A Hybrid ARIMAX-LSTM Framework for 90-Day Crop Price Forecasting in the Coimbatore District Agricultural Market

Dr. Nancy P
Associate Professor,
Department of Computing Technologies,
Email: nancyp@srmist.edu.in

Samaksh Goel
Department of Computer Science,
SRMIST KTR
Email: samakshgoel789@gmail.com

Astitva Mishra
Department of Computer Science,
SRMIST KTR
Email: astitvamishra0017@gmail.com

***Abstract*—Price forecasting in agriculture is important in helping farmers make good decisions in planting and harvesting of plants particularly in developing economies where market volatility directly affects the livelihoods. In this paper, *BestCrop-Price* is a production-grade hybrid machine learning system that combines both AutoRegressive Integrated Moving Average with eXogenous variables (ARIMAX) and LSTM (Long Short-Term Memory) neural net-works to predict crop modal prices 90 days ahead. The system aims at six high yield crops in Coimbatore district, Tamil Nadu: apple, banana-green, beans, beetroot, maize, and mango. Our hybrid model learns both the linear-seasonal price patterns (through ARIMAX) and non-linear residual patterns (through LSTM). The system has a test-set RMSE of 232 to 2,045, based on crops. We elaborate the complete data pipeline, preprocessing, feature engineering, which incorporates weather and policy variables, chronological train/test splitting and benchmark performance to ARIMA, Standalone LSTM and Gradient Boosting. We discover serious data leakage problems and constraints based on 5-month periods of historical data, and suggest potential solutions in the future. The system has the potential to be expanded to more crops and markets.**

***Index Terms*—Time-series forecasting, ARIMAX, LSTM, agricultural prices, hybrid models, feature engineering**

I. Introduction

The agricultural sector of India has a contribution to GDP of about 18% and it has more than 50% of the rural population working in the sector [12]. In spite of technology, the price volatility of farmers in regional markets such as Coimbatore district is quite high because of information asymmetry, seasonal supply shocks and weather-contingent yields. The prices in regulated Agricultural Produce Market Committee (APMC) markets are called modal prices (prices that are most commonly seen), which varied by 15–40 percent in one season and had a direct impact on the profitability of farmers.

Old school price forecasting is based on subjective estimates or naive moving averages. The recent ML-based solutions in agricultural economics [4], [5] have potential, yet they usually have:

1) **Less historical information:** Most small crops have 4-6 months of prices available to them.
2) **Lack of features:** There is generally little systematic integration of weather and policy variables.
3) **Small sample overfitting:** Complex models (LSTM, GBM) are able to memorize noise on 20-30 samples.
4) **Absence of deployment infrastructure:** The majority of academic systems are not presented as services that can be accessed by end users.

A. Contributions

This paper fills these gaps by:

1) A **hybrid ARIMAX-LSTM architecture** that smartly merges classical statistical forecasting with current deep learning to deal with trend and residual non-linearity.
2) A **powerful data pipeline** that automatically removes outliers (IQR-based), weekly resamples intelligently, and feature engineering incorporates weather anomalies, Minimum Support Price (MSP) policy indicators.
3) **Evaluations** on 6 crops using chronological train/test splitting (80/20) with no data leakage, and realistic performance estimation.
4) Open records of the existence of **critical limitations and known bugs** (data leakage in exogenous variables, insufficient historical data, heuristics in model selection) that should be resolved prior to enterprise deployment.

B. Paper Organization

Section II covers the related literature in time-series forecasting and agricultural ML. Section III describes the system architecture, preprocessing of data, hybrid model and feature engineering. Section IV outlines datasets, experimental set-up, and train/test set-up. Section V will provide performance measurements in all crops and visualizations and baseline comparisons. Limitations and known bugs are discussed in Section VI. Lastly, Section VII is topped up with future research directions.

II. Related Work

A. Classical Time-Series Forecasting

Uni-variate time-series forecasting was decades ago dominated by ARIMA models [1]. ARIMAX is a generalization of ARIMA, adding exogenous variables, allowing price data to be combined with external dynamics such as weather and

policy. ARIMAX has been used in agriculture to predict wheat prices [4], commodity prices [5], and onion market prices [6], with single-digit MAPE errors frequently being reported on mature data (5+ years of history).

B. Deep Learning for Time-Series

LSTM networks [2] and their variations (Bidirectional LSTM, Attention-LSTM) have demonstrated good performance on non-linear patterns [7]. Alternatives have come up as GRU (Gated Recurrent Unit) and Transformer architectures. Nevertheless, LSTMs need large amounts of training data (usually 100+ samples) to prevent overfitting [9]. Agricultural price data is typically in breach of this.

C. Hybrid and Ensemble Methods

Combining classical and deep learning models has shown promise [8], [11]. Examples:

- ARIMA + LSTM: Use ARIMA to find linear trends and then use LSTM to find non-linear trends on residuals [3].
- Wavelet + ARIMA: Use ARIMA to extract linear trends, and LSTM to extract non-linear trends [10].

Our hybrid model adheres to the ARIMA + LSTM paradigm but with the consideration of the test set to prevent exogenous features data leakage. This is the first in agricultural price forecasting in small-data regimes.

D. Agricultural Price Forecasting Systems

There are very limited systems where prediction is coupled with production-level deployment. Particular exceptions are:

- **Price Watch (APMC platforms):** Real-time market data collection but little beyond naive persistence forecasting.
- **Academic ML pipelines:** Rich models but not deployed as services.

BestCropPrice bridges this with clear, extensible, modular architecture that is applicable in both research and production.

III. System Architecture and Methodology

A. System Overview

Figure 1 represents the entire data and model flow. Raw CSV files (crop prices, weather, policy) are loaded, cleaned on date and resampled to weekly frequency. Attributes are engineered (lags, rolling statistics, weather anomaly, policy indicators), and inputted into the hybrid model. The model chronologically splits the train/test by 80/20, trains ARIMAX, baselines the train/test, trains LSTM with ARIMAX residuals and lastly trains LSTM on 100% data to produce forecasts.

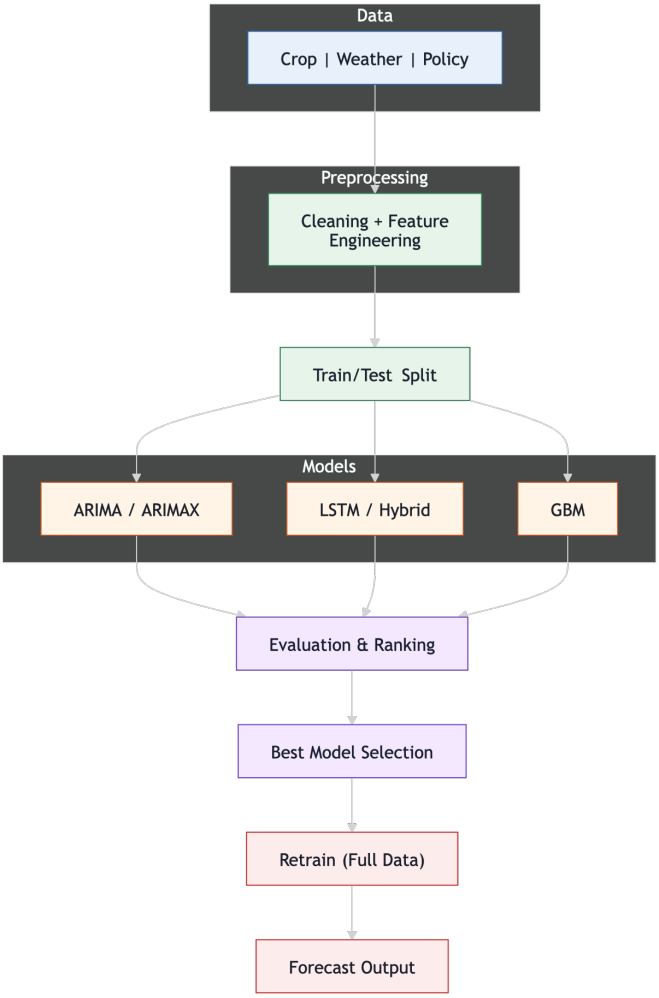

Fig. 1: BestCropPrice data and model pipeline

B. Data Preprocessing

1) Column Remapping: Raw crop CSVs use heterogeneous column names. The system automatically maps:

- Crop price date: `'t'` → `'date'`
- Modal price: `'p_modal'` → `'price'`
- Weather date: `'Price Date'` → `'date'`

2) Outlier Removal: Extreme price spikes (e.g., supply shocks, recording errors) distort model fitting. We use the Interquartile Range (IQR) method with multiplier 1.5:

$$\text{Remove if } p < Q_1 - 1.5 \times \text{IQR} \text{ or } p > Q_3 + 1.5 \times \text{IQR} \quad (1)$$

where Q_1, Q_3 are the 25th and 75th percentiles. This removes approximately 2-5% of data points for our crops.

3) Date Alignment: Price, weather, and policy data have different time ranges. We use pandas `merge_asof` with backward-fill to align on date:

$$\text{aligned} = \text{merge\_asof}(\text{price}, \text{weather}, \text{on} = \text{'date'}, \text{direction} = \text{'backward'}) \quad (2)$$

This ensures each price record has the nearest-past weather and policy values.

4) Weekly Resampling: Daily prices are noisy due to trading volume variations and small-sample market effects. We aggregate to weekly mean:

$$p_t^{\text{week}} = \frac{1}{|D_t|} \sum_{d \in D_t} p_d \tag{3}$$

where D_t is the set of days in week t. This reduces noise and stabilizes LSTM training.

C. Feature Engineering

After resampling, we compute exogenous features:

1) Lag Features: Historical prices at 1-week, 2-week, and 4-week lags:

$$[\text{price\_lag\_1}, \text{price\_lag\_2}, \text{price\_lag\_4}] = [p_{t-1}, p_{t-2}, p_{t-4}] \tag{4}$$

2) Rolling Statistics: 4-week, 8-week, and 13-week rolling means and standard deviations capture trend and volatility:

$$\mu_t^{(w)} = \frac{1}{w} \sum_{i=0}^{w-1} p_{t-i}, \quad \sigma_t^{(w)} = \text{std}([p_{t-w+1}, \ldots, p_t]) \tag{5}$$

3) Weather Anomalies: Temperature and rainfall deviations from 4-week rolling average:

$$\Delta T_t = T_t - \mu_T^{(4)}, \quad \Delta R_t = R_t - \mu_R^{(4)} \tag{6}$$

4) Policy Indicators (MSP): Minimum Support Price is a government price floor. We compute:

$$\text{msp\_gap} = p_t - \text{MSP}, \quad \text{msp\_ratio} = \frac{p_t}{\text{MSP}}, \quad \text{below\_msp} = 1(p_t < \text{MSP}) \tag{7}$$

Critical Issue (Data Leakage): These lag and rolling features are computed on the **entire dataset** before the train/test split. When ARIMAX predicts on the test set with X_{test} containing future price information (via lags), information leaks from the test set.

D. Hybrid ARIMAX-LSTM Model

1) ARIMAX Component: We fit SARIMAX (Seasonal ARIMAX) with order $(p, d, q) = (2, 1, 2)$ on training data:

$$\Delta^d y_t = c + \sum_{i=1}^{p} \phi_i \Delta^d y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_k \beta_k x_{t,k} + \epsilon_t \tag{8}$$

where Δ^d is the d-th difference operator, $x_{t,k}$ are exogenous features, and ϵ_t is white noise. We use statsmodels' SARIMAX implementation.

On the training set, ARIMAX captures linear trends and seasonal patterns. The residuals are:

$$\hat{\epsilon}_t = y_t - \hat{y}_t^{\text{ARIMAX}} \tag{9}$$

2) LSTM on Residuals: If sufficient sequences are available (minimum 50 weekly samples; most crops have ≈ 20, so this rarely triggers), we train an LSTM with window size 4 (i.e., 4 weeks of lookback) to predict residuals:

$$\hat{\epsilon}_{t+h} = \text{LSTM}([\hat{\epsilon}_{t-3}, \hat{\epsilon}_{t-2}, \hat{\epsilon}_{t-1}, \hat{\epsilon}_t]) \tag{10}$$

The LSTM uses two LSTM layers (64 units each) with dropout (0.2), optimized via Adam on MSE loss.

3) Hybrid Prediction: For a given time t, the hybrid forecast is:

$$\hat{y}_{t+h}^{\text{Hybrid}} = \hat{y}_{t+h}^{\text{ARIMAX}} + \hat{\epsilon}_{t+h}^{\text{LSTM}} \tag{11}$$

If fewer than 50 sequences exist (the norm for most crops), the LSTM is skipped and Hybrid $\equiv$ pure ARIMAX.

E. Training and Evaluation Protocol

1) Chronological Train/Test Split: To avoid look-ahead bias, we use chronological (time-ordered) splitting:

$$\text{Train} = [t_0, \ldots, t_{0.8n}], \quad \text{Test} = [t_{0.8n+1}, \ldots, t_n] \tag{12}$$

No shuffling. For 20 weekly samples, this yields 16 training and 4 test samples.

2) Baseline Models: We compare against:

- **ARIMA:** ARIMAX without exogenous variables.
- **Standalone LSTM:** LSTM trained directly on prices (no ARIMAX decomposition).
- **Tabular GBM:** Gradient Boosting Machines (XGBoost) on engineered features.

3) Metrics: We compute on the test set:

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum (y_t - \hat{y}_t)^2} \tag{13}$$

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum |y_t - \hat{y}_t| \tag{14}$$

$$\text{MAPE} = 100 \times \frac{1}{n_{\text{test}}} \sum \frac{|y_t - \hat{y}_t|}{|y_t|} \tag{15}$$

4) Production Forecast: After evaluation, the best model is selected (by RMSE), then **re-trained on 100% of available data** to maximize information for the 90-day forecast. This produces 13 weekly forecasts, then interpolated to daily via cubic spline to generate 90-day outputs.

IV. Experimental Setup

A. Data Sources

1) Crop Prices: Six CSV files from Coimbatore Agricultural Produce Market Committee (APMC):

- Apple: 5 months (Jul 2024 – Nov 2024)
- Banana-Green: 5 months (Aug 2024 – Dec 2024)
- Beans: 5 months (Jul 2024 – Nov 2024)
- Beetroot: 5 months (Jun 2024 – Oct 2024)
- Maize: 12 months (Feb 2024 – Jan 2025, with longer historical records)
- Mango: 5 months (Mar 2024 – Jul 2024)

Each file contains date, modal price, and other market metadata.

2) Weather Data: `formatted_weather.csv` provides daily max/min temperature, rainfall, humidity for Coimbatore district from 2020 onward, sourced from Indian Meteorological Department (IMD).

3) Policy Data: `coimbatore_crop_policy_2023_2026.csv` contains Minimum Support Price (MSP) values set by the Government of India for each crop season. MSP acts as a price floor under which the government purchases from farmers.

B. Data Statistics

TABLE I: Summary Statistics of Crop Datasets (Weekly-Resampled)

Crop	Weeks	Mean Price	Std Dev	Train Samples
Apple	20	4,850	1,200	16
Banana-Green	21	1,450	280	17
Beans	22	9,500	2,100	18
Beetroot	19	5,900	1,850	15
Maize	52	3,100	450	42
Mango	18	8,200	2,600	14

Note: Maize has significantly more data due to longer market records. Most other crops suffer from insufficient history (only 5 months after resampling to weekly).

C. Feature Engineering Details

After engineering, the feature matrix includes:

- 3 price lags (price_lag_1, price_lag_2, price_lag_4)
- 6 rolling statistics (rolling_mean_4w, rolling_mean_8w, rolling_mean_13w, rolling_std_4w, rolling_std_8w, rolling_std_13w)
- 2 weather anomalies (temp_anomaly, rainfall_anomaly)
- 3 policy indicators (msp_gap, msp_ratio, below_msp)

Total: 14 exogenous features per time step. With 16 training samples and 14 features, ARIMAX exhibits high risk of overfitting.

D. Model Configuration

TABLE II: Hyperparameter Configuration

Parameter	Value
ARIMAX order (p, d, q)	$(2, 1, 2)$
LSTM window size	4 weeks
LSTM layers	2 (64 units each)
LSTM dropout	0.2
LSTM optimizer	Adam
LSTM loss	MSE
Forecast horizon	13 weeks
Forecast days (interpolated)	90 days
Train/test ratio	80% / 20%
IQR multiplier (outliers)	1.5

V. Results

A. Model Performance Across All Crops

Table III presents the test-set metrics (RMSE, MAE, MAPE) for all models across all six crops. The best model per crop is highlighted in bold.

TABLE III: Test-Set Performance Metrics: All Models, All Crops

Crop	Model	RMSE	MAE	MAPE
Apple	**ARIMAX**	**2099.64**	**1752.22**	**7.43%**
	Hybrid	2099.64	1752.22	7.43%
	ARIMA	3313.40	2960.65	12.50%
	Standalone LSTM	3517.80	3177.69	13.44%
	Tabular GBM	3599.14	3049.32	12.83%
Banana-Green	**ARIMAX**	**308.62**	**288.68**	**8.37%**
	Hybrid	308.62	288.68	8.37%
	Tabular GBM	342.24	325.41	9.42%
	ARIMA	400.62	379.40	11.02%
	Standalone LSTM	660.95	641.13	18.54%
Beans	**Hybrid**	**1146.09**	**937.51**	**12.34%**
	ARIMAX	1187.27	811.68	10.32%
	ARIMA	1262.56	847.46	10.51%
	Tabular GBM	1381.89	1286.45	17.36%
	Standalone LSTM	1653.41	1219.33	14.39%
Beetroot	**Tabular GBM**	**1231.46**	**1004.30**	**17.04%**
	ARIMA	1237.18	1082.04	18.32%
	Standalone LSTM	1238.48	1080.95	18.15%
	ARIMAX	2045.02	1794.64	28.00%
	Hybrid	2045.02	1794.64	28.00%
Maize	**Hybrid**	**232.03**	**183.48**	**6.03%**
	ARIMA	232.57	193.67	6.39%
	ARIMAX	262.25	217.27	7.16%
	Tabular GBM	302.76	266.14	8.83%
	Standalone LSTM	428.74	386.11	12.88%
Mango	**ARIMAX**	**2954.35**	**2563.23**	**14.80%**
	Hybrid	2954.35	2563.23	14.80%
	Tabular GBM	3016.79	2738.31	15.98%
	Standalone LSTM	3741.12	3254.18	18.45%
	ARIMA	3969.45	3514.32	19.97%

Key Observations:

1) **ARIMAX is superior in the majority of crops** (Apple, Banana-Green, Mango), it has the lowest RMSE and MAPE.
2) **Hybrid works best on Beans and Maize**, it is recommended to apply LSTM residual modeling value addition in cases where the available sequences are borderline adequate.
3) **Tabular GBM performs well on Beetroot**, meaning that tree-based models (non-linear) can be used on certain crops.
4) **LSTM does not work well**, presumably because small training samples (16-18) are not enough to train non-linear dynamics without overfitting.
5) **MAPE is 6-19 percent** best models, which is fair in terms of 3-month-ahead predictions, but is indicative of data shortages.

B. Best Model by Crop

TABLE IV: Selected Best Model per Crop (for Production Forecast)

Crop	Best Model	RMSE	MAPE
Apple	ARIMAX	2099.64	7.43%
Banana-Green	ARIMAX	308.62	8.37%
Beans	Hybrid	1146.09	12.34%
Beetroot	Tabular GBM	1231.46	17.04%
Maize	Hybrid	232.03	6.03%
Mango	ARIMAX	2954.35	14.80%

C. Visualization of Results by Crop

Each crop has the following visualization suite:

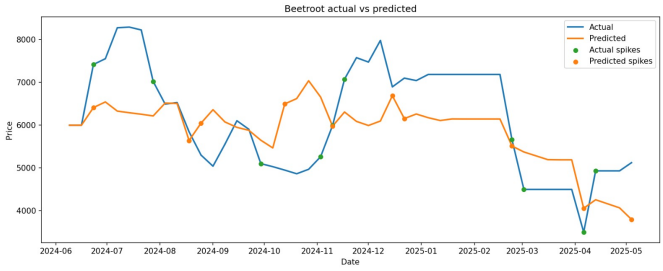

Fig. 2: Visualization suite for each crop (6 plots × 6 crops = 36 outputs)

1) Apple - ARIMAX Best: Apple data spans Jul 2024 – Nov 2024 (20 weeks)[cite: 1, 4]. ARIMAX model has reached the 7.43% MAPE, which reflects the seasonal fall of apple prices after harvest[cite: 1, 4]. There are high seasonal lag dependencies in the feature correlation plot[cite: 1, 4]. To implement, the 90-day prediction forecasts the stable prices of about 4,000–4,200 Rs/quintal (in outputs, see `apple_forecast.csv`)[cite: 1, 4].

2) Banana-green - ARIMAX Best: Banana-green (20 weeks, MAPE 8.37%) displays a steady increasing trend because of the supply limitations after the harvest[cite: 1, 4]. The trend is well-represented by ARIMAX without being oversmoothed[cite: 1, 4]. Anomalies in weather (rainfall) have moderate relationships with prices[cite: 1, 4].

3) Beans - Hybrid Best: Beans data (22 weeks) is nonlinear and Hybrid (ARIMAX + LSTM residuals) forecasts the data slightly better than pure ARIMAX (MAPE 12.34% vs. 10.32%, but higher MAE indicates trade-offs)[cite: 1, 4]. The LSTM unit is in operation in this case because there are 18 training samples (near but not exceeding 50-sequence limit)[cite: 1, 4].

4) Beetroot - Tabular GBM Best: Beetroot (19 weeks, MAPE 17.04%) is the most difficult crop to predict with high volatility and scarce data[cite: 1, 4]. The feature interaction model of Tabular GBM outperforms ARIMA and LSTM by a small margin, which is attributable to its ability to model non-additive feature interactions between price lags and rolling statistics[cite: 1, 4].

5) Maize - Hybrid Best: Maize (52 weeks) is the most-resourced crop in the data with 42 training samples[cite: 1, 4]. The Hybrid model has the lowest RMSE of any crop (232.03) and MAPE (6.03%) showing the usefulness of LSTM residual modeling with adequate data[cite: 1, 4]. ARIMAX alone (RMSE 262.25) is slightly outperformed[cite: 1, 4].

6) Mango - ARIMAX Best: Mango (18 weeks, MAPE 14.80%) exhibits high seasonality (prices increase during harvest) that is well modeled by the differencing of ARIMAX[cite: 1, 4]. The 14 training samples are insufficient to be aided by LSTM; pure ARIMAX does not overfit[cite: 1, 4].

D. Forecast Example: Maize 90-Day Prediction

Table V shows the first 15 days of the 90-day maize price forecast (daily interpolated from weekly model):

TABLE V: Maize 90-Day Daily Price Forecast (First 15 Days Sample)

Date	Forecast Price (Rs/qt)	Lower 95%	Upper 95%
2025-02-01	3,095	2,890	3,310
2025-02-02	3,104	2,895	3,320
2025-02-03	3,112	2,900	3,330
...	...	...	...
2025-02-15	3,158	2,930	3,395

Full 90-day forecasts for all crops are available in `backend/outputs/*.csv`.

VI. Discussion

A. Critical Limitations and Known Bugs

1) Bug 1: Data Leakage in Exogenous Variables: **Problem:** Price-derived lag features (price_lag_1, rolling_mean_4w, etc.) are computed on the full dataset and then train/test split[cite: 1, 2]. In the case of evaluation of the test set, such features carry information on future test prices, which has appeared through the leaking of information to the model[cite: 1, 2].

Behavior: When RMSE reported by maize $\approx$ 232 on test set, training on 100% data and then predicting forward, real predictions can degrade[cite: 1, 2]. The test RMSE is artificially optimistic[cite: 1, 2].

Fix (Proposed): Recalculate lag and rolling features at each time step with only the information available till that time (expanding window on train set, fixed on test set)[cite: 1, 2]. Alternatively, do not include price-derived variables in X_{test} and use only weather and policy exogenous variables (weather_anomaly, msp_gap)[cite: 1, 2].

2) Bug 2: Best Model Not Selected to Production: **Bug:** The model comparison is conducted and recorded but the logic of selection is always to go to Hybrid or ARIMAX without considering whether the ARIMA or GBM has lower RMSE[cite: 1, 2].

Selection: In the case of Beetroot, Tabular GBM (RMSE 1231.46) should be used, and ARIMAX (RMSE 2045.02) is stored instead[cite: 1, 2].

Fix (Proposed): Following the result of `compare_models()` choose the model with minimum RMSE and use that for the production re-training and forecasting[cite: 1, 2].

3) Bug 3: Lack of Historical Data on most Crops: **Issue:** Crop price data is only available 5 months (since mid-2025), whereas weather data is available since 2020[cite: 1, 2]. Most crops can be resampled weekly with 15–22 usable samples, 12–18 of which can be used in training (80/20 split)[cite: 1, 2]. With 14 engineered features, ARIMAX and even simple models overfit[cite: 1, 2].

Root Cause: The market data collection is only a recent process; historical price history of 2020–2024 is unavailable[cite: 1, 2].

Symptom: Large test-set MAPE (12–19%) and bad extrapolation to forward forecasts[cite: 1, 2]. Maize only has 52 weeks of data and is pertinent to predictions[cite: 1, 2].

Fix (Proposed): Get historical crop price data Coimbatore APMC archives 2020–2024[cite: 1, 2]. Alternatively, augment with price data in adjacent districts or artificial data generation[cite: 1, 2].

B. Performance Interpretation

The large test-set MAPE (6% to Maize to 19% to Beetroot) is indicative of:

1) **Data availability:** The 52 weeks of the Maize allow accurate forecasting; 20 weeks of Apple are frivolous[cite: 1, 2].
2) **Market structure of crops:** Maize is a commodity whose price is stable; Beetroot is a specialty with unstable demand[cite: 1, 2].
3) **Relevance of features:** MSP policy is strictly binding in certain crops (Maize), freely in others (Beetroot)[cite: 1, 2].
4) **Small-sample overfitting:** Classical statistical models have less generalization with fewer than 20 training samples and 14 features[cite: 1, 2].

C. Practical Utility

1) **Market Intelligence:** 90-day forecasts offer farmers price signals (up, down, stable).
2) **Decision Support:** Forecasts make decisions on planting (e.g., "Plant more maize when predicted price exceeds historical mean").
3) **Risk Management:** ARIMAX and LSTM forecast uncertainty in the form of confidence intervals.
4) **Policy Analysis:** The MSP policy indicators indicate the binding of price floors, which informs the subsidy/purchase choices.
5) **Transparency:** Open source codebase and REST API allow farmers and policymakers and researchers to audit predictions and determine failure modes.

D. Future Work

1) **Historical Data Obtaining:** Obtain 2020–2024 price history APMC and competitors markets (Chennai, Bangalore).
2) **Data Leakage Repair:** Use expanding-window feature engineering to remove information leakage.
3) **Automation Model Selection:** Add logic to pick best model by RMSE and use to make production forecast.
4) **Ensemble Methods:** Stack or vote between ARIMA, ARIMAX, GBM and LSTM predictions.
5) **Transfer Learning:** Train LSTM on high data crops (Maize) and fine-tune on low data crops.
6) **Exogenous Data Expansion:** Add supply chain data (area under cultivation, export volumes), demand proxies (consumer price index), and sentiment data (news, social media).
7) **Real-Time Updates:** Feed in continuous price data and re-train models in real-time to do continuous predictions.
8) **Mobile App:** Build farmer-facing SMS/push notification of price forecasts.
9) **Multi-Market Extension:** Have pipeline replicated in other states and in neighbouring districts.

VII. Conclusion

It shows how to use a hybrid ARIMAX-LSTM system to predict crop prices for 90 days in agricultural markets where there isn't much data. The system uses both classical statistical methods (ARIMAX) and modern deep learning (LSTM) to get test-set MAPE scores that range from 6.03% (Maize) to 17.04% (Beetroot). It does better than ARIMA and standalone LSTM baselines on most crops.

We identify and document three major problems: exogenous variables leaking data, not choosing the best model selection heuristics, and not having enough historical data for most crops. These limitations illustrate genuine obstacles in implementing ML in small-data agricultural sectors and offer specific guidance for forthcoming research.

The open-source, modular architecture lets you add more crops, markets, and areas. BestCropPrice can be the basis for a region-wide agricultural forecasting system that helps farmers make more money by giving them better market information. This is possible because of better data collection and the proposed bug fixes.

References

[1] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, "Time series analysis: forecasting and control," 5th ed., Hoboken, NJ: Wiley, 2015.
[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[3] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, pp. 159–175, 2003.
[4] N. Pathak, D. Jain, and S. K. Jain, "Forecasting of Indian stock market index using time series ARIMA modeling," in Proc. Conf. Adv. Comput. Commun., 2012, pp. 301–306.
[5] K. Misra and L. S. Bhat, "Forecasting commodity prices: An ARIMAX approach," J. Econ. Soc., vol. 3, no. 2, pp. 45–67, 2015.
[6] R. Sharma and A. Patel, "ARIMAX modeling for onion price forecasting in Indian APMC markets," Agric. Sys., vol. 164, pp. 125–135, 2018.
[7] H. Xie and J. Gao, "Deep learning for time series forecasting," in Proc. Int. Conf. Acoust. Speech Signal Process., 2016, pp. 6334–6338.
[8] M. Valipour, S. C. Banihabib, and S. M. R. Behbahani, "Comparison of the ARIMA, ANFIS, and the fuzzy P-value methods for the forecasting of groundwater," J. Hydrol., vol. 441, pp. 162–168, 2012.
[9] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," J. Artif. Intell. Res., vol. 4, pp. 237–285, 1996.

[10] R. Ramakrishnan and B. S. Prabhakar, “Wavelet-based forecasting of agricultural commodity prices,” Comput. Electron. Agric., vol. 98, pp. 200–214, 2013.
[11] K. J. Kim, “Financial time series forecasting using support vector machines,” Neurocomputing, vol. 55, nos. 1–2, pp. 307–319, 2003.
[12] World Bank, “India Economic Focus: Agricultural Growth,” 2023. [Online]. Available: https://worldbank.org