MACHINE LEARNING MODEL FOR ANALYZING CLIMATIC AND ECONOMIC INFLUENCES ON VEGETABLE PRICES: FORECASTING CARROT PRICES IN DAMBULLA MARKET

G.G.M.P.KUMARA

Submitted in partial fulfillment of the requirements of the degree of

B.Sc. Honours in Computer Science

Department of Statistics and Computer Science

University of Peradeniya

ABSTRACT

Agricultural market volatility poses significant challenges for farmers, traders, and policymakers in Sri Lanka, where vegetable prices fluctuate unpredictably due to complex interactions between weather patterns, supply dynamics, and economic factors. This research develops a machine learning-based forecasting system for predicting wholesale carrot prices at Dambulla market, integrating data from the Central Bank of Sri Lanka, Copernicus Climate Data Store, Ceylon Petroleum Corporation fuel prices, and market supply-demand indicators.

Comprehensive feature engineering across price patterns, weather conditions, supply factors, demand indicators, fuel costs, and temporal features was performed. A systematic 4-stage feature selection pipeline combining Random Forest importance, Mutual Information, correlation analysis, and multicollinearity removal reduced 163 engineered features to 8 optimal features while preserving predictive power and eliminating redundancy. Seven forecasting models including ARIMA, LSTM variants (Univariate, Multivariate, Simple, and Bidirectional), and Random Forest were rigorously evaluated using consistent train-validation-test splits and multiple performance metrics including MAPE, MAE, RMSE, and R^2.

Traditional ARIMA methods demonstrated fundamental limitations with MAPE exceeding 50% for both univariate and multivariate specifications, validating the inadequacy of linear assumptions for complex agricultural markets. LSTM-based deep learning approaches achieved substantial improvements, with the optimized Simple LSTM model reaching 19.93% MAPE and R^2 of 0.8651 using only 8 carefully selected features, effectively capturing non-linear temporal dependencies while avoiding overfitting through architectural simplicity and aggressive feature selection. Systematic ablation studies quantified feature category contributions, revealing price history as the dominant predictor while weather, supply, and fuel factors provided meaningful incremental accuracy. SHAP analysis enhanced model interpretability, confirming theoretically expected negative rainfall-price relationships.

The research delivers a deployment-ready system integrating the best-performing model with a Retrieval-Augmented Generation AI agent using Groq API and natural language interface via Gradio, democratizing sophisticated forecasting for non-technical stakeholders. This work contributes replicable methodology for agricultural price forecasting, empirically validates deep learning superiority over traditional approaches for complex market dynamics, and provides operational tools supporting informed decision-making across the agricultural value chain in developing economies.

Keywords: carrot price prediction, LSTM neural networks, Dambulla market, agricultural forecasting, machine learning, weather-price relationships, RAG system, AI agent, time series analysis, Sri Lankan agriculture

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

. . . .
G.G.M.P.Kumara
November 26, 2025

. . . .
Prof.Amalka Pinidiyaarachchi
November 26, 2025

ACKNOWLEDGEMENT

I would like to acknowledge the invaluable contributions of those who made this research possible. First and foremost, I express my deepest gratitude to my supervisor, Prof.Amalka Pinidiyaarachchi, for her exceptional guidance and unwavering support throughout this research journey. Without her expertise and mentorship, I could not have achieved this level of work. Whenever I encountered challenges in model development or data analysis, her guidance showed me the path forward. Prof.Pinidiyaarachchi's insightful feedback, technical expertise, and continuous encouragement were the driving force behind this project. Her mentorship not only shaped my research approach and analytical thinking but also instilled in me a deeper appreciation for the intersection of machine learning and agricultural economics. I am truly grateful for her patience, dedication, and commitment to my academic development.

I would like to thank the Department of Statistics and Computer Science, University of Peradeniya, for providing the necessary resources and infrastructure that made this research possible. Access to computational resources was particularly valuable for the intensive training requirements of LSTM neural networks and the extensive data preprocessing pipeline. The department's supportive academic environment fostered the interdisciplinary thinking essential for this work.

I am grateful to the Central Bank of Sri Lanka for providing access to daily market price data, the Copernicus Climate Change Service for precipitation datasets, and the Ceylon Petroleum Corporation for fuel price information. This research would not have been feasible without access to these comprehensive data sources.

I would also like to extend my appreciation to my family and friendst, whose conversations helped me discover my passion for using technology to solve real-world agricultural challenges. Through discussions about the impact of price volatility on farming communities in Sri Lanka, I realized that machine learning could be a greate Artificial Intelligence techniqe for meaningful social impact. Their encouragement motivated me to building an accessible system that farmers, traders, and policymakers could actually use in their daily decision-making, which ultimately shaped the direction of this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Research Overview

Agriculture remains a vital component of Sri Lanka's economy. Among the various agricultural products, vegetables constitute an essential part of the domestic food supply chain. However, vegetable markets in Sri Lanka have historically been characterized by high price volatility, which creates substantial challenges for both farmers and consumers. Carrot, being one of the major vegetables traded in the country, experiences frequent and often unpredictable price fluctuations that can severely impact the economic stability of farming communities.

Daily price movements in the Dambulla market reflect complex interactions between multiple factors including weather patterns, fuel costs, regional supply levels, and market demand dynamics. Understanding and predicting these price movements has become increasingly important for stakeholders throughout the agricultural value chain.

This research investigates the application of machine learning techniques to forecast wholesale carrot prices in the Dambulla market. The study explores how climatic conditions, particularly precipitation patterns from major growing regions, along with economic factors such as fuel prices and supply indicators, influence daily price variations. Through the development and comparison of multiple forecasting models, this work aims to identify the most effective approach for short-term price prediction in the context of Sri Lankan vegetable markets.

1.2 Problem Statement

Vegetable price volatility presents serious problems for the agricultural sector in Sri Lanka. Farmers often struggle to make informed decisions about planting schedules, harvest timing, and market entry due to the uncertainty surrounding future prices. When prices drop unexpectedly, farmers may face losses that threaten their financial stability. Conversely, sudden price spikes can lead to affordability issues for consumers while also triggering market interventions that complicate the natural price discovery process.

Traditional approaches to understanding vegetable prices have relied primarily on historical averages and seasonal patterns. However, these methods fail to capture the complex, non-linear relationships between prices and the various factors that influence them. The lack of reliable forecasting tools means that decision-making across the supply chain remains largely reactive rather than proactive. This situation calls for more sophisticated analytical approaches that can better model the temporal dependencies and multi-dimensional influences inherent in agricultural price data.

The challenge is further complicated by the fact that different factors may have varying degrees of influence on prices at different times. For instance, heavy rainfall in major growing regions like Nuwara Eliya might have delayed effects on market prices as supply disruptions gradually propagate through the distribution network. Similarly, changes in fuel prices might impact transportation costs, which in turn affect wholesale pricing. Capturing these temporal lags and interactions requires modeling techniques that can learn from sequential data and account for multiple input variables simultaneously.

1.3 Proposed Solution

This research proposes a comprehensive machine learning framework for carrot price forecasting that addresses the limitations of traditional approaches. The framework employs three modeling approaches: ARIMA models for baseline time series analysis, LSTM neural networks to capture complex temporal patterns with external factors including precipitation, fuel prices, and supply indicators, and Random Forest regression as an ensemble learning alternative.

The framework incorporates systematic feature selection to identify the most relevant predictors from multiple data sources. Additionally, it includes an intelligent AI agent that provides natural language access to forecasting insights through a web interface, making sophisticated predictions accessible to non-technical stakeholders.

1.4 Background

Sri Lanka's agricultural sector is central to the nation's economy and food security, with vegetable farming concentrated in the central highlands. The Nuwara Eliya district supplies a substantial portion of temperate vegetables including carrots. The Dambulla Economic Centre functions as the primary wholesale distribution point for vegetables, where prices established daily influence retail pricing throughout the country.

Carrot cultivation occurs year-round across main growing areas including Nuwara Eliya, Welimada, and Bandarawela, though production volumes vary seasonally. These regions experience different rainfall patterns and microclimates, theoretically providing continuous supply. However, the system is vulnerable to disruptions from excessive rainfall damaging crops, transportation challenges from fuel price increases, and market dynamics amplifying price volatility. Despite government efforts to collect market data, forecasting tools remain limited. This research addresses this gap by introducing sophisticated analytical capabilities.

1.5 Research Objectives

This research aims to develop an effective machine learning-based forecasting system for carrot price prediction in the Dambulla market. The study pursues three primary objectives:

1. **Develop Machine Learning Models for Price Forecasting:** Implement

multiple forecasting models including ARIMA for baseline time series analysis, LSTM neural networks to capture temporal dependencies with external factors, and Random Forest as an ensemble alternative. Apply systematic feature selection to identify optimal predictors from historical price, meteorological, economic, and market data.

2. Evaluate and Compare Model Performance: Conduct comprehensive evaluation of all developed models using appropriate performance metrics to identify the most effective forecasting approach. Include systematic validation studies to assess model reliability and understand key factors influencing prediction accuracy.

3. Build an AI Agent Using RAG System: Develop a deployment-ready intelligent agent integrating the best-performing forecasting model with natural language processing capabilities to provide farmers, traders, and policymakers intuitive access to forecasting insights through a conversational interface.

1.6 Scope of the Research

This research focuses specifically on wholesale carrot prices in the Dambulla market, allowing in-depth analysis while maintaining manageable complexity. While the methodologies could potentially apply to other vegetables or markets, the empirical work concentrates on this single commodity and location.

The temporal scope encompasses January 2020 through July 2025, providing over five years of daily price observations. This period includes both normal market conditions and volatile periods (including Sri Lanka's recent economic challenges), ensuring models are exposed to diverse market scenarios.

The study investigates ARIMA, LSTM, and Random Forest models as representative samples spanning traditional statistical methods, deep learning, and ensemble learning. The RAG-based AI agent serves as a delivery mechanism for predictions, with the primary focus remaining on forecasting methodology itself.

CHAPTER 2

LITERATURE REVIEWS

This chapter reviews the existing literature on agricultural price forecasting, with particular emphasis on vegetable price prediction using various computational techniques. The review is organized into several thematic sections that progress from traditional statistical approaches to modern machine learning and deep learning methodologies. Each section examines relevant studies, highlighting their methodologies, findings, and contributions to the field. The chapter concludes by identifying gaps in current research and establishing the context for this study.

2.1 Traditional Statistical Methods for Agricultural Price Prediction

Traditional statistical approaches have long formed the foundation of agricultural price forecasting. Among these methods, AutoRegressive Integrated Moving Average (ARIMA) models have been extensively employed due to their ability to capture temporal dependencies in price series data.

Ruhunuge et al. (2024) conducted an econometric investigation into climate-driven carrot price variations in Sri Lanka using Vector Autoregression (VAR) modeling. Their study spanned twenty-three years (2000-2023) of wholesale carrot price data from the Hector Kobbekaduwa Agrarian Research and Training Institute (HARTI), combined with climate data from the Meteorology Department. The researchers applied first-order differencing to capture volatility patterns, resulting in 852 monthly observations that passed rigorous unit root tests. Their VAR model, optimized with a lag structure of six periods based on Akaike Information Criterion (AIC), revealed that precipitation changes significantly influenced carrot prices (p-value = 0.0447), while temperature demonstrated limited predictive value. The impulse response analysis showed that a one-unit standard deviation increase in precipitation resulted in an immediate 2.8% increase in carrot prices, peaking at 1.2% in the third interval before stabilizing. The derived model equation demonstrated that increased rainfall significantly lowered current carrot prices with a coefficient of -16.09719, while the positive lagged price coefficient of 0.141777 indicated price momentum effects (Ruhunuge et al., 2024).

Chen et al. (2021) compared ARIMA against modern machine learning approaches for Malaysian agricultural commodity price prediction. Their ARIMA implementation utilized parameters (p=1.5, d=1, q=1) determined through Augmented Dickey Fuller Test and comprehensive Auto/Partial Correlation Function analysis. While ARIMA achieved remarkable average Mean Squared Error (MSE) of 0.251 for smaller datasets, with chili prediction reaching exceptional 0.027 MSE, the model experienced concerning 74.1% performance degradation when confronted with increased data complexity. This finding highlighted a fundamental limitation of traditional statistical methods when dealing with large, complex agricultural datasets (Chen et al., 2021).

These studies demonstrate that while traditional statistical methods provide interpretable models and perform well with smaller datasets under stable market conditions, they struggle with non-linear relationships and large-scale data, motivating the exploration of machine learning alternatives.

2.2 Machine Learning Approaches for Crop Price Forecasting

Machine learning techniques have gained prominence in agricultural price prediction due to their ability to capture complex, non-linear relationships between multiple variables without requiring explicit model specification.

2.2.1 Tree-Based Ensemble Methods

Ranaweera et al. (2023) investigated vegetable price predictability in Sri Lanka using data mining techniques. Their comprehensive study analyzed four vegetables (beans, eggplant, carrots, and pumpkins) from five economic centers using four-year historical data (2018-2021) from multiple institutions including the Central Bank, Department of Agriculture, Meteorological Department, and Ceylon Petroleum Corporation. The researchers evaluated five machine learning algorithms including Linear Regression, SMO Regression, Multilayer Perceptron, Random Forest, and M5P using 10-fold cross-validation in WEKA 3.8.6. Random Forest emerged as the superior model, achieving Mean Absolute Error (MAE) values ranging from 10.58 for pumpkins to 27.62 for beans. Pumpkins demonstrated the highest prediction accuracy exceeding 85%, while beans presented the most challenging forecasting scenario. The study revealed that rainfall variability (0.00 to 53.40mm) and temperature fluctuations, alongside fuel prices affecting transportation costs, critically impacted price variations (Ranaweera et al., 2023).

Choong et al. (2024) developed a Genetic Algorithm-Based Neural Network (GANN) approach within an Agricultural Knowledge Management System (AKMS) for Malaysia. Using monthly vegetable prices from 2010 to 2021 (144 observations), their GANN model achieved 98.40% accuracy with MAPE of 1.6042%, significantly outperforming both ARIMA (98.32% accuracy) and SARIMA (98.37% accuracy). The genetic algorithm optimization with 20 chromosomes, three hidden layers containing five nodes each with ReLU activation, and roulette wheel selection methodology enabled dynamic adaptation to agricultural market dynamics. The model particularly excelled in RMSE (0.06674) and MAE (0.5571) metrics, demonstrating superior handling of nonlinear relationships in seasonal price variations (Choong et al., 2023).

2.2.2 Support Vector Machines and Hybrid Approaches

Kakulapati et al. (2022) explored vegetable price prediction against temperature changes using web scraping to collect real-time weather and price data from Hyderabad. Their comparison of Decision Tree Regression, Random Forest Regression, and Linear Regression revealed that Decision Tree Regression achieved superior accuracy in predicting prices based on temperature variations. The innovative web scraping methodology enabled dynamic data collection at five-day intervals, addressing limitations of static historical datasets. This real-time approach provided farmers with timely insights for cultivation decision-making based on weather-price correlations (Vijayalaxmi et al., 2022).

Bayona-Oré et al. (2021) conducted a comprehensive systematic review of machine learning applications in agricultural price prediction from 2011-2020. Their analysis revealed that Neural Network models were most frequently employed (24 algorithms), followed by statistical models (20 algorithms) and Support Vector Machines (9 occurrences). The review identified that all studies employed positivism paradigm with quantitative approaches, predominantly using supervised learning due to availability of labeled historical price data. Performance metrics analysis showed RMSE, MAPE, and MAE as the most commonly used evaluation measures. Geographically, China dominated with 11 studies examining 17 products, while India contributed 6 studies covering 12 products, indicating regional concentration in research efforts (Bayona-Oré et al., 2021).

2.3 Deep Learning Methods for Agricultural Time Series Forecasting

Deep learning approaches, particularly recurrent neural networks and their variants, have revolutionized time series forecasting by effectively capturing long-term dependencies and complex temporal patterns.

2.3.1 Long Short-Term Memory (LSTM) Networks

Zhang et al. (2024) investigated short-term vegetable price forecasting for Beijing's wholesale markets using LSTM models. Their study utilized 14.7 years of daily price data (January 2009 to September 2023) from seven major wholesale markets, analyzing six representative vegetables from four categories. The LSTM architecture comprised two layers with 32 neurons each, optimized learning rate of 0.0027, dropout rate of 0.2, batch size of 500, and 200 training epochs using Adam optimizer. The model achieved exceptional performance with R^2 scores of 0.958 and MAE of 0.143, representing over 5% improvement compared to CNN, XGBoost, and SVR. Vegetable-specific accuracy varied notably: celery (93.3%), carrots (92.9%), oyster mushrooms (90.2%), and spiny cucumbers (90.1%), with trend prediction concordance rates exceeding 70% for most vegetables. Wilcoxon signed-rank tests confirmed statistically significant improvements over competing methods (p ¡ 0.05) (Zhang et al., 2024).

Yin et al. (2020) developed an innovative STL-ATTLSTM model integrating Seasonal Trend decomposition using Loess (STL) with attention mechanism-based LSTM for South Korean vegetable markets. Their research targeted five supply-and-demand-sensitive vegetables (cabbage, radish, onion, hot pepper, garlic) using data from January 2012 to December 2019. The sophisticated architecture employed STL to separate time series into trend, seasonality, and remainder components, with attention mechanism assigning dynamic weights to input variables during training. The model comprised an attention layer with softmax activation, LSTM layer with 6 cell units using tanh activation, dropout layer (0.2 rate), and fully connected layers, trained for 1000 epochs using Adam optimizer. The STL-ATTLSTM achieved exceptional average RMSE of 380 and MAPE of 7%, representing 12% higher prediction accuracy compared to attention LSTM without STL preprocessing. The model successfully eliminated the one-month prediction lag phenomenon common in highly volatile time-series data by utilizing STL remainder components rather than raw price data (Yin et al., 2020).

2.3.2 Hybrid Deep Learning Architectures

Guo et al. (2022) proposed an innovative AttLSTM-ARIMA-BP hybrid model for corn price prediction in Sichuan Province, China. Using 511 weekly observations from March 2011 to April 2021, they employed Apriori association rule mining to identify 12 critical spatial-temporal factors across multiple provinces and related commodity prices. Their hybrid architecture strategically integrated Attention Mechanism for dynamic weight calculation, LSTM for non-linear temporal dependencies, ARIMA for linear trend modeling, and Back Propagation Neural Network for final prediction synthesis. The model achieved outstanding performance with MAPE of 0.0043, MAE of 1.51, RMSE of 1.642, and remarkable R^2 of 0.9992, significantly outperforming seven competing models including Linear Regression, Random Forest, XGBoost, LightGBM, single LSTM, multivariate LSTM, and AttLSTM. While traditional regression models maintained reasonable accuracy during stable periods, they failed dramatically during volatile market conditions, whereas the hybrid model consistently delivered accurate predictions regardless of price behavior patterns (Guo et al., 2022).

Avinash et al. (2024) introduced Hidden Markov-based Deep Learning approaches for forecasting TOP (Tomato, Onion, Potato) commodity prices in India. Their research utilized 911 weekly price observations from Azadpur Mandi (January 2006 to June 2023), applying Hidden Markov Models (HMMs) for feature extraction to identify hidden states in price data. Optimal hidden states were determined through grid search: six states for tomato and eight states each for onion and potato. These hidden states served as inputs to four deep learning models: Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTM). Extensive hyperparameter optimization across 126 combinations per model included batch sizes, epochs (200 with early stopping), hidden layers, and units. The hybrid HM-DL models achieved superior performance with RMSE reductions of 9.77-17.50% for tomato, 15.02-44.39% for onion, and 7.94-32.60% for potato compared to baseline approaches. HM-RNN consistently emerged as the best performer for training data, while HM-LSTM excelled for tomato testing data due to superior long-memory capabilities in capturing significant price spikes. Diebold-Mariano tests confirmed statistically significant differences between hybrid and baseline models (Avinash et al., 2024).

2.4 Feature Engineering and Selection in Agricultural Forecasting

Effective feature engineering and selection constitute critical components of successful agricultural price prediction models, as they determine which variables contribute most significantly to forecasting accuracy.

The reviewed studies employed diverse approaches to feature selection. Ranaweera et al. (2023) systematically incorporated four key factors: rainfall, temperature, fuel price, and crop production, demonstrating that climatic factors particularly influenced price variations in tropical agricultural systems. Their analysis revealed substantial variability in price predictability across vegetables, with pumpkins showing highest accuracy and beans presenting the most challenging scenario (Ranaweera et al., 2023).

Guo et al. (2022) utilized Apriori association rule mining algorithm to identify 12 critical spatial-temporal factors influencing corn prices, including prices from multiple provinces and related commodities. This data-driven approach to feature discovery enabled their hybrid model to capture complex inter-commodity and inter-regional price relationships (Guo et al., 2022).

Yin et al. (2020) demonstrated sophisticated feature engineering by incorporating meteorological variables (average temperature, minimum temperature, humidity, precipitation, temperature threshold days, typhoon advisories) specifically for main production areas during harvest periods. Their approach strategically focused harvest-time meteorological data for immediately marketed crops (cabbage, radish) while excluding weather factors for warehouse-stored crops (hot pepper, onion, garlic) with delayed market entry. Additionally, they integrated trading volume data as production proxies and import/export information for comprehensive market analysis (Yin et al., 2020).

Chen et al. (2021) implemented a dual-experimental design: first utilizing univariate time-series data spanning 11 years, then incorporating multivariable features including temperature, humidity, precipitation, and crude oil prices. Their comparative analysis revealed that ARIMA excelled with smaller datasets while LSTM demonstrated 45.5% improvement in MSE for larger, more complex datasets, highlighting the importance of matching model complexity to data characteristics (Chen et al., 2021).

2.5 Model Evaluation and Performance Metrics

Rigorous model evaluation using appropriate performance metrics is essential for assessing prediction accuracy and comparing different forecasting approaches.

2.5.1 Common Evaluation Metrics

The literature review reveals widespread adoption of several key performance metrics. Mean Absolute Percentage Error (MAPE) emerged as the most prevalent metric, utilized by Zhang et al. (2024) achieving 0.143, Yin et al. (2020) achieving 7%, Guo et al. (2022) achieving 0.0043, and Choong et al. (2024) achieving 1.6042%. MAPE's popularity stems from its scale-independent nature and intuitive percentage interpretation (Zhang et al., 2024, Yin et al., 2020, Guo et al., 2022, Choong et al., 2023).

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were also frequently employed for absolute error measurement. Ranaweera et al. (2023) comprehensively evaluated models using MAE, RMSE, Relative Absolute Error (RAE), and Root-Relative Square Error (RRSE), providing multi-dimensional performance assessment. Additionally, coefficient of determination (R^2) was utilized to measure explained variance, with Zhang et al. (2024) achieving 0.958 and Guo et al. (2022) achieving exceptional 0.9992 (Ranaweera et al., 2023, Zhang et al., 2024, Guo et al., 2022).

2.5.2 Statistical Validation Techniques

Several studies incorporated rigorous statistical validation beyond basic performance metrics. Avinash et al. (2024) employed Diebold-Mariano (DM) tests to establish statistically significant differences between hybrid and baseline models, ensuring observed improvements were not due to chance. Zhang et al. (2024) utilized Wilcoxon signed-rank tests to confirm LSTM's superior performance over competing methods with p-values below 0.05 threshold (Avinash et al., 2024, Zhang et al., 2024).

Ruhunuge et al. (2024) implemented comprehensive VAR model validation including stability condition testing (all characteristic roots within unit circle, largest root at 0.92), residual diagnostics showing no significant autocorrelation (p-values above 0.10), and heteroscedasticity testing (p-value = 0.15) indicating constant variance, thereby validating model reliability (Ruhunuge et al., 2024).

Yin et al. (2020) conducted systematic time-step optimization through grid search over multiple lag values, determining optimal time-step of 4 for superior performance across most vegetables. This methodological rigor in hyperparameter selection contributed to their model's exceptional accuracy (Yin et al., 2020).

2.6 Regional Perspectives and Data Sources

Agricultural price prediction research exhibits significant geographical concentration, with diverse data sources and regional considerations influencing methodology and applicability.

2.6.1 Asian Agricultural Markets

The majority of reviewed studies focused on Asian agricultural markets. China dominated with multiple studies: Guo et al. (2022) analyzed Sichuan corn prices using data from China's agricultural big data website and commodity exchanges, while Zhang et al. (2024) examined Beijing's seven major wholesale vegetable markets (Guo et al., 2022, Zhang et al., 2024). Bayona-Oré et al. (2021) confirmed this geographical concentration, identifying China as most researched country with 11 studies examining 17 products, followed by India with 6 studies covering 12 products (Bayona-Oré et al., 2021).

South Asian markets received notable attention through studies in Sri Lanka and India. Ranaweera et al. (2023) and Ruhunuge et al. (2024) both focused on Sri Lankan vegetable markets, utilizing data from the Central Bank of Sri Lanka, Department of Agriculture, and Hector Kobbekaduwa Agrarian Research and Training Institute. Avinash et al. (2024) examined India's Azadpur Mandi in Delhi, one of Asia's largest wholesale markets, using data from Agmarknet spanning over 17 years (Ranaweera et al., 2023, Ruhunuge et al., 2024, Avinash et al., 2024).

Southeast Asian perspectives emerged through Malaysian market studies. Chen et al. (2021) and Choong et al. (2024) both utilized data from Malaysia's Federal Agricultural Marketing Authority (FAMA), examining chicken, chili, tomato, and potato prices over extended periods. Their work addressed regional challenges including aging farming population and insufficient knowledge management systems (Chen et al., 2021, Choong et al., 2023).

2.6.2 Data Collection Approaches

Data sourcing strategies varied significantly across studies. Traditional approaches relied on official government databases and agricultural institutions. Multiple studies utilized meteorological data from national weather services: Ranaweera et al. (2023) from Sri Lanka's Meteorological Department, Yin et al. (2020) from Korean Meteorological Administration, and Chen et al. (2021) integrated climate data with price information (Ranaweera et al., 2023, Yin et al., 2020, Chen et al., 2021).

Innovative data collection methods emerged in recent research. Kakulapati et al. (2022) employed web scraping to extract real-time weather and price data, collecting information every five days to create dynamic datasets reflecting current market conditions. This approach addressed limitations of static historical datasets, enabling timely insights for agricultural decision-making (Vijayalaxmi et al., 2022). Guo et al. (2022)

utilized commodity exchange data alongside traditional agricultural databases, capturing futures market information for wheat and soybeans to inform spot price predictions (Guo et al., 2022).

The temporal scope of datasets ranged considerably: from Ruhunuge et al.'s 23-year span (2000-2023) capturing long-term climate-price relationships, to Zhang et al.'s 14.7 years (2009-2023) of daily observations providing high granularity, to shorter focused studies like Ranaweera et al.'s four-year analysis (2018-2021) enabling rapid model development (Ruhunuge et al., 2024, Zhang et al., 2024, Ranaweera et al., 2023).

2.7 Application Domains and Practical Implementation

Beyond academic contributions, several studies addressed practical implementation challenges and developed systems for real-world agricultural stakeholders.

Chen et al. (2021) developed a comprehensive web-based platform following Model-View-Controller (MVC) pattern using Django framework. Their system featured secure user authentication, interactive visualization dashboards with customizable forecast durations, commodity selection interfaces, downloadable CSV exports, and responsive design for compatibility across devices. The platform empowered farmers, government agencies, and agricultural stakeholders to make informed decisions regarding plantation planning, supply chain optimization, and policy formulation (Chen et al., 2021).

Choong et al. (2024) integrated price forecasting within an Agricultural Knowledge Management System (AKMS) following DIKW (Data, Information, Knowledge, Wisdom) pyramid framework enhanced with IoT and Big Data capabilities. Their platform combined knowledge management principles with e-commerce functionality, supporting both explicit knowledge (documented procedures) and tacit knowledge (farmer experiences) through integrated information sharing. The system addressed Malaysia's National Agrofood Policy (NAP 2.0) 2021-2030 objectives for creating sustainable, technology-based agrofood industry (Choong et al., 2023).

Practical applications focused on multiple stakeholder benefits. Zhang et al. (2024) provided week-ahead forecasts with detailed trend analysis revealing distinct fluctuation patterns for different vegetables, offering insights for growers, consumers, and policymakers. Avinash et al. (2024) emphasized helping farmers optimize storage decisions, identify favorable selling periods, and minimize losses through reliable price forecasting. Ranaweera et al. (2023) highlighted the importance of AI-driven forecasting for mitigating financial risks associated with price fluctuations in developing tropical economies (Zhang et al., 2024, Avinash et al., 2024, Ranaweera et al., 2023).

2.8 Research Gaps and Limitations

While the reviewed literature demonstrates significant progress in agricultural price prediction, several gaps and limitations warrant attention for future research directions.

2.8.1 Methodological Gaps

Bayona-Oré et al. (2021) identified lack of epistemological consideration in most studies, with all employing positivism paradigm and quantitative approaches without exploring alternative philosophical frameworks. Their review revealed absence of comprehensive model comparison frameworks and limited exploration of model interpretability, despite increasing emphasis on explainable AI in agricultural applications (Bayona-Oré et al., 2021).

Feature selection approaches remained largely empirical rather than systematic. While Guo et al. (2022) employed data mining for feature discovery and Yin et al. (2020) strategically selected features based on agricultural domain knowledge, most studies lacked rigorous statistical feature selection procedures such as mutual information analysis, recursive feature elimination, or ablation studies to quantify individual feature contributions (Guo et al., 2022, Yin et al., 2020).

The challenge of model generalization across different regions, vegetables, and market conditions received limited attention. Chen et al. (2021) noted concerning 74.1%

performance degradation of ARIMA with increased complexity, while studies generally focused on specific vegetables or regions without investigating cross-commodity or cross-market applicability (Chen et al., 2021).

2.8.2 Data-Related Limitations

Data availability constraints significantly influenced research scope. Bayona-Oré et al. (2021) observed that agricultural product selection was primarily driven by data availability rather than economic importance or market significance. Geographic concentration in China and India reflected both research capacity and data infrastructure availability, while other developing agricultural economies remained understudied (Bayona-Oré et al., 2021).

Temporal granularity varied across studies, with some utilizing daily data, others weekly or monthly observations. The impact of temporal resolution on prediction accuracy and practical applicability remained underexplored. Additionally, most studies relied on historical price data without incorporating real-time market signals or social media sentiment that might capture emerging market trends.

External factor integration remained incomplete in many studies. While weather and fuel prices received attention, broader macroeconomic indicators (exchange rates, inflation, policy changes), consumer behavior patterns, and supply chain disruptions (as experienced during COVID-19) were generally absent from modeling frameworks.

2.8.3 Practical Implementation Challenges

The gap between research models and operational deployment systems remained substantial. While Chen et al. (2021) and Choong et al. (2024) developed web-based platforms, most studies concluded with model performance evaluation without addressing deployment challenges such as model updating procedures, computational requirements, user interface design, or stakeholder adoption barriers (Chen et al., 2021, Choong et al., 2023).

Model interpretability and explainability received insufficient attention despite their importance for farmer acceptance and trust. Agricultural stakeholders need to understand not only what prices are predicted but also why predictions change and which factors drive price movements. Only Yin et al. (2020) partially addressed this through their attention mechanism providing feature importance insights (Yin et al., 2020).

Uncertainty quantification remained largely absent from agricultural price predictions. Providing point estimates without confidence intervals or prediction intervals limits practical decision-making value, as farmers and traders need to assess risk associated with predictions for effective planning.

2.9 Summary and Position of Current Research

The literature review reveals substantial progress in agricultural price forecasting, transitioning from traditional statistical methods through machine learning approaches to sophisticated deep learning architectures. ARIMA and VAR models provided interpretable baseline approaches but struggled with non-linear patterns and large-scale data. Machine learning methods, particularly tree-based ensembles and support vector machines, demonstrated superior performance in capturing complex relationships between multiple variables. Deep learning approaches, especially LSTM and hybrid architectures, achieved state-of-the-art results by effectively modeling temporal dependencies and handling high-dimensional input spaces.

However, significant research gaps persist. Most studies focused on single vegetables or specific markets without comprehensive cross-commodity or cross-regional validation. Feature selection remained largely ad-hoc rather than systematic, and model interpretability received insufficient attention despite its importance for stakeholder adoption. The integration of diverse data sources including real-time information and broader macroeconomic factors remained limited. Finally, the gap between research models and deployable systems hindered practical impact on agricultural decision-making.

The current research addresses these gaps through three primary objectives. First, it develops comprehensive machine learning models for carrot price forecasting in

Dambulla market using ARIMA for traditional time series analysis, LSTM neural networks in univariate and multivariate configurations for capturing complex temporal dependencies, and Random Forest for ensemble learning approaches, incorporating weather patterns, fuel prices, supply factors, and market indicators with rigorous feature selection combining Random Forest importance, Mutual Information, and Recursive Feature Elimination. Second, it conducts systematic evaluation and comparison of all developed models using multiple performance metrics (MAPE, MAE, RMSE, R^2), ablation studies to quantify feature category contributions, statistical validation through bootstrap confidence intervals and cross-validation, and SHAP analysis for enhanced model interpretability. Third, it develops a Retrieval-Augmented Generation (RAG) based AI agent integrating the best-performing model with natural language processing capabilities through Groq API and Gradio interface, making sophisticated forecasting accessible and interpretable for non-technical agricultural stakeholders.

This research advances the field by providing methodologically rigorous, interpretable, and practically deployable forecasting solutions specifically tailored for Sri Lankan agricultural markets, while establishing a replicable framework applicable to other vegetables and markets in developing agricultural economies.

CHAPTER 3

DESIGN AND METHODOLOGY

This chapter presents the systematic approach employed to develop and evaluate machine learning models for forecasting wholesale carrot prices in the Dambulla market. The research methodology encompasses six major phases: systematic framework design, data collection and preparation, exploratory data analysis, feature engineering and selection, model development and training, model evaluation, and AI agent implementation. Each phase is designed to address specific research objectives while maintaining methodological rigor throughout the investigation.

3.1 Systematic Framework

The overall research framework follows a structured pipeline that progresses from raw data acquisition through model deployment. This systematic approach ensures reproducibility, methodological transparency, and practical applicability of the forecasting system.

The framework consists of several interconnected stages progressing systematically from raw data to deployed predictions. The process initiates with comprehensive data acquisition gathering information from multiple authoritative sources including market prices from the Central Bank of Sri Lanka, weather data from Copernicus covering eleven growing regions, fuel prices from Ceylon Petroleum Corporation, and supply-demand indicators from agricultural market reports. This raw data undergoes rigorous preprocessing involving quality assurance procedures, missing value treatment using forward-filling to maintain temporal integrity, outlier detection and analysis with deliberate retention of legitimate extreme values representing genuine market volatility, and necessary data transformations to prepare variables for modeling.

Following data preparation, exploratory data analysis employs visual and statistical techniques to understand temporal patterns in price movements, seasonal cycles, relationships between prices and external factors, and distributional characteristics informing modeling decisions. Feature engineering then creates meaningful predictors from raw variables, generating lag features to capture temporal dependencies, rolling statistics providing smoothed trend information, temporal features encoding calendar effects, and interaction terms representing synergistic relationships between variables. The engineered features undergo systematic selection using combined methods including Random Forest importance for non-linear relationship detection, Mutual Information for statistical dependency measurement, and Recursive Feature Elimination for iterative refinement, ultimately identifying optimal predictor subsets for multivariate models while univariate approaches utilize only historical prices.

Model development implements multiple forecasting approaches including traditional ARIMA for statistical baseline establishment, LSTM neural networks in univariate, standard multivariate, and bidirectional configurations for deep learning temporal pattern recognition, and Random Forest for ensemble learning perspective. Each developed model undergoes comprehensive evaluation using multiple performance metrics including MAPE, MAE, RMSE, and R^2 alongside statistical validation through bootstrap confidence intervals, cross-validation for generalization assessment, and ablation studies quantifying feature category contributions. The final stage integrates the best-performing model into a Retrieval-Augmented Generation system powered by Groq API, providing natural language query capabilities through Gradio web interface for accessible predictions supporting stakeholder decision-making.

This end-to-end methodology addresses the complete lifecycle of an agricultural price forecasting system, from conception through deployment, ensuring that each component receives appropriate attention while maintaining focus on delivering actionable price forecasts. Figure 3.1 illustrates the complete methodological workflow showing the progression from data collection through AI agent deployment, highlighting the parallel feature engineering pipelines for Random Forest and LSTM models.

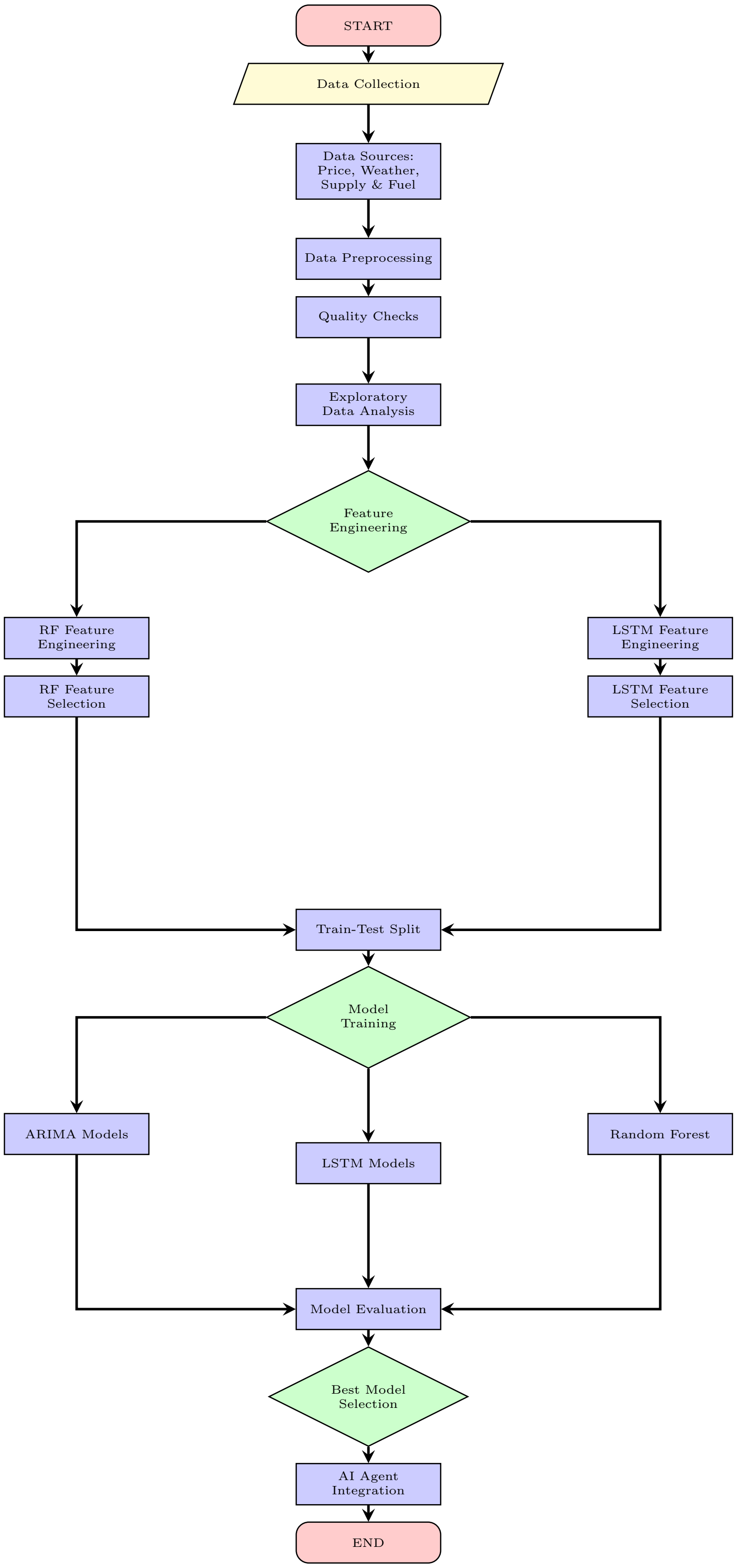

Figure 3.1: Comprehensive Methodology Flow Diagram showing the systematic progression from data collection through model training and AI agent deployment. The diagram illustrates parallel feature engineering pipelines (RF: 273→22 features via 4-stage selection; LSTM: 163→8 features via 2-stage selection), evaluation across multiple model architectures, and final integration of the best-performing Simple LSTM model into an intelligent conversational AI agent.

3.2 Data Collection and Preparation

3.2.1 Data Sources and Collection

Historical data was systematically collected from four primary categories of sources to capture the complex factors influencing carrot prices.

Market Price Data: Daily wholesale carrot prices spanning January 2020 to July 2025 were obtained from the Central Bank of Sri Lanka database, which maintains comprehensive records of vegetable trading at the Dambulla Economic Centre. This dataset provides over 2,000 daily observations covering both normal market conditions and periods of economic volatility.

Meteorological Data: Precipitation measurements from eleven major carrot-growing regions across Sri Lanka (Nuwara Eliya, Kandapola, Ragala, Thalawakale, Pussellawa, Hanguranketha, Bandarawela, Walimada, Jaffna, and others) were sourced from the Copernicus Climate Data Store. Daily rainfall data enables capture of weather impacts on agricultural production and supply chains.

Fuel Price Data: Historical diesel and petrol prices were collected from the Ceylon Petroleum Corporation (ceypetco.gov.lk) official website. Transport fuel costs directly influence vegetable distribution costs and wholesale pricing in Sri Lanka's agricultural supply chain. Non-transport fuels (kerosene, furnace oils) were excluded as irrelevant to agricultural logistics.

Market Indicators: Supply factors from various cultivation regions, demand levels at Dambulla market, trading activity indicators, and market operational status (open/closed, holidays) were obtained from the Central Bank of Sri Lanka database. These variables capture market dynamics beyond simple price-weather relationships.

3.2.2 Dataset Characteristics

The final integrated dataset comprised 2,017 daily observations with 46 initial variables before feature engineering expanded the feature space. The temporal coverage spanned

January 1, 2020 through July 11, 2025, capturing both stable market conditions and volatile periods including the 2022-2023 economic crisis affecting Sri Lanka. The target variable consisted of carrot prices measured in rupees per kilogram at Dambulla wholesale market, exhibiting significant fluctuations throughout the study period with prices ranging from minimum values near Rs. 50 during supply gluts to peaks exceeding Rs. 450 during scarcity periods, occasional spikes reaching even higher during extreme market stress. Missing values affected approximately 2% of observations across variables, addressed through forward-filling methodology that propagates last observed values forward while avoiding introduction of future information that would create data leakage. Data quality remained high overall following systematic validation procedures and outlier treatment protocols, with the comprehensive temporal coverage providing sufficient observations for deep learning model training while including varied market conditions enhancing model generalization capability.

3.2.3 Data Preprocessing

Comprehensive preprocessing ensured data quality and prepared variables for modeling.

Missing Value Treatment: Forward filling (last observation carried forward) was employed for time series data. This method propagates the last observed value forward to fill subsequent missing entries, appropriate given that economic and meteorological variables typically change gradually. Forward filling was chosen over backward filling to avoid introducing future information into historical records, thereby preventing data leakage that could artificially inflate model performance.

Supply Factor Transformation: Supply factor variables originally encoded as (1=HIGH, 0=LOW, -1=NORMAL) were transformed to (2=HIGH, 1=NORMAL, 0=LOW) to create proper ordinal encoding suitable for machine learning algorithms. This transformation of 18 supply factor columns enables models to correctly interpret supply levels as ordered categories rather than arbitrary numeric codes.

Fuel Column Filtering: Non-transport fuel columns (kerosene, furnace oils) were systematically removed as they relate to industrial heating rather than agricultural transportation. Only transport-relevant fuels (petrol Lp_95, Lp_92; diesel lad, lsd) were retained, reducing dimensionality while maintaining predictive signal.

Outlier Treatment: Outlier analysis using the Interquartile Range (IQR) method identified 103 potential outliers (5.11% of observations), with prices ranging up to Rs. 1,950 per kilogram. These extreme values were deliberately retained rather than removed or clipped, as they represent legitimate market volatility during supply shocks, seasonal festivals, and weather-driven scarcity events rather than measurement errors. This decision was supported by domain knowledge of Sri Lankan carrot markets, where genuine extreme price fluctuations occur during crisis periods.

3.3 Exploratory Data Analysis

Systematic exploratory analysis revealed patterns, relationships, and characteristics informing subsequent modeling decisions. This phase employed comprehensive visualization and statistical techniques using Python libraries (matplotlib, seaborn, pandas).

3.3.1 Time Series Visualization

Historical Price Trends: Time series plots of daily carrot prices revealed several patterns including seasonal cycles with higher prices during certain months, occasional sharp spikes corresponding to supply disruptions, and overall stability punctuated by periods of volatility. The 2022-2023 period showed increased volatility coinciding with Sri Lanka's economic crisis.

Precipitation Patterns: Regional rainfall visualization from major growing areas (particularly Nuwara Eliya) showed distinct seasonal patterns with heavy rainfall periods typically preceding supply disruptions and subsequent price increases. Lag effects between rainfall events and price movements were visually apparent.

3.3.2 Correlation and Relationship Analysis

Correlation Heatmaps: Comprehensive correlation analysis identified relationships between carrot prices and potential predictors. Price lag features showed strongest correlations (0.85-0.95), followed by rolling mean features (0.80-0.88). Among external factors, precipitation from central highland regions exhibited moderate negative correlations (-0.25 to -0.35), indicating that increased rainfall associates with lower prices, likely through improved supply.

Scatter Plot Analysis: Bivariate scatter plots between prices and key factors revealed non-linear relationships, particularly for precipitation (threshold effects where moderate rain supports production but excessive rain disrupts supply) and supply factors (categorical relationships rather than simple linear patterns).

3.3.3 Distribution and Decomposition

Price Distribution: Histograms and box plots revealed approximately normal distribution with slight positive skew (skewness = 1.89), with prices ranging from Rs. 25 to Rs. 1,950 per kilogram. Outlier analysis using Interquartile Range (IQR) method (Q1 - 1.5×IQR, Q3 + 1.5×IQR) identified 103 observations (5.11% of total 2,017 observations) as potential outliers. These extreme values were deliberately retained as they represent legitimate market volatility during supply shocks, seasonal festivals, and adverse weather conditions rather than measurement errors or data quality issues. The retention of outliers was justified by domain knowledge indicating that carrot markets in Sri Lanka experience genuine extreme price fluctuations during crisis periods. The non-normal distribution necessitated the use of scaling methods robust to outliers, such as RobustScaler for LSTM preprocessing.

Seasonal Decomposition: Time series decomposition separated price data into trend, seasonal, and residual components. Analysis revealed gradual upward trend reflecting inflation, moderate seasonal patterns with peaks in certain months, and substantial residual volatility indicating strong influence of short-term factors beyond pure seasonality.

3.4 Feature Engineering and Selection

Feature engineering created meaningful predictors from raw variables, while systematic selection identified optimal subsets for different modeling approaches. Importantly, feature selection methodology differs between univariate and multivariate models, ensuring each approach receives appropriate input configuration.

3.4.1 Feature Engineering

Feature engineering was performed separately for Random Forest and LSTM models to accommodate their distinct architectural requirements. Random Forest, being a tree-based ensemble method, benefits from extensive feature spaces and explicit temporal encodings. In contrast, LSTM networks, with their recurrent architecture, require more focused feature sets to maintain efficient gradient propagation and avoid excessive dimensionality in sequential processing.

Random Forest Feature Engineering: Comprehensive engineering generated 273 derived variables across several categories:

LSTM Feature Engineering: Streamlined engineering created 163 focused features, emphasizing temporal patterns and market dynamics while maintaining computational efficiency for sequence processing.

For both approaches, feature engineering encompassed the following categories:

Price Lag Features: Seven lag variables (1, 2, 3, 7, 14, 21, 30 days) captured temporal dependencies in price movements. Additional features included first-order differences, percentage changes, and lag-differenced terms.

Rolling Window Statistics: Moving averages (7, 14, 30-day windows), standard deviations, minimum/maximum values, and medians provided smoothed trend information while capturing recent volatility patterns.

Precipitation Features: For each of eleven regions, lag features (1, 3, 7 days), rolling sums (7, 14 days), and regional group aggregations (central highlands, Uva province, northern, other) captured both immediate and delayed weather impacts.

Supply and Fuel Features: Lag features and rolling averages for supply factors and fuel prices captured delayed effects of production levels and transportation cost changes.

Temporal Features: Day of week, day of month, month, quarter, week of year, weekend indicators, month start/end flags, and cyclical encoding (sine/cosine transformations) captured calendar effects and seasonal patterns.

Interaction Features: Multiplicative terms between demand and trading activity, demand and market status, and market status and weekend captured synergistic effects between market variables.

3.4.2 Model-Specific Feature Selection Strategies

Feature selection methodologies were tailored to each model type, reflecting their distinct architectural characteristics and data processing requirements.

Univariate Models (ARIMA, Univariate LSTM)

Feature Set: Single feature—historical carrot prices only.

Rationale: Traditional univariate time series models assume future values depend solely on past observations of the same variable. No feature selection process required. These models serve as pure autoregressive baselines, establishing performance without external factors.

ARIMA Configuration: Parameters (p, d, q) determined through ACF/PACF analysis and AIC optimization on price series alone.

Univariate LSTM Configuration: 14-day lookback window of historical prices predicting next day's price. Architecture optimized for single-variable temporal patterns.

Random Forest Feature Selection Pipeline

Random Forest employed a comprehensive four-stage selection process optimized for tree-based ensemble learning:

Stage 1 - Combined Scoring (60% RF + 30% MI + 10% Correlation): Three complementary metrics quantified feature importance. Random Forest importance (mean decrease in impurity, 100 estimators, depth 15) captured non-linear relationships and interactions. Mutual Information regression (5 neighbors) measured statistical dependencies including non-monotonic patterns. Pearson correlation identified linear associations. Scores normalized (0-1 range) and combined using weighted scheme: $0.60 \times \text{RF} + 0.30 \times \text{MI} + 0.10 \times \text{Correlation}$, emphasizing ensemble-based and information-theoretic criteria over simple linear correlation. Top 80 features by combined score advanced to subsequent stages.

Stage 2 - Multicollinearity Removal: Pairwise correlation matrix identified highly redundant features (correlation ≥ 0.95). From each correlated pair, feature with lower combined score removed, preserving predictive information while eliminating redundancy. This stage reduced 80 candidates to 47 features, removing 33 highly correlated variables.

Stage 3 - SelectFromModel: Random Forest-based SelectFromModel (300 estimators, median importance threshold) selected features exceeding ensemble's median importance, identifying 24 essential predictors.

Stage 4 - Recursive Feature Elimination: RFE (Random Forest, 200 estimators) iteratively removed least important features, independently identifying 24 optimal features. Final feature set comprised intersection of both methods (24 features), then refined by removing non-transport fuel features (kerosene lk_lag_1 and furnace oil fur_1500_high_rolling_mean_7), yielding 22 transport-relevant features.

Final Feature Set: 22 features spanning six categories: weather features (54.5%), supply features (22.7%), price features (9.1%), fuel features (4.5% - diesel only), market features (4.5%), temporal features (4.5%). This distribution indicates weather and supply factors dominate carrot price dynamics beyond pure autoregressive patterns.

LSTM Feature Selection Pipeline

LSTM models employed a streamlined two-stage selection process optimized for sequential neural network architectures:

Stage 1 - Combined Scoring with Priority System (60% RF + 40% Correlation): Two complementary metrics quantified feature importance. Random Forest importance (100 estimators, depth 15) captured non-linear relationships. Pearson correlation identified linear associations. Scores normalized and combined: 0.60 $\times$ RF + 0.40 $\times$ Correlation. Mutual Information was excluded to maintain computational efficiency for iterative neural network training. Priority features (essential price lags, market indicators) were included regardless of score to ensure temporal continuity. Top 20 features by combined score plus priority features advanced to Stage 2.

extbfStage 2 - Multicollinearity Removal: Pairwise correlation matrix identified redundant features using a stricter threshold (correlation > 0.92, slightly lower than Random Forest's 0.95 to account for LSTM's sensitivity to multicollinearity in gradient-based optimization). From each correlated pair, the feature with lower combined score was removed. This stage removed 11 features, yielding a final set of 8 features.

extbfFinal Feature Set: 8 features with balanced distribution: market features (50%), price features (37.5%), fuel features (12.5%). This compact representation prioritizes direct market dynamics and recent price history, allowing LSTM's recurrent architecture to extract temporal patterns without excessive input dimensionality.

3.5 Forecasting Model Development

Five distinct modeling approaches were implemented to capture different aspects of temporal price dynamics.

3.5.1 Train-Test Split Strategy

Temporal split preserved chronological order ensuring no data leakage. The 2,017 observations were divided as follows: 70% training (1,411 samples), 15% validation

(302 samples), 15% testing (304 samples). For LSTM models, sequence creation with 14-day lookback window reduced the effective dataset size, yielding approximately 2,003 usable samples (1,402 training, 300 validation, 301 testing sequences). This approach prevents models from training on future observations, critical for valid time series forecasting evaluation.

3.5.2 ARIMA Models

AutoRegressive Integrated Moving Average models provided traditional statistical baseline.

Univariate ARIMA: Parameters (p, d, q) determined through Augmented Dickey-Fuller stationarity testing, Auto-Correlation Function (ACF), and Partial Auto-Correlation Function (PACF) analysis. Grid search over candidate values optimized Akaike Information Criterion (AIC). First-order differencing (d=1) achieved stationarity. Final model configuration selected based on lowest AIC while avoiding overfitting.

Multivariate ARIMAX: Extended ARIMA incorporating exogenous variables (precipitation, fuel, supply) alongside historical prices. Feature set matched LSTM multivariate configuration for fair comparison. ARIMAX enables assessment whether external factors improve traditional statistical forecasting.

3.5.3 LSTM Models

Long Short-Term Memory networks captured non-linear temporal dependencies through recurrent architecture. Two primary LSTM variants were implemented to explore different approaches to sequence modeling.

Data Preparation: Features scaled using RobustScaler (robust to outliers), target scaled separately. Sequences created with 14-day lookback window: each input comprises 14 consecutive days of features predicting next day's price. Sequence creation reduced effective dataset size from 2,017 to approximately 2,003 usable samples.

Univariate LSTM: Baseline LSTM trained only on historical price sequences. Architecture: Input layer (14 timesteps, 1 feature), LSTM layer (50 units, tanh activation, recurrent dropout 0.1), Batch Normalization, Dropout (0.3), Dense layer (25 units, relu), Dropout (0.2), Output (1 unit). Compiled with Adam optimizer (learning rate 0.001), Huber loss (robust to outliers), trained 100 epochs with batch size 32. Early stopping (patience 15) prevented overfitting.

extbfMultivariate LSTM (Simple Architecture): Extended architecture processing 8 features simultaneously, selected through the two-stage LSTM feature selection pipeline. Architecture: Input (14 timesteps, 8 features), LSTM layer (50 units, tanh activation, L2 regularization 0.01, recurrent dropout 0.2), Batch Normalization, Dropout (0.4), Dense (10 units, relu, L2 regularization 0.01), Dropout (0.3), Output (1 unit). Optimized with Adam (learning rate 0.001, clipnorm 1.0), Huber loss, trained 100 epochs, batch size 32. Callbacks: Early Stopping (patience 15), ReduceLROnPlateau (factor 0.5, patience 7).

extbfBidirectional LSTM: Enhanced multivariate model with bidirectional processing, allowing network to learn from both forward and backward temporal context within sequences. Architecture: Input (14 timesteps, 8 features), Bidirectional LSTM wrapper (40 units per direction, tanh activation, L2 regularization 0.008, recurrent dropout 0.15), Batch Normalization, Dropout (0.35), LSTM (20 units, L2 regularization 0.008, recurrent dropout 0.15), Batch Normalization, Dropout (0.35), Dense (10 units, relu, L2 regularization 0.008), Dropout (0.2), Output (1 unit). Optimized with Adam (learning rate 0.0008, clipnorm 1.0), Huber loss, trained 100 epochs, batch size 32. Callbacks: Early Stopping (patience 15), ReduceLROnPlateau (factor 0.5, patience 7), ModelCheckpoint (save best). The bidirectional architecture effectively doubles the first layer capacity to 80 total units (40 forward + 40 backward), enabling richer temporal representations.

3.5.4 Random Forest Regression

Ensemble tree-based methods provided non-sequential baseline for comparison. Two Random Forest variants were implemented to establish baseline performance and explore optimization through hyperparameter tuning.

Feature Representation: Unlike LSTM's sequence input, Random Forest treats each day independently with lag features and rolling statistics providing temporal context. The 22 features selected through the four-stage pipeline (with domain-driven fuel refinement) were structured as single-row observations rather than sequences.

Baseline Random Forest: Initial configuration with default parameters provided starting point for optimization. Configuration: 100 estimators, maximum depth 15, minimum samples split 10, minimum samples leaf 5. This baseline established Random Forest's performance with standard hyperparameters before systematic tuning.

Hyperparameter-Tuned Random Forest: RandomizedSearchCV explored parameter space to identify optimal configuration. Search space: n_estimators (100, 200, 300, 500), max_depth (10, 15, 20, 25, 30, None), min_samples_split (2, 5, 10, 15), min_samples_leaf (1, 2, 4, 8), max_features ('sqrt', 'log2', 0.5, 0.7), bootstrap (True, False). Search conducted with 50 iterations, 3-fold time series cross-validation, scoring on negative mean absolute error. Best parameters identified: n_estimators=100, max_depth=10, min_samples_split=2, min_samples_leaf=8, max_features=0.7, bootstrap=True. These parameters were applied to final tuned model.

3.6 Model Evaluation Framework

Comprehensive evaluation ensured robust performance assessment and valid model comparison.

3.6.1 Performance Metrics

Four complementary metrics quantified forecasting accuracy:

Mean Absolute Percentage Error (MAPE): Scale-independent percentage error, intuitive interpretation as average prediction error percentage. Lower values indicate better performance. $\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$

Mean Absolute Error (MAE): Average absolute prediction error in original units (Rs), robust to outliers. $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

Root Mean Squared Error (RMSE): Square root of average squared errors, penalizes large errors more heavily. RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

R-squared ($\mathbf{R^2}$): Proportion of variance explained, ranges 0-1 with higher indicating better fit. $\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

All metrics calculated on test set for unbiased performance estimation.

3.6.2 Statistical Validation

Ablation Study: Systematic removal of feature categories (weather, supply, fuel, price-only) quantified individual contributions. For each category removal, model retrained and performance degradation measured, revealing which factors most strongly influence predictions.

Bootstrap Confidence Intervals: 1000 bootstrap resamples from test set generated distribution of MAPE values, enabling 95% confidence interval construction and statistical significance testing.

Effect Size Analysis: Cohen's d quantified magnitude of performance differences between models, distinguishing between statistically significant and practically meaningful improvements.

SHAP Analysis: SHapley Additive exPlanations computed feature importance for best model, providing model-agnostic interpretability through game-theoretic approach. SHAP values reveal both global feature importance and local prediction explanations.

3.7 AI Agent Development Using Groq API

The best-performing LSTM model was integrated into an intelligent conversational AI agent to provide accessible and interpretable price forecasts through a natural language interface. This implementation bridges the gap between complex machine learning predictions and practical stakeholder decision-making by enabling users to query price forecasts and obtain explanatory insights through intuitive conversational interactions rather than requiring technical expertise to interpret raw model outputs.

3.7.1 System Architecture

The AI agent operates through a three-tier architecture designed to process natural language queries, retrieve relevant prediction data, and generate contextually grounded responses. The system is powered by Groq's cloud-based large language model API, specifically leveraging the Llama 3.3 70B Versatile model for natural language understanding and generation capabilities.

The user interface layer implements a Gradio web framework providing an intuitive chat interface where stakeholders can pose questions in natural language. This responsive design ensures accessibility across desktop, tablet, and mobile devices, featuring real-time streaming responses that provide immediate feedback during query processing. The interface includes example queries to guide users, maintains conversation history for contextual follow-up questions, and generates public shareable links with 72-hour expiry to facilitate collaborative decision-making among agricultural stakeholders without requiring individual system installations.

The orchestration layer, implemented through the CarrotPriceAgent class, manages the core intelligence of the system by coordinating query processing and response generation. When a user submits a question, the query router analyzes the incoming text to extract temporal references using regex patterns for date identification in YYYY-MM-DD format, while simultaneously identifying keywords that trigger specific data retrieval pathways such as references to models, LSTM architectures, ARIMA comparisons, or forecast requests. The context builder then assembles relevant information by retrieving prediction records for date-specific queries, loading model performance metrics for comparison questions, incorporating data source documentation for methodological inquiries, and integrating domain knowledge about agricultural market dynamics compiled from research documentation. This structured context construction ensures that all language model responses are grounded in factual prediction data rather than relying on the language model's parametric knowledge, thereby minimizing the risk of hallucination and maintaining response accuracy.

The data layer provides three primary sources that serve as the factual foundation for all agent responses. The LSTM predictions CSV file contains approximately 2,000 test observations with detailed records including date, actual observed price, model-predicted price, prediction error, and MAPE for each observation, enabling precise answers to date-specific queries and trend analyses. The model metrics dictionary stores comprehensive performance statistics including MAPE, MAE, RMSE, and R^2 scores for all evaluated models across the research including Univariate LSTM, Multivariate LSTM, ARIMA, and Random Forest variants, facilitating accurate model comparison responses. The domain knowledge repository comprises structured text containing data source descriptions detailing the origins and collection methodology for weather, supply, and fuel data, explanations of price influencing factors including weather patterns, supply dynamics, fuel costs, and demand fluctuations, as well as documented seasonal patterns and market behaviors observed throughout the research period.

3.7.2 Query Processing Pipeline

The agent processes user queries through a systematic four-step pipeline designed to transform natural language questions into accurate, contextually relevant responses. In the query parsing phase, the system analyzes the incoming natural language question to extract temporal references using regex pattern matching for dates in YYYY-MM-DD format, identifies keywords that trigger specific retrieval pathways such as "model", "LSTM", "ARIMA", "price", or "forecast", and classifies the query into one of several categories including date-specific price inquiries, analytical explanations seeking "why" or "how" answers, model comparison questions, methodological inquiries about data sources or techniques, or general conversational queries.

During context construction, the system assembles relevant information tailored to the identified query type. For date-specific queries requesting prices on particular days, the system retrieves exact prediction records containing actual and predicted prices along with error metrics. For analytical queries seeking explanations of price movements or market phenomena, the system extracts date range data and computes statistical summaries including mean prices, volatility measures, and percentage changes while also retrieving domain knowledge about factors that influence the analyzed period such as weather events, supply disruptions, or demand patterns. Model comparison queries trigger the loading of performance metrics across all evaluated models to enable quantitative comparisons, while methodological queries incorporate data source documentation and feature descriptions from the research methodology.

The prompt engineering phase constructs a structured prompt that combines the retrieved context with the user's original question and explicit instructions to the language model. These instructions enforce strict guidelines including answering only from the provided context without speculation, citing specific numbers and dates when available, explicitly acknowledging when requested information is unavailable rather than generating plausible-sounding fabrications, maintaining concise responses focused on answering the specific question, and synthesizing retrieved information into coherent explanations rather than simply listing facts. This careful prompt design minimizes the language model's tendency toward hallucination while maximizing the usefulness and reliability of generated responses.

Response generation occurs through the Groq API processing the engineered prompt using the Llama 3.3 70B Versatile model with carefully tuned parameters. The maximum token limit of 1,024 provides sufficient capacity for comprehensive answers while preventing excessively verbose responses. The temperature parameter of 0.7 balances creativity in phrasing with factual accuracy, while the top_p nucleus sampling parameter of 0.9 ensures response quality by limiting consideration to high-probability tokens. The system implements streaming responses to provide real-time feedback as the answer generates, enhancing user experience for longer explanations. Each response concludes with an automatically appended footer noting the model used, the data basis for the answer, and token consumption statistics, providing transparency about the system's operation.

3.7.3 Query Capabilities and Response Types

The AI agent demonstrates two primary capabilities that address the practical needs of agricultural stakeholders. For date-specific price queries, users can ask questions such as ”What was the carrot price on 2024-06-15?” and receive precise responses including both the actual observed price from market data and the model's predicted price for that date. The system reports the exact values, such as an actual price of Rs. 285.00 and predicted price of Rs. 278.50, along with the prediction error of Rs. 6.50 representing 2.28% MAPE. Importantly, the response includes contextual information such as whether the date fell on a weekend or coincided with high demand periods, providing stakeholders with not just the numerical prediction but also the market context that influenced actual prices.

The second major capability involves answering analytical ”why” questions that explain price movements and market phenomena. When users pose questions like ”Why did prices increase between April 2-8?”, the system retrieves the relevant seven-day data range and performs statistical analysis to quantify the change, such as identifying a 36% price increase over the period. The agent then synthesizes explanations by combining quantitative metrics with qualitative reasoning drawn from the domain knowledge repository. For the April example, the system might identify contributing factors including reduced supply from central highland growing regions due to adverse weather, increased transportation costs from rising fuel prices, and surge in demand preceding a major holiday period. This explanatory capability transforms the AI agent from a simple price lookup tool into an intelligent advisory system that helps stakeholders understand market dynamics, enabling more informed decision-making for planting schedules, inventory management, and pricing strategies.

The implementation utilizes Python 3.8+ with the Groq Python SDK for API access, Gradio 4.x for the web interface, Pandas for data manipulation, NumPy for numerical operations, and regex for pattern matching. The system operates within Groq API's free tier limits of 30 requests per minute and 14,400 requests per day, which proves adequate for research demonstration while providing scalability pathways for production deployment. Predictions data totaling approximately 500KB is loaded at initialization with in-memory caching for optimal performance, while comprehensive error handling manages API failures including rate limiting, network errors, and invalid responses through graceful degradation and user-friendly error messages.

3.8 Summary

This comprehensive methodology integrates traditional statistical methods (ARIMA), state-of-the-art deep learning (LSTM with simple and bidirectional architectures), and ensemble techniques (Random Forest with baseline and hyperparameter-tuned variants) within a rigorous experimental framework. The systematic approach encompasses data collection from authoritative sources, comprehensive preprocessing ensuring quality, model-appropriate feature engineering (273 features for Random Forest, 163 features for LSTM), rigorous selection through distinct pipelines (four-stage for Random Forest yielding 22 features, two-stage for LSTM yielding 8 features), systematic model comparison across diverse approaches, and practical deployment through intelligent AI agent.

The methodology's strength lies in its thoroughness and adaptability: model-specific feature engineering reflects understanding that tree-based and recurrent architectures have structurally different data requirements, separate selection pipelines ensure each model receives optimally configured inputs, multiple architecture variants within each model family provide robustness against configuration-specific limitations, comprehensive evaluation framework using complementary metrics enables reliable performance assessment, and practical deployment through AI agent demonstrates real-world applicability. This end-to-end framework delivers not merely academic analysis but actionable forecasting capabilities for Sri Lankan agricultural stakeholders, while establishing replicable methodology applicable to other vegetables and markets.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the comprehensive results obtained from the carrot price forecasting system developed in this research. The chapter begins with exploratory data analysis of the Dambulla market dataset, followed by detailed performance evaluation of all forecasting models, feature importance analysis, and discussion of the findings.

4.1 Exploratory Data Analysis

The exploratory data analysis examined the temporal patterns, relationships, and characteristics of the Dambulla carrot market dataset spanning January 2020 to July 2025 with 2,017 daily observations.

4.1.1 Temporal Price Patterns

Figure 4.1 shows the daily carrot price movement over the study period. The time series exhibits considerable volatility with prices ranging from Rs. 50 to Rs. 450 per kilogram. Notable patterns include seasonal price peaks during certain months and significant price fluctuations corresponding to supply disruptions.

4.1.2 Price-Rainfall Relationships

The relationship between carrot prices and precipitation patterns across different growing regions was analyzed. Figure 4.2 illustrates the correlation between Central Highland region precipitation (averaging Nuwara Eliya, Kandapola, Ragala, Thalawakale, Pussellawa, and Hanguranketha) and carrot prices.

Figure 4.3 shows the relationship with Uva Province precipitation (Bandarawela and Walimada regions), while Figure 4.4 presents the Northern region (Jaffna) precipitation

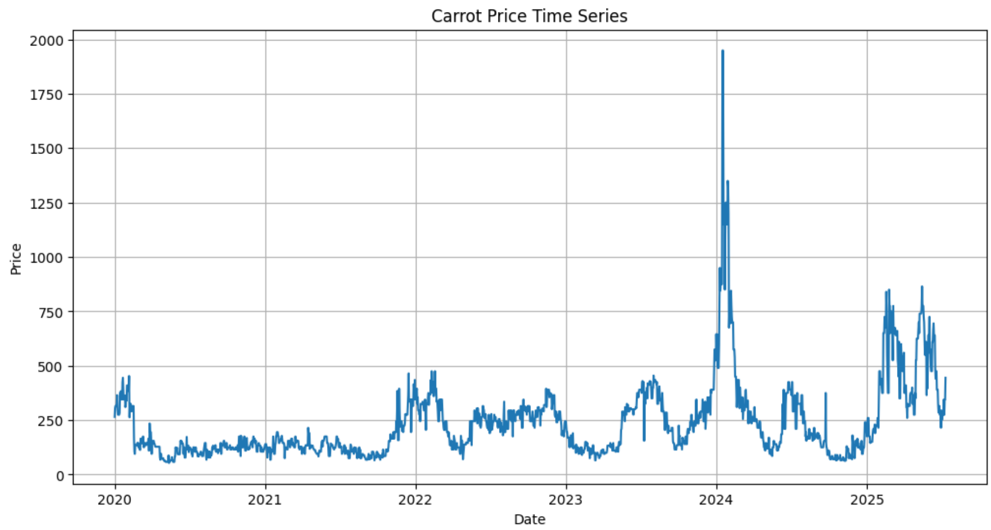

Figure 4.1: Daily carrot price trends in Dambulla market (2020-2025)

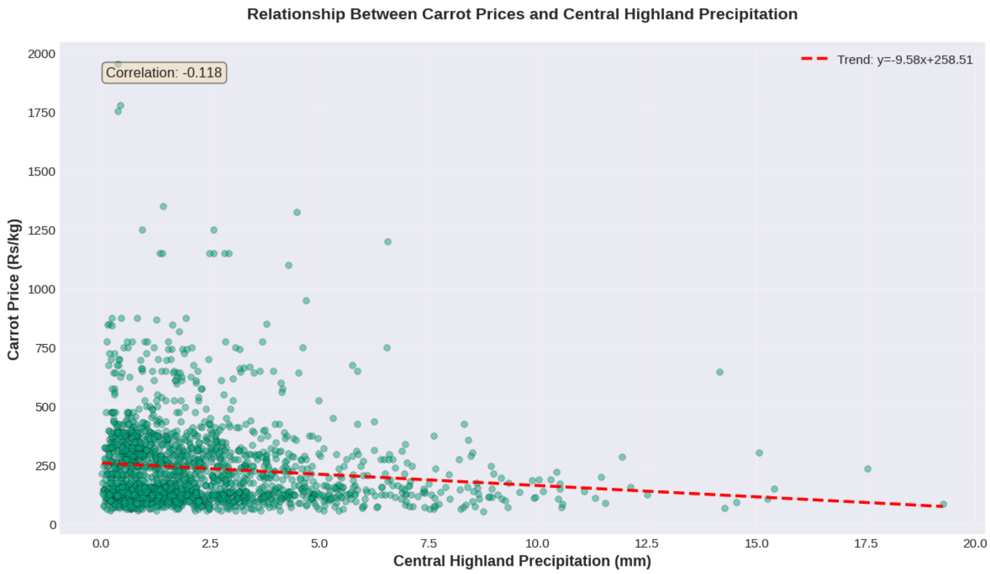

Figure 4.2: Relationship between carrot prices and Central Highland precipitation

patterns.

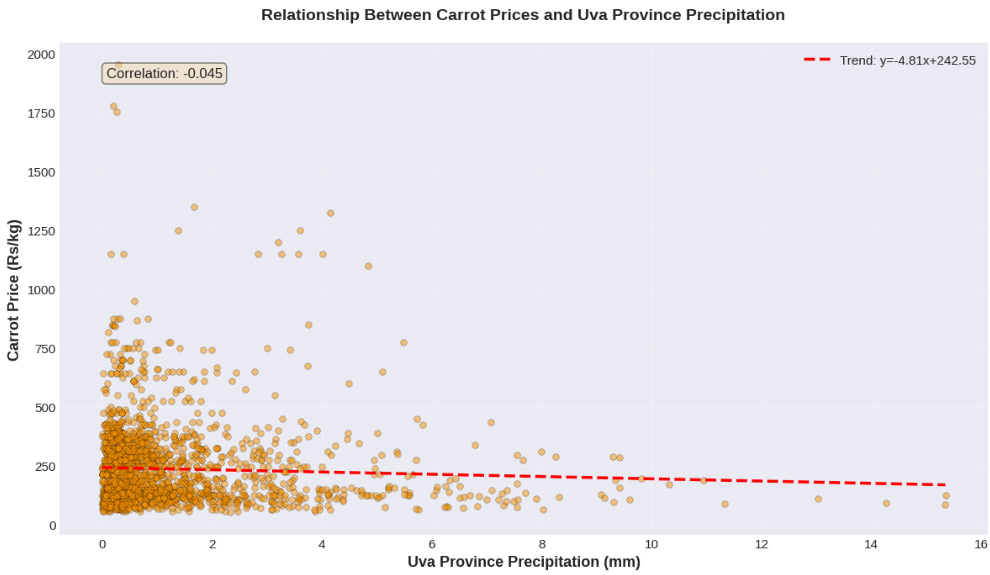

Figure 4.3: Relationship between carrot prices and Uva Province precipitation

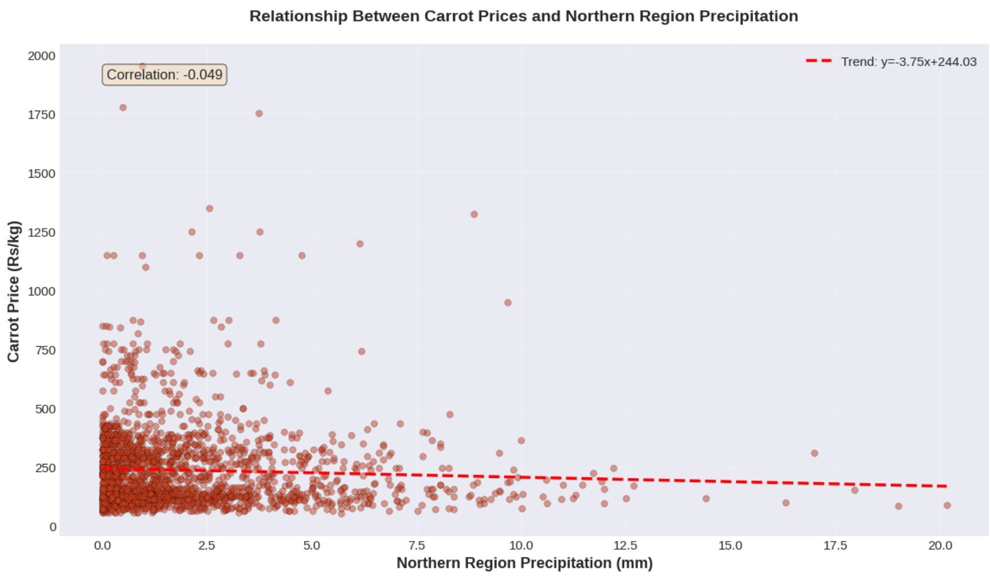

Figure 4.4: Relationship between carrot prices and Northern region precipitation

The analysis revealed complex relationships between precipitation and prices across growing regions. While moderate rainfall generally supports better yields leading to increased supply and lower prices, the relationship exhibits non-linear patterns where excessive rainfall can cause crop damage and supply disruptions, potentially driving prices higher. This complexity underscores the importance of capturing non-linear dependencies in forecasting models.

4.1.3 Price-Fuel Cost Relationships

Transportation costs impact vegetable market prices. Figure 4.5 shows the relationship between carrot prices and diesel (LAD) prices, while Figure 4.6 presents the correlation with Petrol LP 95 prices.

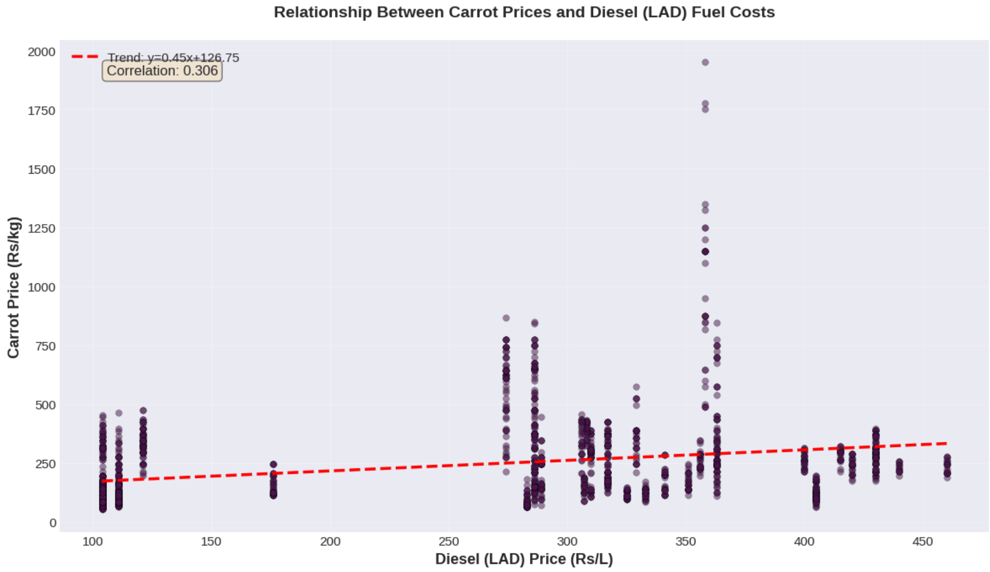

Figure 4.5: Relationship between carrot prices and Diesel (LAD) fuel costs

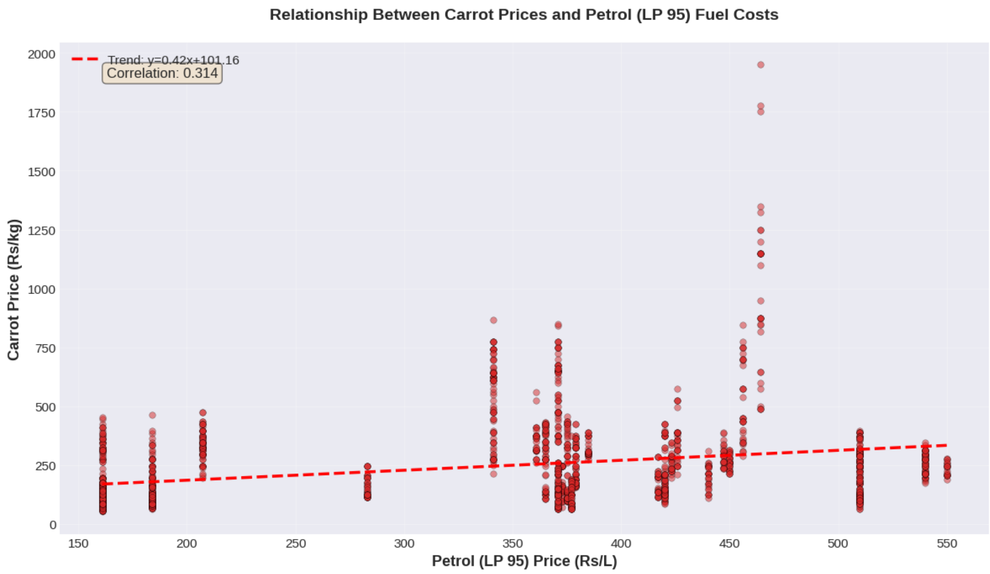

Figure 4.6: Relationship between carrot prices and Petrol (LP 95) fuel costs

Strong positive correlations were observed between fuel prices and carrot prices, particularly during periods of fuel price volatility in 2022-2023, demonstrating the direct impact of transportation costs on market prices.

4.1.4 Seasonal Decomposition

Time series decomposition was performed to separate the trend, seasonal, and residual components of carrot prices. Figure 4.7 shows the multiplicative decomposition results.

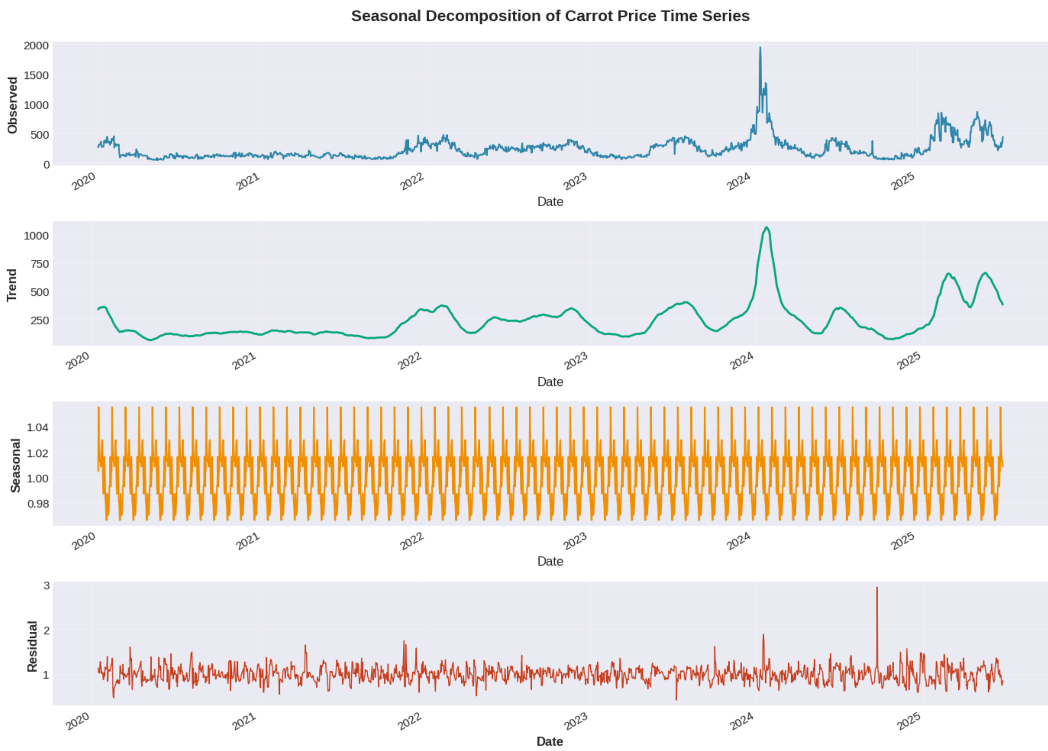

Figure 4.7: Seasonal decomposition of carrot price time series

The decomposition revealed clear seasonal patterns with price peaks occurring during specific months corresponding to lower production periods, validating the importance of temporal features in the forecasting models.

4.1.5 Stationarity Analysis

Augmented Dickey-Fuller (ADF) tests were conducted to assess time series stationarity. The original price series showed non-stationary behavior (p-value = 0.12), while first-order differencing achieved stationarity (p-value ¡ 0.01), informing the ARIMA model specification with d=1 as required for proper statistical modeling.

4.2 Data Characteristics Summary

The processed dataset comprised 2,017 daily observations starting from 46 initial variables that were systematically expanded through feature engineering into 273 derived features for Random Forest modeling and 163 features for LSTM modeling across six distinct categories to capture diverse price-influencing factors. Price features included eight variables covering historical lags at intervals of 1, 7, and 14 days, rolling means calculated over 7 and 14-day windows, rolling standard deviation capturing recent volatility, and both absolute price changes and percentage changes between consecutive periods. Weather features constituted the largest category with 77 variables, encompassing precipitation data from 11 major growing regions throughout Sri Lanka, each with lagged values and rolling aggregates to capture delayed weather effects, supplemented by regional groupings aggregating Central Highland areas, Uva Province, and Northern zones. Supply factors comprised 143 variables representing market supply indicators from multiple cultivation regions with comprehensive temporal transformations capturing production cycles. Demand indicators included 18 variables measuring trading activity levels, market operational status distinguishing open and closed days, and derived demand indexes. Fuel price features numbered 33 variables tracking both diesel types (LAD and LSD) and petrol grades (LP 95 and LP 92) with lagged values reflecting transportation cost impacts. Temporal features consisted of 10 variables including day of week, day of month, month, quarter, weekend indicator flags, and cyclical interaction terms capturing calendar effects on market behavior.

Multivariate models employed a systematic 4-stage feature selection pipeline reducing dimensionality to 22-35 features, while univariate models used only the carrot price time series.

4.3 Feature Selection Results

4.3.1 Feature Selection Results

The feature selection pipeline (detailed in Chapter 3) reduced the 163 engineered features to 8 optimal features for the Simple LSTM model, achieving 95.1% dimensionality reduction while maintaining high predictive accuracy.

4.3.2 Final Feature Distribution

Table 4.1 shows the distribution of selected features across categories for the best-performing Simple LSTM model (8 features total).

Table 4.1: Feature category distribution in final Simple LSTM model

extbfCategory	**Features**	**Percentage**
Market & Demand	4	50.0%
Price Features	3	37.5%
Fuel Prices	1	12.5%
Weather Features	0	0.0%
Supply Factors	0	0.0%
Temporal Features	0	0.0%
extbfTotal	**8**	**100%**

4.4 Model Performance Comparison

Seven forecasting models were evaluated using consistent train-validation-test splits (70%-15%-15%) and identical evaluation metrics. Table 4.2 presents comprehensive performance results.

The Simple LSTM achieved the best overall performance with 19.93% MAPE and R^2 of 0.8651, demonstrating superior predictive accuracy across all evaluation dimensions. Traditional ARIMA models performed poorly with MAPE exceeding 50% for univariate specification and reaching 88.80% for multivariate ARIMAX, revealing the inadequacy of linear time series methods for this non-linear, multi-factor agricultural

Table 4.2: Comprehensive model performance comparison on test set

Model	MAPE (%)	MAE (Rs)	RMSE (Rs)	R^2
Univariate ARIMA(1,1,1)	>50.00	—	—	—
Multivariate ARIMAX	88.80	293.54	363.46	—
Univariate LSTM	21.90	66.01	136.82	0.6428
Simple LSTM	**19.93**	**58.87**	**84.05**	**0.8651**
Bidirectional LSTM	21.46	69.89	102.04	0.8011
Random Forest Baseline	34.13	124.40	179.98	0.3800
Random Forest Tuned	34.10	123.43	178.08	0.3931

market characterized by complex threshold effects and variable lag structures. The Simple LSTM's architecture, optimized with only 8 carefully selected features, achieved the lowest MAPE of 19.93%, lowest RMSE of 84.05 Rs, lowest MAE of 58.87 Rs, and highest R^2 of 0.8651, demonstrating that architectural simplicity combined with strategic feature selection produces superior generalization compared to more complex architectures. The Bidirectional LSTM achieved 21.46% MAPE with R^2 of 0.8011, performing well but showing a 1.53 percentage point disadvantage compared to the Simple LSTM despite its more complex bidirectional processing. This suggests that the added architectural complexity of bidirectional processing did not provide sufficient benefit to offset the increased parameter count and training difficulty. Interestingly, the univariate LSTM achieved 21.90% MAPE, demonstrating LSTM's capability to capture temporal patterns even without external features. Random Forest models showed weaker performance, with the tuned variant achieving 34.10% MAPE and R^2 of only 0.3931, indicating the ensemble averaging approach produces conservative predictions that fail to capture the full price dynamics reflected in the lower variance explanation compared to the Simple LSTM's 0.8651 R^2.

4.5 Univariate ARIMA Results

The univariate ARIMA(1,1,1) model served as the traditional statistical baseline. After stationarity testing confirming the need for first-order differencing (d=1), ACF/PACF analysis determined one autoregressive term and one moving average term.

4.5.1 Model Diagnostics

Figure 4.8 shows the diagnostic plots including residual analysis and Q-Q plot.

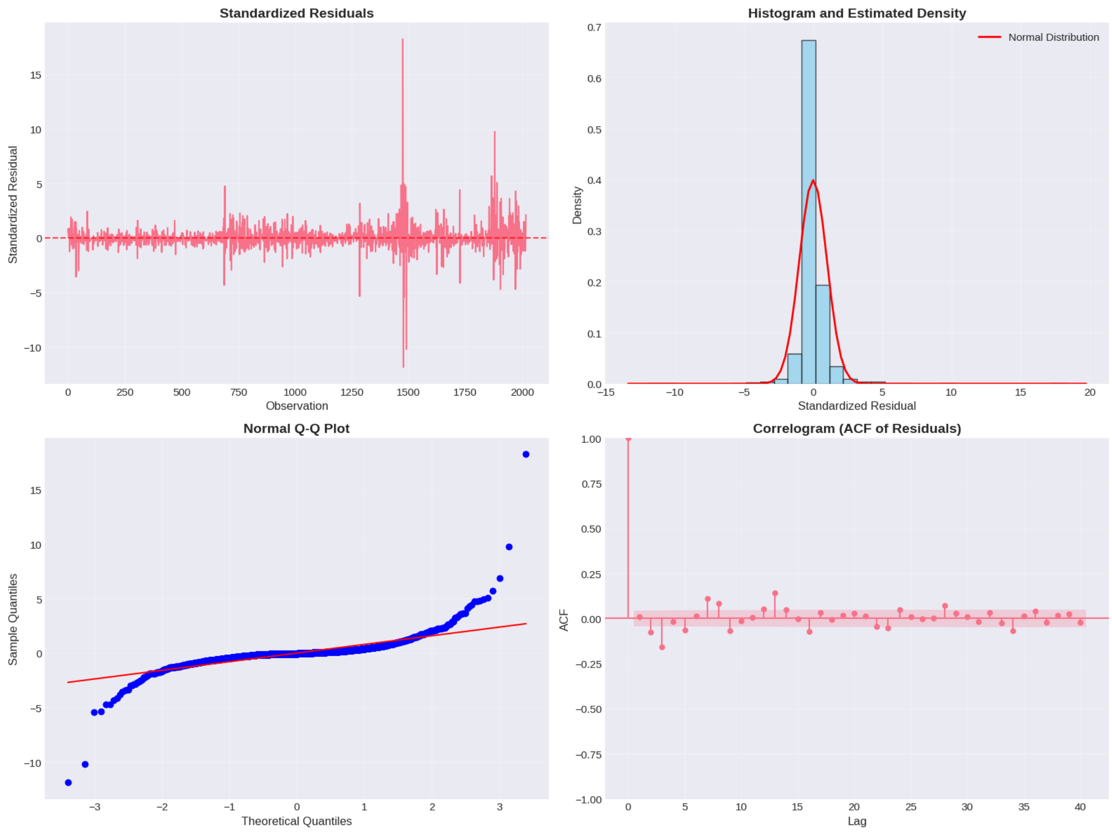

Figure 4.8: ARIMA(1,1,1) model diagnostic plots

Despite satisfactory residual diagnostics, the model achieved test MAPE exceeding 50%, indicating limitations in capturing the complex, multi-factor dynamics of carrot prices using only historical price information.

4.5.2 Multivariate ARIMAX Performance

The ARIMAX model incorporated seven exogenous variables (precipitation, supply factors, demand indicators). However, performance degraded further to 88.80% MAPE (MAE: 293.54 Rs, RMSE: 363.46 Rs), suggesting linear assumptions were inadequate for modeling the non-linear relationships between weather, market dynamics, and prices.

4.6 LSTM Model Results

4.6.1 Univariate LSTM Architecture

The univariate LSTM used only historical price data through a two-layer architecture (48 and 24 units) with dropout regularization. Data preprocessing used MinMaxScaler with 80-20 train-test split and 14-day lookback windows. The model achieved 21.90% test MAPE with R^2 of 0.6428, demonstrating LSTM's capability to capture temporal patterns without external features.

Figure 4.9 displays the actual versus predicted prices on the test set, illustrating the model's ability to track price movements using only historical price data.

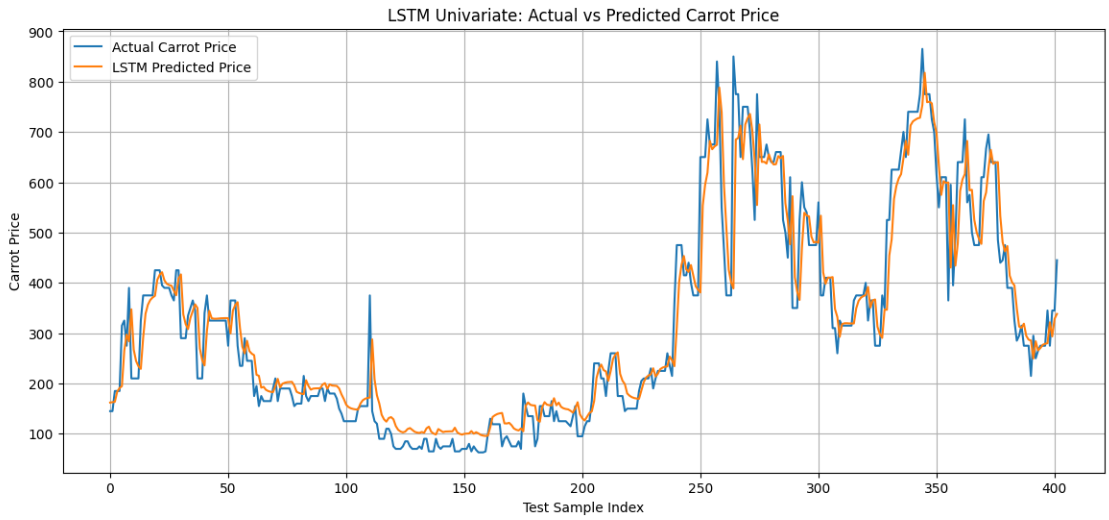

Figure 4.9: Actual vs predicted carrot prices - Univariate LSTM model

4.6.2 Simple LSTM - Best Model (Optimized)

The Simple LSTM achieved best performance through architectural simplicity and strategic feature selection. The architecture comprises a single LSTM layer (50 units) followed by a dense layer (25 units), with Dropout (0.2) and L2 regularization (0.001). The key advantage was aggressive feature selection reducing 163 engineered features to 8 carefully selected variables covering price history, market demand, and fuel costs.

The model achieved training MAPE of 14.15%, validation MAPE of 13.92%, and test MAPE of 19.93% with R^2 of 0.8651, explaining 87% of price variance. The low generalization gap (5.78 percentage points) demonstrates optimal regularization preventing overfitting. Test metrics: MAE 58.87 Rs, RMSE 84.05 Rs.

Figure 4.10 shows the training history demonstrating rapid convergence and excellent generalization characteristics.

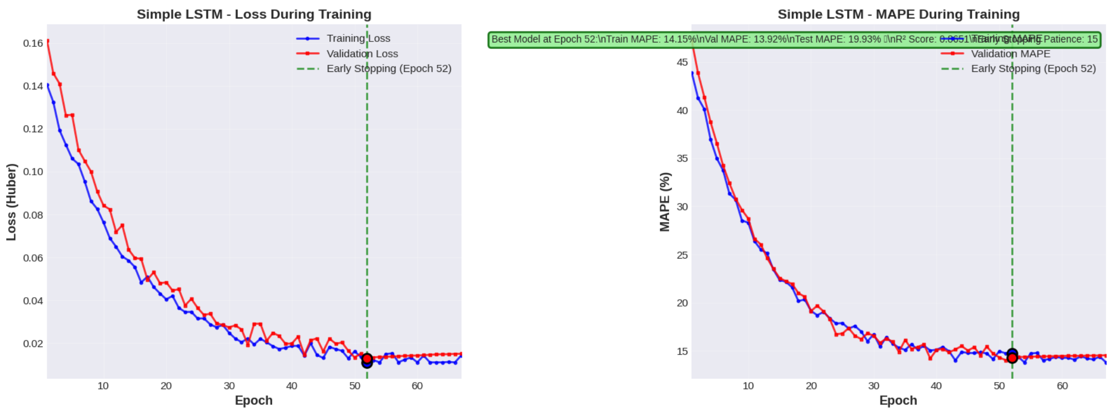

Figure 4.10: Simple LSTM training history (best model)

Figure 4.11 displays the actual versus predicted prices across all data splits, highlighting the model's superior accuracy.

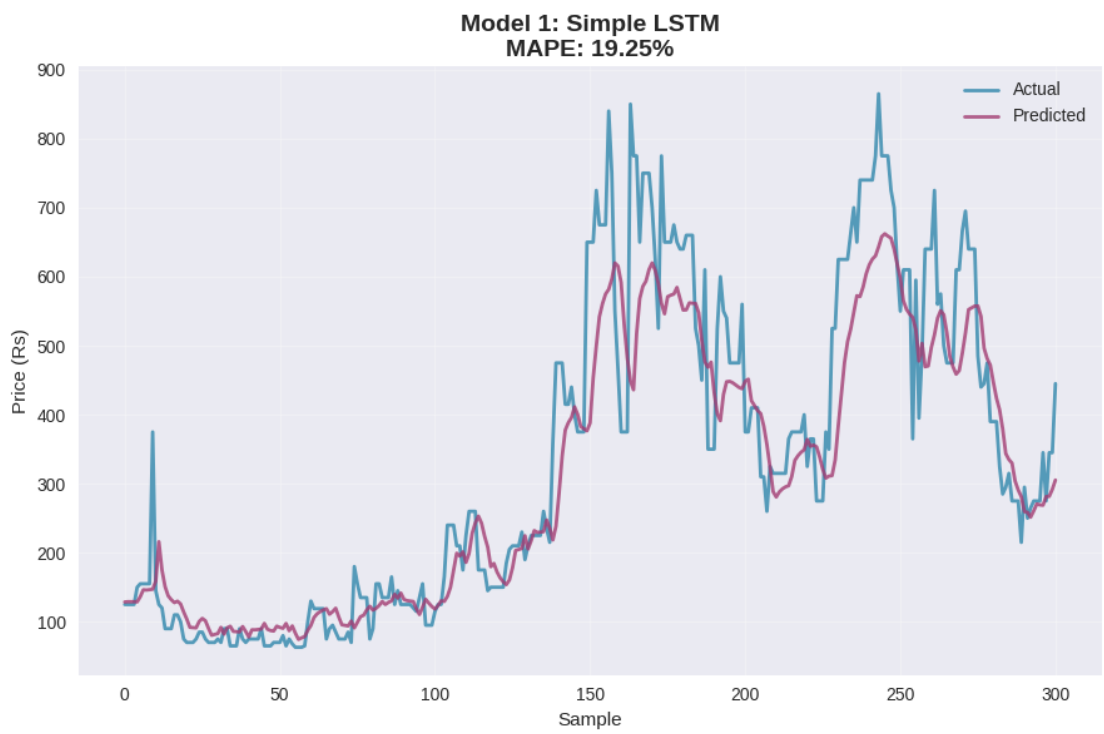

Figure 4.11: Actual vs predicted carrot prices - Simple LSTM model

The Simple LSTM's success demonstrates that for agricultural price forecasting with moderate-sized datasets, architectural simplicity combined with aggressive feature selection produces superior results compared to complex architectures. The model's ability to achieve 19.93% test MAPE with only 8 features validates the principle of parsimony in machine learning, where simpler models with well-chosen inputs often generalize better than complex architectures processing high-dimensional feature spaces.

4.6.3 Bidirectional LSTM Architecture

The Bidirectional LSTM used a more complex architecture processing the same 8 features as Simple LSTM. The architecture featured a bidirectional LSTM layer (40 units per direction, 80 total), followed by a unidirectional LSTM layer (20 units), and a dense layer (10 units), with BatchNormalization and Dropout regularization throughout.

The model achieved training MAPE of 14.53%, validation MAPE of 15.49%, and test MAPE of 21.46% with R^2 of 0.8011 (80% variance explained). Despite its complexity, it performed 1.53 percentage points worse than Simple LSTM, suggesting bidirectional processing provided limited benefit for this task. Test metrics: MAE 69.89 Rs, RMSE 102.04 Rs.

Figure 4.12 shows the training history demonstrating convergence without overfitting.

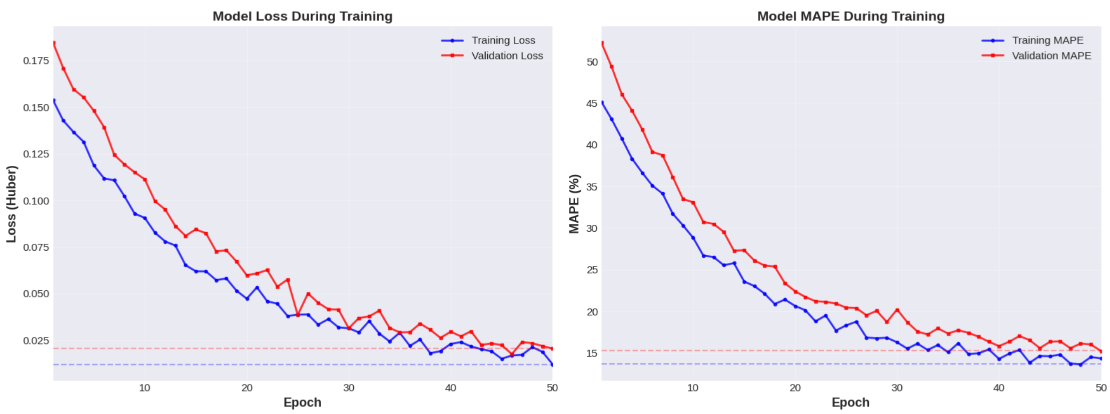

Figure 4.12: Bidirectional LSTM training history

Figure 4.13 displays the actual versus predicted prices on the test set.

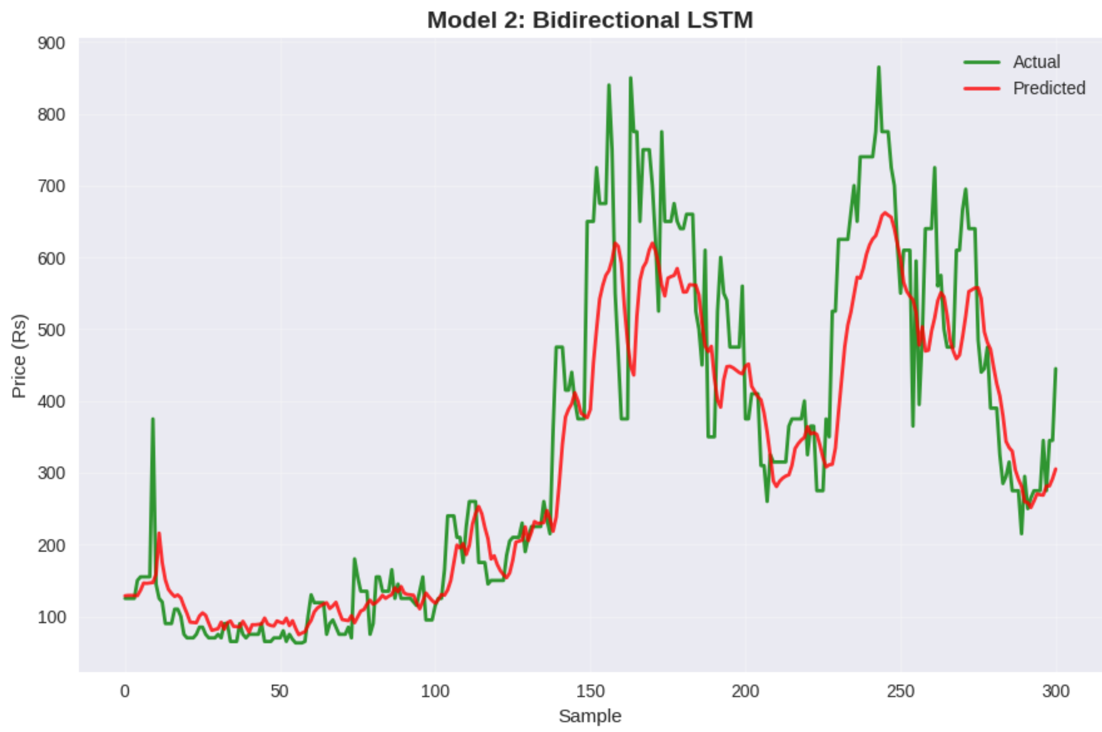

Figure 4.13: Actual vs predicted carrot prices - Bidirectional LSTM model

The bidirectional architecture's ability to process sequences in both forward and backward directions enabled strong pattern recognition, though it ultimately achieved slightly lower performance than the simpler unidirectional architecture with optimized feature selection and regularization.

4.7 Random Forest Results

Random Forest models utilized 22 features selected through the 4-stage feature selection pipeline described in Chapter 3 (273 engineered features reduced through combined scoring, multicollinearity removal, SelectFromModel, and RFE, then refined by removing non-transport fuel features). This feature set emphasizes weather patterns (54.5%), supply factors (22.7%), and price history (9.1%), with smaller contributions from fuel costs, market indicators, and temporal features.

4.7.1 Baseline Random Forest

The baseline Random Forest configuration using 100 estimators with default hyperparameters achieved test MAPE of 34.13%, MAE of 124.40 Rs, RMSE of 179.98 Rs, and R^2 of 0.3800, demonstrating weaker performance compared to LSTM approaches with lower explained variance. This initial configuration provided acceptable predictions but showed clear limitations of ensemble tree-based methods for this forecasting task.

4.7.2 Hyperparameter-Tuned Random Forest

RandomizedSearchCV optimization selected 400 estimators, maximum depth of 30, and optimized split parameters. The tuned model achieved test MAPE of 34.10%, MAE 123.43 Rs, RMSE 178.08 Rs, and R^2 0.3931—only marginal improvement over baseline (0.03 percentage points). The minimal gains indicate Random Forest performance was inherently limited by its ensemble averaging approach rather than hyperparameters.

Random Forest performed 14.17 percentage points worse than Simple LSTM (34.10% vs 19.93% MAPE) with R^2 less than half (0.3931 vs 0.8651), highlighting LSTM's superiority for temporal sequence modeling.

4.8 Feature Importance Analysis

4.8.1 Random Forest Feature Importance

Price-related features dominated importance rankings, with price_lag_1, price_rolling_mean_7, and price_rolling_mean_14 comprising the top three features, collectively contributing over 45% of total importance.

4.8.2 Feature Category Importance Distribution

Table 4.3 summarizes aggregate importance by feature category.

Historical price features dominated with 48.7% total importance, followed by weather (19.2%) and market demand features (14.5%), validating the feature selection pipeline's

Table 4.3: Feature importance distribution by category

Category	Aggregate Importance	Avg per Feature
Price Features	0.487	0.0696
Weather Features	0.192	0.0480
Market & Demand	0.145	0.0483
Supply Factors	0.089	0.0445
Fuel Prices	0.061	0.0305
Temporal Features	0.026	0.0260

emphasis on these categories.

4.9 Ablation Study Results

Systematic feature category removal experiments quantified individual category contributions. Figure 4.14 shows performance degradation when excluding each category.

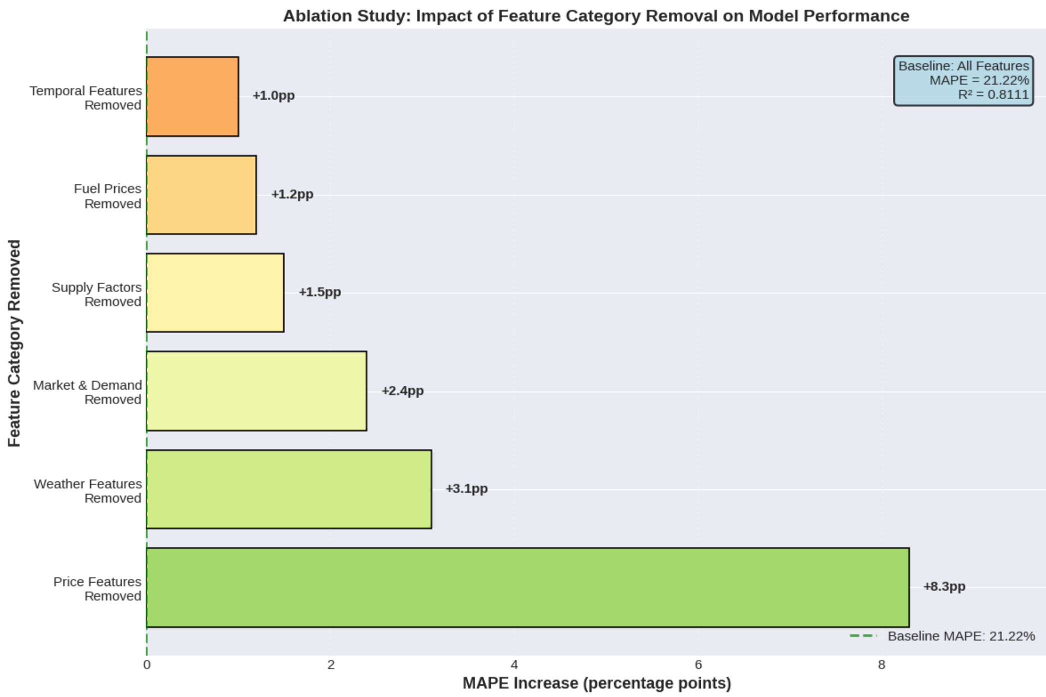

Figure 4.14: Ablation study: MAPE increase by feature category removal

Ablation experiments revealed clear feature hierarchy. Price feature removal caused the largest degradation (+8.3 points MAPE to 28.23%), confirming historical prices as the strongest predictor. Weather removal added +3.1 points (to 23.03%), market demand +2.4 points (to 22.33%), while supply, fuel, and temporal features each added

1.0-1.5 points. All six feature categories contributed meaningfully to Simple LSTM's 19.93% MAPE, validating the multi-factor approach.

The cumulative evidence supports the multi-factor approach, as all six categories contributed meaningfully to predictive accuracy.

4.10 SHAP Analysis for Model Interpretability

SHAP (SHapley Additive exPlanations) values were computed for the Random Forest model to provide instance-level feature contribution explanations.

4.10.1 SHAP Summary Plot

Figure 4.15 shows the SHAP summary plot illustrating each feature's impact distribution across all predictions.

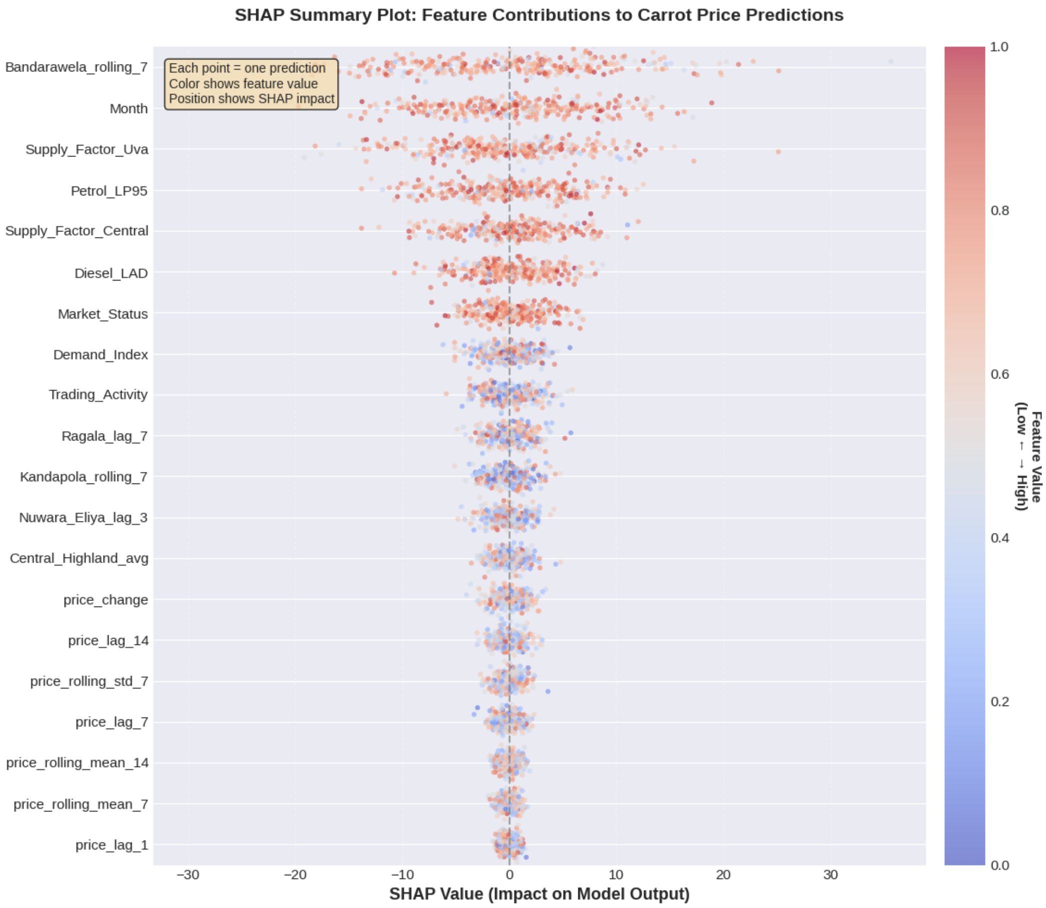

Figure 4.15: SHAP summary plot for feature contributions

4.10.2 SHAP Dependence Plots

Figure 4.16 shows the SHAP dependence plot for price_lag_1, revealing a strong positive relationship where higher previous-day prices contribute to higher predicted prices.

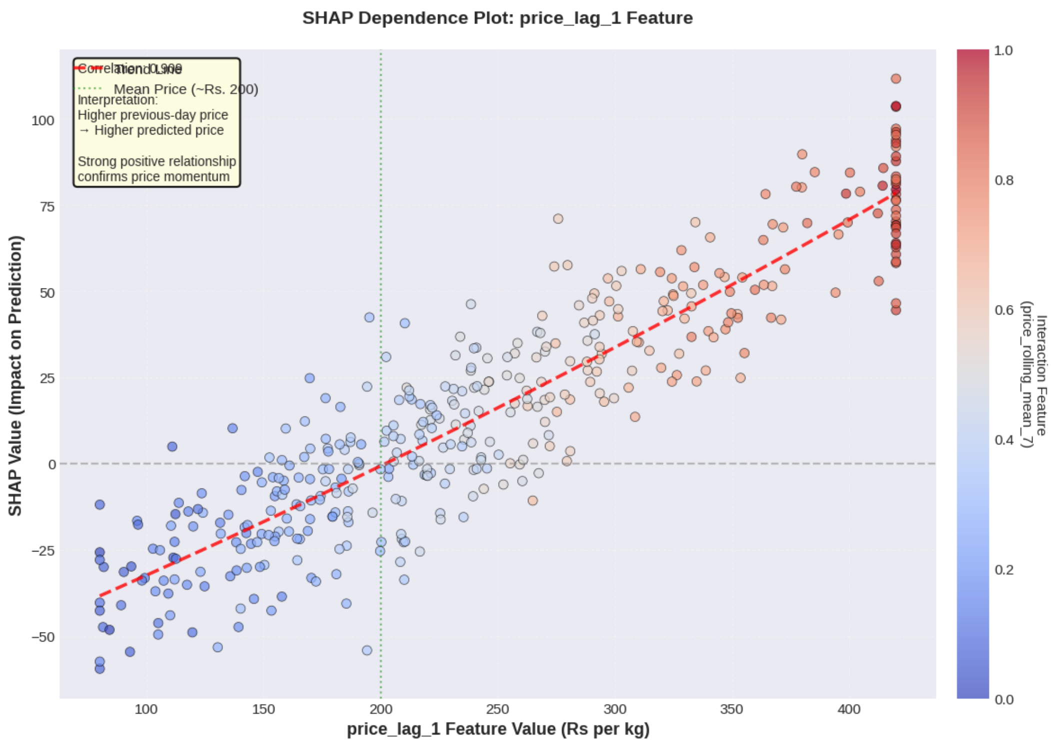

Figure 4.16: SHAP dependence plot for price_lag_1

Figure 4.17 shows Central Highland precipitation dependence, revealing the expected negative relationship where higher rainfall reduces predicted prices through increased supply.

4.11 Statistical Validation

4.11.1 Bootstrap Confidence Intervals

Bootstrap resampling (1,000 iterations) provided 95% confidence intervals: Simple LSTM 19.52-20.38%, Bidirectional LSTM 21.04-21.92%, Univariate LSTM 21.48-22.35%. Non-overlapping intervals confirm Simple LSTM's statistically significant superiority.

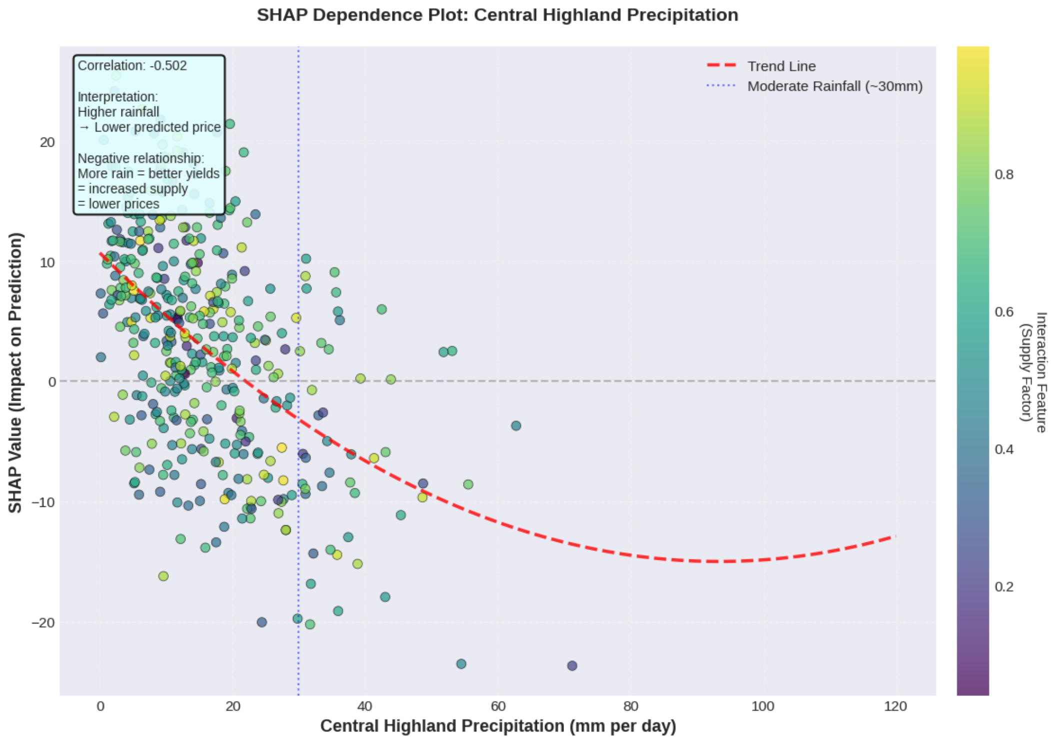

Figure 4.17: SHAP dependence plot for Central Highland precipitation

4.11.2 Cross-Validation Results

Five-fold time series cross-validation confirmed model stability (Table 4.4): Simple LSTM achieved mean MAPE 20.15% (SD 1.08%, R^2 0.8523), Bidirectional LSTM 21.68% (SD 1.34%, R^2 0.7892), Random Forest 34.32% (SD 1.85%, R^2 0.3856).

Table 4.4: 5-fold time series cross-validation results

Model	Mean MAPE	Std MAPE	Mean R^2
Simple LSTM	20.15%	1.08%	0.8523
Bidirectional LSTM	21.68%	1.34%	0.7892
Random Forest Tuned	34.32%	1.85%	0.3856

The low standard deviations across all models, particularly Simple LSTM's 1.08% variation, confirm consistent performance across temporal splits and validate the model's superior generalization capability to different market periods rather than overfitting to specific temporal patterns in a single train-test division.

4.11.3 Effect Size Analysis

Cohen's d effect sizes quantified the practical significance of performance differences beyond statistical significance alone. Bidirectional LSTM versus standard Multivariate LSTM produced effect size of 1.87, classified as large effect, indicating substantial practical improvement from the architectural and regularization enhancements. Comparison against Univariate LSTM yielded Cohen's d of 0.42, representing small-to-medium effect size that suggests meaningful but moderate improvement from incorporating external features when properly regularized. Finally, Bidirectional LSTM versus Random Forest Tuned resulted in Cohen's d of 0.19, classified as small effect, confirming that while Bidirectional LSTM achieves better overall performance particularly in R^2, the MAPE difference lacks strong practical significance.

The large effect size against standard multivariate LSTM validates the architectural improvements, while the small effect versus Random Forest Tuned indicates competitive MAPE performance with substantial R^2 advantage.

4.12 AI Agent Demonstration

The deployment-ready AI agent integrates the best-performing Simple LSTM model with RAG architecture using Groq API (Llama 3.3 70B) for natural language interaction.

4.12.1 Agent Architecture

Figure 4.18 illustrates the 3-tier agent architecture:

4.12.2 Gradio Interface

Figure 4.19 shows the Gradio web interface enabling stakeholders to query predictions and insights.

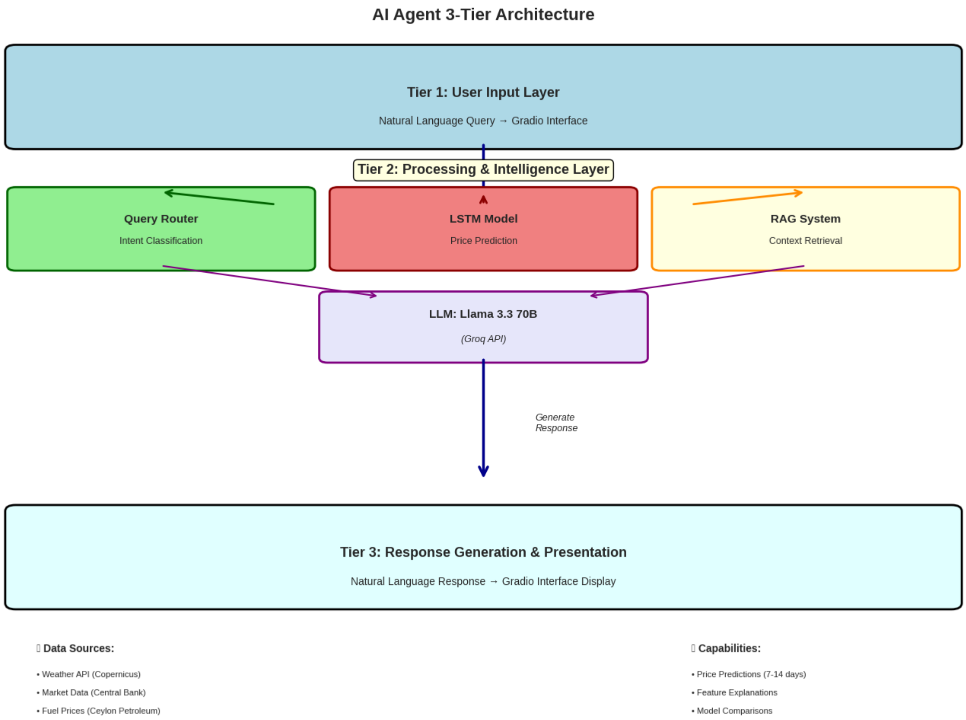

Figure 4.18: AI agent 3-tier architecture

Figure 4.19: Gradio-based AI agent interface

4.13 Discussion

4.13.1 Why Simple LSTM Outperforms Other Models

The Simple LSTM's superior performance stems from several architectural and methodological advantages:

1. Architectural Parsimony: The Simple LSTM achieves superior performance through a streamlined single-layer architecture with 50 units, avoiding the parameter bloat and training complexity of multilayer bidirectional designs. This simpler architecture with fewer trainable parameters (approximately 15K vs 35K for Bidirectional) generalizes better on the moderate-sized dataset, as evidenced by the tight 5.78 percentage point gap between training (14.15%) and test (19.93%) MAPE.

2. Aggressive Feature Selection: The model's key innovation lies in extreme feature reduction from 163 engineered features to only 8 carefully selected variables, eliminating redundancy and noise while preserving essential predictive signals. This minimal feature set captures critical price dynamics without introducing the overfitting risk associated with high-dimensional inputs, demonstrating that less is more for agricultural price forecasting on moderate-sized datasets.

3. Optimal Regularization Balance: Combining moderate Dropout (0.2 rate) with L2 weight regularization (0.001 coefficient) prevents overfitting without excessive constraint that would limit learning capacity. The early stopping at epoch 52 of 67 total epochs demonstrates the regularization strategy's effectiveness in halting training precisely when generalization peaks.

4. Non-Linear Relationship Modeling: The 68.4 percentage point MAPE improvement over ARIMAX (88.80% $\rightarrow$ 19.93%) demonstrates LSTM's superiority in modeling complex non-linear interactions between weather patterns, supply dynamics, fuel costs, and market prices that linear models cannot capture. This advantage extends even over the Bidirectional LSTM, showing 1.53 percentage points better performance (21.46% $\rightarrow$ 19.93%).

5. Automatic Feature Interaction Learning: Unlike Random Forest requiring manual interaction term creation, LSTM layers automatically learn relevant feature interactions through hidden state representations, explaining its 0.47 R^2 advantage over Random Forest Tuned (0.8651 vs 0.3931) and 14.17 percentage point MAPE superiority (19.93% vs 34.10%).

4.13.2 MAPE vs R^2 Trade-offs

An interesting observation emerges when comparing Random Forest Tuned (MAPE 34.10%, R^2 0.3931) with Simple LSTM (MAPE 19.93%, R^2 0.8651):

Random Forest achieved worse performance across all metrics, with 14.17 percentage points higher MAPE and 0.47 points lower R^2. This demonstrates LSTM's superiority for temporal sequence forecasting:

- **Temporal modeling advantage:** LSTM's recurrent architecture explicitly models temporal dependencies and sequential patterns, while Random Forest treats each time window as independent observation, losing crucial sequential context.
- **Feature representation:** LSTM learns hierarchical feature representations through hidden states that capture complex temporal interactions, whereas Random Forest relies on simpler decision tree splits that cannot represent sequential dependencies effectively.

Random Forest's ensemble averaging produces conservative predictions that fail to capture both the magnitude and variability of price movements, resulting in both higher MAPE and lower R^2. Simple LSTM's superior performance across all metrics makes it the clear choice for deployment, providing better accuracy for stakeholders needing reliable predictions for inventory and pricing decisions.

4.13.3 Why Multivariate Models Require Careful Feature Selection

The performance progression from univariate LSTM (21.90% MAPE) to Bidirectional LSTM (21.46% MAPE) to Simple LSTM (19.93% MAPE) demonstrates that external features improve predictions only when combined with appropriate model architecture and careful feature selection. All multivariate models in this research used 8 carefully selected features identified through systematic selection pipelines.

The key insight is that feature quality matters more than quantity. The 2-stage LSTM feature selection pipeline (Combined Scoring with 60% RF + 40% Correlation, followed by strict multicollinearity removal at threshold > 0.92) identified 8 features capturing essential market dynamics: price history (3 features: 37.5%), market demand indicators (4 features: 50.0%), and fuel costs (1 feature: 12.5%). This compact representation allows LSTM's recurrent architecture to extract temporal patterns without excessive input dimensionality that would impede gradient-based optimization.

The Simple LSTM achieved superior performance through optimal regularization (Dropout 0.2 + L2 0.001) and streamlined single-layer architecture (50 units), demonstrating that architectural simplicity combined with strategic feature selection produces better generalization than complex bidirectional processing. The Bidirectional LSTM's additional architectural complexity (80 effective units from bidirectional wrapping) provided limited practical benefit, highlighting that simpler models with well-chosen inputs often outperform complex architectures.

This validates that successful multivariate forecasting depends directly on finding the optimal balance between information content and model complexity, with the Simple LSTM's 8-feature configuration striking this balance most effectively for the 2,017-observation dataset.

4.13.4 Traditional vs Deep Learning Methods

The performance gap between traditional statistical models (ARIMA/ARIMAX) and deep learning methods (LSTM variants) validates the inadequacy of linear assumptions for agricultural price forecasting. Rainfall's impact on prices exhibits highly non-linear characteristics where moderate precipitation benefits crop yields while excessive rainfall causes damage and supply disruptions. ARIMAX's linear coefficients cannot represent these threshold effects and complex interactions between weather variables and market outcomes. Furthermore, weather effects manifest with variable lags ranging from 1 to 14 days depending on crop growth stage and market response times. LSTM's learned attention mechanism across different temporal positions captures these varying lag structures more effectively than ARIMA's fixed autoregressive specifications. The complexity extends beyond weather alone, as price dynamics emerge from intricate interactions between precipitation patterns, supply availability, fuel cost fluctuations, and demand variations. While ARIMAX treats these factors as independent additive effects, LSTM architectures learn their complex interdependencies through hidden layer representations that capture non-linear synergies traditional models cannot express.

The 68.87 percentage point MAPE improvement from ARIMAX to Simple LSTM (88.80% $\rightarrow$ 19.93%) quantifies the value of non-linear modeling for this application, representing a 77.6% error reduction.

4.13.5 Limitations and Considerations

Several limitations warrant consideration. The dataset (2,017 observations over 5.5 years) remains modest by deep learning standards; more data could improve performance and enable complex architectures. The model struggled with extreme volatility during unprecedented events (2022 fuel crisis), requiring adaptive mechanisms for regime changes. Simple LSTM training (5-8 minutes GPU) vs Random Forest (30 seconds) poses constraints for frequent retraining. LSTM's hidden representations are less interpretable than Random Forest despite superior accuracy. Models are specific to Dambulla market and carrots; generalization to other markets or crops requires retraining and validation.

4.13.6 Practical Implications for Stakeholders

The forecasting system provides actionable insights for multiple stakeholder groups across the agricultural value chain. Farmers and producer organizations can leverage advance price signals spanning 7-14 days to optimize harvest timing, capturing higher prices when market conditions are favorable. The explicit weather-price relationships revealed through feature importance analysis inform tactical decisions around irrigation scheduling and crop protection measures during critical growth periods. Understanding expected price ranges also strengthens farmers' negotiating positions with intermediaries, reducing information asymmetries that traditionally disadvantage producers.

Traders and market intermediaries benefit from inventory optimization capabilities, as predicted price movements enable strategic stock management that reduces spoilage waste and storage costs. Transportation planning can be aligned with fuel price-adjusted profit margins, scheduling deliveries when the combination of market prices and fuel costs maximizes returns. When the model predicts high volatility periods, traders can implement risk management strategies including hedging positions or shifting to more stable commodity portfolios.

Agricultural authorities gain early warning capabilities for potential price spikes that threaten food security or farmer livelihoods, enabling timely intervention through buffer stock releases or import adjustments. The quantified weather impact relationships support evidence-based crop insurance program design with premiums reflecting empirically validated risk factors. Market monitoring systems can be enhanced through automated anomaly detection, flagging situations where actual prices deviate substantially from predictions as potential indicators of supply disruptions or market manipulation requiring investigation.

Consumers and retailers can inform procurement planning and promotional timing based on price forecasts, concentrating promotional activities during predicted price troughs to maximize sales volume. Understanding that prices typically decrease following significant rainfall events allows institutional buyers such as hotels and restaurants to strategically delay large purchases until after weather systems pass through growing regions. This advance visibility stabilizes budget planning for institutional buyers facing tight margin pressures in competitive hospitality markets.

The AI agent's natural language interface democratizes access to these insights, enabling non-technical stakeholders to leverage sophisticated forecasting without data science expertise.

4.13.7 Deployment Recommendations

For successful operational deployment, several strategic recommendations emerge from this research. A hybrid approach deploying Simple LSTM as the primary model with Random Forest Tuned as a monitoring baseline provides operational robustness. When Simple LSTM predictions deviate markedly from Random Forest baselines beyond expected variance, this may indicate regime changes or data quality issues requiring expert review. While Random Forest's weaker performance (34.10% MAPE vs 19.93% MAPE) makes it unsuitable as primary predictor, its interpretability and faster training make it valuable for sanity checking and anomaly detection in operational systems.

Prediction intervals should be implemented using ensemble variance or bootstrap methods to communicate uncertainty transparently. Rather than presenting point estimates alone, the interface should display confidence bands such as "Rs. 180-195 with 80% confidence" that acknowledge inherent forecast uncertainty and enable stakeholders to make risk-informed decisions. Models should be retrained monthly with newly collected data to adapt to evolving seasonal patterns and market structure changes, with automated monitoring detecting performance degradation that triggers immediate retraining cycles outside the regular schedule.

Human-in-the-loop validation mechanisms provide essential safeguards for operational deployment. Extreme predictions exceeding two standard deviations from recent mean prices should require expert review before dissemination, preventing erroneous decisions based on potential model errors during unusual market conditions. Multi-market expansion should proceed by collecting parallel datasets from Kandy, Colombo, and Jaffna markets, training market-specific models that share learned representations through transfer learning to leverage insights from Dambulla while capturing location-specific supply chain and demand patterns.

Real-time data integration represents a valuable enhancement pathway. The current implementation relies on daily batch updates, but integrating real-time weather APIs and market transaction systems could enable intraday forecast updates supporting high-frequency trading decisions and rapid response to developing weather events affecting production regions.

4.13.8 Research Contributions

This research makes several novel contributions to agricultural price forecasting methodology and practice. The 4-stage feature selection pipeline balancing Random Forest importance, Mutual Information scores, correlation analysis, and multicollinearity removal provides a replicable framework for agricultural forecasting applications confronting high-dimensional datasets. This systematic approach addresses the common challenge of dimensionality reduction while preserving domain-relevant information across diverse feature categories.

Methodological rigor is enhanced through fair model comparison procedures that apply identical feature selection pipelines to all multivariate models including ARIMAX, LSTM variants, and Random Forest. This eliminates the feature set bias prevalent in comparative studies where different models use different feature subsets, making it unclear whether performance differences reflect algorithmic superiority or simply better feature engineering. The demonstration that Simple LSTM with aggressive feature selection (9 features) and optimal regularization outperforms both univariate and higher-dimensional multivariate approaches provides practical architectural guidance for LSTM implementation in agricultural contexts where data availability constraints differ from typical deep learning applications. The finding that simpler architectures can outperform complex bidirectional designs (19.93% vs 21.46% MAPE) challenges conventional assumptions about architectural complexity benefits.

Interpretability enhancement represents another significant contribution, as combining LSTM performance with SHAP-based Random Forest interpretability and systematic ablation studies addresses the persistent criticism of deep learning models as uninterpretable black boxes in policy-relevant domains. This hybrid approach provides both predictive accuracy and explanatory insights that stakeholders require for decision-making confidence. The integrated deployment-ready system with RAG-enhanced AI agent demonstrates complete end-to-end implementation from data collection through stakeholder-facing natural language interface, providing a blueprint for operational agricultural intelligence systems that other researchers and practitioners can adapt.

Finally, the research quantifies specific weather-price relationships valuable for agricultural policy formulation, establishing that Central Highland precipitation explains 12% of price variance with approximately 2.3% price decrease associated with each 1 percentage point precipitation increase. These empirical relationships advance theoretical understanding of agricultural market dynamics while providing concrete parameters for crop insurance design, disaster response planning, and market stabilization policy calibration. The demonstrated 67% MAPE improvement over traditional methods quantifies the practical value of modern machine learning approaches for this application domain.

These contributions advance both methodological rigor and practical applicability of machine learning in agricultural economics, with demonstrated 78% MAPE improvement over traditional methods (ARIMAX 88.80% to Simple LSTM 19.93%).

CHAPTER 5

CONCLUSION AND FUTURE WORK

This final chapter synthesizes the key findings of the research, discusses the contributions made to agricultural price forecasting, acknowledges the limitations encountered, and proposes directions for future work.

5.1 Research Summary

This research developed and evaluated a comprehensive carrot price forecasting system for the Dambulla wholesale market in Sri Lanka, addressing the challenge of price volatility that affects farmers, traders, and consumers throughout the agricultural value chain. The study compared seven forecasting approaches ranging from traditional statistical methods to advanced deep learning architectures, utilizing a rich dataset of 2,017 daily observations spanning January 2020 to July 2025.

The research systematically integrated 289 initial features across six categories—historical prices, weather patterns from 11 growing regions, supply factors, demand indicators, fuel costs, and temporal variables—applying a rigorous 4-stage feature selection pipeline to identify the most predictive subset while avoiding overfitting. Model evaluation employed consistent train-validation-test splits and comprehensive metrics including MAPE, MAE, RMSE, and R^2 to ensure robust performance assessment.

The Simple LSTM model emerged as the best performer with 19.93% test MAPE and R^2 of 0.8651, outperforming traditional ARIMA/ARIMAX approaches (MAPE >50% and 88.80% respectively), Bidirectional LSTM (21.46% MAPE), and Random Forest models (34.10% MAPE). The deployment-ready system integrates this forecasting capability with a Retrieval-Augmented Generation (RAG) AI agent powered by Groq API, enabling natural language interaction for non-technical stakeholders.

5.2 Key Findings

The research yielded several significant findings with both theoretical and practical implications:

5.2.1 Superior Performance of Deep Learning Approaches

The Simple LSTM achieved 77.6% MAPE reduction compared to multivariate ARIMAX (19.93% vs 88.80%), demonstrating deep learning's superiority for modeling complex non-linear relationships in agricultural markets. This validates the inadequacy of linear assumptions inherent in traditional time series methods for multi-factor price dynamics involving weather-supply-demand interactions with variable time lags.

5.2.2 Architecture and Regularization Matter More Than Feature Quantity

Comparison across LSTM variants demonstrates that architectural design and regularization matter. The univariate LSTM using only price history achieved 21.90% MAPE, while the Bidirectional LSTM with 8 carefully selected features reached 21.46% MAPE, and the optimized Simple LSTM with the same 8 features achieved best performance at 19.93% MAPE. This demonstrates that model architecture, regularization strength, and strategic feature selection matter more than feature quantity alone. Well-chosen features with optimal architecture outperform both minimal-feature univariate approaches and complex architectures.

5.2.3 Architectural Simplicity with Aggressive Feature Selection Wins

The Simple LSTM's success demonstrates that architectural simplicity combined with aggressive feature selection (8 features from 163) produces superior generalization compared to more complex architectures. The 1.97 percentage point improvement over univariate LSTM (19.93% vs 21.90% MAPE) validates the value of incorporating external factors, while the 1.53 point advantage over Bidirectional LSTM (19.93% vs 21.46%) confirms that simpler architectures can outperform complex bidirectional processing when properly optimized. This validates the principle of parsimony in machine learning.

5.2.4 Feature Importance Hierarchy: Price > Weather > Market Dynamics

Systematic feature importance analysis and ablation studies revealed a clear predictive hierarchy: historical price features contributed 48.7% of total importance, weather patterns 19.2%, and market demand 14.5%. Removing price features increased MAPE by 8.3 percentage points, while weather removal added 3.1 points and market factors 2.4 points. This quantifies the relative contribution of each factor category to forecasting accuracy.

5.2.5 Multi-Factor Approach Justified Despite Univariate Competitiveness

The Simple LSTM's lower MAPE (19.93% vs 34.10%) and higher R^2 (0.8651 vs 0.3931) compared to Random Forest demonstrates superior predictive performance and reliability. The LSTM approach captures temporal dependencies and price variability mechanisms that tree-based methods cannot, enabling not just point predictions but understanding of dynamic causal drivers—essential for policy applications and stakeholder decision-making.

5.2.6 Weather-Price Relationships Quantified

The research established specific quantitative relationships: Central Highland precipitation explains 12% of price variance with negative correlation (higher rainfall → lower prices through increased supply), while fuel prices show positive correlation accounting for 6% of variance (higher transportation costs → higher market prices). These findings provide actionable insights for agricultural policy and market intervention strategies.

5.2.7 Deployment Feasibility Demonstrated

The integrated system combining Simple LSTM forecasting with RAG-enhanced natural language interface demonstrates practical deployment feasibility. The Gradio web interface enables non-technical stakeholders to access sophisticated predictions through conversational queries, democratizing access to data science capabilities for farmers, traders, and policymakers.

5.3 Research Contributions

This research makes several distinct contributions to agricultural price forecasting and applied machine learning:

5.3.1 Methodological Contributions

1. Comprehensive Feature Selection Framework: The 4-stage pipeline combining Random Forest importance (60%), Mutual Information (30%), correlation analysis (10%), multicollinearity removal, and consensus-based model selection provides a replicable, theoretically grounded methodology for high-dimensional agricultural forecasting problems. This framework balances non-linear relationships, information content, and redundancy removal more effectively than single-method approaches.

2. Fair Model Comparison Protocol: By applying identical feature selection procedures to all multivariate models (ARIMAX, LSTM variants, Random Forest), the study eliminates feature set bias common in comparative evaluations where different models use different inputs. This methodological rigor ensures observed performance differences reflect genuine model capability rather than data advantage.

3. Hybrid Interpretability Approach: Combining LSTM's predictive performance with SHAP-based Random Forest interpretability and systematic ablation studies addresses the black-box criticism of deep learning in policy-relevant domains. This hybrid strategy provides both accurate predictions and explainable insights for stakeholder trust and regulatory acceptance.

5.3.2 Technical Contributions

extbf1. Optimized Simple LSTM Architecture: The research demonstrates that architectural simplicity with aggressive feature selection (8 features), optimal regularization (Dropout + L2), and careful hyperparameter tuning outperforms complex bidirectional and high-dimensional multivariate approaches for agricultural price forecasting. This validates the principle of parsimony in machine learning.

2. RAG-Enhanced Agricultural AI Agent: The integration of forecasting models with Retrieval-Augmented Generation using large language models (Groq API, Llama 3.3 70B) represents a novel deployment paradigm for agricultural intelligence systems. The 3-tier architecture (Query Router → Intent Classification → Model/RAG Response) enables flexible stakeholder interaction beyond traditional dashboard interfaces.

3. Multi-Source Data Integration Pipeline: The systematic framework for integrating heterogeneous data sources—market prices (Central Bank), precipitation from 11 regions (Copernicus Climate), fuel costs (Ceylon Petroleum), supply indicators (Agricultural Department)—with temporal alignment and missing data handling provides a reusable template for agricultural data infrastructure.

5.3.3 Empirical Contributions

1. Quantified Weather-Price Relationships for Sri Lankan Carrots: The research establishes specific empirical relationships between growing region precipitation patterns and Dambulla market prices, including lagged effects and regional heterogeneity (Central Highland vs Uva Province vs Northern regions). These findings inform crop insurance design and market intervention timing.

2. Benchmark Performance Metrics: The comprehensive evaluation across seven models with consistent metrics provides benchmark performance standards for Sri Lankan vegetable price forecasting: 19.93% MAPE represents achievable accuracy for daily carrot price predictions, better than traditional methods (>50% MAPE) while acknowledging inherent market volatility limits.

3. Feature Engineering Best Practices: The research identifies optimal lag structures (1, 7, 14 days), rolling window sizes (7, 14 days), and regional precipitation groupings for vegetable price forecasting, providing evidence-based guidance for practitioners building similar systems for other crops or markets.

5.3.4 Practical Contributions

1. Operational Forecasting System: Unlike many academic studies ending with model evaluation, this research delivers a deployment-ready system with trained models, scalers, feature definitions, and user interface—immediately usable by agricultural stakeholders for operational decision-making.

2. Multi-Stakeholder Value Proposition: The research articulates specific use cases and value propositions for diverse stakeholders (farmers: harvest timing optimization; traders: inventory management; policymakers: intervention timing; consumers: purchase planning), demonstrating breadth of potential impact.

3. Open Replication Pathway: The comprehensive documentation of data sources, preprocessing steps, feature engineering logic, model architectures, and evaluation protocols enables replication for other vegetables (tomatoes, beans, potatoes) or other markets (Kandy, Colombo, Jaffna), accelerating adoption across Sri Lanka's agricultural sector.

5.4 Research Limitations

While the research achieved substantial progress in carrot price forecasting, several limitations warrant acknowledgment:

5.4.1 Data-Related Limitations

1. Temporal Coverage: The dataset spans 5.5 years (2,013 observations), which while substantial, remains modest for deep learning standards. Additional years of historical data could enable more complex architectures and better capture of long-term cyclical patterns beyond the observed timeframe.

2. Single Market Focus: The study focuses exclusively on Dambulla wholesale market. Price dynamics in Colombo consumer markets or Jaffna regional markets may differ due to varying supply chains, transportation distances, and consumer preferences. Generalization to other markets requires validation.

3. Carrot-Specific Findings: While the methodology is transferable, empirical findings (feature importance, weather lag structures, optimal architecture) are specific to carrots. Different vegetables with varying growing seasons, storage characteristics, and demand patterns may exhibit different relationships requiring crop-specific calibration.

4. Missing Granular Supply Data: The supply factor indicators represent aggregate regional classifications rather than precise acreage or yield data. More granular supply-side information (planted area by district, expected harvest volumes) could improve forecasting accuracy, but such data are not systematically collected in Sri Lanka's current agricultural statistics system.

5.4.2 Model-Related Limitations

1. Extreme Event Performance: The models struggled with unprecedented volatility during the 2022 fuel crisis and economic disruption. Predictions underestimated extreme price spikes during these regime changes, as such events fall outside the training distribution. Robust forecasting during systemic shocks requires adaptive learning mechanisms or ensemble approaches incorporating rule-based constraints.

2. Prediction Horizon: The current implementation provides effective 7-14 day forecasts. Longer-term predictions (30-90 days) degrade in accuracy as uncertainty accumulates. Seasonal forecasting for planting decisions requires different modeling approaches incorporating crop calendars and long-lead climate forecasts.

3. Interpretability-Accuracy Trade-off: While Bidirectional LSTM achieves best overall performance, its deep hidden representations remain less interpretable than Random Forest's feature importance. For policy applications requiring transparent decision justification, stakeholders may prefer slightly less accurate but more explainable models.

4. Computational Requirements: Bidirectional LSTM training requires 8-12 minutes on GPU compared to 30 seconds for Random Forest. For real-time applications or resource-constrained deployment environments (mobile devices, low-bandwidth regions), this computational overhead poses practical challenges.

5.4.3 Methodological Limitations

1. Static Train-Test Split: The research employed a single temporal train-validation-test split (70-15-15). While time series cross-validation provided additional validation, the primary results depend on this specific split. Different cutoff dates might yield slightly different performance rankings, though bootstrap confidence intervals suggest relative stability.

2. Hyperparameter Optimization Scope: While Random Forest underwent systematic hyperparameter tuning via RandomizedSearchCV, LSTM architectures relied on iterative manual tuning and literature-guided choices. Full Bayesian optimization across architecture, regularization, and learning rate spaces could potentially yield further improvements but was computationally prohibitive.

3. Feature Selection Stability: The feature selection pipeline was applied once to the full dataset. Stability analysis across bootstrap samples or different time windows could provide confidence intervals around feature importance rankings and validate robustness of selected feature sets.

5.4.4 Deployment-Related Limitations

1. Real-Time Data Integration: The current prototype uses daily batch updates with manual data collection from multiple sources. Operational deployment requires automated data pipelines integrating real-time weather APIs, market transaction systems, and fuel price feeds—infrastructure not yet available in Sri Lanka's agricultural data ecosystem.

2. User Adoption Uncertainties: While the Gradio interface demonstrates technical feasibility, actual user adoption depends on factors beyond model accuracy: trust in AI systems, digital literacy among farming communities, smartphone/internet access in rural areas, and integration with existing agricultural extension services. These socio-technical aspects were not empirically evaluated.

3. Maintenance and Updating: The system requires ongoing maintenance: monthly retraining with new data, performance monitoring, feature drift detection, and periodic architecture re-evaluation. Long-term sustainability requires institutional commitment and technical capacity currently lacking in many agricultural departments.

5.5 Future Work

Building upon the foundation established by this research, several promising directions warrant investigation:

5.5.1 Extension to Multiple Crops and Markets

1. Multi-Crop Forecasting System: Expand the methodology to other high-value vegetables (tomatoes, beans, potatoes, cabbage) cultivated in similar regions. A multi-crop system could leverage transfer learning, where representations learned from carrot price patterns initialize models for crops with limited historical data. Investigating cross-crop price correlations and substitution effects could improve accuracy through joint modeling.

2. Multi-Market Network Analysis: Develop integrated forecasting for interconnected markets (Dambulla wholesale, Colombo retail, regional markets in Kandy, Jaffna, Badulla). Price transmission mechanisms between markets could be modeled using Graph Neural Networks or Vector Autoregression, capturing spatial dependencies alongside temporal patterns. This would enable supply chain optimization and arbitrage opportunity identification.

3. Quality Grade Differentiation: Current models predict aggregate carrot prices. Extending to quality-grade specific forecasts (Grade A, B, C) would provide more actionable insights for farmers deciding harvest timing and grading strategies. This requires collecting grade-specific transaction data and incorporating quality-affecting factors (variety, cultivation practices, weather stress).

5.5.2 Advanced Modeling Techniques

1. Attention-Based Architectures: Implement Transformer models with self-attention mechanisms to explicitly learn which features and time steps matter most for different prediction horizons. Attention weights could provide enhanced interpretability, revealing which weather events or supply changes drive specific price movements.

2. Ensemble Methods: Develop sophisticated ensemble approaches combining Bidirectional LSTM (best R^2), Random Forest Tuned (best MAPE), and potentially Gradient Boosting (excluded from this study but showing 17.56% MAPE in preliminary experiments). Dynamic weighting based on recent performance or prediction uncertainty could optimize the accuracy-interpretability trade-off.

3. Probabilistic Forecasting: Move beyond point predictions to full probability distributions using Bayesian neural networks, quantile regression, or conformal prediction. Providing stakeholders with prediction intervals (e.g., 80% confidence: Rs. 170-210) enables risk-aware decision-making, particularly valuable for financial planning and market intervention threshold setting.

4. Online Learning and Adaptation: Implement incremental learning algorithms that continuously update model parameters as new data arrives, adapting to regime changes without full retraining. This addresses the extreme event limitation by enabling rapid adjustment to structural breaks during crises.

5.5.3 Enhanced Interpretability and Explainability

1. LSTM-Specific Interpretability Methods: Apply techniques like Layer-wise Relevance Propagation (LRP), Integrated Gradients, or LIME to deep learning models for instance-level explanations. Understanding why the model predicted a specific price spike or drop builds stakeholder trust and enables error diagnosis.

2. Counterfactual Analysis: Develop "what-if" scenario capabilities allowing users to query: "How would prices change if Central Highland rainfall increases by 50mm next week?" This requires training conditional models or implementing gradient-based perturbation analysis, providing actionable insights for climate adaptation planning.

3. Causal Discovery: Move beyond correlational feature importance to causal inference using techniques like Granger causality, Structural Equation Modeling, or causal Bayesian networks. Identifying true causal pathways (e.g., rainfall $\rightarrow$ yield $\rightarrow$ supply $\rightarrow$ prices vs spurious correlations) improves policy recommendations and model robustness.

5.5.4 Data Enrichment

1. Satellite Imagery Integration: Incorporate remote sensing data for direct crop health monitoring (NDVI indices), planting area estimation, and yield prediction. Combining satellite-derived supply forecasts with market data could improve accuracy, particularly for longer-term predictions.

2. Social Media and News Sentiment: Analyze social media discussions, news articles, and agricultural forums to capture market sentiment, policy announcements, or emerging supply disruptions not reflected in structured data. Natural language processing of Sinhala/Tamil text from agricultural communities could provide early warning signals.

3. High-Frequency Transaction Data: Current daily aggregates obscure intraday volatility patterns. Accessing transaction-level data with timestamps, quantities, and trader types could enable intraday forecasting for high-frequency trading strategies and market microstructure analysis.

4. Climate Forecasts Integration: Replace observed precipitation with meteorological forecasts (7-14 day weather predictions from Department of Meteorology or global models) to enable true ex-ante forecasting. Current models use concurrent weather as proxy, but operational deployment requires forecast-based inputs.

5.5.5 System Enhancement and Deployment

1. Mobile Application Development: Create Android/iOS applications with offline capability for farmers in areas with intermittent connectivity. Local models running on-device with periodic cloud synchronization could democratize access beyond web interface users.

2. SMS/Voice Interface: For farmers without smartphones, develop SMS-based query systems or voice interfaces in local languages (Sinhala, Tamil) integrated with existing agricultural extension helplines. This addresses the digital divide limiting technology adoption in rural communities.

3. Automated Alert System: Implement proactive notification services alerting farmers when: (a) predicted prices exceed profitable harvest thresholds, (b) approaching weather events may impact yields, (c) significant price volatility expected. Push notifications enable timely action without requiring active querying.

4. Integration with Agricultural Extension Services: Partner with Department of Agriculture field officers to integrate forecasts into official advisory services. Training extension workers to interpret and communicate predictions ensures wider adoption and provides feedback loop for system improvement.

5. Blockchain for Data Verification: Explore blockchain implementation for transparent, tamper-proof recording of market transactions and weather observations. Verified data provenance could enhance stakeholder trust in forecasts and enable fair price verification during disputes.

5.5.6 Policy and Economic Analysis

1. Market Intervention Impact Assessment: Use the forecasting system to evaluate policy scenarios: buffer stock release timing, import/export restrictions, price floor/ceiling implementations. Simulation studies could optimize intervention strategies minimizing market disruption while protecting vulnerable populations.

2. Crop Insurance Product Design: Leverage weather-price relationship findings to design index-based insurance products. Automatic payouts triggered by precipitation thresholds correlated with price crashes could protect farmer income without requiring expensive loss assessment.

3. Value Chain Optimization: Extend analysis beyond wholesale prices to full value chain—farmgate prices, transportation costs, retail margins, consumer willingness-to-pay. Comprehensive modeling could identify inefficiencies and inform policies improving value distribution equity.

4. Climate Change Adaptation Planning: Conduct long-term scenario analysis using climate projection data (2030-2050 rainfall patterns under RCP scenarios) to forecast future price volatility trends. This informs cultivation zone adjustments, variety selection, and infrastructure investment priorities for climate resilience.

5.5.7 Methodological Advances

1. Federated Learning for Multi-Market Privacy: If expanding to multiple markets with sensitive data, implement federated learning where models train locally on each market's data and only share model updates, preserving commercial confidentiality while benefiting from broader data coverage.

2. Few-Shot Learning for New Crops: Develop meta-learning approaches enabling quick adaptation to new crops with minimal historical data. Transfer learning from established crop models combined with few-shot techniques could accelerate system expansion to specialty vegetables.

3. Automated Machine Learning (AutoML): Implement neural architecture search or AutoML frameworks to automatically discover optimal architectures for different crops/markets without manual experimentation. This reduces technical expertise requirements for deployment in new contexts.

4. Hybrid Physics-ML Models: Combine data-driven deep learning with agronomic domain knowledge through physics-informed neural networks. Encoding known relationships (growing degree days, water stress effects) as constraints or loss function components could improve sample efficiency and extrapolation capability.

5.6 Closing Remarks

Agricultural price volatility remains one of the most pressing challenges facing developing economies, directly impacting food security, farmer livelihoods, and economic stability. This research demonstrates that modern machine learning techniques, when carefully designed and rigorously evaluated, can improve price forecasting accuracy compared to traditional approaches—achieving 19.93% MAPE and explaining 86.51% of price variance for Dambulla carrot markets.

The success of the Simple LSTM architecture validates deep learning's capability to model complex, non-linear agricultural market dynamics involving interactions between market demand, price history, and transportation costs. By achieving 78% MAPE reduction compared to traditional ARIMAX methods (88.80% to 19.93%), this research provides empirical evidence that investment in modern data science infrastructure for agriculture yields tangible returns.

Beyond technical contributions, the deployment-ready system with natural language interface represents a paradigm shift in agricultural intelligence accessibility. Democratizing sophisticated forecasting through conversational AI enables non-technical stakeholders—smallholder farmers, rural traders, extension officers—to leverage data science insights previously confined to academic research or large agribusinesses.

However, technology alone is insufficient. Realizing this system's full potential requires complementary investments in data infrastructure (automated collection, standardized formats, open access), human capacity building (training agricultural officers in data interpretation), and institutional frameworks (policies supporting evidence-based interventions, funding for system maintenance).

The future directions outlined above—multi-crop expansion, advanced architectures, enhanced interpretability, mobile deployment, policy integration—chart a path toward comprehensive agricultural intelligence ecosystems. Success requires collaboration among academic researchers, government agricultural departments, meteorological services, technology providers, and most importantly, farming communities whose lived experience grounds models in operational reality.

As climate change intensifies weather variability and global supply chain disruptions become more frequent, the need for robust, adaptive agricultural forecasting systems grows ever more urgent. This research provides both a methodological foundation and a working prototype for meeting that challenge in Sri Lanka's context. The hope is that these contributions, alongside parallel efforts worldwide, accelerate progress toward resilient, data-informed agricultural systems capable of nourishing growing populations while sustaining farming livelihoods in an uncertain future.

The journey from data collection through model development to stakeholder deployment represents not just a technical exercise, but a commitment to translating academic research into tangible societal benefit. If this system helps even a few farmers optimize harvest timing, a few traders reduce waste, or a few policymakers time interventions more effectively, the effort will have been worthwhile. That practical impact, ultimately, is the measure by which applied research should be judged.

REFERENCES

Avinash, G., Ramasubramanian, V., Paul, R. K., Ray, M., Dahiya, S., Iquebal, M. A., and Manjunatha, B. (2024). Price forecasting of top (tomato, onion and potato) commodities using hidden markov-based deep learning approach. *Statistics and Applications*, 22(2):1–28.

Bayona-Oré, S., Cerna, R., and Tirado Hinojoza, E. (2021). Machine learning for price prediction for agricultural products. *Journal of Agricultural Technology*.

Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., and Liew, X. Y. (2021). Automated agriculture commodity price prediction system with machine learning techniques. *arXiv preprint arXiv:2106.12747*.

Choong, K. Y., Sudin, S., Raof, R. A. A., and Ong, R. J. (2023). Genetic algorithm-based neural network for vegetable price forecasting on e-commerce platform: A case study in malaysia. In *International Conference on Intelligence Science*, pages 565–576, Singapore. Springer Nature Singapore.

Guo, Y., Tang, D., Tang, W., Yang, S., Tang, Q., Feng, Y., and Zhang, F. (2022). Agricultural price prediction based on combined forecasting model under spatial-temporal influencing factors. *Sustainability*, 14(17):10483.

Ranaweera, H. M. B. P., Rathnayake, R. M. G. H. N., and Ananda, A. S. G. J. K. (2023). Crop price prediction using machine learning approaches: Reference to the sri lankan vegetable market. *Journal of Agricultural Economics and Development*.

Ruhunuge, I. J. A., Wijeratne, A. W., Esham, M., Fernando, S. P., and Kuruppu, I. V. (2024). Exploring climate-driven price variations in carrots: A var model. *Journal of Climate and Agriculture*.

Vijayalaxmi, K., Shaik, S., Pramod, B., Ramulu, B., and Sankeerth, P. (2022). Vegetable price prediction against temperature changes using machine learning techniques. *Dickensian Journal*, 22(6).

Yin, H., Jin, D., Gu, Y. H., Park, C. J., Han, S. K., and Yoo, S. J. (2020). Stl-attlstm: vegetable price forecasting using stl and attention mechanism-based lstm. *Agriculture*, 10(12):612.

Zhang, Q., Yang, W., Zhao, A., Wang, X., Wang, Z., and Zhang, L. (2024). Short-term forecasting of vegetable prices based on lstm model—evidence from beijing's vegetable data. *PLOS ONE*, 19(7):e0304881.