et al. [11] highlighted the wider application of ML in smart agriculture that demonstrates ML uses in yield prediction, crop monitoring, and software optimization and discusses issues of data integration and scaling. Other methods such as artificial neural networks (ANNs) are known to be more precise than linear models in price forecasting in farmlands; a study has established that the price of commodities such as oil and potatoes such as groundnut and potatoes in the Indian markets is more accurately predicted by neural network models as compared to ARIMA. One particularly promising kind of DL method that could be applied in time-series prediction is known as a recurrent neural network (RNN)—Long Short-Term Memory (LSTM) [12]. The LSTM networks seek to capture the time variations of the process and have been demonstrated to be superior to the classical models in the prediction of agricultural prices.

On the basis of such developments, this paper proposes to forecast the price of soybeans in India with the assistance of an enriched dataset where the past market prices and arrivals are not the only sources of information, but they are supplemented with the weather data and other pertinent contextual variables. In a bid to determine the modeling method that has the highest predictive performance in this task, we will compare various methods, such as ensemble tree-based models and state-of-the-art DL models. The main worth of the research is that it creates a powerful and accurate model of predicting soybean prices in the Indian agricultural market, relying on the approaches of ML and DL.

The contributions that this research makes in particular are as follows:

1) To generate a multi-source farm data through the combination of daily soybean prices and market delivery with contextual data, provided by official sources like Agmarknet and India Meteorological Department (IMD), to predict soybean prices effectively, including January 2015 to June 2025.
2) To conduct and compare various predictive models, such as ML algorithms (random forest (RF), XGBoost), time-series and DL models (ARIMA, LSTM, Gated Recurrent Unit (GRU)), and an AgroNET (Hybrid LSTM-GRU) architecture, to assess their usefulness in predicting soybean prices.
3) To achieve feature engineering with effective performance in a model, the use of lagged variables, seasonal indicators, and exogenous inputs (e.g., rainfall, production, exports) together with hyper-parameter optimization to achieve maximum prediction accuracy.
4) To assess model accuracy using standard evaluation metrics, such as root mean squared error (RMSE), mean squared error (MSE), and R-squared (R^2), and to compare model outputs in terms of forecasting precision and generalization on unseen data.
5) To enhance model interpretability using explainable artificial intelligence (XAI) through the local interpretable model-agnostic explanation (LIME) technique, identifying the key drivers of soybean price predictions and providing transparent, actionable insights for farmers and agri-businesses.

To appropriately tackle the issues of predicting the price of soybeans and introduce our research, this paper is subdivided into the following sections: Section 1 describe the introduction and Soybean price issues, Section 2 reviews the related literature

about crop price prediction through the use of statistical, ML, and DL methods. Section 3 explains data collection and preprocessing, such as Agmarknet and IMD, and the combination of the trade and production information. Section 4 includes the exploratory data analysis (EDA), which shows the most important trends, correlations, and insights of visualization, methodology and Algorithms used. Section 5 describes the performance of several ML and DL models including RF, XGBoost, LSTM, GRU, ARIMA, and Proposed Hybrid LSTM—GRU(AgroNET). Lastly, Section 6 summarizes the paper on its main findings and indicates the prospects of what can be done in order to enhance and implement predictive systems in the agricultural markets.

2. Literature Review

2.1. Scope and selection criteria

The reviewed studies are those published recently using ML to predict the price of agricultural crops and, as a point of reference, the yields of crops. We chose the literature with the focus on the Indian agricultural setting and plants of significant importance; thus, we had a variety of methods and results. The major selection criteria were (1) related to crop price forecasting, (2) applied ML or hybrid techniques, and (3) crop-specific and region-based study. With the help of these criteria, we filtered the list of possible works that can be used to reflect the variety of phenomena: short-term and long-term forecasting horizons, price or yield prediction goals, case studies on specific crops or regions of study, and the variety of ML and DL algorithms and features that are used in this area.

2.2. Price forecasting vs yield prediction

The majority of identified studies are about crop price forecasting, which is focused on predicting the price fluctuations in the market of agricultural products. Such works are commonly based on time-series data of previous prices and similar variables in predicting the near future prices (e.g., next week or month) to enable farmers to make informed decisions. Indicatively, several recent studies in India investigate ML-based prediction of prices of staple crops and vegetables. The particular short-term price forecasting model of Indian crops is the target in Ahmed [12]. There are also works that consider the price and yield: Singh and Janu [13] wrote about the ML methods of making a prediction of price and agricultural yields simultaneously because the agricultural yield (production) determines supply and, consequently, prices. In comparison, crop yield prediction research seeks to predict levels of crop production (yield) typically seasonally or annually. Yield prediction is a long-term issue that may entail agronomic and weather characteristics. To give an example, Ghildiyal et al. [14] created a convolutional neural network (CNN) to predict crop yield, and Agarwal et al. [15] optimized crop yield with the help of ML. Normally, yield-oriented models or those that are price oriented make use of soil data, rainfall, and farming practices or market price history and perhaps other economic factors. Others combine them: Prity et al. [16] introduce a crop combination and a multimodal approach for a market prediction system, where the agronomic data are combined with the price trends using ML to recommend crops and predict their yields and market prices. The objective of the price forecasting models is to reflect the dynamics in the short-term of the market, whereas yield prediction models are used to predict longer-term agricultural output, but both of them use similar ML methods.

2.3. Forecasting horizon: short term vs long term

Short-term forecasting typically means the one who is predicting prices in the short term (days, weeks, and a few months ahead) usually to help make tactical decisions such as where and when to sell produce. A large number of price prediction research is short term in nature. As an illustration, the arecanut price forecasting in Kerala forecasted the monthly prices of the coming month using 10 years of data. Models they tried were the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model, Holt–Winters seasonal model, and an LSTM neural network, with the LSTM producing the best fit (lowest RMSE) to short-term price patterns. Likewise, Mithuna and Terence [17] also studied and analyzed the monthly price of coconuts in Tamil Nadu (2010–2022) and developed models to predict future prices; in their findings, the ensemble tree models made very high predictions ($R^2 = 0.97$) of the one-month-ahead price of coconuts.

Long-term forecasting represents long-range forecasting, such as predicting prices or yields, on a multi-month or annual basis, which in itself is more difficult to predict as more uncertainties compound over time. Few papers directly address multi-year price forecasting, though a few are on annual trends or adaptive approaches to keep the model working over time. Celik and Celik [18] point out the challenge of effective long-term predictions in developing economies because of data quality problems and structural adjustments. They suggest the context-based model selection and regular retraining to keep the accuracy of crops such as tomato and maize across various market-years. On the yield side, long-term involves the prediction of yield at the end of the season months prior to harvesting. Study by Oikonomidis et al. [19] fits into this category by forecasting annual/6 monthly yields using ML models, effectively months in advance of harvest. In general, short-term price forecasts (e.g., next month's price) are more common and often rely on recent price trends, whereas long-term projections (next year's price or yield) may require incorporating macro-level inputs (climate forecasts, policy changes, etc.) and often use hybrid or adaptive ML strategies to remain reliable.

2.4. Crop-specific and regional focus

The research on certain crops or types of commodities is the center of many studies, which is often driven by economic significance or because of the distinct price behavior of the crops. The models can be used to capture the patterns of crops by working at the individual crop level. The examples in our survey are corn prices forecasted with multiple algorithms and extra features, coconut prices in a local market forecasted with ensemble ML models, arecanut with time-series and LSTM forecasts, and brinjal prices in various markets in Odisha forecasted with various ML models. Different crops have different challenges—for example, perishable produce such as brinjal or tomato have large seasonal price changes and models that can adapt to seasonality well are needed, whereas plantation crops such as coconut have slower changes but can be influenced by weather events.

One common insight across regions and crops is the need to include relevant features beyond just past prices. Many studies that achieved high accuracy did so by incorporating additional inputs: for example, economic indicators and global events for corn prices, or weather and yield information for integrated price models. In a comparative study, Banerjee and Mondal [20] took

into account rainfall and previous crop yields as features to predict future crop prices using a decision tree model. Such inclusion of climate and yield data can improve the model's ability to anticipate supply-driven price changes. Indeed, Zhang et al. [21] found a strong correlation between planted area and price for garlic, suggesting production area data can inform price forecasts. Overall, crop- and region-specific models tend to perform better when they are tailored with domain knowledge—capturing local seasonality, crop calendars, and relevant exogenous factors.

2.5. Machine learning techniques

A wide range of ML algorithms have been explored for crop price (and yield) forecasting. Below, we group the approaches into major categories and highlight findings:

2.5.1. Traditional time-series models vs ML

Several studies compare classical statistical forecasting methods (ARIMA, SARIMA, exponential smoothing, etc.) with newer ML models. It is generally observed that ML approaches can outperform purely statistical models for complex agricultural data. Paul et al. [22] note that while ARIMA has been widely used for commodity prices, ML techniques have recently shown superior accuracy for vegetable price series in India. For example, in forecasting brinjal (eggplant) prices, they found that all tested ML models outperformed the ARIMA baseline in most cases. Another study on onion/tomato/potato prices combined exponential smoothing (ETS) with ML, yielding hybrids that significantly improved error metrics over standalone ARIMA or ETS. That said, classical models are still used as components or benchmarks in many works—for example, ARIMA is employed as a feature extractor or benchmark in hybrid frameworks.

2.5.2. Regression and tree-based models

Simple regression models (linear or polynomial regression) provide a baseline but often struggle with nonlinear price patterns. Tree-based models have gained popularity due to their ability to capture nonlinear relationships. Decision tree and ensemble tree algorithms like RF and gradient boosting machines (GBM) (e.g., XGBoost) are frequently used. These models handle irregular time series and multi-factor inputs well, and several studies report them as top performers. An example is a case study by Krishna et al. [23] on the prediction of prices of crops (using linear regression, decision tree, RF, Support Vector Machine (SVM), AdaBoost); an RF model was identified as the most accurate based on the lowest Mean Absolute Error (MAE)/RMSE, and they used Shapley Additive Explanations (SHAP) values to explain the effect of features on prices. Mohanty et al. [24] created the full ML model with the use of the yield, import, and demand data to forecast prices; in their testing process, the Decision Tree Regressor was the most efficient in predicting final prices compared to the time-series models or ANN. The appeal of tree-based methods lies in their interpretability and the ability to capture interactions: for example, RF was noted as effective in explaining corn price influences from economic indicators and supply–demand dynamics.

Tree-based models are widely used for agricultural price prediction due to their ability to capture nonlinear relationships, handle multi-factor inputs, and provide interpretability, often outperforming regression and time-series models; however, they may struggle to model long-term temporal dependencies and seasonality without additional feature engineering and can be computationally intensive or prone to overfitting on limited or noisy data.

2.5.3. Artificial neural networks (ANN) and deep learning

Neural network-based models have shown great promise in capturing complex temporal patterns. Feed-forward ANNs (e.g., multi-layer perceptron) were used in earlier works for price prediction and often outperformed linear models. A notable variant is the Generalized Regression Neural Network (GRNN)—Paul et al. [22] found GRNN to be the best performer for brinjal price forecasting across markets, outshining RF, Support Vector Regression (SVR), GBM, and ARIMA in terms of error metrics. In fact, GRNN achieved consistently lower forecasting errors (the second-best was RF in a few cases), while SVR, GBM, and ARIMA were "not up to the mark" in that study. For sequence modeling, RNN especially LSTM is widely used for agricultural price time series. LSTMs can learn seasonal patterns and long-term dependencies in price data. Multiple studies report LSTM as a top model: for example, in the Kerala arecanut price study, an LSTM network fit the data best (beating SARIMA) for monthly price forecasting. Similarly, Chaitra and Meena [25] implemented both a regression model and an LSTM for crop price prediction (using inputs like crop type, nutrient content, MSP, variety, location) and found that the LSTM exhibited an outstanding fit, surpassing the regression model and other approaches. LSTM's advantage is evident especially for crops with strong seasonal price cycles (e.g., tomato), where it can adapt to the time-dependent patterns better than static models.

DL has also been applied to yield forecasting and spatial data. Newer creative models even integrate the DL structures. Bhardwaj et al. [26] suggested that a hybrid between Graph Neural Networks (GNNs) and CNNs could be used to take advantage of geospatial relationships in crop prices. Their DL algorithm, which was trained using the price data of potatoes and tomatoes in Indian markets, is said to have made errors that were two times smaller than the previous ones and was capable of making price predictions, as far as 30 days into the future. This is an indication of the tendency to use more sophisticated AI (spatiotemporal networks) to get even more accurate prices.

Neural network–based models such as GRNN, LSTM, and recent CNN/GNN hybrids have demonstrated superior performance in agricultural price forecasting by effectively capturing nonlinear, seasonal, temporal, and spatial patterns, consistently outperforming traditional statistical and standalone ML models. However, their effectiveness is limited by high data and computational requirements, reduced interpretability, and potential instability in data-scarce or highly volatile market environments.

2.5.4. Hybrid models and integrated frameworks

Various studies have concluded that the combination of various modeling methods is more effective than a single model. Hybrid models are usually a combination of a statistical component to model trend/seasonality and an ML component to model nonlinear patterns of the residual. Purohit et al. [27] suggested a number of hybrid models, including Additive ETS LSTM, Multiplicative ARIMA-SVM, etc., to forecast the prices of vegetables in a month. These hybrids performed better than two standalone statistical models and different individual ML models in predicting tomato, onion, and potato prices with the lowest errors (MAE, RMSE) in general. The other example is the combination of ML models with ARIMA to extract features and forecast, which was demonstrated to enhance forecasting performance in staple food crop prices [28]. This is based on the fact that the linear structure of the price series that is autoregressive can be modeled by ARIMA or ETS, and the remaining nonlinear structures

are then forecasted by the ML models, leading to more accurate forecasting.

There are some frameworks that combine several sub-models of various sub-tasks. A good example is Mohanty et al. [24]; they developed a pipeline comprising four blocks: yield prediction, supply estimation (predicted yield + carryover + imports), demand prediction, and lastly, the price prediction. The yield they relied on was computed using classical time series (AutoRegressive (AR), ARIMA, etc.), simple regression with demand, and thereafter experimented with the three methods of the price based on the inputs of the supply and demand of the price. The time-series model performed worse than the decision tree in their results, with the latter showing the smallest RMSE on price. This holistic focus highlights that a focused application of domain variables such as yield and import on price and the application of ML can be used to increase the degree of forecasting reliability. On the same note, Celik and Celik [18] focus on the data preprocessing and data quality characteristics in addition to weather and arrival quantities and construct strong price models. They additionally indicate an adaptive choice of models to the context, such as an automatic selection of an LSTM in a highly seasonal setting or a tree model in case data are noisy, to ensure stability when used long term.

Hybrid forecasting approaches consistently outperform single models by combining statistical methods (e.g., ARIMA/ETS) to capture linear trends and seasonality with ML models to learn nonlinear residual patterns, resulting in lower forecasting errors. Recent studies also show that integrating domain-driven variables (yield, supply, demand) and adaptively selecting ML models further enhances price prediction accuracy and long-term robustness.

2.5.5. Multimodal and multi-module fusion

Recent work on multimodal and multi-module fusion shows that combining heterogeneous signals and specialized learning blocks can substantially improve forecasting accuracy and robustness. TriChronoNet [29] advances electricity price prediction by fusing complementary temporal representations through a multi-module fusion design, demonstrating how separating and recombining different temporal dynamics (e.g., short/long horizons and seasonality-like effects) can better capture complex market behavior than single-stream models. In a different smart-city domain, hyper-relational interaction modeling [30] extends multimodality beyond simple feature concatenation by explicitly modeling higher-order relations among multiple input modalities and agents, improving trajectory prediction under real-world interaction complexity. Together, these studies highlight a key multimodal lesson for forecasting: performance gains often come not only from adding more data sources but also from architectures that learn cross-modal interactions and align signals across time and context.

Multimodality in agri-commodity forecasting is increasingly achieved by combining visual/sensing data with market signals as a proxy of supply-side conditions. TThaker et al. [31] show that CNNs have the capability to utilize structured inputs to predict wheat futures, which implies that deep feature extractors have the ability to learn nonlinear trends and regime changes that are applicable to the commodity market. In addition, Yewle et al. [32] propose RicEns-Net, a deep ensemble model for rice yield prediction that integrates multimodal data, including Seasonal AutoRegressive (SAR), optical satellite imagery (Sentinel-1/2/3), and meteorological variables. Applied to field data from Vietnam's Mekong Delta, the study reduces over 100 features to 15 key predictors to improve generalization and minimize overfitting. The results show that multimodal fusion combined with ensemble DL significantly improves crop yield prediction accuracy compared to conventional approaches. These approaches provide incentive to multimodal agricultural price systems, which integrate financial indicators (e.g., futures/indexes), agronomic sensing (satellite vegetation/health proxies), and local market variables where the main contribution is enhanced generalizability, particularly in the face of shocks, through the integration of price forecasts based upon market dynamics, as well as realistic crop conditions.

These studies show that multimodal fusion frameworks, which integrate diverse data sources through multi-module or interaction-aware architectures, consistently outperform unimodal models by capturing complementary temporal, spatial, and contextual information. These approaches enhance forecasting accuracy, robustness to volatility, and generalizability by jointly modeling market dynamics alongside real-world physical and environmental signals.

3. Data Collection and Preprocessing

The paper makes use of a wide range of data gathered between January 2015 and June 2025. The main source of information on prices and the market is the Agmarknet portal (Agmarknet, Ministry of Agriculture and Farmers Welfare) of the Government of India, which gathers daily wholesale market information in mandis around the country. Agmarknet provided the market information on soybean daily, including modal price (average price or price at which the majority of prices were made, in RS per quintal), maximum and minimum price, total arrival (quantity of soybean entering into the market, usually in tons), location of market (district and state), and variety of soybean. The most important data sources and variables are presented in Table 1.

Table 1
Data sources used for soybean price prediction

Data source	Variables collected (units)	Frequency/granularity
Agmarknet[1] (Ministry of Agriculture)	Daily modal price (₹/quintal), minimum price, maximum price; daily arrivals (tons); market name, district, state; variety of soybean	Daily records per market
Patel et al. [33] (IMD rainfall data)	Rainfall amount (mm); percentage of normal rainfall (for each district)	Monthly or seasonal, per district
Agricultural Statistics (Govt. reports)	Soybean area under cultivation in Maharashtra (lakh hectares); soybean total production in India (lakh tons)	Yearly (2015–2025)
Trade Data[2] (Ministry of Commerce and Industry)	Soybean export value (₹ crore); soybean import value (₹ crore)	Yearly (2015–2025)

[1] https://agmarknet.gov.in/home
[2] https://tradestat.commerce.gov.in/eidb/commodity_wise_import

The dataset includes the records of various markets, including the major producers of soybeans. The date and location are recorded on each record, and hence, the spatial and temporal trends can be analyzed. We also added weather information from the IMD to the market information. Specifically, we employed rain statistics at the district level comprising the actual rainfall (in mm) and the percentage of normal (long-term average) rainfall. Monsoon rainfall is an important parameter in soybean production since it is largely rainfed in India, and therefore, it is an important aspect in influencing market supply. We combined the IMD data and the market data based on matching the district and time of each record. We have also included annual agricultural statistics including total area under soybean plantation (in lakh hectares) in Maharashtra and total production of soybean in India (in lakh tons) in each year 2015–2025. These are the numbers obtained by government agricultural reports. Furthermore, to capture the influence of trade on domestic prices, we included soybean import and export values (in crore) for India on an annual basis. Imports and exports affect the domestic supply–demand balance (e.g., high exports or low imports can tighten domestic supply, raising prices). These trade metrics were merged into the dataset by year. Table 2 shows the features considered in soybean price prediction.

Table 3 shows the combined sample data from various sources showing daily soybean market arrivals, price variations, and corresponding rainfall statistics.

4. Data Analysis

We had carried out a significant amount of EDA prior to constructing prediction models to get to know the nature of the soybean price data and factors surrounding it. Through this analysis, issues of distributional properties, trends, seasonal trends, and correlations between variables were identified.

4.1. Exploratory data analysis (EDA)

4.1.1. Instruments price distribution

The price distribution will help to capture the range between minimum, maximum, and the modal prices in the market of soybeans in Maharashtra and provide an insight into price changes and market volatility in this region. Figure 2 shows the spatial heterogeneity of the district-average modal soybean prices (₹/quintal) in the study period (2015–2025), indicating that the major agricultural markets in Maharashtra possess a high level of spatial heterogeneity.

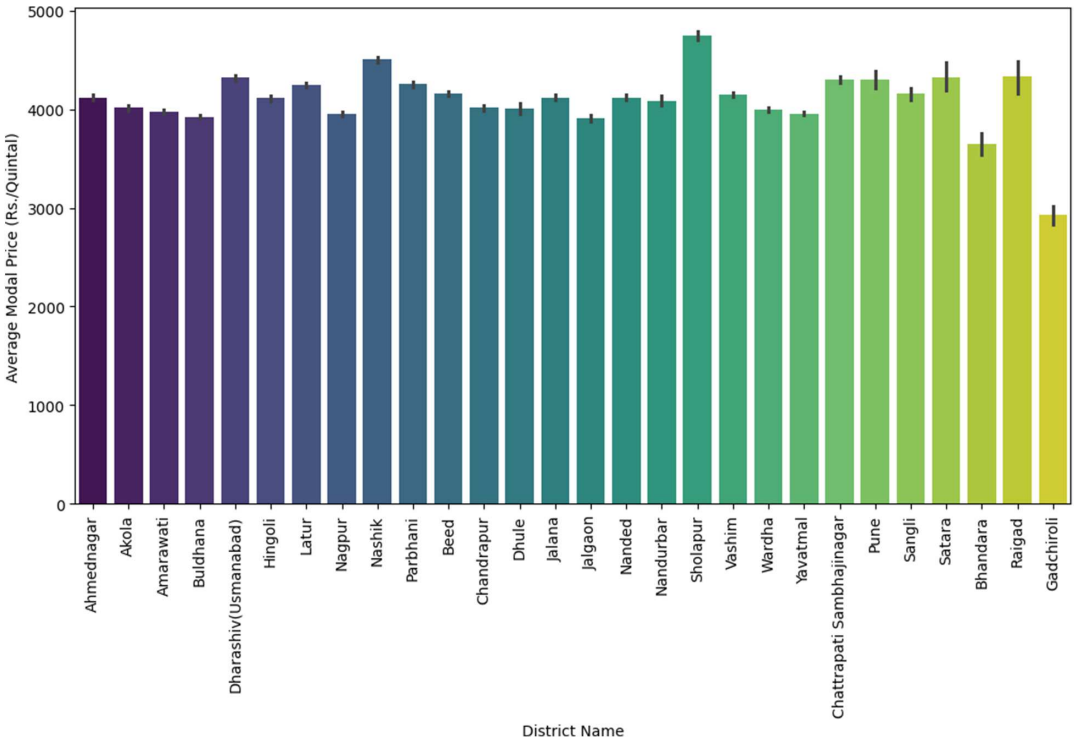
Table 2
Feature considered for soybean price predictions

S. no.	Feature name	S. no.	Feature name
1	State name	10	Reported date
2	District name	11	Normal rainfall (mm)
3	Market name	12	Actual rainfall (mm)
4	Variety	13	% to normal rainfall (%)
5	Group	14	Area under cultivation for soybean (lakh hectare) in Maharashtra
6	Arrivals (tons)	15	Soybean total production in India (lakh tons)
7	Min Price (₹/quintal)	16	Soybean export by India (lakh ₹)
8	Max Price (₹/quintal)	17	Soybean import in India (lakh tons)
9	Modal Price (₹/quintal)		

Table 3
Sample data (combine) from various sources showing daily soybean market arrivals, price variations (minimum, maximum, and modal prices), and corresponding rainfall statistics across different districts in Maharashtra

District name	Market name	Variety	Arrivals (tons)	Min price (Rs./quintal)	Max price (Rs./quintal)	Modal price (Rs./quintal)	Reported date	Normal rainfall (mm)	Actual rainfall (mm)
Ahmednagar	Karjat	Yellow	1.0	4000	4500	4000	2025-03-21	1075.3	1251.0
Akola	Akola	Yellow	206.9	3400	4100	4000	2025-03-21	1075.3	1251.0
Amravati	Varud (Rajura Bazar)	Yellow	0.1	3805	3805	3805	2025-03-21	1075.3	1251.0
Buldhana	Deoulgaon Raja	Oil	3.0	3700	3700	3700	2025-03-21	1075.3	1251.0
Dharashiv (Usmanabad)	Tuljapur	Other	6.0	4000	4000	4000	2025-03-21	1075.3	1251.0
Hingoli	Hingoli	Other	40.0	3550	4050	3800	2025-03-21	1075.3	1251.0

Figure 2
Average modal price of soybean across districts (2015–2025) in India



Some of the districts have a set of prices that are higher than the general average, with the deviations being very high in areas like Solapur. These trends indicate that there are structural variations across the markets that may be occasioned by improved infrastructure, the existence of aggregation centers, and the existence of good connectivity to export channels. The error bars attached to it, which are standard deviations, show different degrees of price stability in the districts, which shows the way local supply–demand dynamics differ. In general, this geographical difference shows the significance of considering the regional factors in the price forecasting models to represent the location-specific market behavior.

4.1.2. Correlation analysis

We examined the correlations among key numerical variables (price, arrivals, rainfall, production, etc.) to gauge linear relationships. Figure 3 shows the correlation matrix for the main variables after preprocessing.

From Figure 3, we observe several notable correlations:

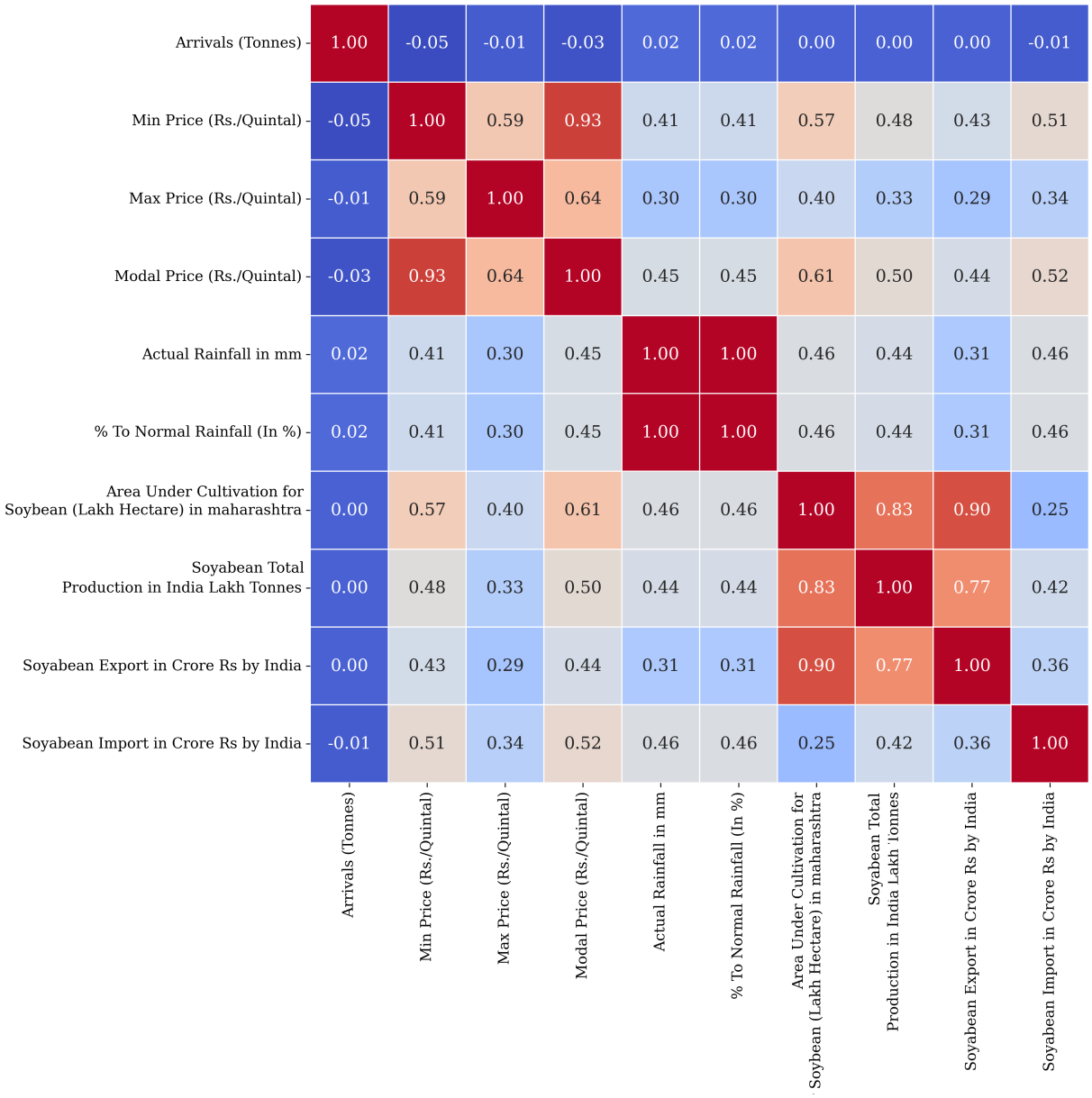
Different price measures (Minimum, Maximum, and Modal prices) are positively correlated with each other, as expected. In particular, the modal price is very strongly correlated with the minimum price (corr ≈ 0.93) and moderately with the maximum price (corr ≈ 0.64). This is intuitive since on any given day, the min, modal, and max prices in a market move together to some extent. The high correlation (0.93) between modal and min price suggests that the modal price does not diverge wildly from the lower end of prices.

Daily arrivals (supply volume) show a very weak negative correlation with prices (corr with modal price ≈ –0.03). The negative value agrees with the intuition about the economy (an increase in supplies tends to reduce prices), whereas the value is close to zero, which shows that on a day-to-day basis, arrivals are virtually not linearly related to price. This means that the daily frequency of simple supply volume is not a good predictor of price without any other factor or lag. It may be because it does not match timings; however, cost formation might rely on anticipations and inventories and not merely on the inflows for that day.

Rainfall variables show a moderate positive correlation with price. The actual rainfall (in mm) has corr ≈ 0.45 with modal price, and the "% of normal rainfall" similarly has a positive correlation. At first glance, this might seem counterintuitive—one might expect better rainfall (which leads to higher production) to lower prices. However, this correlation is likely reflecting year-to-year co-movement rather than a direct causal relationship in the short term. During the period in question, years with good monsoon ironically saw reduced yields due to flooding and pest issues, and global factors drove prices up in 2021. Additionally, rainfall here is aggregated at an annual or seasonal level; higher rainfall years coincided with higher price years in this short sample, hence the positive correlation. It underscores that correlation does not imply a simple causal effect, and the relationship between rainfall and price is complex (timing, distribution of rain, etc., matter).

The correlation analysis shows that there are strong positive relationships among minimum, maximum, and modal soybean prices, which show similar price formations across markets. Market arrivals show poor correlations with prices, indicating that short-run arrivals are not the sole causes of price fluctuations. On the contrary, weather variables, production indicators, and trade-related factors exhibit moderate correlations with prices, and they have an indirect but significant impact on the market. The fact that domestic prices have a positive relationship with export values indicates that high domestic prices are related to high export earnings. On the same note, a positive relationship with imports demonstrates that high domestic prices are often coupled with high-import spending that may be attributed to shortages in supply, which requires imports of soybeans or other products. These

Figure 3
Correlation table of important variables of the soybean data (2015–2025)



are trade-based correlations that show that the global market integration involves trade flows that are sensitive to the domestic price movements and can lead to price stability in the long-term. For modeling purposes, the inclusion of these indicators helps capture longer-term supply–demand adjustments. The key observations derived from the correlation analysis are summarized in Table 4.

4.1.3. Pairwise scatter plot analysis

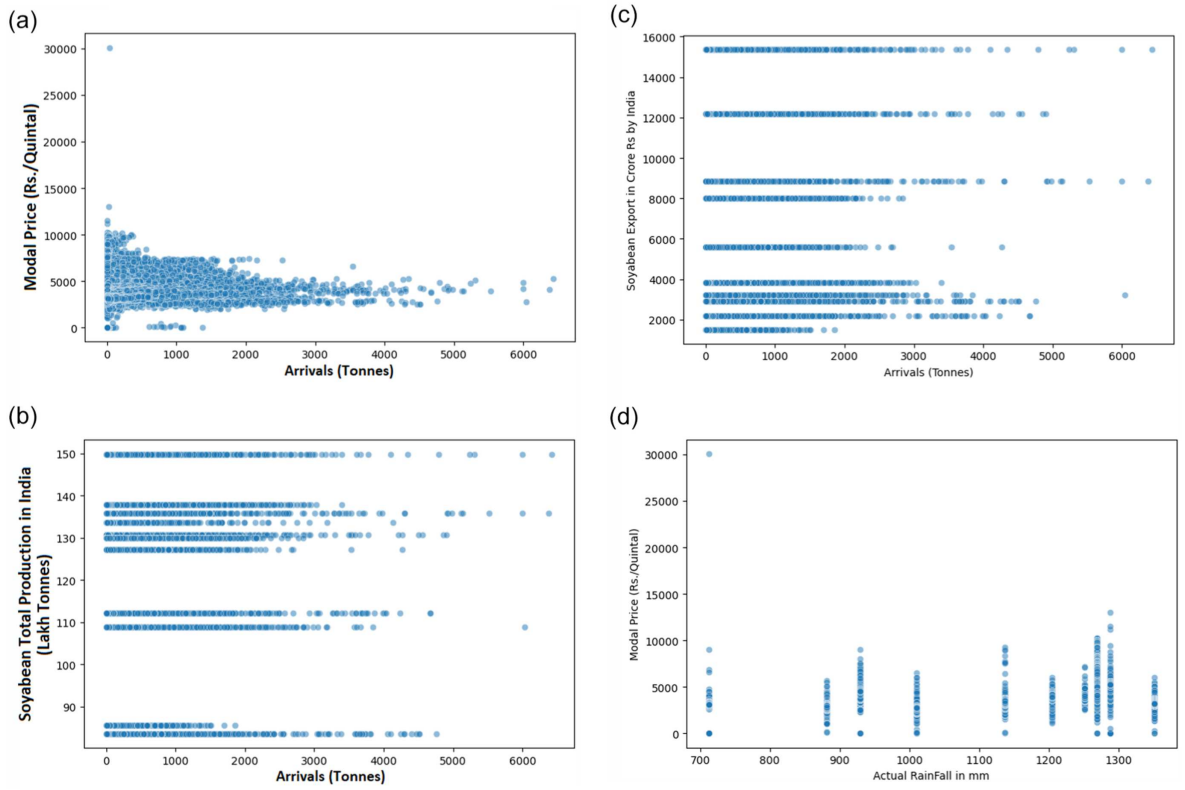
In addition to the correlation matrix, we performed pairwise scatter plot analysis to visually inspect relationships (each paired with a Pearson correlation annotation). Most pairwise plots (e.g., arrivals vs price, rainfall vs price, production vs price) reinforced the observation that linear correlations are weak in isolation. As an example, the relationship between modal price and arrivals was a broad cloud of scatter without a definite trend, which is in line with the close to zero correlation. Minimum price vs modal price, on the other hand, had a more or less linear relationship (with a bunch around the line, the modal price was approximately equal to min price plus some days when the modal price was greater). This only goes to confirm that modal price movements are in line with the daily price range. Such scatter plots also showed what we have actually merged: in variables that vary annually (such as production, area, trade), the plot of daily price vs these appeared like a series of parallel vertical bands (each year offering a fixed level of production and a band of prices that year). This implies that the model can learn year-specific effects of those variables. Figure 4 presents the pairwise scatter plot analysis illustrating the relationships between arrivals, prices, rainfall, and various soybean market indicators. Figure 4 contains a set of scatter plots with correlation coefficients, illustrating relationships among key agricultural variables:

1) Arrivals (Tons) vs Modal Price (₹/Quintal) – Corr: −0.029

Table 4
Correlation observations between various features including rainfall, cultivation area, and soybean market prices

Description	Correlation
% To normal rainfall (In %) is highly overall correlated with actual rainfall in mm	High correlation
Actual rainfall in mm is highly overall correlated with % to normal rainfall (In %)	High correlation
Area under cultivation for soybean (lakh hectare) in Maharashtra is highly overall correlated with max price (₹./quintal) and 4 other fields	High correlation
Max price (₹./quintal) is highly overall correlated with area under cultivation for soybean (lakh hectare) in Maharashtra and 3 other fields	High correlation
Min price (₹./quintal) is highly overall correlated with area under cultivation for soybean (lakh hectare) in Maharashtra and 3 other fields	High correlation
Modal price (₹./quintal) is highly overall correlated with area under cultivation for soybean (lakh hectare) in Maharashtra and 3 other fields	High correlation

Figure 4
Scatter plot analysis showing relationships between key agricultural variables. (a) Arrivals (Tons) to modal price (RS/Quintal). (b) Arrivals (Tons) to soybean total production in India. (c) Arrivals (Tons) to soybean export in crore RS. (d) Actual rainfall to modal price

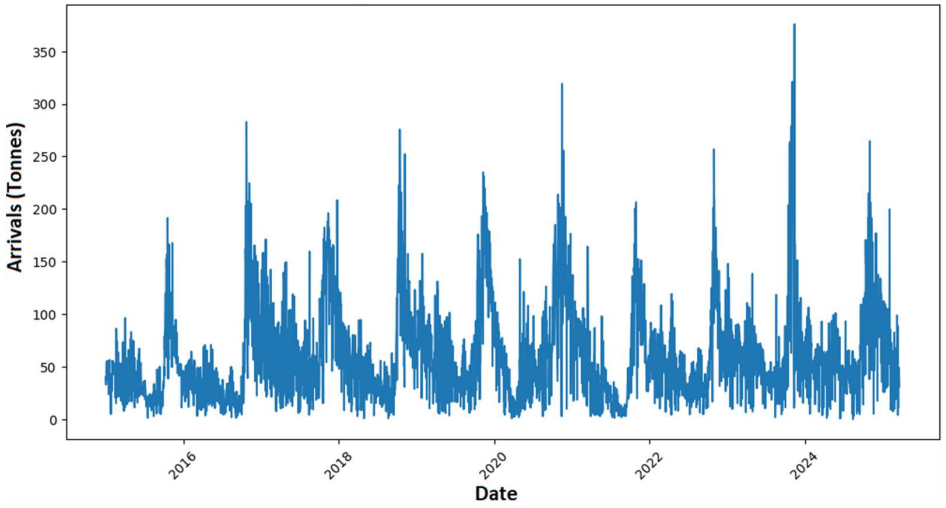


2) Arrivals (Tons) vs Soybean Total Production in India (Lakh Tons) – Corr: 0.003
3) Arrivals (Tons) vs Soybean Export in Crore Rs by India – Corr: 0.003
4) Actual Rainfall (mm) vs Modal Price (₹./Quintal) – Corr: 0.450

4.1.4. Temporal analysis

1) Arrivals and price over time: We examined how soybean arrivals and prices varied over the 10-year period. Figure 5 shows the time-series plot of monthly soybean arrivals, as expected, the data exhibited seasonality tied to the crop calendar. Arrivals in the mandis tend to surge during the harvest months. Soybean is harvested around October–November in the Kharif season; indeed, we observed that each year, arrivals peak in the last quarter (Q4) of the calendar year. For instance, in 2021, there was a pronounced spike in arrivals during November, after which arrivals dropped in the lean months. Prices often show an inverse seasonal pattern—they tend to soften or dip when arrivals flood the markets post-harvest and rise during the lean supply period before the next harvest. However, this pattern can be overshadowed by larger trends and shocks. In our data, the usual seasonal price dip after harvest was not very pronounced in years like 2021 because overriding factors kept prices high.

Figure 5
Time-series plot showing monthly soybean arrivals (in tons) from 2015 to 2025



Geographical and variety analysis: We analyzed how prices varied across different districts and by soybean variety. Table 5 shows the modal price across different districts in India. Figure 6 shows the relationship between Rainfall (mm) and price (Rs./Quintal) across different districts. The market, focused on the markets in Maharashtra, gave us a dataset that enables us to compare average prices in markets across the state. We could find some spatial variation: for example, major producing districts such as Latur or Nagpur have slightly lower average prices than those in the districts, which were further away from the production centers. Nevertheless, the divergences were not colossal; the general market is generally highly integrated, and the disparity in price is mostly within 10–15% of each other after considering the quality and logistics. The average price in the various districts was plotted, and it was found that the majority of the districts are around the average price, and there are a few outliers. In terms of soybean variety, the data were overwhelmingly composed of the yellow soybean variety, which is the standard commercial variety, as shown in Figure 7. We tallied the variety distribution: out of all records, > 99% were "Yellow," with a very small number labeled "Black" or "Other". The dominance of yellow soybean is depicted in the variety distribution chart, essentially a bar chart with Yellow dwarfing the others. Because the minority varieties were so scarce, it is hard to draw conclusions about price differences by variety from this dataset. For the modeling phase, this means variety does not contribute much to predictive power, since effectively all data is one category.

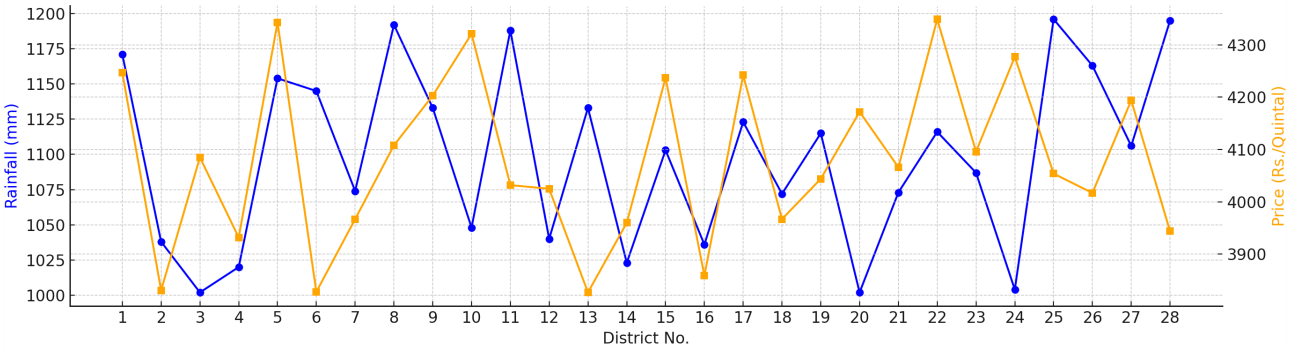
2) Outlier analysis: To complement the histogram and boxplot view of price distribution, we also used violin plots and box plots, grouping data by year and by some key factors. For instance, we plotted the distribution of daily prices for each year (2015–2025) in a series of violin plots as shown in Figure 8. This visualization confirmed that 2021 had a distinctly higher price distribution—its median and bulk of prices were much greater than those of 2020 or 2022. Year 2022 also showed a wide spread. By 2023 and 2024, the distribution shifted back downward. We also grouped prices by the annual production and trade values. For example, one violin plot had the x-axis as "Soybean Production (India, lakh tons)" with one violin for each year's production figure, showing the price distribution for days corresponding to that production level.

Similarly, "area under cultivation" as the x-axis grouped the years. These plots depicted the inverse relation trend: the price distribution violin was higher at years with lower production (e.g., 2021) and lower at years with higher production (e.g., 2023). This again implies that the poor years of harvest meant that there was an escalation in prices. The other interesting plot was the price distribution vs the export values, which showed that just

Table 5
Modal price (₹. /quintal) across different districts

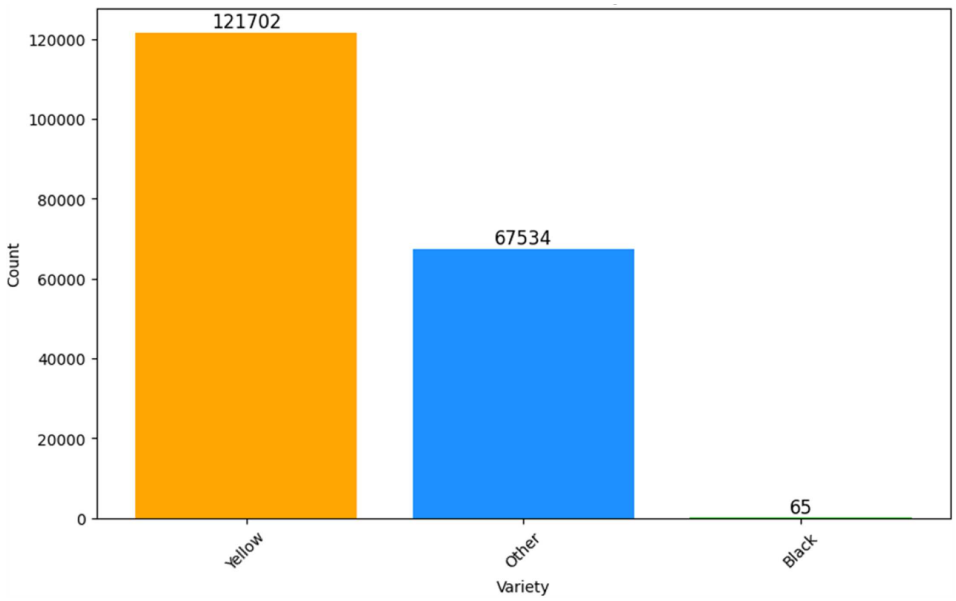
District no.	District	District no.	District	District no.	District	District no.	District
1	Ahmednagar	9	Dhule	17	Nandurbar	25	Solapur
2	Akola	10	Gadchiroli	18	Nashik	26	Wardha
3	Amravati	11	Gondia	19	Osmanabad	27	Washim
4	Beed	12	Hingoli	20	Parbhani	28	Yavatmal
5	Bhandara	13	Jalgaon	21	Pune		
6	Buldhana	14	Jalna	22	Raigad		
7	Chandrapur	15	Nagpur	23	Sangli		
8	Sambhajinagar	16	Nanded	24	Satara		

Figure 6
Line chart illustrating the relationship between Rainfall (mm) and Price (Rs./Quintal) across different districts



District No.	District	District No.	District	District No.	District	District No.	District
1	Ahmednagar	9	Dhule	17	Nandurbar	25	Solapur
2	Akola	10	Gadchiroli	18	Nashik	26	Wardha
3	Amravati	11	Gondia	19	Osmanabad	27	Washim
4	Beed	12	Hingoli	20	Parbhani	28	Yavatmal
5	Bhandara	13	Jalgaon	21	Pune		
6	Buldhana	14	Jalna	22	Raigad		
7	Chandrapur	15	Nagpur	23	Sangli		
8	Sambhajinagar	16	Nanded	24	Satara		

Figure 7
Distribution of soybean varieties in the dataset



like the highest export year, the price distribution would be high, and the lowest export year (2020) had lower prices. The fact that such qualitative observations make sense in economic logic and add to the rationale as to why such exogenous characteristics might be included in making sure that the model prefigures price changes. Lastly, we mention the treatment we gave to the outliers: we retained the outliers in the dataset, but we were conscious that their presence might have an effect on the model training.

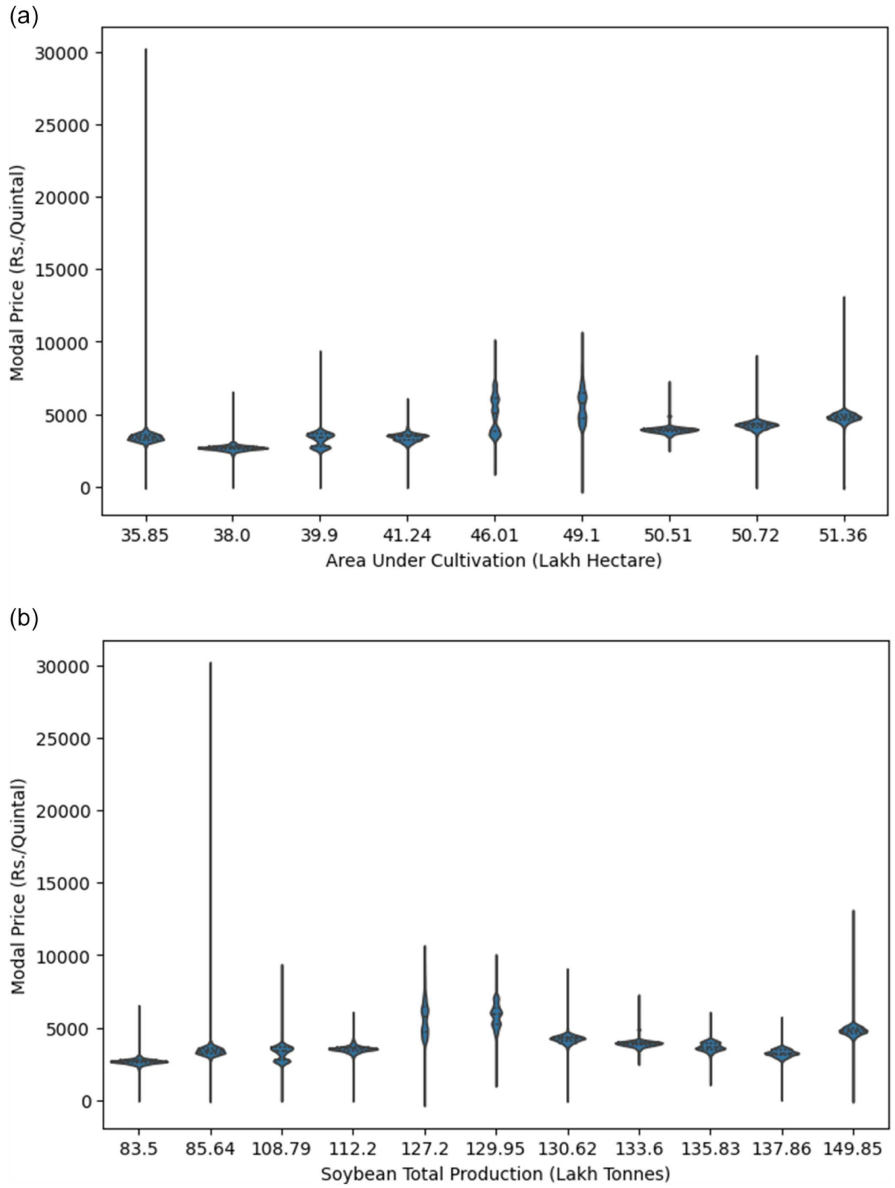
4.2. Methodology

4.2.1. System architecture

The methodology presents a comprehensive and structured pipeline for agricultural price forecasting using both classical ML and DL techniques (shown in Figure 9).

The process of methodology starts with data preprocessing, which entails the treatment of missing values, outliers, integrating rainfall and agricultural indicators, and normalization and

Figure 8
Relationships between key agricultural variables. (a) Seasonality cultivation area vs soybean price. (b) Total production vs soybean price

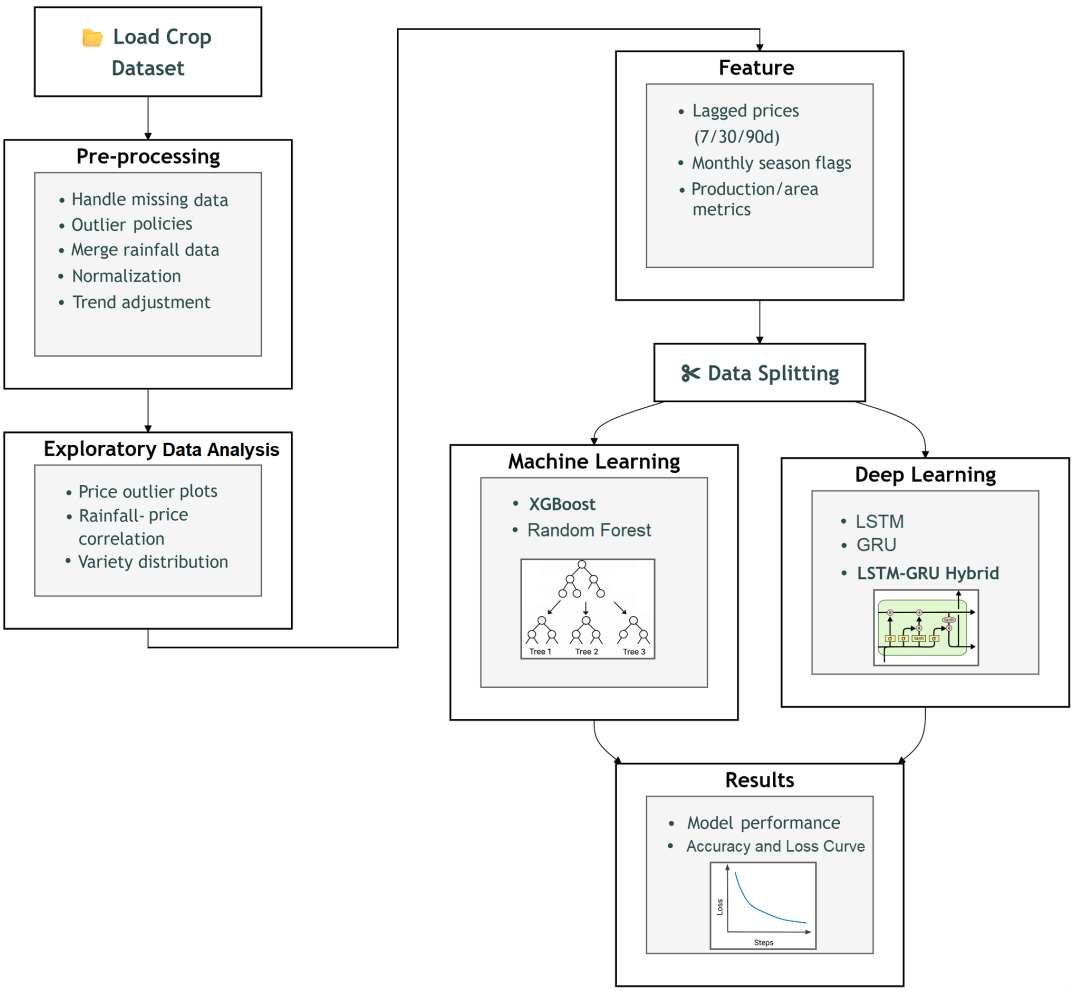


time-sensitive splitting. This is then preceded by EDA to reveal the trends and correlations among rainfall, types of soybean, and price changes. The method of feature engineering also includes lag features, seasonality, and annual agricultural information to add to predictive ability. The ML Models section takes advantage of RF and XGBoost, which have been optimized using hyperparameter optimization and feature importance analysis. The LSTM, GRU, and AgroNET (Hybrid LSTM-GRU) architecture DL models are utilized to extract the temporal dependencies in the data. Lastly, the AgroNET (Hybrid LSTM-GRU) algorithm is a combination of the two models in order to achieve better accuracy with a defined structure and training routine and loss. This end-to-end approach makes it strong in forecasting with a combination of domain-specific features and sophisticated modeling techniques. The proposed methodology has a detailed flow presented in Figure 9, which depicts every step in the preprocessing of data and exploratory analysis, feature engineering, model development, and final evaluation.

4.2.2. Materials and methods

1) Preprocessing: The data obtained was preprocessed before analysis. The Agmarknet price data were cleaned first, in which we had to deal with missing or erroneous data. There were a few records with lost price or quantity details, which were either imputed or dropped when unreliable. We also checked for obvious data entry errors or outliers in prices/quantities—for instance, any record where the price was off by an order of magnitude compared to typical values was flagged. In the soybean price data, the vast majority of daily modal prices fell within a reasonable range (roughly ₹3000–₹6000 per quintal during 2015–2025), but a few extreme values (above ₹15,000 or more) were noted. We retained these in the dataset for analysis, but we gave special attention to them during EDA (outlier analysis) to ensure they did not unduly skew the model training. Next, we merged the datasets. The daily market data was augmented with corresponding rainfall data by matching on district and month. Since daily

Figure 9
System architecture showing the in-depth process of the proposed agricultural price forecasting process



rainfall data at the district level was not always available, we used monthly total rainfall and monthly normal rainfall to compute a percentage of normal rainfall for that month in that district and merged that value with all daily records of that month for the given district. This means every daily record for a district in, say, August 2021 would carry the same "% of normal rainfall" value, indicating how that month's rainfall compared to the long-term norm. Annual variables were merged by year: each daily record in 2020 has attached the 2020 values for these annual indicators, and so on. Although this introduces repeated values across daily observations, it enables the ML models to capture year-specific contextual effects without introducing temporal leakage.

2) One-hot encoding: Categorical variables like district and soybean variety were handled appropriately. The district was not one hot encoded directly; instead, we considered modeling approaches that could handle location as a grouping. The soybean variety in the data was mostly "Yellow" with a small fraction labeled "Black" or "Other". We encoded variety as a categorical feature. However, given that over 99% of the records were yellow soybeans, the variety did not play a major role in the modeling and served mainly to confirm that the dataset was homogeneous in terms of crop type.

3) Feature normalization: We then standardized or normalized features where appropriate. The continuous features had different scales (e.g., rainfall in mm vs price in ₹, vs percentage rainfall, vs tonnage of production). Tree-based models (RF, XGBoost) are not sensitive to feature scaling, but neural network models benefit from normalization for faster convergence. We applied a Min–Max scaling to the features for the neural network models and similarly scaled the target variable when training LSTM/GRU models and then later converted predictions back to original units for evaluation.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where:

- X = original value
- $\boldsymbol{X_{min}}$ = minimum value in the feature
- $\boldsymbol{X_{max}}$ = maximum value in the feature
- $\boldsymbol{X'}$ = normalized value

4.2.3. Feature selection

1) Feature importance using RF: RF computes feature importance based on the mean decrease in impurity (MDI). Each

decision tree in the forest splits the data using features that reduce impurity. The importance of a feature is measured by how much it decreases impurity on average across all trees. The RF handles high-dimensional data well and gives a clear ranking of influential features. Figure 10 displays the relative importance of various features in predicting the target variable, based on the MDI across the trees in an RF. Feature importance is calculated using

$$Importance(f_i) = \sum_{splits\ using\ f_i} \Delta impurity \tag{2}$$

2) Feature importance using XGBoost: XGBoost uses gradient boosting, where trees are built sequentially to correct the errors of previous trees. Feature importance in XGBoost can be calculated in several ways; the most common is based on the F-score, which counts the number of times a feature is used in a split. XGBoost captures nonlinear relationships and interactions between features more effectively than RF.

Figure 11 shows how frequently each feature was used to split data across all trees in the XGBoost model, indicating their relative influence on predictions.

3) Train test split: The dataset was subsequently divided into training and testing sets (80% training, 20% testing) to be used in developing the model. To mitigate look-ahead bias, we collaborated with a time-conscious splitting strategy: we trained/validated the model using the data of 2015–2022 as an internal validation split, and we used the 2023–2025 data as a separate, independent test set to assess the performance of the forecasting. This time division is important in the sense that our models are continually making predictions of future prices based on historic data in a way that is reminiscent of a real forecasting situation. Training data, when fed to models that take samples, were also shuffled, but in sequential models (LSTM/GRU), we kept the time sequence of each training sequence.

4) Classification and price prediction: Five models were used in the classification and price prediction of the trends of the soybean market, which included RF, XGBoost (XGB), LSTM, GRU, and a proposed AgroNET model. RF and XGB were the strong ensemble-based ML baselines that are characterized by their interpretability and capability to work with structured data. Both LSTM and GRU are RNNs, which are used to extract the temporal links of the time-series price data, and LSTM is oriented toward long-term memory storage, whereas GRU is a simpler neural network that is faster. To improve predictive performance, the proposed AgroNET model is a hybrid of the two architectures, with the advantages of sequence learning provided by LSTM and the efficiency of computational performance by GRU. Each model was trained on historical data, which comprised the arrival, rainfall, and production-related factors, and they were tested on the basis of standard performance indicators to determine how accurate and predictive the model was. The modeling details are provided in the Algorithm section.

In order to compare the effectiveness of the proposed and baseline models in the prediction of soybean prices, a number of performance measures were used that include RMSE, MSE, Mean Absolute Percentage Error (MAPE), and R^2. These measures were employed to evaluate the models in their capacity to represent the actual tendencies of the price data. RMSE and MSE will give information on the size of the error in prediction, and the smaller the value, the higher the performance of the model, and R^2 indicates how much of the variation of the target variable can be attributed to the model. Also, the accuracy of classification was applied to determine the accuracy of the price groups forecast. The loss curves and accuracy plots obtained in the course of model training also assisted in getting a clear picture of convergence and overfitting patterns over epochs. These overall assessments made it possible to effectively compare conventional ML models (RF and XGB) and DL networks (LSTM, GRU, and the proposed AgroNET model) and to note the high performance of the hybrid model in the process of regression and classification. The

Figure 10
Feature importance derived from the random forest model

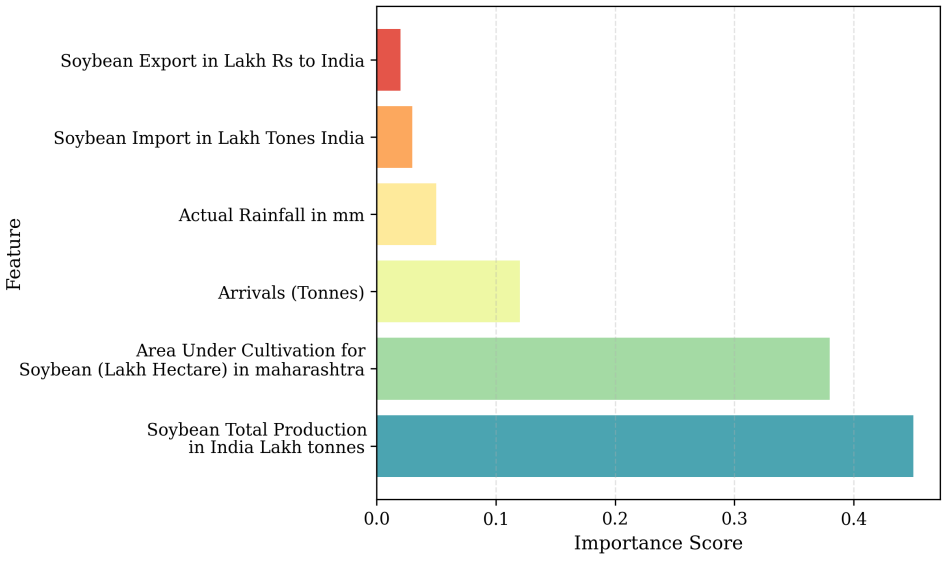
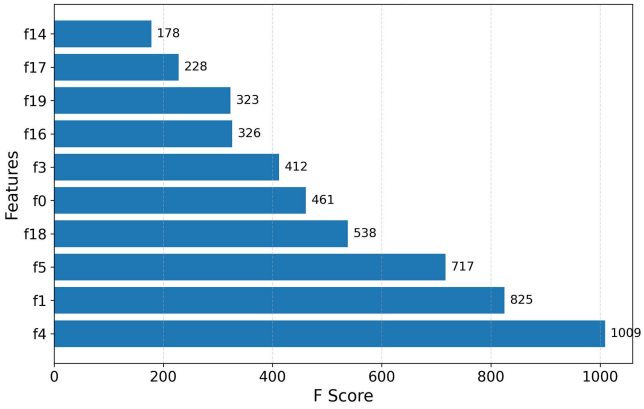


Figure 11
Feature importance based on the XGBoost model



Results section provides comprehensive results and performance evaluation parameters that are discussed there.

4.3. Algorithms

In this section, we describe the predictive modeling approaches used for soybean price forecasting, along with the experimental setup and evaluation metrics. We implemented five different models ranging from ML ensembles to DL networks, allowing a comprehensive comparison of techniques: RF, XGBoost, LSTM, GRU, and AgroNET (Hybrid LSTM-GRU). All model development was done using Python. Below, we explain each model and how it was configured in our study:

4.3.1. Machine learning models

1) Random forest (RF): RF is an ensemble learning method that builds numerous decision trees and averages their predictions to improve generalization. For our regression task, we used the Random Forest Regressor from scikit-learn. We fed the RF model with a feature set that included current and recent past information. Because RF has no built-in notion of sequence, we had to explicitly provide any relevant lagged features or time indicators. The RF model was trained with an ensemble size of 100 trees, and we limited the depth of trees to prevent overfitting (max_depth ~ 10). We used the default criterion of MSE for splits. The model was trained on the training set (2015–2022), and we validated using out-of-bag error as well as a separate validation set.
2) XGBoost: XGB is an optimized gradient boosting library that often yields state-of-the-art results for structured data. We used the XGBoost Regressor with a similar feature set to the RF. Gradient boosting builds trees sequentially, where each new tree corrects the errors of the ensemble so far. We performed a hyper-parameter search for XGBoost, as shown in Table 6. One advantage of XGBoost is that it provides feature importance scores; after training, we examined these to see which features it found most predictive. Typically, we expected lagged price and perhaps month or production to be high importance features.

4.3.2. Deep learning models

1) ARIMA: The ARIMA model is a classical statistical time-series forecasting approach designed to model linear temporal dependencies. Unlike neural networks, ARIMA relies on historical values and past error terms to predict future observations, making it suitable for datasets with strong trends and temporal structure. In this study, ARIMA was configured for one-step-ahead forecasting, where the model predicts the next day's soybean price based solely on past price values. The ARIMA model is defined by three parameters: p (autoregressive order), d (degree of differencing), and q (moving average order). First-order differencing ($d = 1$) was applied to remove non-stationarity in the price series, as confirmed through stationarity tests. The autoregressive component (p) captures the influence of previous price values, while the moving average component (q) models the impact of past forecast errors. Based on autocorrelation function and partial autocorrelation function analysis, as well as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) minimization, ARIMA (2,1,2) was selected as the optimal configuration. This setup effectively balances model complexity and forecasting accuracy. Model training involved estimating parameters using maximum likelihood estimation. Residual diagnostics were performed to ensure the absence of autocorrelation, confirming that the model residuals approximate white noise.
2) LSTM neural network: The LSTM model is a recurrent neural network capable of learning long-term dependencies. We set up the LSTM to perform sequence-to-one prediction: given a sequence of past days, predict the next day's price. We constructed input sequences of length T (we experimented with $T = 30$ days and $T = 60$ days; 30 days was used in the final model as it was sufficient to capture a one-month pattern and kept the model size smaller). Each input sequence consisted of features for each day in the window. The features for the LSTM included past daily modal price values, past daily arrivals, and past daily rainfall or weather indicator. We also appended static or slow-changing features to each time step of the sequence. Another approach could

Table 6
XGBoost hyper-parameters

Hyper-parameter	Values tried	Selected value
learning_rate	0.05, 0.1, 0.2	0.1
max_depth	6, 8, 10	6
n_estimators	~50+ (with early stopping)	~50
early_stopping_rounds	Used	Used
missing_values_handling	Default (built-in)	Default
data_normalization	Not required	Not required
feature_importance	Used for analysis	Yes

have been to input those separately, but we chose to repeat them in the sequence for simplicity. We used a single LSTM layer with a certain number of units (e.g., 50 units), followed by a dense output layer that produces the price prediction. We also experimented with more LSTMs, but only one layer was found to be enough given the size and complexity of the data. Optimizer used to do the model training was Adam with MSE as the loss function. Figure 12(a) simplifies the architectural diagram of the LSTM model, whereas the architecture of the LSTM model, in detail, per-layer, is pictured in Figure 13(a).

3) GRU neural network: The GRU model was adapted from the LSTM, with the exception that it used GRU cells. The input format we used was the same (30-day sequences). GRUs can, in most cases match, the performance of LSTMs using fewer parameters. We also initialized the GRU layer to have about 50 units in our case. The procedure (optimizer, loss, epochs) used was the same as that of LSTM. We wanted to see if there was any notable difference between LSTM and GRU on our data. We expected them to be fairly close in results, as is common, but one might edge out the other depending on data patterns. The simplified architectural diagram of the GRU model is shown in Figure 12(b), while the detailed layer-wise architecture is illustrated in Figure 13(b).
4) AgroNET (Hybrid LSTM-GRU) model: We designed a custom hybrid network that combines both LSTM and GRU units in one architecture. The motivation for this hybrid is to leverage any complementary learning abilities of the two RNN types—perhaps the LSTM might capture certain patterns and the GRU others. One simple way to hybridize is stacking (e.g., an LSTM layer feeding into a GRU layer, or vice versa). We experimented with a structure where the first layer is LSTM and the second layer is GRU. Specifically, the hybrid model takes the input sequence and passes it through an LSTM layer (e.g., 50 units), and then the sequence output of that is fed to a GRU layer (e.g., another 30 units) and finally to a dense output neuron. We found that stacking in this order worked slightly better than the reverse (GRU then LSTM) for our data, though we did not exhaustively test all possibilities. The hybrid network was trained with the same regime as the single RNNs. Because the hybrid has more parameters, we used a bit stronger regularization: a dropout of 0.2 was applied on the LSTM output before feeding it to the GRU to prevent overfitting given the limited data.

Figure 12
Architectural comparison of recurrent neural network models used in the study. (a) LSTM architecture consisting of stacked LSTM layers followed by dropout and dense layers. (b) GRU architecture using stacked GRU layers with dropout and dense layers. (c) AgroNET (Hybrid LSTM-GRU) model with a bidirectional LSTM layer followed by GRU, dropout, and dense layers

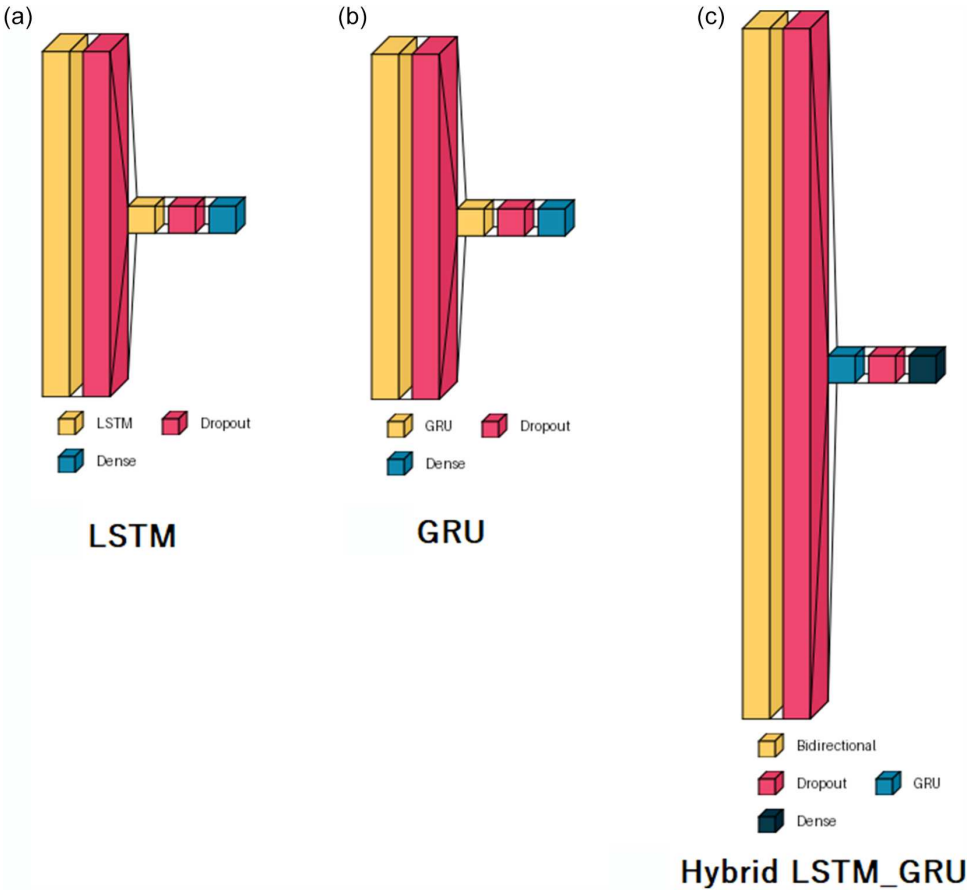
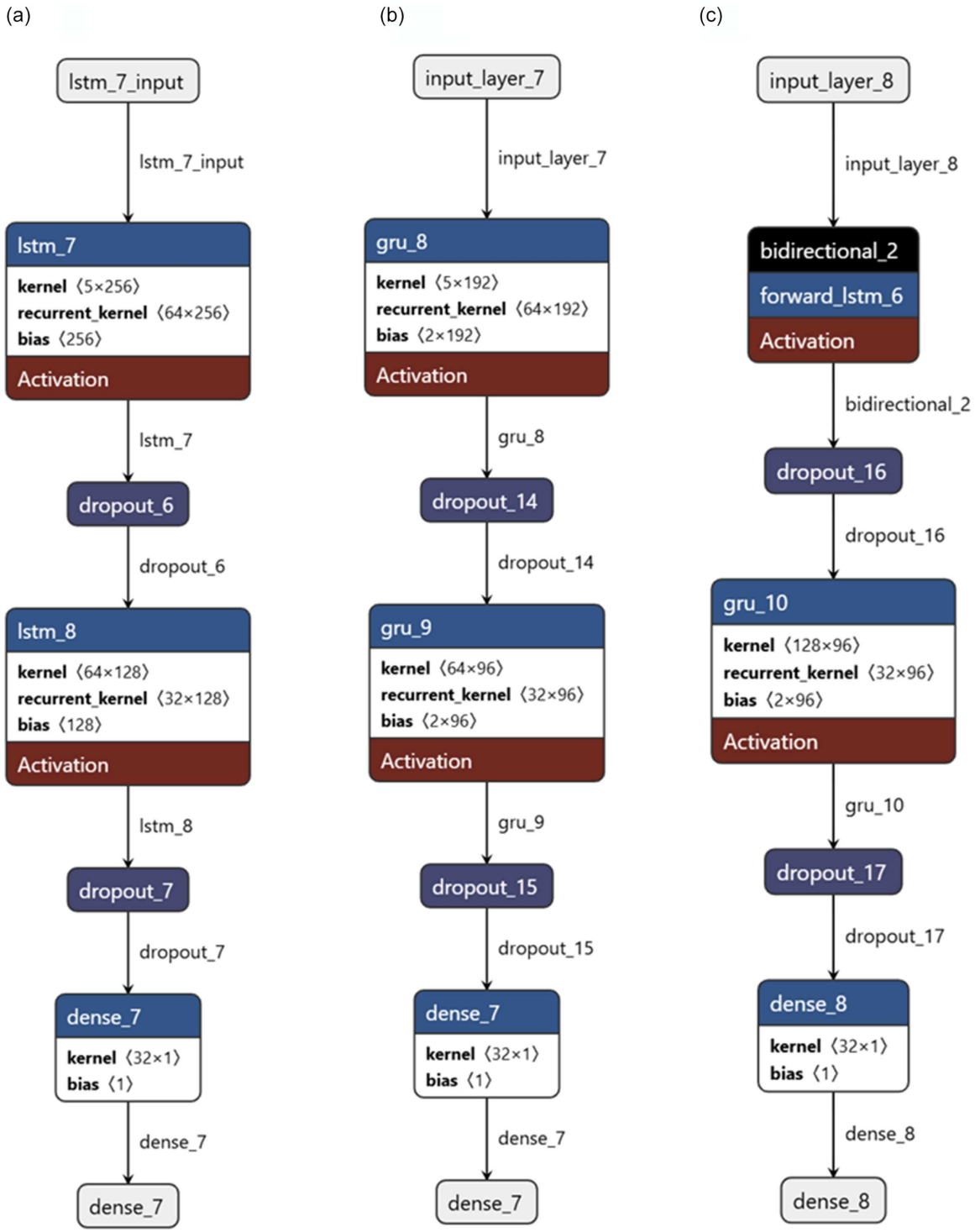


Figure 13
Detailed layer-wise architecture of three recurrent models used in the study. (a) LSTM model. (b) GRU model. (c) AgroNET (Hybrid LSTM-GRU) model



Algorithm: Hybrid Bidirectional LSTM-GRU Model (AgroNET)

Input:
Multivariate time-series data $\boldsymbol{X\_train} \in \mathbb{R}^{n}\frown\approx\frown$, where:
$\boldsymbol{n}$ = number of training samples,
$\boldsymbol{t}$ = number of timesteps,
$\boldsymbol{d}$ = number of input features.
Corresponding target values $\boldsymbol{y\_train} \in \mathbb{R}$^n
Validation set ($\boldsymbol{X\_test}$, $\boldsymbol{y\_test}$)
Output:
Predicted values $\boldsymbol{y\_pred}$ for $\boldsymbol{X\_test}$.
Procedure:
Step 1: Network Initialization
Bidirectional LSTM Layer
Units: 64
Output: 2 hidden sequences (forward & backward) concatenated.
for each direction;

$$\boldsymbol{h_t^{(f)} = LSTM(x_t, h_{t-1}^{(f)})}$$

$$\boldsymbol{h_t^{(b)} = LSTM(x_t, h_{t+1}^{(b)})}$$

Dropout Layer
Dropout rate = 0.2 (Prevents overfitting)
GRU Layer
Units: 32
Operates on the full sequence from the Bidirectional LSTM
Reset / Gate Update;
Update gate: $\boldsymbol{z_t = \sigma(W_z x_t + U_z h_{t-1})}$
Reset gate: $\boldsymbol{r_t = \sigma(W_r x_t + U_r h_{t-1})}$
Hidden state:

$$\boldsymbol{\tilde{h}_t = tanh(W_h x_t + U_h(r_t \bigodot h_{t-1})}$$

$$\boldsymbol{h_t = (1 - z_t) \bigodot h_{t-1} + (z_t \bigodot \tilde{h}_t)}$$

Dense Layer
Fully connected layer with 1 output neuron

$$\boldsymbol{\hat{y} = W_{dense} h_t + b}$$

Output is a scalar prediction for regression.
Step 2: Model Compilation
Loss Function: MSE

$$\boldsymbol{MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Optimizer: Adam with learning rate $\alpha = 0.001$ (Adaptive updates using momentum and squared gradients.)
Step 3: Training Procedure

- Epochs: 20
- Batch Size: 32
- Validation Monitoring: EarlyStopping

Monitor validation loss ($\boldsymbol{val\_loss}$)
Stop training if no improvement in 5 consecutive epochs.
Step 4: Prediction
Use the trained model to generate predictions on $\boldsymbol{X\_test}$:

$$\boldsymbol{\widehat{y_{test}} = Model(X_{test})}$$

5. Result and Discussion

To evaluate the robustness of different predictive approaches, we analyzed the performance of several ML and DL models across multiple agricultural commodities, including RF, XGBoost, LSTM, GRU, ARIMA, and Proposed Hybrid LSTM—GRU (AgroNET). The data split was 80/20, with experiments being done in 80% of the training and 20% of the testing. Standardized EDA, data preprocessing/engineering, and time-sensitive data partitioning were used in the experimental setup to enforce realistic forecasting conditions. All the models were assessed based on the standard performance measures including MSE, RMSE, R^2, MAPE, and the accuracy and loss curves. In addition, k-fold cross-validation was employed to ensure robust performance evaluation, with results reported as the mean across folds to reduce bias arising from a single train–test split and to better assess model generalization. The details of the experimental design, the algorithm configuration, and the evaluation strategies are described in the sections below.

5.1. Experimental setup

Our model training and analysis were run on Google Colab, and we used the benefit of the neural network training on the GPU. To achieve reproducibility, we set random seeds in libraries (numpy, TensorFlow) where possible, but in some cases (i.e., ensemble methods), some randomness still exists. The experimental setup is described in Table 7.

5.2. Performance parameters

We evaluated model performance using several standard metrics for regression. Table 8 shows various performance parameters such as MSE, RMSE, R^2, and MAPE.

5.3. K-fold cross-validation

K-fold cross-validation is a widely used validation technique to assess the generalization ability of ML models. In this approach, the dataset is divided into k equal-sized folds. During each iteration, k−1 folds are used for training the model, while the remaining fold is used for validation. This process is repeated k times so that each fold serves as the validation set exactly once. The final model performance is reported as the average (and standard deviation) of evaluation metrics across all folds. K-fold cross-validation reduces the bias associated with a single train–test split and provides a more reliable estimate of model robustness and stability. In our result analysis, the dataset was partitioned into k = 10 mutually exclusive folds, where in each iteration, k−1 folds were used for training, and the remaining fold was used for validation.

5.4. Result analysis

5.4.1. Machine learning models result analysis

In order to visualize the performance of the ML models, we plotted actual and predicted prices of the test set. Figure 14(a, b) shows sample scatter diagrams of the models, indicating the similarity of the predictions to the real observed prices. The distributions of the points in these two instances are around the diagonal red line, which depicts the ideal prediction scenario, and it implies that the models are capable of capturing the overall price patterns. The model of XGBoost (Figure 14(a)) indicates

Table 7
Experimental setup

Category	Tool/library/resource	Purpose/usage
Platform	Google Colab	Cloud-based environment for running code with GPU/CPU support
Programming Language	Python 3.10	Core language for implementation
Hardware	NVIDIA Tesla K80/T4 (GPU)	Accelerated training for deep learning models (LSTM/GRU)
	CPU (Colab default)	Used for traditional ML models (random forest, XGBoost)
Data Format	CSV files	Storage format for input datasets
Libraries	pandas	Data loading, preprocessing, and manipulation
	numpy	Numerical computations and setting random seeds
	matplotlib	Data visualization
	sklearn	Implementation of random forest and evaluation metrics
	xgboost	Extreme gradient boosting model implementation
	tensorflow/keras	Deep learning framework for LSTM and GRU models

Table 8
Performance parameters

Performance parameter	Formula	
Mean squared error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	(3)
Root mean squared error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{MSE}$	(4)
R-squared (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$ - $\overline{\boldsymbol{y}}$: Mean of the actual values - The numerator is the **residual sum of squares (RSS)** - The denominator is the **total sum of squares (TSS)**	(5)
MAPE	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$ - y_i is the actual (true) value, - $\hat{y}_i$ is the predicted value, - n is the total number of observations.	(6)

that there is a tighter grouping of points around the diagonal, and thus, the model may have relatively more accuracy and be closer to the real values. The model of RF (Figure 14(b)) is also trend-following, but it is more dispersed around the prediction line, indicating a relatively greater error. Generally, the two models are effective, although XGBoost has more accuracy and predictability in terms of soybean prices.

5.4.2. Deep learning models result analysis

In the given work, we have compared the performance of two ML models of RF, XGBoost, and three DL models of LSTM, GRU, and AgroNet in predicting soybean prices based on accuracy and loss curves as the main metrics of their performance.

The Receiver Operating Characteristic (ROC) curve of the AgroNet model of predicting the price of soybeans is shown in Figure 15. The plots of the ROC curves give the impression of the capability of the model to differentiate between positive and negative classes. As revealed, the curve is steep at the upper-left corner, which is a high sensitivity (true positive rate) with a relatively low number of false positives. This Area Under the Curve (AUC) of 0.96 proves the high level of discriminatory power of AgroNet. An AUC of nearly 1.0 indicates a good classification performance, showing that AgroNet is always able to distinguish between a correct and an incorrect prediction with a high level of reliability. This higher ROC performance is consistent with the accuracy and loss curve performance, which once again confirms AgroNet to work as the best model over LSTM and GRU because of the nonlinear, intricate behavior of soybean price prediction.

The LSTM model exhibited a steady convergence pattern as the training and validation loss reduced with the number of epochs (Figure 16(a)). Nonetheless, the validation accuracy leveled off at 87.88 %, which means that the model could be used to capture temporal dependencies, but its ability to generalize was limited to some extent. GRU model also converged more quickly than LSTM, and slightly better validation accuracy was found in the range of 88.89% as indicated in Figure 16(b). Its training and validation loss were in close relation, which is indicative of better generalization and fewer chances of overfitting. This implies that GRU is more effective at processing chronology patterns of

Figure 14
(a) Actual vs predicted prices using XGBoost. (b) Actual vs predicted prices using random forest

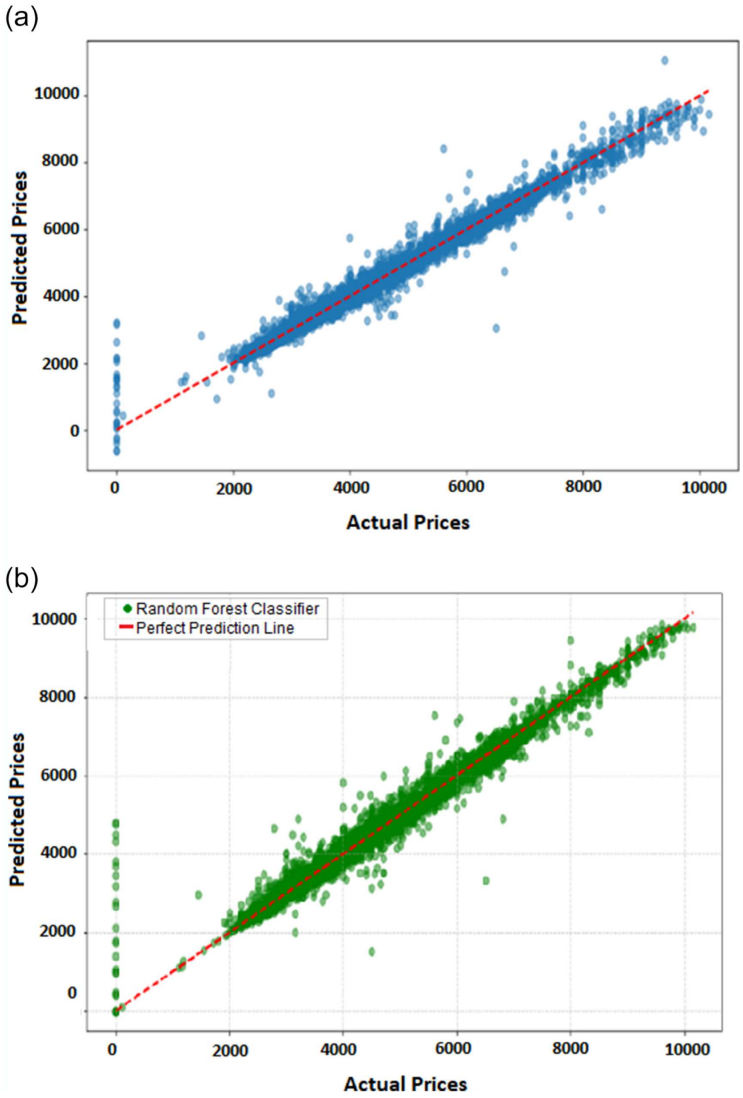


Figure 15
AgroNET model ROC AUC curve

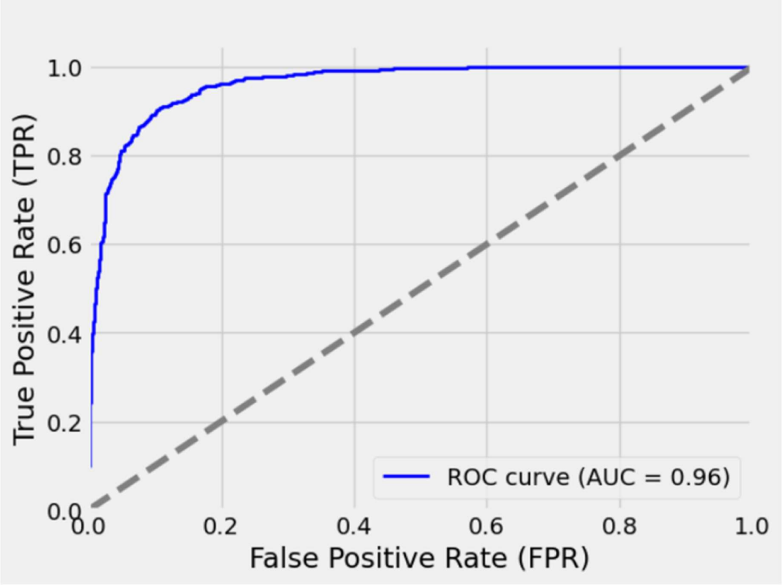
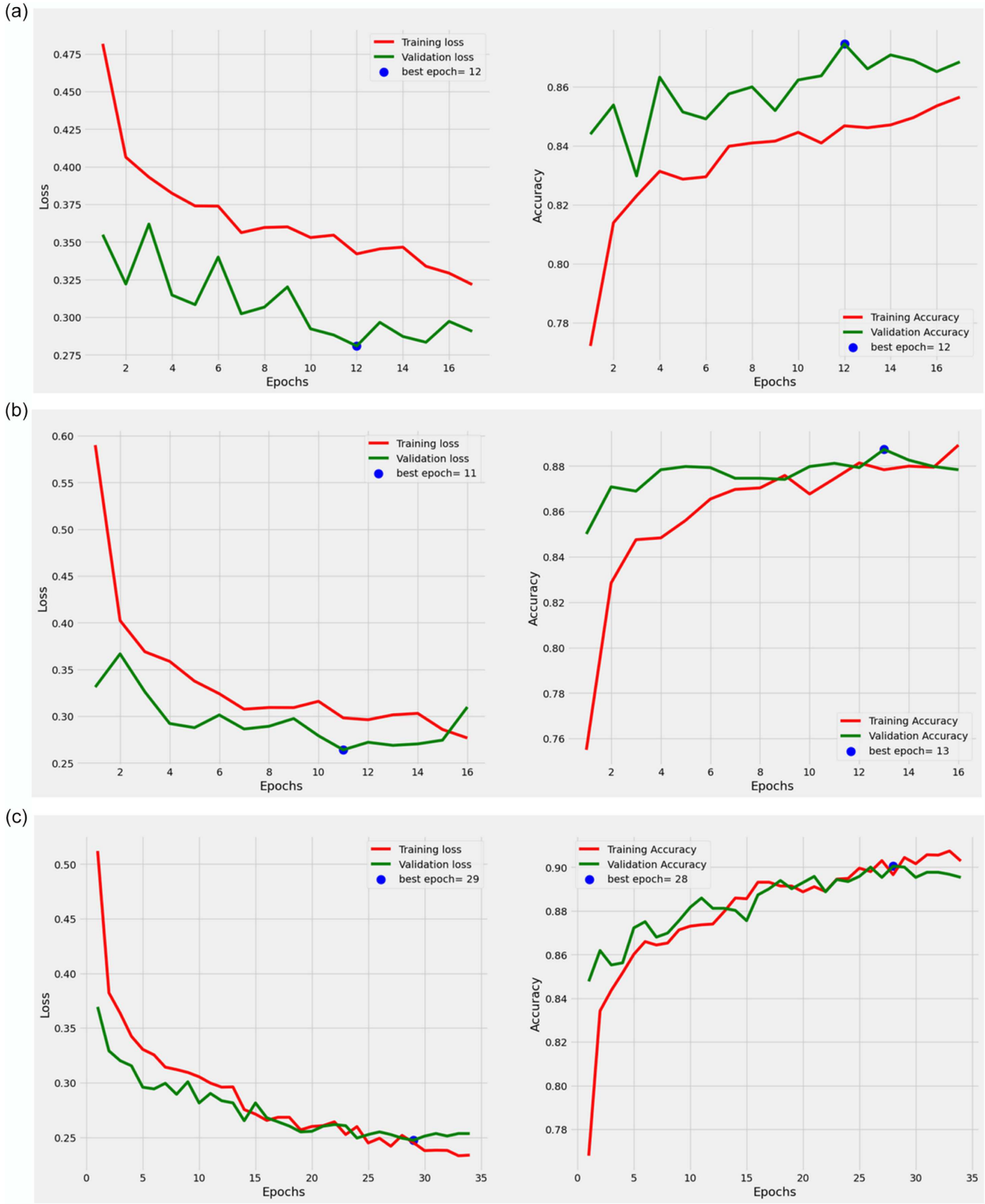


Figure 16

(a) LSTM model accuracy and loss curve. (b) GRU model accuracy and loss curve. (c) AgroNET model accuracy and loss curve



the soybean price data compared to LSTM. On the other hand, the AgroNET model has evidently performed better compared to LSTM and GRU, as evidenced in Figure 16(c). It had stable loss curves with minimal variation, and the validation loss was the same as the training loss, which indicated its high stability. Most importantly, AgroNET obtained the best validation accuracy of about 90.91%, which is more than the other two models. It shows that AgroNET architecture is especially suited to the modeling of nonlinear and dynamic patterns of agricultural price time-series data.

5.4.3. Comparative analysis of models

ARIMA models comparison: Table 9 summarizes the performance of all ARIMA model variants. ARIMA (2,1,2) outperformed alternative configurations by achieving the lowest AIC and BIC values, indicating superior goodness-of-fit with minimal complexity. Unlike ARIMA (1,2,1), which suffered from over-differencing, the selected model preserved temporal information while effectively capturing short-term dependencies. Additionally, ARIMA (2,1,2) produced white-noise residuals and lower RMSE and MAPE values, confirming its robustness and suitability for soybean price forecasting.

1) Models comparison: Table 10 summarizes the performance metrics of all the models under both standard evaluation and k-fold cross-validation. This table compares the five modeling methods based on prediction error, where the performance comparison of the models reveals that the traditional ML methods including RF and XGB gave reasonable predictive accuracy; however, they were worse off than the DL models in both single-split and k-fold evaluations. GRU was the most accurate with fewer errors than LSTM among the sequential models, which demonstrates its ability to learn temporal dependencies consistently across folds. The AgroNET (hybrid LSTM-GRU) model was the most successful in general, and it achieved the highest R^2 and the lowest error values on average over k-fold validation, which proves its superiority in predicting the price of soybean. Figure 17 shows the comparative results of various models in bar graph form for RMSE, MSE, R^2, and MAPE using k-fold mean values. As evident in the results, unlike in the cases with RF and XGBoost, where the performance of the two models still stays at the base level, sequential models, that is, LSTM and GRU, show significant increases in predictive accuracy even under a stricter k-fold validation. Despite the slight reduction in model performance during a k-fold cross-validation, the AgroNET (Hybrid LSTM-GRU) model is a consistently robust model and is better than any other method.

From the results, we can make several observations:

The RF model on the test data presented a somewhat an RMSE of around 120 Rs/quintal and R^2 of about 0.85 with a corresponding MAPE of about 11.8%. This indicates that the combination of the decision trees explained almost 85% of the variation in the prices of soybeans, which is fairly satisfactory as a baseline model. The average error of some 120 also means that RF predictions missed the actual price by about 120 Rs on average, which is approximately 3% on average over the normal Price Range (PR) range (some 4000 Rs/quintal). When using the k-fold cross-validation, the RMSE rose only slightly to 125 ± 8, and MAPE to 12.4 ± 1.1%, which suggests that it is not very sensitive to data splits. One of the advantages of the RF model was the use of lagged price information as well as seasonal indicators, which clarified the repetitive seasonal patterns and production factors in 2024, which affected the prices of soybeans.

XGBoost outperformed RF, and the RMSE values were approximately 105 Rs/quintal, R^2 was about 0.87, and the MAPE was lower (approximately 10.3). This increase indicates that the gradient boosting model has better reflected underlying relationships, which might have been caused by the fact that it can model complex nonlinear interactions. XGBoost still showed a consistent performance with RMSE of 112 ± 7 and MAPE of 10.9 ± 0.9% under k-fold validation, indicating that it generalized better than RF. XGBoost also proved to be more useful in

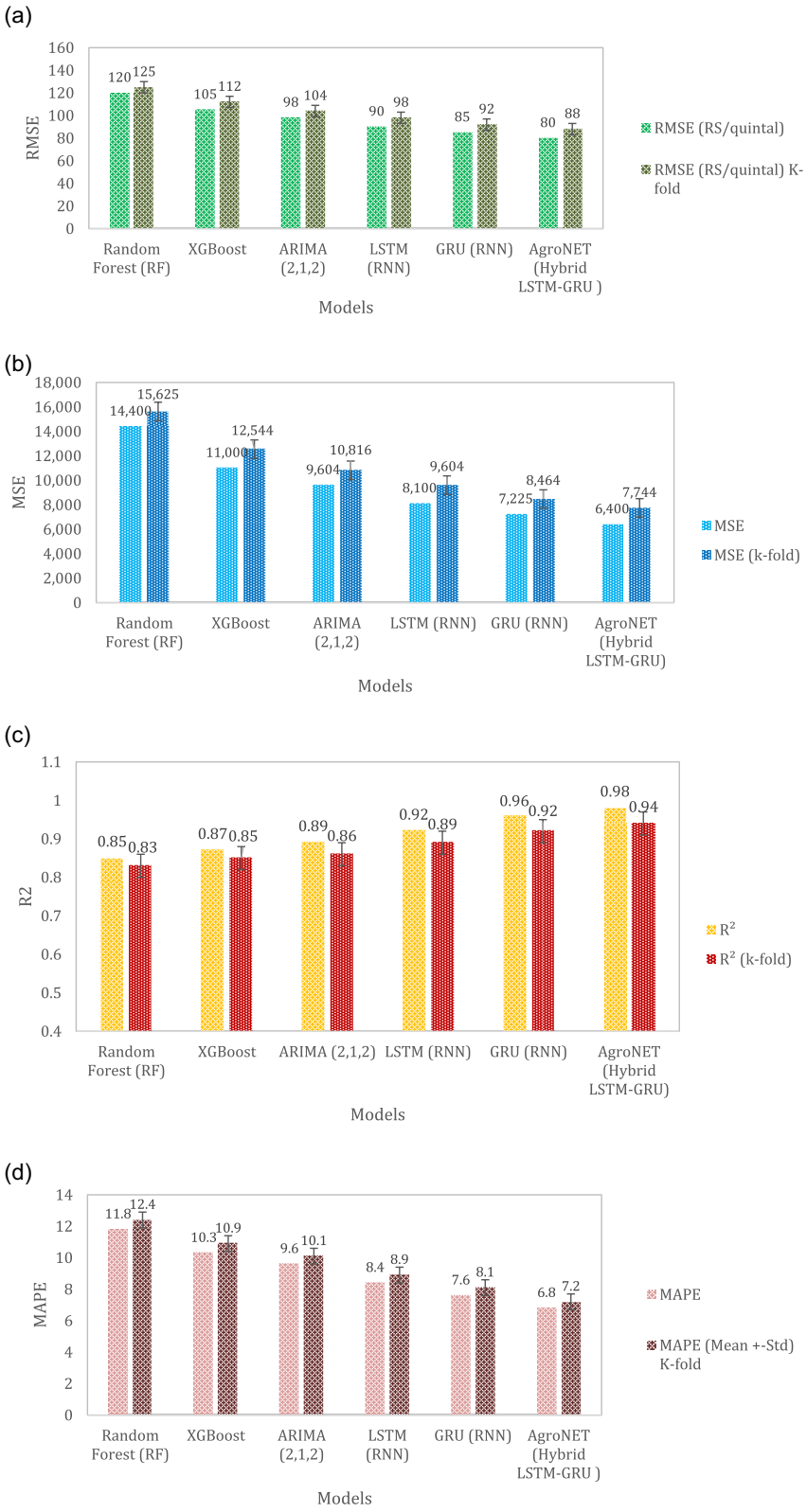
Table 9
Performance comparison of the ARIMA model

Model	p	d	q	AIC ↓	BIC ↓	RMSE (R/quintal) ↓	MAPE (%) ↓	Residual diagnostics
ARIMA (1,1,1)	1	1	1	1426.8	1442.3	102	10.2	Minor autocorrelation
ARIMA (1,2,1)	1	2	1	1441.5	1456.7	108	11.1	Over-differenced
ARIMA (2,1,2)	2	1	2	1409.2	1429.6	98	9.6	White-noise residuals

Table 10
Performance parameters comparison table

Model	RMSE (RS/quintal)	RMSE (RS/quintal) k-fold	MSE (RS/quintal)^2	MSE (Rs/quintal)^2 k-fold	R^2	R^2 (mean ± std) k-fold	MAPE	MAPE (mean ± std) – k-fold
Random Forest (RF)	120	125 ± 8	14,400	15,625 ± 1,020	0.85	0.83 ± 0.03	11.8	12.4 ± 1.1
XGBoost	105	112 ± 7	11,000	12,544 ± 890	0.87	0.85 ± 0.02	10.3	10.9 ± 0.9
ARIMA (2,1,2)	98	104 ± 6	9,604	10,816 ± 820	0.89	0.86 ± 0.03	9.6	10.1 ± 0.8
LSTM	90	98 ± 6	8,100	9,604 ± 760	0.92	0.89 ± 0.02	8.4	8.9 ± 0.7
GRU	85	92 ± 5	7,225	8,464 ± 640	0.96	0.92 ± 0.01	7.6	8.1 ± 0.6
AgroNET (Hybrid LSTM-GRU)	80	88 ± 4	6,400	7,744 ± 520	0.98	0.94 ± 0.02	6.8	7.2 ± 0.5

Figure 17
(a) RMSE (RS/quintal) comparison of models. (b) MSE (RS/quintal)^2 comparison of models. (c) R^2 comparison of models. (d) MAPE comparison of models



terms of extreme price changes, explicitly following mini-spikes and dips that are experienced in 2024. The analysis of feature importance showed that the price of the previous day was the most significant predictor, and the second was the seasonal (month-of-year) factors and annual factors like production and imports.

The ARIMA (2,1,2) model further reduced the prediction error, achieving an RMSE of approximately 98 Rs/quintal, $R^2 \approx 0.89$, and MAPE of about 9.6%, demonstrating the effectiveness of classical time-series modeling in capturing linear temporal dependencies. The model performed consistently under k-fold cross-validation, with RMSE of 104 ± 6 and MAPE of $10.1 \pm 0.8\%$, indicating robust and stable forecasting behavior. ARIMA benefited from explicitly modeling trend and short-term autocorrelations, making it more accurate than tree-based models, although it remained limited in capturing nonlinear dynamics compared to DL approaches.

The LSTM model further minimized the error, achieving RMSE values of approximately 90 Rs/quintal, $R^2 \approx 0.92$, and MAPE around 8.4%, confirming that long-term temporal dependencies were effectively learned. Under k-fold validation, the RMSE increased modestly to 98 ± 6, with MAPE of $8.9 \pm 0.7\%$, reflecting good generalization. The LSTM demonstrated strong capability in predicting next-day prices using sequences of the preceding 30 days. In particular, LSTM was more successful in capturing consistent upward or downward trends, where tree-based models occasionally under-predicted due to the absence of internal temporal memory.

The GRU model performed similarly to LSTM, with a slight advantage in terms of error reduction (RMSE $\approx$ 85, MAPE $\approx$ 7.6%, $R^2 \approx 0.96$). Under k-fold cross-validation, GRU maintained strong performance with RMSE of 92 ± 5 and MAPE of $8.1 \pm 0.6\%$. Both LSTM and GRU captured nearly 96–98% of price variability, with GRU offering faster convergence and reduced computational complexity. This supports the established understanding that GRUs can achieve comparable predictive performance to LSTMs while requiring fewer parameters and training epochs.

The AgroNET (Hybrid LSTM-GRU) model achieved the best overall performance, with an RMSE of approximately 80 Rs/quintal, $R^2 \approx 0.98$, and the lowest MAPE of about 6.8%, indicating extremely high prediction fidelity. Even under k-fold cross-validation, AgroNET remained robust, with RMSE of 88 ± 4, $R^2 = 0.94 \pm 0.02$, and MAPE of $7.2 \pm 0.5\%$, demonstrating consistent generalization under stricter validation. This may have worked because the hybrid architecture incorporated complementary temporal representations, which were learned by the LSTM and GRU layers. Even though the relative performance relative to single LSTM/GRU models is moderate (about a 10% decrease in RMSE between 90 and 80), it at least shows that hybrid model design is beneficial in terms of improvement in the accuracy of price prediction of soybean.

XAI visualization (LIME): The explainable XAI model of price prediction is conducted with the help of the LIME model that gives clear, instance-based, explanations of the ensemble predictor based on identifying the major factors that affect the individual price predictions. The local black-box approximation of the black-box model as a surrogate, followed by the need to reveal the direction and strength of feature contributions, offers much greater insight into the importance of various historical price indicators in the market and the sustaining effect of these other supply, trade, and climatic variables, which makes the black-box model more transparent and more useful to farmers.

Figure 18 gives a local explanation (based on LIME) of the individual price of a soybean, and this shows the value of various input features in influencing the modal price prediction of the price of soybean either positively or negatively. The presence of features (green) or the absence of them (red) raises or lowers the price predicted, and the length of the bar indicates the strength of the effect. As can be seen in the visualization, Max Price (Rs./Quintal) and Min Price (Rs./Quintal) have a dominant influence on the prediction, implying that they greatly rely on the recent market price limits, whereas the Arrivals (Tons), Market Name, and Variety variables have a moderating effect. Considerations of bigger supply and agro-economic related factors including the imports/exports of soybean, area under cultivation, and actual rainfall have less, though not negligible, influence as they are indirectly related to the price dynamics.

Figure 18
LIME-based local feature contribution for soybean price prediction

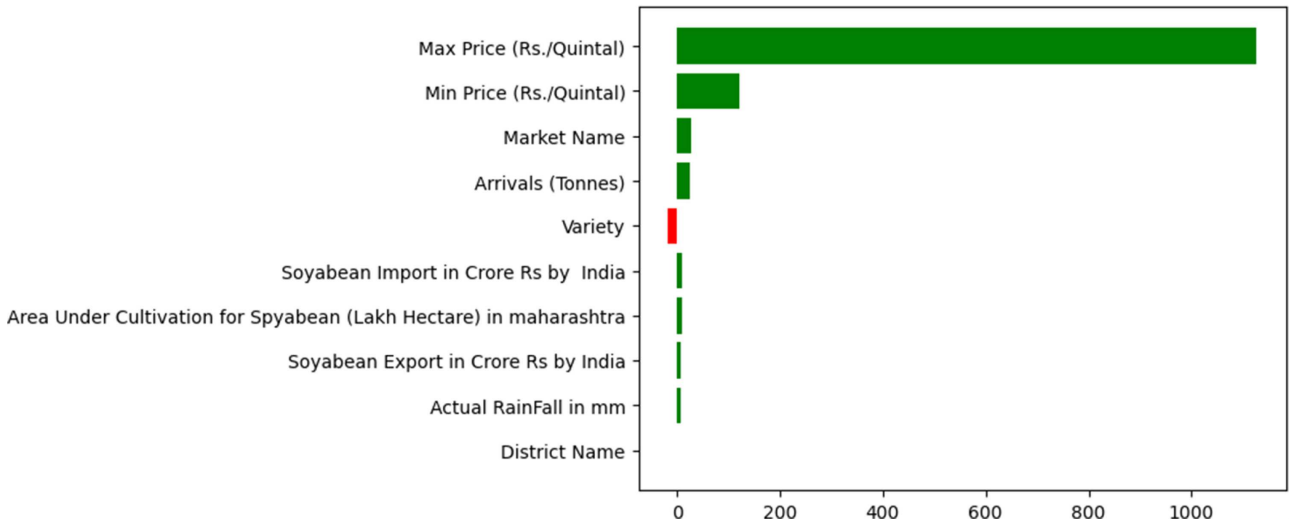
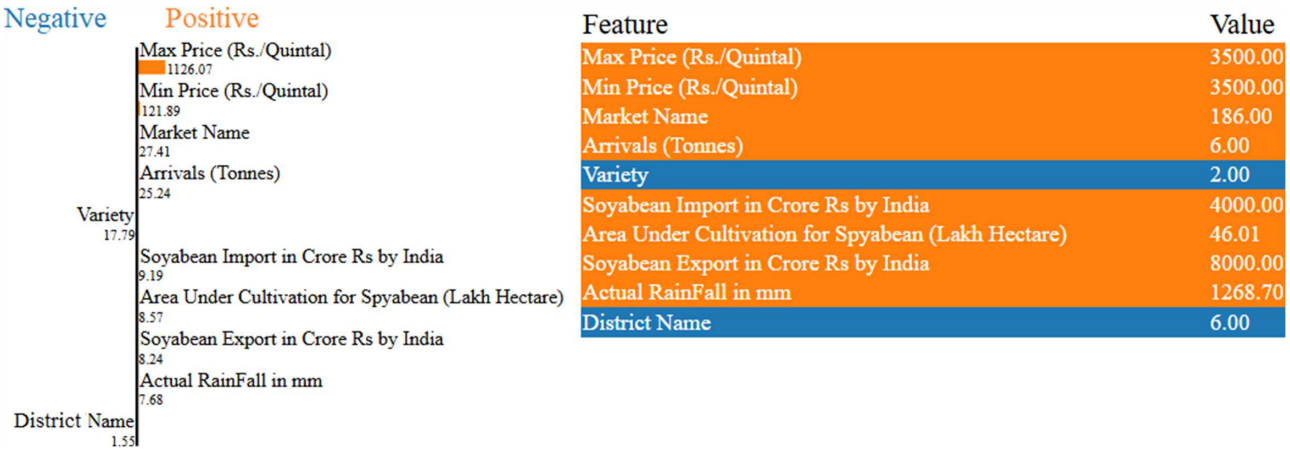


Figure 19
LIME feature importance and corresponding feature values for a sample instance



As far as interpretability is concerned, Figure 19 confirms the fact that the variables based on price and market level have the greatest influence on the model's decision-making, then there are the supply-side ones, and lastly, there are variables that rely on the locations (e.g., district), which in this case play a minor role. When using LIME, transparency and instance-level understanding can be provided by approximating the complex ensemble model locally with an interpretable surrogate, which provides a translation of the black-box prediction into understandable signals. The stakeholder needs are directly addressed through this XAI analysis, as the factors contributing to price changes and the direction are clearly identified, therefore increasing the level of trust and practical use for the farmers.

5.5. Key contributions of the study

1) Demonstrating a successful integration of disparate data (market, weather, macro indicators) for price forecasting, providing a thorough comparison between different modeling paradigms on the same problem, and introducing an AgroNET, a hybrid DL architecture (LSTM-GRU) in the context of commodity price prediction. The high accuracy achieved suggests that the approach could be employed in operational forecasting systems.
2) Enhancing model transparency through XAI (LIME)-based explanations, enabling clear interpretation of key factors driving price predictions and increasing trust and usability for farmers and agri-businesses.

Given that the model was trained on historical data, deploying it would involve continuously updating with new incoming data (e.g., as 2025 data comes in) and retraining or using the model to predict short-term future prices. However, we also recognize the limitations of our study. The model's performance, while strong for the period tested, is dependent on the patterns present in that period. Structural changes could reduce the model's effectiveness if not accounted for by new data. Also, our model predicts one day ahead; while this is useful for near-term decisions, many agricultural decision-makers need medium to long-term forecasts. Extending the horizon of prediction is a challenge that may require different architectures. Additionally, the data could be enhanced—for instance, incorporating global price indices as an input might improve predictive power since global and local prices are linked.

6. Conclusion

This study has offered a price forecast of one of such crops, that is, the soybean crop price in the Indian market, by conducting a thorough analysis of soybean crop price prediction in India, based on an integrated dataset (2015–2025) and using various predictive analysis methods. To provide the models with an abundant context, we gathered and combined data from several sources each day, including market prices and Agmarknet arrivals, weather data (rainfall) from IMD, and annual production and trade statistics. The main patterns that were discovered through the exploratory analysis include the influence of seasonality, a significant price increase in 2021, the price series correction afterward, and the outlier events in the price series. Six predictive models were designed, including RF, XGBoost, ARIMA (2,1,2), LSTM, GRU, and AgroNET (Hybrid LSTM-GRU network), and all models were evaluated using k-fold cross-validation. We found that the time-dependent patterns of the data could be better represented in DL models compared to traditional ML and statistical models and, as a result, yielded a higher level of accuracy. Among the non-neural approaches, ARIMA and XGBoost achieved reasonable predictive performance under k-fold validation, with RMSE values of approximately 104 ± 6 and 112 ± 7 Rs/quintal, respectively, and corresponding MAPE values close to 10%, but their accuracy was lower than that of the DL models.

AgroNET model performed best overall, with a k-fold RMSE of about 88 ± 4 Rs/quintal, $R^2 \approx 0.94 \pm 0.02$, and the lowest MAPE of approximately $7.2 \pm 0.5\%$, indicating strong predictive accuracy and robustness across validation folds. The best non-neural approach exhibited a slightly larger error (~112 Rs/quintal RMSE) and lower R^2 (~0.85) under k-fold validation, namely, the XGBoost model. The exogenous traits like rainfalls, arrivals, and production also aided in the enhancement of the model performance in that the models were able to change their forecasts within the framework of the broader supply–demand dynamics, which is in line with the existing literature on agricultural price forecasting. Also, to deal with model transparency, the XAI analysis with the help of the LIME technique was included in order to analyze the predictions of the proposed models. The models based on the LIME obviously determine the relevant factors of price variations of soybeans, making the forecasting system more interpretable, credible, and applicable to farmers and agri-businesses.

For future work, we suggest exploring the following:

1) Multi-step forecasting: Develop models to predict weekly or monthly average prices, which might involve using the daily model iteratively or training separate models with aggregated data.
2) Advanced DL models: Newer architectures like the Transformer or N-BEATS have shown excellent performance in time-series tasks. Testing these on soybean prices, especially with exogenous inputs, could further improve accuracy.
3) Wider generalization: This study is currently limited to soybean price prediction in India using a crop-specific dataset. Future research can extend the framework to other regions and crops by incorporating region- and crop-dependent features, as well as exploring transfer learning and multi-crop modeling to improve generalizability.

Future research should move beyond historical data–driven modeling by explicitly incorporating structural risk factors such as policy changes, global supply chain disruptions, pandemics, and extreme climate anomalies into agricultural forecasting frameworks. Integrating scenario-based simulations and policy-aware variables and incorporating more granular weather variables, socio-economic indicators, and climate stress indicators with ML models can improve robustness and real-world reliability.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data are available from the corresponding author upon reasonable request.

Author Contribution Statement

Vilas Damodhar Ghonge: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Yogesh Kulkarni:** Conceptualization, Methodology, Investigation, Resources, Supervision.

References

[1] Chawla, A., & Singh, K. (2024). *A summer of discontent for India's soybean farmers*. Centre for Economic Data & Analysis (CEDA) https://ceda.ashoka.edu.in/a-summer-of-discontent-for-indias-soybean-farmers/

[2] Mello, F. D. D. C., Kumar, P., & Sperandio Nascimento, E. G. (2024). Advancements in soybean price forecasting: Impact of AI and critical research gaps in global markets. *Economies*, *12*(11), 310. https://doi.org/10.3390/economies12110310

[3] Nayak, G. H., Alam, M. W., Singh, K. N., Avinash, G., Kumar, R. R., Ray, M., & Deb, C. K. (2024). Exogenous variable driven deep learning models for improved price forecasting of TOP crops in India. *Scientific Reports*, *14*(1), 17203. https://doi.org/10.1038/s41598-024-68040-3

[4] Rana, H., Farooq, M. U., Kazi, A. K., Baig, M. A., & Akhtar, M. A. (2024). Prediction of agricultural commodity prices using big data framework. *Engineering, Technology & Applied Science Research*, *14*(1), 12652–12658. https://doi.org/10.48084/etasr.6468

[5] Liu, X., Zhou, W., Gao, Z., Zhang, D., & Ma, K. (2025). The prediction of soybean price in China based on a mixed data sampling–support vector regression model. *Mathematics*, *13*(11), 1759. https://doi.org/10.3390/math13111759

[6] Torsoni, G. B., de Oliveira Aparecido, L. E., dos Santos, G. M., Chiquitto, A. G., da Silva Cabral Moraes, J. R., & de Souza Rolim, G. (2023). Soybean yield prediction by machine learning and climate. *Theoretical and Applied Climatology*, *151*(3), 1709–1725. https://doi.org/10.1007/s00704-022-04341-9

[7] Sharma, R. K., Kaur, J., Feng, G., Huang, Y., Kumar, C., Wang, Y., . . . , & Dhillon, J. (2025). Maize and soybean yield prediction using machine learning methods: A systematic literature review. *Discover Agriculture*, *3*(1), 64. https://doi.org/10.1007/s44279-025-00215-6

[8] Liu, J., Zhang, B., Zhang, T., & Wang, J. (2023). Soybean futures price prediction model based on EEMD-NAGU. *IEEE Access*, *11*, 99328–99338. https://doi.org/10.1109/ACCESS.2023.3314329

[9] Srichaiyan, P., Tippayawong, K. Y., & Boonprasope, A. (2025). Forecasting soybean futures prices with adaptive AI Models. *IEEE Access*, *13*, 48239–48256. https://doi.org/10.1109/ACCESS.2025.3546786

[10] Santos, L. B., Gentry, D., Tryforos, A., Fultz, L., Beasley, J., & Gentimis, T. (2024). Soybean yield prediction using machine learning algorithms under a cover crop management system. *Smart Agricultural Technology*, *8*, 100442. https://doi.org/10.1016/j.atech.2024.100442

[11] Sowmya, B. J., Meeradevi, A. K., Supreeth, S., Pradeep Kumar, D., Ravi Kumar, B. N., Rohith, S., . . . , & Patil, A. U. (2025). Leveraging machine learning for intelligent agriculture. *Discover Internet of Things*, *5*(1), 33. https://doi.org/10.1007/s43926-025-00132-6

[12] Ahmed, A. I. (2025). A machine learning approach to short-term price forecasting of agricultural crops in India. In *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies*, 325–331. https://doi.org/10.1109/CE2CT64011.2025.10939904

[13] Singh, A., & Janu, N. (2024). Machine learning techniques for forecasting prices and agricultural yields. In *2024 2nd International Conference on Artificial Intelligence Trends and Pattern Recognition,1-5*. https://doi.org/10.1109/ICAITPR63242.2024.10960036

[14] Ghildiyal, S., Deogaonkar, A., Bhandari, N. S., Bisht, M., Vichoray, C., & Naval, N. (2024). Development of convolutional neural network model for crop yield prediction. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things, 1130-1135*. https://doi.org/10.1109/ICoICI62503.2024.10696315

[15] Agarwal, J., Vaswani, S., Sharma, A., Kaushik, D., & Bhardwaj, D. (2023). Optimization of crop yield using machine learning. In *2023 3rd International Conference on Technological Advancements in Computational Sciences*, 469–474. https://doi.org/10.1109/ICTACS59847.2023.10389927

[16] Prity, F. S., Hasan, M. M., Saif, S. H., Hossain, M. M., Bhuiyan, S. H., Islam, M. A., & Lavlu, M. T. H. (2024).

Enhancing agricultural productivity: A machine learning approach to crop recommendations. *Human-Centric Intelligent Systems*, *4*(4), 497–510. https://doi.org/10.1007/s44230-024-00081-3

[17] Mithuna, R., & Terence, J. S. (2023). Crop price prediction for coconut in Tamil Nadu using machine learning. In *2023 International Conference on Circuit Power and Computing Technologies*, 605–609. https://doi.org/10.1109/ICCPCT58313.2023.10245891

[18] Celik, B. A., & Celik, S. (2025). Hybrid forecasting of agricultural commodity prices: Integrating machine learning, time series, and stochastic simulation models. *Borsa Istanbul Review*, *25*(6), 1440–1462. https://doi.org/10.1016/j.bir.2025.10.004

[19] Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied Artificial Intelligence*, *36*(1), 2031822. https://doi.org/10.1080/08839514.2022.2031823

[20] Banerjee, S., & Mondal, A. C. (2023). An ingenious method for estimating future crop prices that emphasises machine learning and deep learning models. *International Journal of Information Technology*, *15*(8), 4291–4313. https://doi.org/10.1007/s41870-023-01474-6

[21] Zhang, H., Xiao, F., He, W., Chai, Z., & Ewe, H.-T. (2022). Multiyear automated mapping and price analysis of garlic in main planting areas of China using time-series remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 5222–5233. https://doi.org/10.1109/JSTARS.2022.3186298

[22] Paul, R. K., Yeasin, M., Kumar, P., Kumar, P., Balasubramanian, M., Roy, H. S., . . . , & Gupta, A. (2022). Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. *PLOS One*, *17*(7), e0270553. https://doi.org/10.1371/journal.pone.0270553

[23] Krishna, P. A., Narayana, G. V., Kotha, S. K., & Pattnayak, D. (2025). Machine learning based agricultural price forecasting for major food crops in India using environmental and economic factors. *Biology and Life Sciences Forum*, *54*(1), 7. https://doi.org/10.3390/blsf2025054007

[24] Mohanty, S., Nanjundan, P., & Kar, T (Eds.). (2024). *Artificial intelligence in forecasting: Tools and techniques* (1st ed.). USA: CRC Press. https://doi.org/10.1201/9781003399292

[25] Chaitra, B., & Meena, K. (2023). Forecasting crop price using various approaches of machine learning. In *2023 International Conference on Innovations in Engineering and Technology*, 1–5. https://doi.org/10.1109/ICIET57285.2023.10220616

[26] Bhardwaj, M. R., Pawar, J., Bhat, A., Deepanshu, Enaganti., Sagar, I., K, & Narahari, Y. (2023). An innovative deep learning based approach for accurate agricultural crop price prediction. In *2023 IEEE 19th International Conference on Automation Science and Engineering*, 1–7. https://doi.org/10.1109/CASE56687.2023.10260494

[27] Purohit, S. K., Panigrahi, S., Sethy, P. K., & Behera, S. K. (2021). Time series forecasting of price of agricultural products using hybrid methods. *Applied Artificial Intelligence*, *35*(15), 1388–1406. https://doi.org/10.1080/08839514.2021.1981659

[28] Shao, Y. E., & Dai, J.-T. (2018). Integrated feature selection of ARIMA with computational intelligence approaches for food crop price prediction. *Complexity*, *2018*(1), 1910520. https://doi.org/10.1155/2018/1910520

[29] He, M., Jiang, W., & Gu, W. (2024). TriChronoNet: Advancing electricity price prediction with multi-module fusion. *Applied Energy*, *371*, 123626. https://doi.org/10.1016/j.apenergy.2024.123626

[30] Lu, Y., Wang, W., Bai, R., Zhou, S., Garg, L., Bashir, A. K., . . . , & Hu, X. (2025). Hyper-relational interaction modeling in multi-modal trajectory prediction for intelligent connected vehicles in smart cities. *Information Fusion*, *114*, 102682. https://doi.org/10.1016/j.inffus.2024.102682

[31] Thaker, A., Chan, L. H., & Sonner, D. (2024). Forecasting agriculture commodity futures prices with convolutional neural networks with application to wheat futures. *Journal of Risk and Financial Management*, *17*(4), 143. https://doi.org/10.3390/jrfm17040143

[32] Yewle, A. D., Mirzayeva, L., & Karakuş, O. (2025). Multimodal data fusion and deep ensemble learning for accurate crop yield prediction. *Remote Sensing Applications: Society and Environment*, *38*, 101613. https://doi.org/10.1016/j.rsase.2025.101613

[33] Patel, G., Das, R., Das, S., Mukherjee, I., & Mausam, I. M. D. (2025). Innovative trend analysis of long-term rainfall variation over West Bengal, India. *Mausam*, *76*(2), 373–386. https://doi.org/10.54302/mausam.v76i2.6136

How to Cite: Ghonge, V. D., & Kulkarni, Y. (2026). Multi-Source Data Fusion and Machine Learning for Soybean Crop Price Forecasting in India. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62027619