*Biological Forum – An International Journal* **16(8): 63-68(2024)**

**ISSN No. (Print): 0975-1130**
**ISSN No. (Online): 2249-3239**

# Comparative Analysis of Statistical Model and Machine Learning Algorithms in Forecasting Black Pepper Price of Kerala

***Muhammed Irshad M.[1], Kader Ali Sarkar[2*], Digvijay Singh Dhakre[2] and Debasis Bhattacharya[3]***
*[1]Research Scholar, Department of Agricultural Statistics,*
*Palli Siksha Bhavana, Visva-Bharati University, Sriniketan (West Bengal), India.*
*[2]Assistant Professor, Department of Agricultural Statistics,*
*Palli Siksha Bhavana, Visva-Bharati University, Sriniketan (West Bengal), India.*
*[3]Professor, Department of Agricultural Statistics,*
*Palli Siksha Bhavana, Visva-Bharati University, Sriniketan (West Bengal), India.*

*(Corresponding author: Kader Ali Sarkar*)*
*(Received: 25 May 2024; Revised: 10 June 2024; Accepted: 01 July 2024; Published: 14 August 2024)*
*(Published by Research Trend)*

**ABSTRACT: Black pepper is one of the most widely used spices in the world and it has played a significant role in global trade and culinary traditions. This study aims to develop effective models for forecasting weekly black pepper prices. Weekly price data of black pepper from 2007 to 2020 was collected, with 95% of the data used for model building and the remaining 5% used for model validation. The study developed a conventional linear model, ARIMA (2, 1, 1), and an advanced machine learning nonlinear model, NNAR (5, 3), based on proper model selection criteria. These models were then compared to determine the best one, and it was found that NNAR (5, 3) is the most effective model for forecasting the weekly black pepper prices in Kerala.**

**Keywords:** ARIM, ANN, Price Forecasting, Machine Learning, Black Pepper Price.

## INTRODUCTION

Black pepper (*Piper nigrum*), often referred to as the "King of Spices," and "Black Gold", is one of the most widely used spices in the world, known for its pungent flavour and health benefits. Originating from the tropical forests of Kerala, India, black pepper has a rich history dating back thousands of years and has played a significant role in global trade and culinary traditions. According to the Department of Economics and Statistics, Government of Kerala, the state produced a total of 27,654 tonnes of pepper from an area of 73,732 hectares in year 2022-23. In the dynamic landscape of global trade and industry, the accurate prediction of commodity prices is crucial for informed decision-making, risk management, and economic planning. By examining historical trends, market dynamics and technological advancements, this study aims to develop effective models for forecasting black pepper prices. These models will empower stakeholders in the pepper industry to navigate challenges and seize emerging opportunities in a rapidly evolving economic environment. Time series forecasting has evolved over the years, with various methods and models developed to improve accuracy and handle complex patterns in temporal data. The foundation for ARIMA was laid in the 1970s with the work of George E.P. Box and Gwilym M. Jenkins. They introduced the ARIMA model as an extension of the Autoregressive (AR) and Moving Average (MA) models. While ARIMA models have limitations in capturing the patterns and nonlinear relationships in pepper price data, Artificial Neural Networks (ANNs) offer improvements by utilizing machine learning capabilities. The concept of ANNs dates back to the 1940s and 1950s, with pioneering work by McCulloch and Pitts (1943). However, practical applications were initially limited by computational constraints. Neural networks experienced a resurgence in the 1980s and 1990s due to advancements in parallel processing and improvements in training algorithms, particularly the back propagation algorithm.

Many recent works are being carried out in these areas. Ranjbar *et al.* (2006) presented an Artificial Neural Network (ANN) model designed for short-term electricity price forecasting in such markets. The proposed model is a four-layered perceptron neural network with an input layer, two hidden layers, and an output layer. Kulkarni and Haidar (2009) proposed a multilayer feed forward neural network model to forecast crude oil spot price direction for up to three days ahead. The optimal ANN structure was determined, and various data pre-processing methods were tested. Futures prices provide additional information for short-term spot price direction, aiding risk management for investors.

Adebiyi *et al.* (2014) proposed a stock price predictive model using ARIMA, utilizing data from the NYSE and NSE, and shows that ARIMA has strong potential for short-term prediction, competing well with existing techniques. Goyal and Kundu (2020) introduced ARIMA model for improving forecasting accuracy of Indian stock indices. By using Sensex's closing stock data, it develops a model that provides enhanced predictive accuracy, making ARIMA a suitable choice for accurate forecasting. Kathayat and Dixit (2021)

proposed forecasts wholesale paddy prices for the 2020-21 agricultural year in five major states using the ARIMA model. Results show price ranges for each state, with northern states exhibiting more significant variations. The best-fit ARIMA models were identified for each state, and these forecasts can aid stakeholders in making informed decisions about production, marketing, and consumption.

In recent years, a significant amount of research has focused on the use of conventional linear models, such as ARIMA, and various growth models for forecasting price trends, including those in the black pepper market of Kerala. While these traditional methods have been extensively studied and applied, their accuracy has often been limited due to the inherent complexity and non-linearity present in price time series data. This limitation poses a challenge, as accurate forecasting is crucial for stakeholders in the agricultural sector, including farmers, traders, and policymakers.

Given these limitations, there has been a growing interest in exploring advanced neural network models, particularly those within the machine learning domain. Unlike conventional linear models, neural networks are capable of capturing complex, non-linear patterns in the data, leading to more accurate and reliable forecasts. These models offer a promising alternative, especially in the context of highly volatile and unpredictable markets such as that of black pepper in Kerala.

Despite the potential advantages of these advanced machine learning models, the literature still lacks a comprehensive comparison between the traditional ARIMA models and the newer neural network approaches in the specific context of black pepper price forecasting in Kerala. This gap highlights the need for further research that not only explores the effectiveness of neural networks in capturing non-linear patterns but also evaluates their performance relative to conventional models. Addressing this gap could provide valuable insights into the most effective forecasting methods for this critical agricultural commodity, ultimately supporting better decision-making and market strategies.

## MATERIALS AND METHODS

**Data description.** Weekly black pepper price data starting from 2007 to 2020 obtained from Directorate of Economics and Statistics, Government of Kerala. To construct the model, 95 percent of the data has been used for model building and the remaining data reserved for model validation purposes.

**Autoregressive Integrated Moving Average (ARIMA) model.** The ARIMA model, a popular linear time series forecasting method, integrates autoregression (AR), differencing (I), and moving averages (MA) to effectively capture patterns and trends in time series data for predicting future values. By combining AR and MA polynomials, the ARIMA model forms a powerful polynomial representation, denoted as ARIMA ($p$, $d$, $q$), where $p$ represents the autoregressive order, $d$ is the differencing order, and $q$ signifies the moving average order. This model is applied across all data points in the time series, offering a versatile and efficient approach to time series forecasting.

$$y_t = \mu + \sum_{i=1}^{p}(\sigma_i y_{t-1}) + \sum_{j=1}^{q}(\theta_j \varepsilon_{t-1}) + \varepsilon_t \qquad (1)$$

where

$\mu$ : represents the mean value of the time series data.

$p$: denotes the number of autoregressive lags.

$\sigma$ : signifies the autoregressive coefficients (AR).

$q$: stands for the number of lags in the moving average process.

$\theta$ : represents the moving average coefficients (MA).

$\varepsilon$: denotes the white noise in the time series data.

Integration factor (I) indicates the number of differences ($d$) needed to make the data stationary, which is calculated as per the below equation,

$\Delta_y = y_t - y_{t-1}$, where $\Delta$ is the difference operator.

**Artificial Neural Network (ANN) model.** An Artificial Neural Network (ANN) is a non-linear computational model inspired by the structure and functioning of the human brain, designed for pattern recognition and information processing tasks. Most important feature of ANN is that it doesn't require any assumptions to be satisfied in the preprocessing stage of data, rather it is data driven model and non-parametric model. Comprising interconnected nodes organized into layers. The architecture of an ANN primarily consists of three layers: the input layer, a single hidden layer, and the output layer. Single hidden layer feed forward network is the most popular for time series modelling and forecasting.

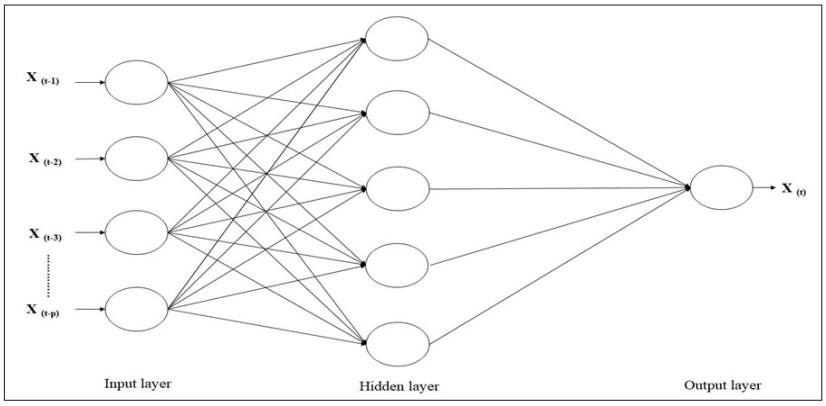


**Fig. 1**. Neural network structure.

ANN model performs a nonlinear functional mapping between the input and output which characterized by a network of three layers of simple processing units connected by acyclic links. The relationship between the output ($X_t$) and the inputs ($X_{t-1}$, $X_{t-2}$…, $X_{t-p}$) can be mathematically represented as follows:

$$X_t = f(X_{t-1}, X_{t-2}, \ldots..., Xt-p, w) + \varepsilon_t \quad (2)$$

Where $w$ is the vector of all parameters and $f$ is a function of network structure and connection weights.

## RESULTS AND DISCUSSION

**Descriptive statistics.** The descriptive statistics (Table 1) for the weekly pepper price data reveal several key characteristics. The minimum price observed is 100, while the maximum price reaches 840, indicating a wide range of price fluctuations. The mean price is 389.5, with a standard deviation of 197.71, highlighting considerable variability in the weekly prices. The Coefficient of Variation (CV) is 50.76%, which reflects a relatively high level of dispersion around the mean price. The skewness of 0.31 suggests a slight positive skew, indicating that the distribution of prices has a longer tail on the right side. The kurtosis value of 1.98 is close to 3, suggesting that the distribution of prices is relatively normal but with slightly fewer extreme values than a normal distribution.

**Table 1: Descriptive statistics of weekly pepper price data of Kerala.**

| Series | Min | Max | Mean | St. Dev. | CV (%) | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Weekly pepper price | 100 | 840 | 389.5 | 197.71 | 50.76 | 0.31 | 1.98 |

**Time series plot**

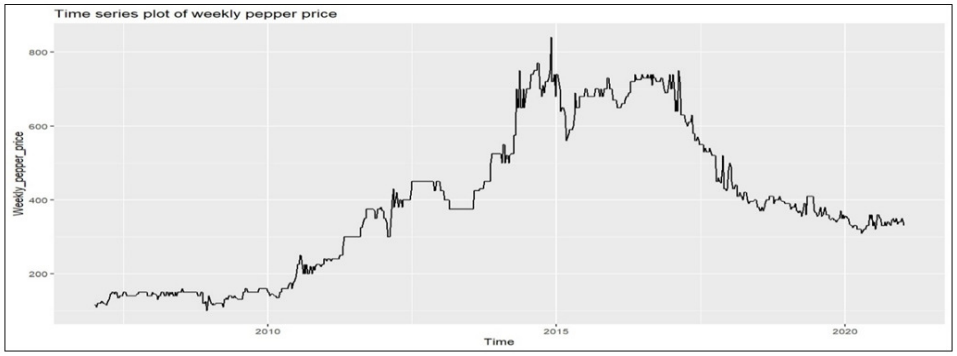


**Fig. 2.** Time series plot of weekly black pepper price of Kerala.

Fig. 2 represents the historical data trends of weekly black pepper prices in Kerala, depicted through a time series plot. The plot reveals a pattern characterized by multiple fluctuations rather than a consistent monotonic trend. This indicates that the prices have experienced numerous ups and downs over the years. However, towards the latter part of the period under study, there appears to be a discernible decreasing trend in the prices. This suggests that despite the volatility in the earlier years, black pepper prices in Kerala have generally trended downward in the more recent years of the dataset.

**Test for stationarity.** The stationarity of the raw data was tested using the Augmented Dickey-Fuller (ADF) test (Dickey & Fuller 1979 ; Phillips and Perron 1988) test (Table 2). The results showed a test statistic (*d*) of -0.5 with a p-value of 0.98 for the ADF test, and a test statistic (*Z*) of -1.48 with a p-value of 0.97 for the PP test. These high p-values indicate that the null hypothesis of non-stationarity cannot be rejected, suggesting that the raw data is not stationary. However, after differencing the data, the tests were repeated, yielding a test statistic of -9.51 with a p-value less than 0.01 for the ADF test and a test statistic of -802.9 with a p-value less than 0.01 for the PP test. These results indicate strong evidence against the null hypothesis of non-stationarity, confirming that the differenced data is stationary.

**Table 2: Test for stationarity on raw data.**

| ADF test | | PP test | |
|---|---|---|---|
| ***d*** | **p-value** | ***Z*** | **p-value** |
| -0.5 | 0.98 | -1.48 | 0.97 |
| **Test for stationarity on differenced data** | | | |
| **ADF test** | | **PP test** | |
| ***d*** | **p-value** | ***Z*** | **p-value** |
| -9.51 | <0.01 | -802.9 | <0.01 |

**ACF and PACF plots.** Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are presented below in Fig. 3-5. The ACF plot of weekly black pepper prices shows a long-term dependency extending up to 170 lags, suggesting that machine learning techniques may outperform the conventional ARIMA model. Given that the data is non-stationary, as confirmed by the Augmented Dickey-Fuller (ADF) test, we performed differencing before re-analysing the ACF and PACF plots to identify

the AR and MA orders. However, Figs. 4 and 5 reveal that the plots are inconclusive, making it challenging to determine the appropriate AR and MA orders for the model.

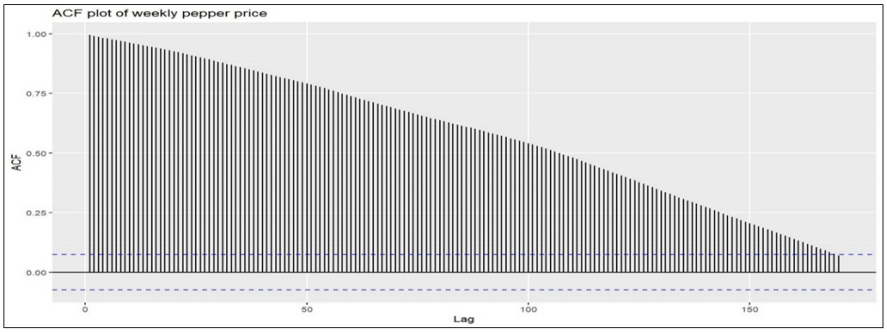


**Fig. 3.** ACF plot of weekly black pepper price (raw data) of Kerala.

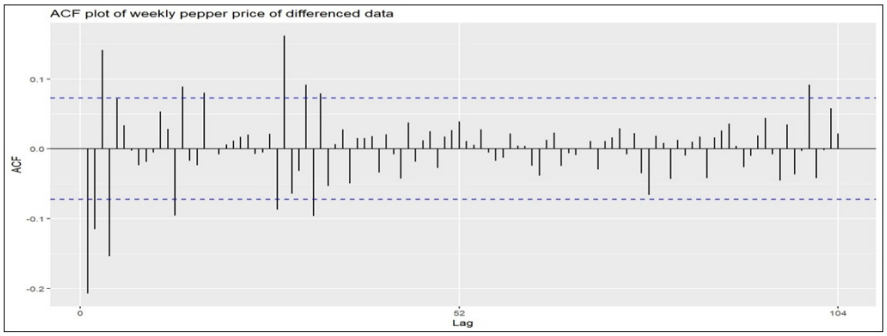


**Fig. 4.** ACF plot of differenced weekly black pepper price of Kerala.

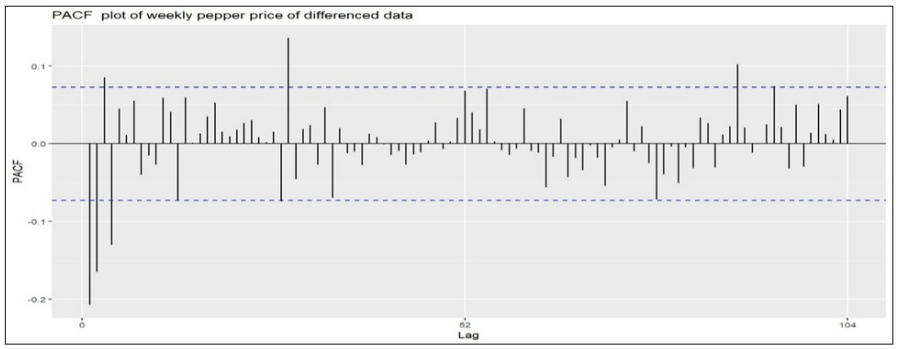


**Fig. 5.** PACF plot of differenced weekly black pepper price of Kerala.

**Model selections; ARIMA.** Due to the inconclusiveness of the ACF and PACF plots in determining the AR and MA orders for the ARIMA model, a trial-and-error method was employed. To select the appropriate AR and MA orders, orders up to two were tested, and their Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) values were calculated, values are given in Table 3. The model with the lowest AIC and BIC values was chosen as the optimal model, which in this case is the ARIMA (2, 1, 1) model. This selection process ensures that the chosen model balances goodness-of-fit with model complexity, providing the best performance for forecasting purposes.

**Table 3: ARIMA model selection based on AIC and BIC.**

| Sr. No. | Model | AIC | BIC |
|---|---|---|---|
| 1. | ARIMA (1, 1, 0) with drift | 5982.44 | 5996.06 |
| 2. | ARIMA (1, 1, 1) with drift | 5972.75 | 5990.91 |
| 3. | ARIMA (1, 1, 2) with drift | 5952.57 | 5975.26 |
| 4. | ARIMA (2, 1, 1) with drift | 5948.24 | 5970.94 |
| 5. | ARIMA (2, 1, 2) with drift | 5950.24 | 5977.48 |
| 6. | ARIMA (1, 1, 0) without drift | 5980.69 | 5989.77 |
| 7. | ARIMA (1, 1, 1) without drift | 5971.11 | 5984.73 |
| 8. | ARIMA (1, 1, 2) without drift | 5950.94 | 5969.1 |
| 9. | **ARIMA (2, 1, 1) without drift** | **5946.56** | **5964.71** |
| 10. | ARIMA (2, 1, 2) without drift | 5948.56 | 5971.25 |

After selecting the ARIMA (2, 1, 1) model, the coefficients for the AR (autoregressive) and MA (moving average) orders were estimated, as shown in Table 4. The table includes the parameter estimates, standard errors, and p-values for each coefficient. The AR (1) coefficient is -0.91 with a standard error of 0.08 and a p-value of less than 0.01, indicating it is highly significant. Similarly, the AR (2) coefficient is -0.3 with a standard error of 0.04 and a p-value of less than 0.01, also indicating high significance. The MA (2) coefficient is 0.71 with a standard error of 0.07 and a p-value of less than 0.01, showing it is significant as well. The low p-values (all less than 0.01) for each coefficient confirm their statistical significance, validating the chosen ARIMA (2, 1, 1) model.

**Model selections; ANN.** Table 5 presents the optimal configuration of hyper parameters for an Artificial Neural Network (ANN) model applied to weekly black pepper price. The model under consideration is denoted as NNAR (5, 3). Here, "5" refers to the number of lagged observations included as inputs to the model, and "3" signifies the number of nodes in the hidden layer. The network type (5-3-1) further details the structure of the ANN, indicating that there are 5 input nodes, 3 hidden nodes, and 1 output node. In total, this configuration results in 22 parameters that need to be estimated during the training process. This optimized setup is crucial for achieving effective forecasting performance by balancing complexity and computational efficiency.

**Residual analysis.** Table 6 displays the results of the Ljung-Box test (Box, 1978) applied to the residuals of two different forecasting models: the ARIMA (2, 1, 1) model and the NNAR (5, 3) model.

**Table 4: ARIMA (2, 1, 1) coefficients.**

| Parameter | Estimate | Standard error | p-value |
|---|---|---|---|
| ar (1) | -0.91 | 0.08 | <0.01 |
| ar (2) | -0.3 | 0.04 | <0.01 |
| ma (2) | 0.71 | 0.07 | <0.01 |

**Table 5: Optimum number of hyper parameters in ANN.**

| Model | Lag | Number of nodes in the hidden layers | Network type | Number of parameters |
|---|---|---|---|---|
| NNAR (5, 3) | 5 | 3 | (5-3-1) | 22 |

**Table 6: Ljung-Box test of residuals.**

| ARIMA (2, 1, 1) model | | |
|---|---|---|
| ***Q**** | **Degrees of freedom** | **p-value** |
| 6.1 | 7 | 0.53 |
| **NNAR (5, 3) model** | | |
| ***Q**** | **Degrees of freedom** | **p-value** |
| 10.13 | 10 | 0.42 |

The Ljung-Box test is used to check for autocorrelation in the residuals, which indicates whether the model has adequately captured the underlying patterns in the data. For the ARIMA (2, 1, 1) model, the test statistic ($Q^*$) is 6.1 with 7 degrees of freedom and a p-value of 0.53. For the NNAR (5, 3) model, the test statistic ($Q^*$) is 10.13 with 10 degrees of freedom and a p-value of 0.42. In both cases, the high p-values suggest that there is no significant autocorrelation in the residuals, implying that both models have effectively captured the data patterns without leaving significant autocorrelations unaccounted for.

**Model validation.** Table 7 provides a comparison of model validation metrics for the ARIMA (2, 1, 1) and NNAR (5, 3) models based on their performance on training and testing data. The metrics used for this comparison are Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

For the ARIMA (2, 1, 1) model, the RMSE and MAPE on the training data are 17.65 and 2.7, respectively, while on the testing data, these values are 26.63 and 7.04. In contrast, the NNAR (5, 3) model demonstrates better performance, with an RMSE and MAPE of 16.72 and 2.5 on the training data, and significantly lower values of 14.26 and 3.3 on the testing data. This indicates that the NNAR (5, 3) model not only fits the training data more accurately but also generalizes better to unseen data compared to the ARIMA (2, 1, 1) model.

**Fitted vs Actual plots.** Fig. 6 and 7 illustrate the Fitted vs. Actual plots for the weekly black pepper prices in Kerala, showcasing the results after applying the ARIMA and ANN models, respectively. Both plots indicate that the models were well-fitted to the training data. However, when applied to the testing dataset, the ANN model demonstrates a slight advantage over the ARIMA model. This superiority of the ANN model is further supported by the lower Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) values, as detailed in Table 7.

**Table 7: Model Validation.**

| Sr. No. | Model | Training Data | | Testing Data | |
|---|---|---|---|---|---|
| | | RMSE | MAPE | RMSE | MAPE |
| 1. | ARIMA (2, 1, 1) | 17.65 | 2.7 | 26.63 | 7.04 |
| 2. | NNAR (5, 3) | 16.72 | 2.5 | 14.26 | 3.3 |

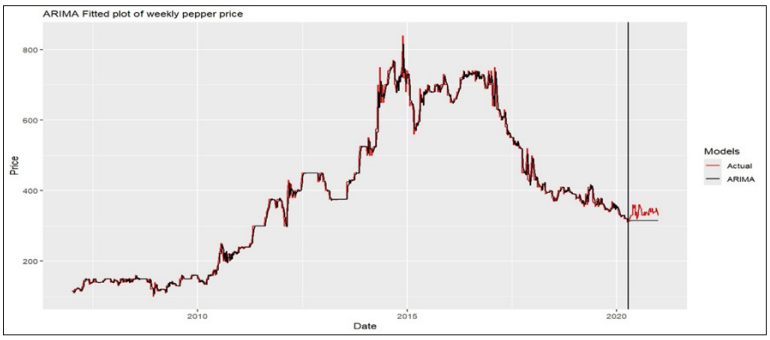


**Fig. 6**. Fitted Vs Actual plot of ARIMA (2, 1, 1) model.

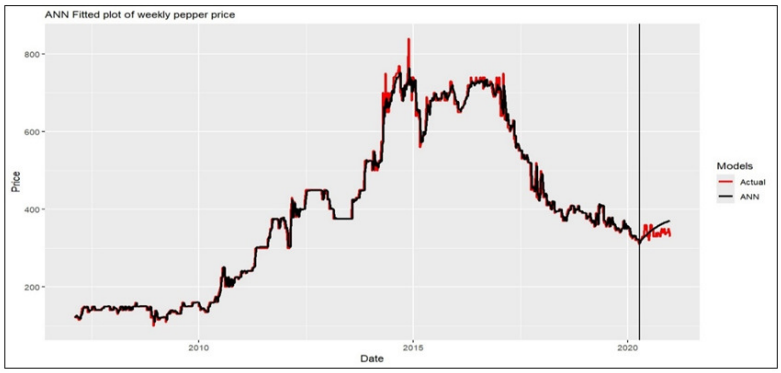


**Fig. 7.** Fitted Vs Actual plot of NNAR (5, 3) model.

## CONCLUSIONS

This study aims to develop a robust model for forecasting the price of black pepper in Kerala, India. Recognizing the increasing relevance of machine learning techniques, the study undertakes a comparative analysis of conventional linear model, specifically ARIMA, and advanced nonparametric non-linear machine learning model ANN. The comparison reveals that the NNAR (5, 3) model outperforms ARIMA (2, 1, 1) model in terms of forecasting accuracy. Consequently, the study proposes the NNAR (5, 3) model - a specific type of neural network autoregressive model as the best model for forecasting the weekly price of black pepper in Kerala. The NNAR (5, 3) model effectively captures the complex, non-linear patterns in the data, offering a significant improvement in predictive performance over traditional forecasting methods.

**Acknowledgement.** We would like to express our sincere gratitude to the Department of Agricultural Statistics, Palli Siksha Bhavana, Visva-Bharati University, for invaluable contributions to the completion of this research paper. We are especially thankful to the Directorate of Economics and Statistics, Government of Kerala, for providing the data, which was essential to our study.

## REFERENCES

Adebiyi, A., Adewumi, A., & Ayo, C. (2014). Stock price prediction using the ARIMA model. *Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time Series Analysis: Forecasting and Control* (4th ed.). John Wiley & Sons.

Dickey, D., & Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*, 427-431.

Goyal, A., & Kundu, A. (2020). ARIMA and Indian Stock Market Forecasting. *Xi'an Jianzhu Keji Daxue Xuebao/Journal of Xi'an University of Architecture & Technology*, *12*, 60–70.

Kathayat, B., & Dixit, A. K. (2021). Paddy price forecasting in India using ARIMA mode. *Journal of Crop and Weed*, *17*(1), 48–55.

Kerala State Planning Board, Government of Kerala (2023). *Economic review 2023* (Vol. 1).

Kulkarni, S., & Haidar, I. (2009). Forecasting Model for Crude Oil Price Using Artificial Neural Networks and Commodity Futures Prices. *International Journal of Computer Science and Information Security, 29*(1).

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297-303.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Phillips, P. C. B., & Perron, P. (1988). Testing for unit roots in time series regression. *Biometrika, 75*, 335-346.

Ranjbar, M., Soleymani, S., Sadati, N., & Ranjbar, A. M. (2006). Electricity Price Forecasting Using Artificial Neural Network. *2006 International Conference on Power Electronic, Drives and Energy Systems*, 1–5.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464.

**How to cite this article:** Muhammed Irshad M., Kader Ali Sarkar, Digvijay Singh Dhakre and Debasis Bhattacharya (2024). Comparative Analysis of Statistical Model and Machine Learning Algorithms in Forecasting Black Pepper Price of Kerala. *Biological Forum – An International Journal, 16*(8): 63-68.