Price Forecasting on a Large Scale Data Set using Time Series and Neural Network Models

Preetha K G[1*], K R Remesh Babu[2], Sangeetha U[3], Rinta Susan Thomas[4], Saigopika[5], Shalon Walter[6], and Swapna Thomas[7]

[1,4,5,6,7]Rajagiri School of Engineering & Technology, Kochi, India
[e-mail: preetha_kg@rajagiritech.edu.in; rintasusanthomas@gmail.com; saigopikaprasanthi@gmail.com, waltershalon@gmail.com; swapnathomas99@gmail.com]
[2,3]Government Engineering College, Sreekrishnapuram, Palakkad, Kerala, India
[e-mail: remeshbabu@yahoo.com; sangeethau2013@gmail.com]
[*]Corresponding author: Preetha K G

Received February 2, 2022; revised June 14, 2022; accepted October 18, 2022; published December 31, 2022

Abstract

Environment, price, regulation, and other factors influence the price of agricultural products, which is a social signal of product supply and demand. The price of many agricultural products fluctuates greatly due to the asymmetry between production and marketing details. Horticultural goods are particularly price sensitive because they cannot be stored for long periods of time. It is very important and helpful to forecast the price of horticultural products which is crucial in designing a cropping plan. The proposed method guides the farmers in agricultural product production and harvesting plans. Farmers can benefit from long-term forecasting since it helps them plan their planting and harvesting schedules. Customers can also profit from daily average price estimates for the short term. This paper study the time series models such as ARIMA, SARIMA, and neural network models such as BPN, LSTM and are used for wheat cost prediction in India. A large scale available data set is collected and tested. The results shows that since ARIMA and SARIMA models are well suited for small-scale, continuous, and periodic data, the BPN and LSTM provide more accurate and faster results for predicting well weekly and monthly trends of price fluctuation.

Keywords: ARIMA, SARIMA, RNN, LSTM, BPN.

1. Introduction

Vegetable product prices in India are volatile and fluctuate frequently, depending on a variety of factors such as season, weather, government policy, region, and so on [1]. Horticultural crop price forecasting is important for farmers to make informed decisions and to make their agribusiness more profitable. The prediction analysis for farmers in agriculture is important and essential for developing a crop recommendation system based on price forecasting for agricultural commodities. Currently, there is no recommendation system [2] available to decide their cost of production. Because of the lack of appropriate knowledge and data availability, the many-sided quality of product expectation is strong, influencing the essence of forecasting.

For over a decade, researchers have been working on predicting the price of horticultural crops. The objective of this paper is the study and implementation of existing time series cost prediction methods and also explores the various possible neural network methods for better accuracy. The proposed method studies and evaluate various time series models and neural network models for price prediction. The most commonly used forecasting model is Auto Regressive Integrated Moving Average (ARIMA)[3]. A seasonal Auto Regressive Integrated Moving Average (SARIMA) [4] is the extended model of ARIMA which includes seasonal parameters. Due to the unavailability of continuous and periodic data these models are not suitable for large-scale data. Neural network models are well suited for large-scale periodic data which can easily predict daily, weekly, monthly and yearly trends of price fluctuation which would be helpful for farmers to decide their cropping plan. It is clear that the existing models have flaws in long term prediction accuracy. This research gap is identified in this paper, thus, the paper tries two univariate time series models, ARIMA and SARIMA for small scale periodic data for wheat price prediction in INDIA. Later the multivariate Neural network models such as BPN [5], RNN [6] extended LSTM are used for long term prediction as it supports the planting time and harvesting time. The work reported here makes the following key contributions.

- The proposed method evaluates time series models ARIMA and SARIMA for small scale periodic prediction.
- It also evaluates the neural network models BPN and LSTM for long term predictions of wheat price.
- Various results are established through experimental evaluations, and the observations reported.

The performance of the different models are also compared in terms of their accuracy and the neural network models significantly improves the performance with respect to other time series models. This also ensures to provide a platform for customers to get an idea of the product price in the market.

The rest of this paper is organized as follows. Section 2 discusses the related studies in the crop prediction area. Section 3 describes the various time series analysis models and neural network models for cost prediction. The results are reported in section 4 and the conclusion is given in section 5

2. Related Works

According to a study of the related literature, the idea of horticultural product cost prediction was introduced in the early 1990s, and research in this field advanced dramatically during the

2000s. The researchers use a time series approach based on Box–Jenkins approach in modelling and forecasting the demand in a food company reported in [7]. The main aim of the time series analysis is to study the observations and to create a model to describe the structure of the data and then forecast the values of time series in the future. The significance of time series forecasting is applied in variety of fields. It is critical to construct an effective model in the field of applied sciences with the goal of increasing forecasting precision. The paper illustrated the usage of historical data to aid future demand forecasting and a mathematical model ARIMA was developed from the historical data. The authors claimed that ARIMA model gives more accurate results for prediction with respect to the parameters Akaike criterion, Schwarz Bayesian criterion, maximum likelihood and standard error. Researchers used an ARIMA model to forecast monthly rainfall in the Khordha district of Odisha, India [8]. Rainfall from 1901 to 1982 was used to train the model, and rainfall from 1983 to 2002 was used for testing and validation. The model is chosen based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The Nash–Sutcliffe efficiency (NSE) and coefficient of determination methods were used to validate the model's efficiency.

In [9], a seasonal autoregressive integrated moving average model was used to forecast cucumber prices when accounting for seasonal variations. The experimental results show that the SARIMA model accurately predicts cucumber market prices in previous months. But it has a 17 percent average fitting error. Authors in [10] developed a SARIMA model to forecast daily and monthly solar radiation in Seoul, South Korea, based on hourly solar radiation data from the Korean Meteorological Administration for the past 37 years. The model performance was tested and the results show daily solar radiation can be represented using the ARIMA model, while monthly solar radiation can be represented using the SARIMA model with 12 lags. In paper [11], a robust forecasting model using a Long Short-Term Memory (LSTM) neural network and Support Vector Regression (SVR) were constructed to predict phone prices in European markets. For these two methods, they conducted a comparative analysis of time series forecasting models. The authors claimed that SVMs were the best among all other models.

The methods of ARIMA and SARIMA were used to estimate Turkey's primary energy demand from 2005 to 2020 in the study [12]. The result was found to be more accurate than the sum of the individual results. The findings yielded several valuable observations about energy management, as well as policy recommendations. In [13], Wang et al. proposed an efficient model for dealing with non-stationary and extremely noisy time series using a De-noising -dependent Backpropagation (WDBP) neural network. They used the model to create an efficient algorithm for predicting stock indexes from 1993 to 2009. To demonstrate the output boost obtained, the results were compared to single backpropagation. In [14], a model was created for predicting the number of dengue cases in Ribeirao Preto, So Paulo, Brazil, using time series analysis techniques. SARIMA was used to construct the algorithm. The findings showed that prediction was efficient, and that it could help in the implementation of preventive and control measures.

Zhang et al. created an interesting data-driven model to forecast water quality in [15] They developed a predictive model to forecast the trend of dissolved oxygen using a combination of Kernel Principal Component Analysis (kPCA) and Recurrent Neural Network (RNN) to resolve the limitations of ANN, such as the inability to retain and use the accumulated temporal information. The results show better accuracy than FFNN with respect to noise sensory data. Back-propagation neural networks were used to create a model to predict long-term tidal levels in [16]. In analyzing the disposal and movement of sediments, as well as other offshore applications, an accurate prediction is critical. This neural approach delivered more reliable

results than traditional harmonic methods. Various computational components of traditional LSTM variants are discussed in [17]. The researchers looked at eight different LSTM variants for three different tasks: speech recognition, handwriting recognition, and polyphonic music modeling. They used a powerful functional analysis of variance method to assess the value of hyperparameters. The findings of 5400 experiments were then summarized.

The existing univariate time series models, ARIMA and SARIMA in the literature have faults in terms of long-term forecast accuracy. This research gap is addressed in this work and neural network models are found to be suited for large-scale periodic data, which can assist farmers in making agricultural decisions accurately.

The authors in [18] compare the performance of long short-term memory(LSTM) and ARIMA for time series analysis. When it comes to forecasting time series data, the accuracy of ARIMA and LSTM as representative approaches is impressive. These two strategies were put to the test on a set of financial data, with the findings showing that LSTM outperformed ARIMA. The authors claimed that LSTM increased prediction by an average of 85 percentage than ARIMA. Researchers in [19] reports the methods for passenger flow prediction in the metro. For prediction, the LSTM, gated recurrent unit (GRU), and back propagation network (BPN) models are used. For short-term prediction, the authors claimed that BPN is superior to LSTM and GRU is better in both short and long-term forecasting. A study on horticulture price forecasting methods reported by Girish et al in [20]. Various methods for price prediction listed in the paper. Advantages and limitations of various methods are highlighted. The authors concluded that neural network methods are more promising for horticulture price prediction.

ARIMA and SARIMA models are ideal for small-scale periodic data because it includes continuous and periodic data. It performed well with average monthly data but not with daily data. The need for a well-designed and efficient system for forecasting market volatility for large scale data is clearly brought out from the analysis of the literature section.

3. Proposed Method

During an analysis of the literature, it was observed that ARIMA and SARIMA models are primarily used to forecast agricultural product prices. But ARIMA and SARIMA model has a weakness in that it cannot detect nonlinear patterns and can only be used for stationary results. Instead, Deep learning methods such as BPN and LSTM are more suitable for large scale data and accurately forecast market fluctuations on a regular, weekly, and monthly basis. In this work, wheat price in INDIA is predicted for a short term (daily) and long term (weekly and monthly) basis using ARIMA, SARIMA, BPN and LSTM models. Results are compared and the best model is selected for wheat price forecasting.

3.1 Data Set

The historic data regarding the stock price of wheat in the US [21] dated from 14-10-2009 to 12-03-2018 was used to build both time series and deep learning models for forecasting wheat prices on daily, weekly, and monthly basis. Weather details [22] are also used to improve the accuracy of the neural network models and to compare them within multivariate models; a dataset of weather details was created for this purpose. The complete information of training and testing data is summarized in **Table 1**. For LSTM and BPN, the same range of data is used for training and testing to calculate the validation loss.

Table 1. Training and Testing Data

Model	Temporal scale	Training data	Testing data
ARIMA	Daily	15/10/2009 – 31/01/2018	01/02/2018 – 12/03/2018
	Weekly	21/10/2009 – 03/07/2017	30/06/2017 – 07/03/2018
	Monthly	21/10/2009 – 03/07/2017	30/06/2017 – 07/03/2018
SARIMA	Daily	15/10/2009 – 31/01/2018	01/02/2018 – 12/03/2018
	Weekly	15/10/2009 – 26/08/2017	27/08/2017 – 18/03/2018
	Monthly	01/11/2009 – 30/09/2015	01/10/2015 – 01/03/2018
LSTM	Daily	14/10/2009 – 24/04/2017	28/10/2009 – 24/04/2019
	Weekly	18/10/2009 – 30/04/2017	02/05/2010 – 30/04/2017
	Monthly	01/10/2009 – 01/04/2017	01/12/2010 – 01/04/2017
BPN	Daily	14/10/2009 – 12/03/2018	14/10/2009 – 12/03/2018
	Weekly	18/10/2009 – 18/03/2018	18/10/2009 – 18/03/2018
	Monthly	01/10/2009 – 01/03/2018	01/10/2009 – 01/03/2018

3.2 Data Preprocessing

Data preprocessing is an essential step because the historical data may have some quality issues. After data collection, data cleaning was performed, which included adjustments, deletion, calibration, and unification. Missing values are added, and null values are removed. ARIMA and SARIMA time series models does not tolerate the variability in data. Other than the normal data preprocessing steps, the time series model requires the non-variability in data. To test this, initially, the close price is plotted as shown in **Fig. 1**. From the figure it is clearly observed that the available historical data is not stationary. To make it stationary, logarithmic differencing is used. The result after differencing is shown in **Fig. 2**. It is clearly shown that the data become stationary after logarithmic differencing and data ready for ARIMA and SARIMA model. To substantiate the results, the Augmented Dickey–Fuller (ADF) test is conducted and the result is depicted in **Fig.4 and 5**. The high value of ADF and P- value indicate the non-stationary characteristics of data which is illustrated in **Fig. 4**. The low value of ADF and P-value in **Fig. 4** indicate the remarkable improvement in data characteristics after applying logarithmic differencing. The variance in data is noticeably low and the data become stationary. Label encoding and scaling are applied before the data is fed into neural network models.

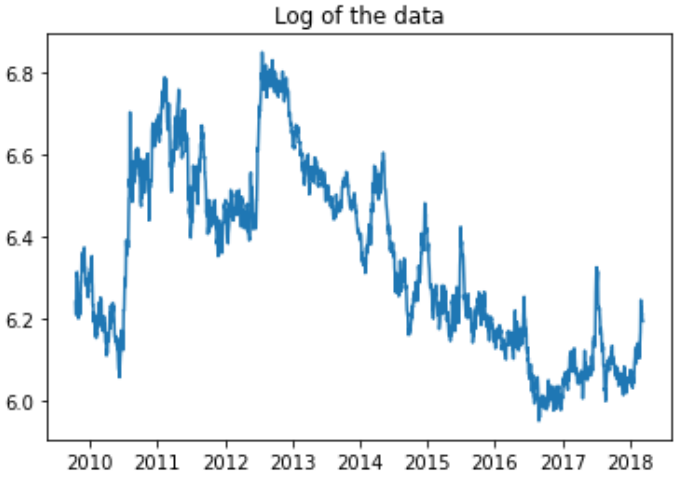


Fig. 1. Initial close prize of Wheat

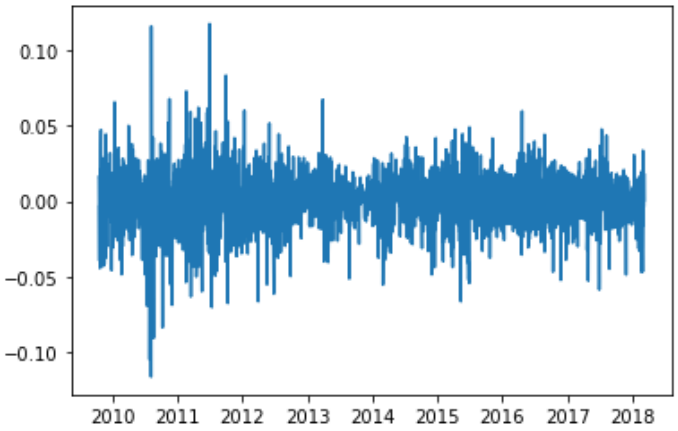


Fig. 2. Close price of Wheat after logarithmic differencing

```
1. ADF :  -2.051190601777192
2. P-Value :  0.26458826190955387
3. Num Of Lags :  15
4. Num Of Observations Used For ADF Regression and Critical Values Calculation : 2256
5. Critical Values :
         1% :  -3.4332519309441296
         5% :  -2.8628219967376647
         10% :  -2.567452466810334
```

Fig. 3. Initial Augmented Dickey–Fuller result (Data not stationary)

```
1. ADF :  -16.546237075126665
2. P-Value :  1.9634908621366017e-29
3. Num Of Lags :  9
4. Num Of Observations Used For ADF Regression and Critical Values Calculation : 2261
5. Critical Values :
         1% :  -3.4332455062745577
         5% :  -2.862819159865148
         10% :  -2.567450956377989
```

Fig. 4. Augmented Dickey–Fuller result – data become stationary after logarithmic differencing

3.3 Model Description

A time series is a set of numerical data points arranged in a logical order. It follows the movement of the selected data points over a predetermined time span, with data points collected at regular intervals. We use a model to forecast future values based on previously observed values in this approach. The time series model gathers observations over a period, with each observation representing a distinct time, and then forecasts future production based on past events. This approach can be used for the data is true, continuous, and discrete numeric or discrete symbolic. The most used time series forecasting model is the Auto Regressive Integrated Moving Average (ARIMA). To measure the frequency of events over time, the ARIMA model is used. The Autocorrelation and Partial Autocorrelation functions are used to evaluate the structure of the ARIMA model. Using a set of lagged observations, this model comprehends past data or forecasts future data in a sequence. ARIMA model can be used to predict future values from a time series based on its own past values, considering lags and

lagged forecast errors. The ARIMA model is denoted by a standard notation (p, q, d). p is the number of autoregressive terms required, d is the order of differencing required to make the data stationary, and q is the number of lagged forecast errors. To determine the value of the parameters p, q, and d, the Autocorrelation Function and Partial Autocorrelation Function are used.

Seasonal Autoregressive Integrated Moving Average (SARIMA) is an ARIMA model extension that can handle univariate time series data with seasonal components. SARIMA model is denoted by (p, q, d) (P, Q, D)m. Where p, q, d is the order of non-seasonal components P, Q, D are the order of seasonal components and m is the periodic term. The order of the SARIMA model from preprocessed data is determined by using Auto Correlation Function (ACF) plot and the seasonal value is determined by the Partial Auto Correlation Function (PACF) plot with respect to the correlation between the data and the number of lags. The ACF and PACF plots for daily price data are shown in **Fig. 5**. The x-axis in the figure represents the number of lags and the y axis represents the correlation coefficient. It is observed that the SARIMA order from the given data is 12 as indicated in the figure. Similarly, the order of the weekly and monthly models is determined.

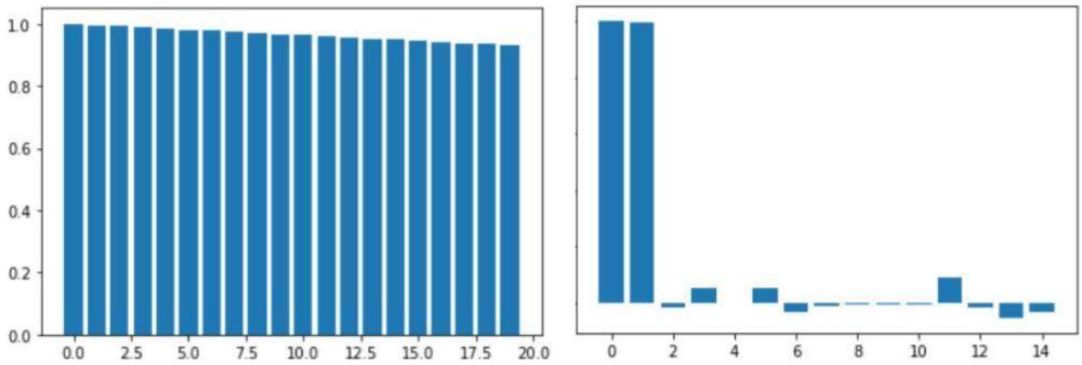


Fig. 5. ACF and PACF graph of daily price data

Recurrent Neural Network (RNN) is a class of Artificial Neural Network. The relation between the output and input nodes acts as a directed graph in a RNN, which has been commonly used to process sequential data. The RNN can remember previous information and use it to calculate the current output, i.e., the nodes between the hidden layers are linked, and the hidden layer's input includes not only the output of the input layer but also the hidden layer's output at the time of calculation. LSTM neural network is a form of RNN that excels at modeling long-range dependencies. Unlike RNNs, the LSTM architecture consists of memory blocks rather than hidden modules. A memory block is made up of one or more memory cells that are multiplicatively modulated by nonlinear sigmoidal gates. LSTM is capable of dealing with problems efficiently involving multiple input variables in an efficient way, which is critical in time series forecasting. This model overcomes the difficulty of adapting traditional linear approaches to multivariate and multiple input forecasting problems. The LSTM has the added benefit of being able to deal with uncertain period lags between critical events in a time sequence. The key concept behind LSTM is to remember previous data and use it to evaluate current performance. The RNN with LSTM units is trained on a collection of training data in supervised manner, and it uses an optimization algorithm to calculate the gradients that are needed during the optimization process, in order to vary each weight of the LSTM network in proportion to derive the error with respect to the training data. In addition to all other parameters in time series model, weather data is also used in the LSTM

model. In the proposed model a 3-layer feed-forward neural network with the first layer as the input layer, a second layer as the hidden layer for raising the model's complexity, and a third layer as the output layer, of which all are connected by weights is used. The structure of the proposed LSTM model is represented in **Fig. 6**. The historical data of price for the previous 14 days is provided to the input layer in order to predict the price for future days, weeks, or months. The method begins by dividing the dataset into 70 percent training and 30 percent testing. The training dataset, the number of epochs and the number of neurons are the inputs of the LSTM. A prediction function in LSTM predicts the next step in the dataset. A loss function and an optimization strategy must be specified when constructing a model. It employs an optimization algorithm to calculate the gradients that are needed during the optimization process in order to vary and LSTM network weight in proportion to the corresponding weight and derive the error. The Adam optimizer is used for optimization, with a learning rate of 0.01 and an output function of root mean squared error. After a sequence of various epochs for the monthly, weekly and daily models, the model is trained to correctly match the data and accurately predict future prices. The details of the input and output layer of the LSTM model for daily, weekly and monthly prediction as follows.

- Daily Average Price Prediction: The first and second LSTM input layer is set to 64 and 32 nodes respectively and the output layer is set to one node to predict the future price of a day considering the price of the previous 14 days.
- Weekly Average Price Prediction: The first and second LSTM input layer is set to 64 and 32 nodes respectively and the output layer is set to one node to predict the future price of a week considering the price of the previous 14 weeks.
- Monthly Average Price Prediction: The first and second LSTM input layer is set to 64 and 32 nodes, respectively and the output layer is set to one node to predict the future price of a month considering the price of the previous 14 months.

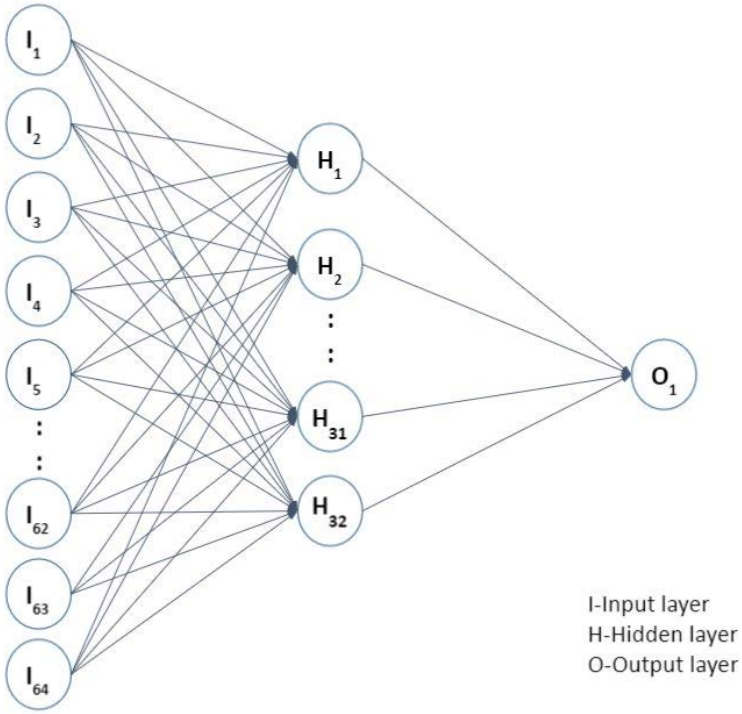


Fig. 6. Structure of the LSTM Model

Back Propagation Neural Network (BPN) is a type of neural network that is commonly used to solve complex problems. The BPN model, unlike mathematical models, does not require the data to be stationary. Many variables can influence price data, and it can have a sudden increase or decrease due to a variety of circumstances. These effects are difficult to forecast reliably, and the BPN model comes into play at these times. The BPN model is trained using a scaled dataset that is split into training and test sets. The BPN model has 5 dense layers, the first of which is the input layer, the second, third, and fourth of which are hidden layers that add more complexity to the model and improve its efficiency by forward and backward propagating across all layers. The structure of the BPN model is represented in **Fig. 7**. The BPN model, as well as the output function mean squared error, are compiled by the Adam optimizing algorithm. After a sequence of various epochs, the model is trained to correctly suit the data and forecast future prices. The following are the descriptions of the BPN model's input and output layers for daily, weekly, and monthly prediction.

- Daily Average Price Prediction: The input layer and all the 3 hidden layers of the neural network have 500 nodes each and the output layer is set to one node to predict the future price of a day.
- Weekly Average Price Prediction: The input layer and all the 3 hidden layers of the neural network have 500 nodes each and the output layer is set to one node to predict the future price of a week.
- Monthly Average Price Prediction: The input layer and all the 3 hidden layers of the neural network have 300 nodes each and the output layer is set to one node to predict the future price of a month.

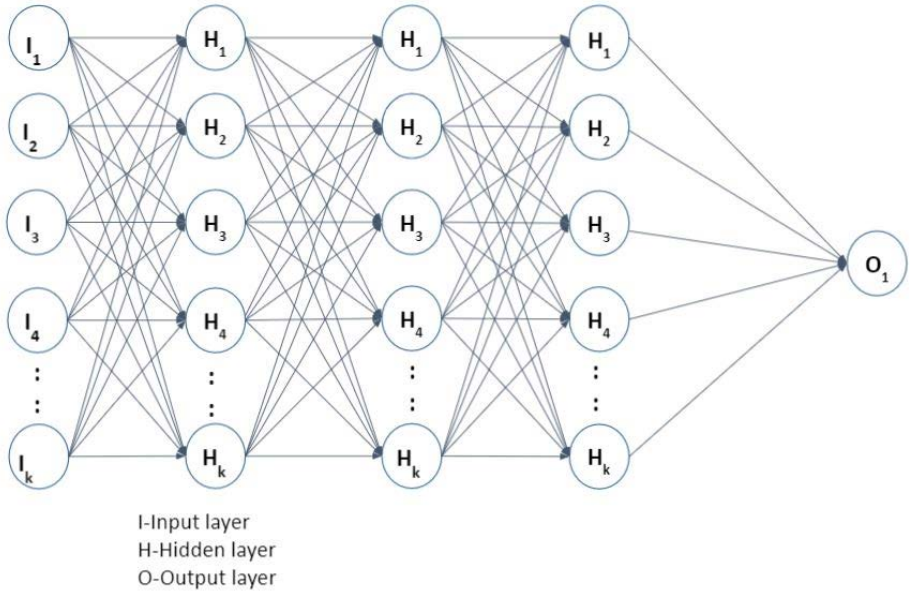


Fig. 7. Structure of the BPN model

The overall architecture of the proposed method is represented in **Fig. 8**. The different prediction systems used in this paper are the time series model ARIMA and SARIMA and the neural network models LSTM and BPN. To train the model and predict future price, time series models use the dataset's feature price, while neural network models use all of the features,

including weather information for the specific location. After it has been trained, each prediction system goes through an evaluation layer that evaluates the system's accuracy. The models were well fitted with the data, resulting in a very low loss value, indicating that the system is ready for price prediction in the future. The prices from the previous N days from the scaled dataset are fed into the neural network, and the output layer predicts prices for the coming days, weeks, or months. The activation function RELU is used by the input and hidden layers to evaluate the output, while the activation function sigmoid is used by the output layer of the BPN model and the activation function linear is used by the output layer of the LSTM model.

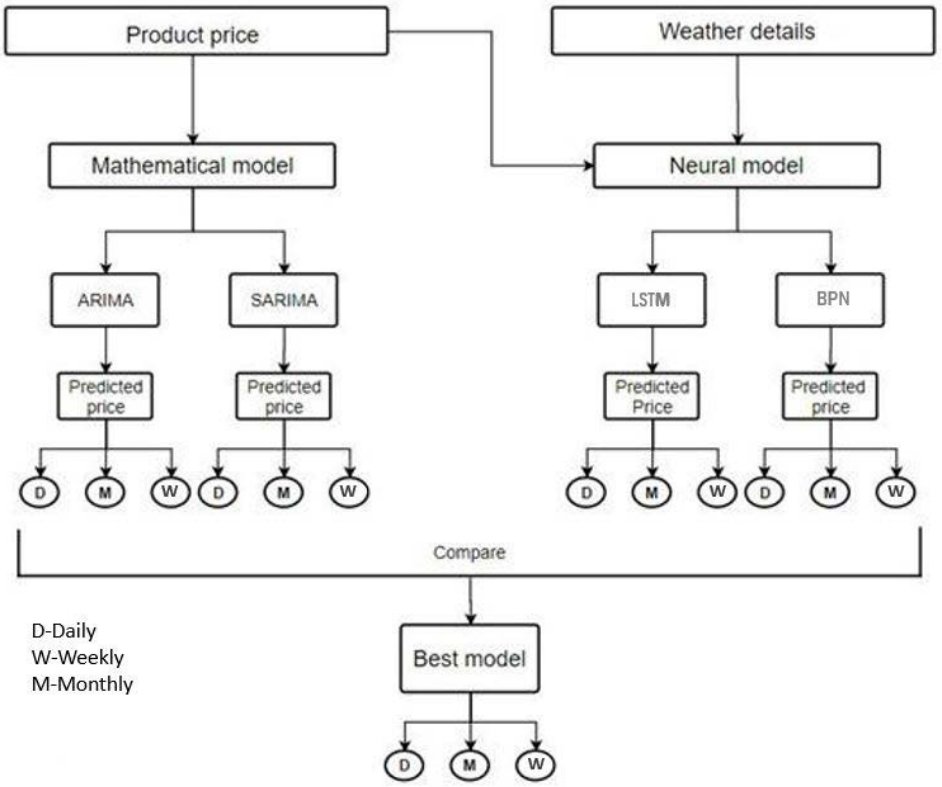


Fig. 8. Overall architecture of the proposed method

4. Results and Discussions

By comparing the experimental and simulated wheat prices over the same time span, the accuracy of the various developed models ARIMA, SARIMA, LSTM, and BPN was assessed.

4.1 ARIMA Model

Daily, weekly and monthly average wheat price is forecasted using ARIMA model. **Fig. 9** represents the daily predicted price and the actual price of wheat from 14-10-2009 to 12-03-2018. The actual price represented in orange colour and the predicted price represented in blue colour. Weekly average price and the forecasting price for the same time period is represented in **Fig. 10**. Orange colour represent the actual price and the blue colour represent the predicted

price. **Fig. 11** gives the monthly average price and the predicted price for wheat for the same time period. The negligible difference between actual and the forecasted price for the daily, weekly and monthly values in ARIMA model underscores the accuracy of the model.
The ARIMA parameters that were selected are as follows:

- Daily close price prediction corresponded to ARIMA (0,0,0) with an AIC value of -11594.461 and a BIC value of -11588.733.
- Weekly close prediction corresponded to ARIMA (0,0,1) with an AIC value of -1728.158 and a BIC value of -1719.989.
- Monthly close price prediction corresponded to ARIMA (0,0,0) with an AIC value of -253.752 and a BIC value of -251.137.

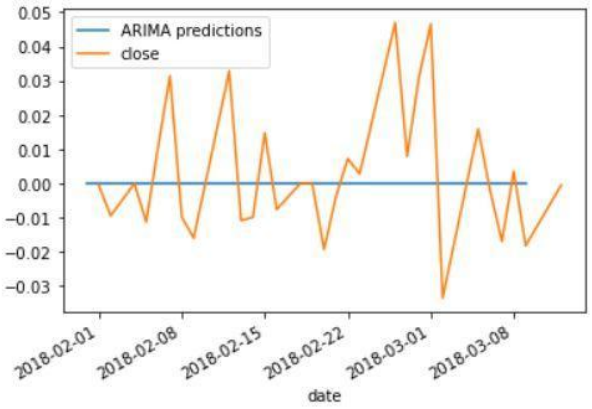


Fig. 9. Daily Price Prediction using ARIMA Model

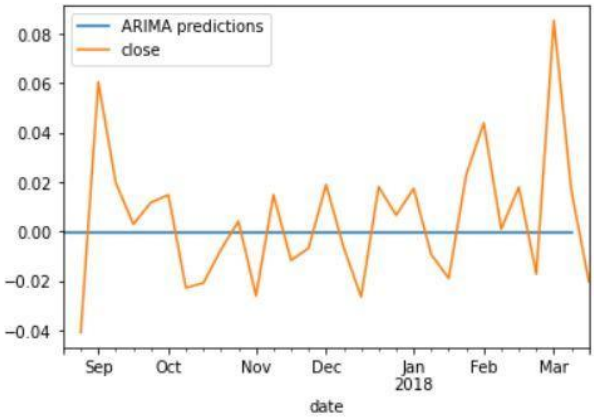


Fig. 10. Weekly Price Prediction using ARIMA Model

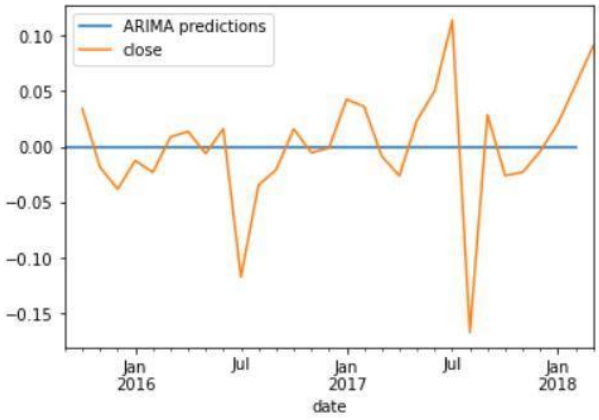


Fig. 11. Monthly Price Prediction using ARIMA Model

4.2 SARIMA Model

The seasonal aspect of the dataset is also taken into account in the SARIMA model to produce more reliable predictions. Data from 14-10-2009 to 12-03-2018 were taken for training and testing. **Fig. 12** depicts a graph that compares the daily expected wheat price to the actual price for the same time span. Orange colour line in picture represents the actual price and the blue colour represents the SARIMA prediction. **Fig. 13** depicts the weekly prediction for wheat cost. Monthly average price and the forecasting price for the same time period is represented in **Fig. 14**. The graph shows that there is very little difference between the actual price and the predicted one, which shows the accuracy of the method.

The following are the SARIMA parameters that were chosen:

- Daily close price prediction corresponded to ARIMA (0,0,0) (1,0,0)12 with an AIC value of -11594.461 and a BIC value of -11588.733.
- Weekly close prediction corresponded to ARIMA (0,0,1) (1,0,0)12 with an AIC value of -1577.214 and a BIC value of -1569.186.
- Monthly close price prediction corresponded to ARIMA (0,0,0) (1,0,0)12 with an AIC value of -253.752 and a BIC value of -251.137.

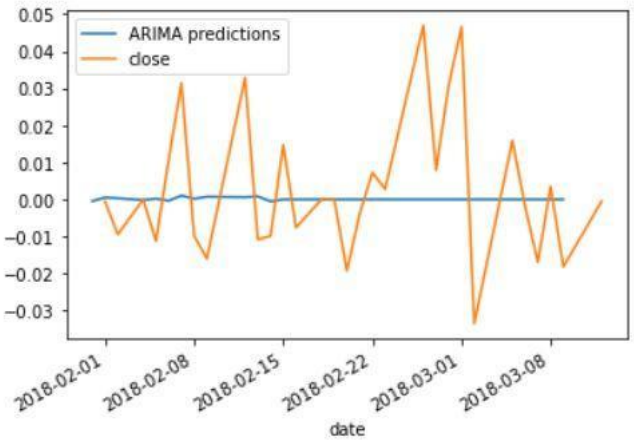


Fig. 12. Daily Price Prediction using SARIMA Model

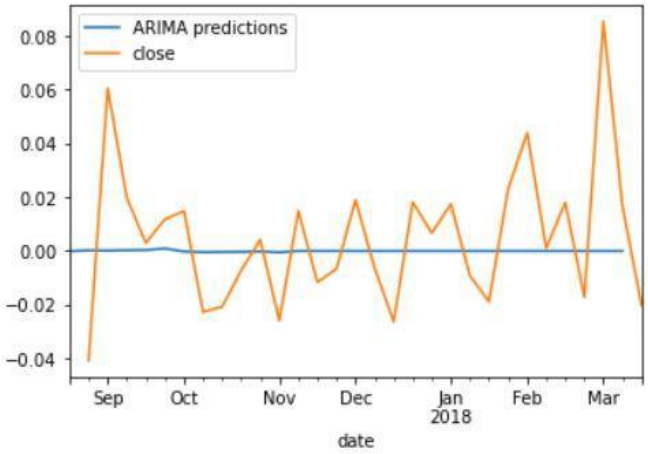


Fig. 13. Weekly Price Prediction using SARIMA Model

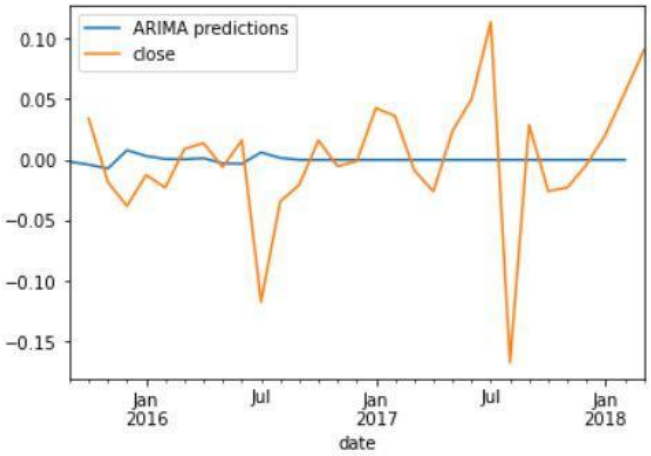


Fig. 14. Monthly Price Prediction using SARIMA Model

4.3 LSTM Model

For daily prediction the previous 14 days prices are used in the LSTM input layer to estimate for the next day, the LSTM model makes training predictions on a daily basis and future predictions for the next 60 days starting from the date the dataset ends. For weekly prediction the previous 14 weeks prices are used in the LSTM input layer to estimate the price for the next week, the LSTM model makes training predictions on a weekly basis and future predictions for the next 60 weeks starting from the date the dataset ends. For monthly prediction the previous 14 months prices are used in the LSTM input layer to estimate the price for the next month, the LSTM model makes training predictions on a monthly basis and future predictions for the 5 weeks starting from the date the dataset ends. **Fig. 15** shows the graph plotted for daily comparison of actual price and the predicted price during the training. The blue line in the figure shows the actual price and the orange line shows the prices predicted by LSTM and the red line shows the future price prediction. **Fig. 16** shows a weekly comparison of wheat prices by LSTM on actual and anticipated prices. **Fig. 17** is the graph plotted for comparing the prices predicted during training and actual price for the same month,

as well as the future prices predicted starting from the date the dataset finishes. The blue line shows the actual price in the dataset, the orange line shows the prices predicted by the LSTM model on monthly basis during training for the dates in the dataset and red line shows the future price predictions. All the statistics show a small variation between the anticipated and actual price, demonstrating the model's accuracy.

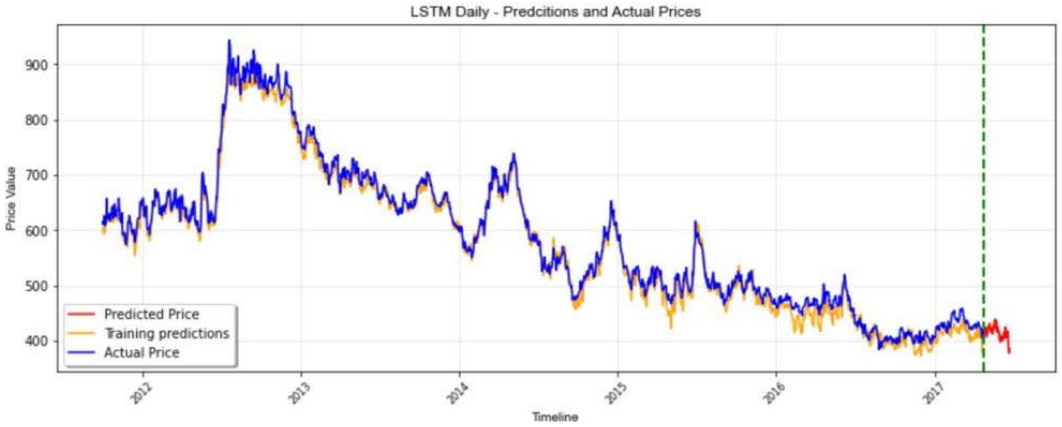


Fig. 15. Daily Price Prediction using LSTM Model

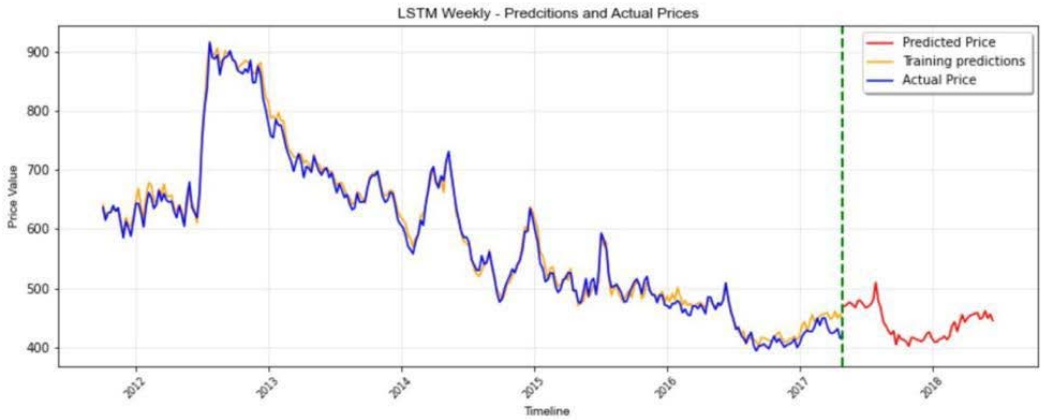


Fig. 16. Weekly Price Prediction using LSTM Model

Fig. 17. Monthly Price Prediction using LSTM Model

4.3 BPN Model

For daily prediction the previous N days prices are used in BPN input layer to estimate the price for the next day, the BPN model makes training predictions on a daily basis and the future predictions for the next 60 days starting from the date that the dataset ends. For weekly prediction the previous N weeks prices are used in the BPN input layer to estimate the price for the next week, the BPN model make the training predictions on a weekly basis and future predictions for the next 30 weeks starting from the data the dataset ends. For monthly prediction the previous N months prices are used in the BPN input layer to estimate the price for the next month, the BPN model makes training predictions on a monthly basis and future predictions for the next 10 months starting from the data the dataset ends. **Fig. 18, 19** and **20** show the graph plotted for comparing the prices predicted during training and actual price for the same day, as well as the future prices predicted starting from the date the dataset ends. The blue line shows the actual price in the dataset, the orange line shows the prices predicted by the BPN model on a daily, weekly, or monthly basis during training for the dates in the dataset and the red line shows the future price predictions. All the figures above indicate the model's accuracy by showing the smallest difference between the predicted and actual price.

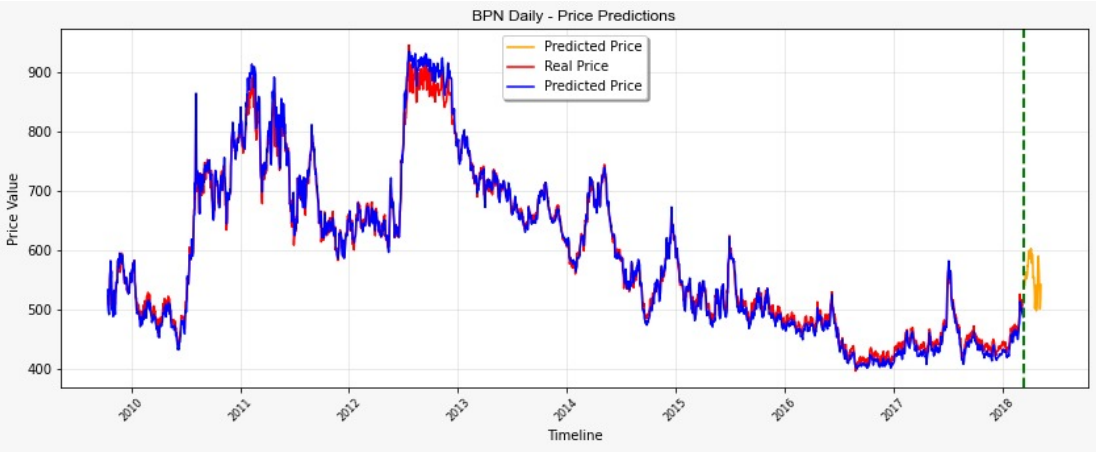


Fig. 18. Daily Price Prediction using BPN Model

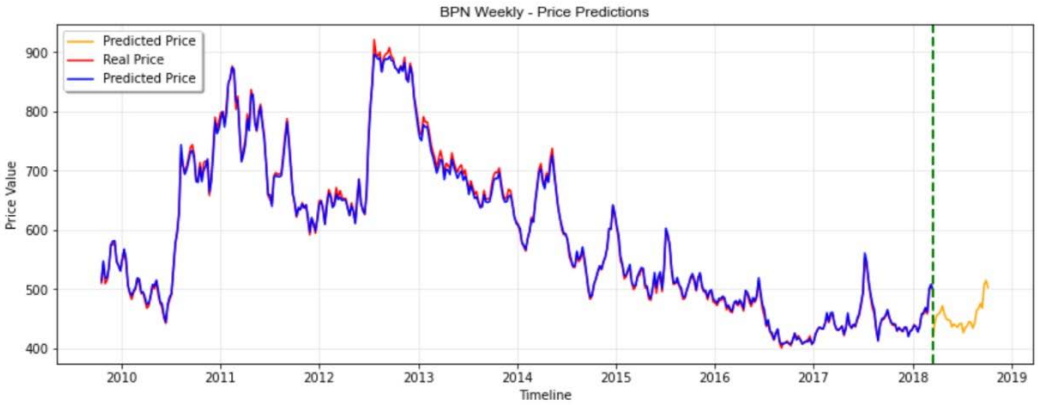


Fig. 19. Weekly Price Prediction using BPN Model

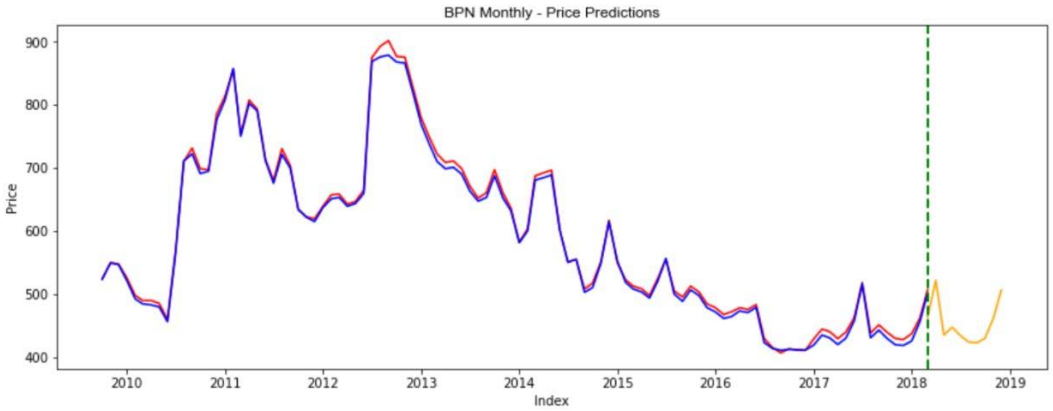


Fig. 20. Monthly Price Prediction using BPN Model

The Akaike information criterion (AIC) is a statistical measure that can be used to compare time series models. It indicates how well a model matches the available data compared to other models. As a result, when a certain model is used to depict the process that generated the data, it provides an estimate of the information lost. A model balances the goodness of fit and the complexity in this technique.
AIC is calculated using equation (1)

$$AIC = -2\left(l/n\right) + 2(k/n) \tag{1}$$

where n number of data; k estimated parameters, and the likelyhood function l can be calculated using equation (2)

$$l = -\left(n/2\right)(1 + ln(2\pi) + ln\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2\right] \tag{2}$$

The minimum AIC value shows better goodness of fit.

In this case, we utilized Bayesian information criterion (BIC) to evaluate time series models in comparison. The BIC penalizes more than the AIC when we employ a larger number of parameters (k). The equation for calculating BIC is

$$BIC = -2\left(l/n\right) + k.ln(n)/n \tag{3}$$

Model with low BIC value is selected as best model.

ARIMA (0, 0, 0) with an AIC of -11594.461 and a BIC of -11588.733 was chosen as the ARIMA model for daily close price prediction. ARIMA (0, 0, 1) has an AIC value of -1728.158 and a BIC value of -1719.989 for weekly price prediction and ARIMA (0,0,0) has an AIC value of -253.752 and a BIC value of -251.137 for monthly price prediction. The SARIMA model outperformed the ARIMA model, and the SARIMA (0, 0, 0) (1, 0, 0)12 model was chosen for daily, weekly, and monthly price prediction. The AIC and BIC values for the selected daily model are -11594.461 and -11588.733, respectively; the AIC and BIC values for the weekly model are 1577.214 and - 1569.186, respectively; and the AIC and BIC values for the monthly prediction model are -253.752 and -251.137 respectively.

The neural network model LSTM, which was implemented in a multivariate fashion for close price prediction, was well fitted with the data and was able to capture non-linearity in the data. The validation loss and the training loss of LSTM model is given in **Table 2**. The BPN model used for close price prediction, its training and validation losses are given in **Table 3**. When the training and validation losses are compared, there is a negligible difference, indicating the model is well-suited for daily, weekly, and monthly price prediction.

Table 2. Training Loss and Validation Loss of LSTM Model

Prediction	Training Loss	Validation Loss
Daily Price Prediction	0.0155	0.0315
Weekly Price Prediction	0.019	0.0250
Monthly Price Prediction	0.00198	0.0172

Table 3. Training Loss and Validation Loss of BPN Model

Prediction	Training Loss	Validation Loss
Daily Price Prediction	0.0013	0.0012
Weekly Price Prediction	0.000255	0. 000534
Monthly Price Prediction	0.0004741	0.0001713

The MSE values determined from training loss and validation loss are used to compare the performance of ARIMA, SARIMA, LSTM, and BPN models. Results are listed in **Table 4**, which shows that conventional mathematical models ARIMA and SARIMA perform better in short-term forecasting, such as daily forecasting. Both of these models are linear and univariate. Due to its ability to understand seasonal factors, the SARIMA model has slightly higher accuracy. The weekly average price is not accurately predicted by the mathematical model. The BP network and LSTM forecasting results are more reliable. The LSTM and BPN multivariate neural network models generated more accurate weekly and monthly price forecasts, suggesting that they are appropriate for long-term forecasting. The LSTM approach outperforms the other two approaches, for monthly prediction demonstrating the supremacy of deep learning algorithms. In monthly and weekly average price forecasting tests, it did not outperform the BPN by a significant margin. However, as the training data scale grows larger, the LSTM method's efficiency improves even further. The lower MSE value underscores the efficiency of the BPN model.

Table 4. Daily, Weekly, and Monthly Error value of each Model

Model	Daily MSE	Weekly MSE	Monthly MSE
ARIMA	0.0003	0.0007	0.0027
SARIMA	0.0003	0.0007	0.0026
LSTM	0.0155	0.019	0.0019
BPN	0.0012	0.0005	0.0004

5. Conclusion

The paper has developed and evaluated mathematical models such as ARIMA and SARIMA and neural network models LSTM and BPN for wheat price forecasting in India. The large data set is collected from available sources for training and evaluating the models. The models are trained for forecasting the daily, weekly, and monthly price for wheat. After training the models are tested with the dataset and the results are compared. The training and validation loss for each model is calculated. It is noticed that the difference between the training and

validation losses is minimal for LSTM and BPN models, implying that the model is well-suited for daily, weekly, and monthly price prediction.

The results show that the ARIMA and SARIMA models can accurately predict the pattern over a short period of time, but not over a long period of time. It is observed that neural network models performed better in long-term prediction and also include nonlinear multivariate data. As a result, these models can be used to build forecasting models for price prediction, assisting farmers and economic managers in making informed decisions and taking necessary actions to increase benefit.

References

[1] Nochai, Rangsan, and TitidaNochai, "ARIMA model for forecasting oil palm price," in *Proc. of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and applications*, pp. 1-7, 2006. Article (CrossRef Link)

[2] Ho, Siu Lau, and Min Xie, "The use of ARIMA models for reliability forecasting and analysis," *Computers & industrial engineering*, vol. 35, no. 1-2, pp. 213-216, 1998. Article (CrossRef Link)

[3] Gikungu, Susan W., Anthony G. Waititu, and John M. Kihoro, "Forecasting inflation rate in Kenya using SARIMA model," *American Journal of Theoretical and Applied Statistics*, vol. 4, no. 1, pp. 15-18, 2015. Article (CrossRef Link)

[4] Vagropoulos, Stylianos I., G. I. Chouliaras, Evaggelos G. Kardakos, Christos K. Simoglou, and Anastasios G. Bakirtzis, "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting," in *Proc. of 2016 IEEE International Energy Conference (ENERGYCON)*, pp. 1-6, 2016. Article (CrossRef Link)

[5] Chen, Chang-Shian, Boris Po-Tsang Chen, Frederick Nai-Fang Chou, and Chao-Chung Yang, "Development and application of a decision group Back-Propagation Neural Network for flood forecasting," *Journal of hydrology*, vol. 385, no. 1-4, 173-182, 2010. Article (CrossRef Link)

[6] Bouktif, Salah, Ali Fiaz, Ali Ouni, and Mohamed Adel Serhani, "Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting," *Energies*, vol. 13, no. 2, p. 391, 2020. Article (CrossRef Link)

[7] Fattah, Jamal, Latifa Ezzine, Zineb Aman, Haj El Moussami, and Abdeslam Lachhab, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management*, vol. 10, p.1847979018808673, 2018. Article (CrossRef Link)

[8] Swain, S., S. Nandi, and P. Patel, "Development of an ARIMA model for monthly rainfall forecasting over Khordha district, Odisha, India," *Recent Findings in Intelligent Computing Techniques*, pp. 325-331, 2018. Article (CrossRef Link)

[9] Luo, Chang Shou, Li Ying Zhou, and Qing Feng Wei, "Application of SARIMA model in cucumber price forecast," *Applied Mechanics and Materials*, vol. 373, pp. 1686-1690, 2013. Article (CrossRef Link)

[10] Alsharif, Mohammed H., Mohammad K. Younes, and Jeong Kim, "Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea," *Symmetry*, vol. 11, no. 2, p. 240, 2019. Article (CrossRef Link)

[11] Bakir, Houda, GhassenChniti, and Hédi Zaher, "E-Commerce price forecasting using LSTM neural networks," *International Journal of Machine Learning and Computing*, vol. 8, no. 2, 169-174, 2018. Article (CrossRef Link)

[12] Ediger, Volkan Ş., and SertacAkar, "ARIMA forecasting of primary energy demand by fuel in Turkey," *Energy policy*, vol. 35, no. 3, pp. 1701-1708, 2007. Article (CrossRef Link)

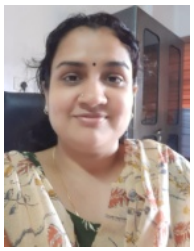
[13] Wang, Jian-Zhou, Ju-Jie Wang, Zhe-George Zhang, and Shu-Po Guo, "Forecasting stock indices with back propagation neural network," *Expert Systems with Applications*, vol. 38, no. 11, pp. 14346-14355, 2011. Article (CrossRef Link)

[14] Martinez, Edson Zangiacomi, and Elisângela Aparecida Soares da Silva, "Predicting the number of cases of dengue infection in Ribeirão Preto, São Paulo State, Brazil, using a SARIMA model," *Cadernos de saude publica*, vol. 27, pp. 1809-1818, 2011. Article (CrossRef Link)

[15] Zhang, Yi-Fan, Peter Fitch, and Peter J. Thorburn, "Predicting the trend of dissolved oxygen based on the kPCA-RNN model," *Water*, vol. 12, no. 2, p. 585, 2020. Article (CrossRef Link)
[16] Lee, Tsong-Lin, "Back-propagation neural network for long-term tidal predictions," *Ocean Engineering*, vol. 31, no. 2, pp. 225-238, 2004. Article (CrossRef Link)
[17] Greff, Klaus, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber, "LSTM: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222-2232, 2017. Article (CrossRef Link)
[18] Siami-Namini, Sima, Neda Tavakoli, and Akbar SiamiNamin, "A comparison of ARIMA and LSTM in forecasting time series," in *Proc. of 2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pp. 1394-1401, 2018. Article (CrossRef Link)
[19] Sha, Shouwei, Jing Li, Ke Zhang, Zifan Yang, Zijian Wei, Xueyan Li, and Xin Zhu, "RNN-based subway passenger flow rolling prediction," *IEEE Access*, vol. 8, pp. 15232-15240, 2020. Article (CrossRef Link)
[20] Hegde, Girish, Vishwanath R. Hulipalled, and J. B. Simha, "A Study on Agriculture Commodities Price Prediction and Forecasting," in *Proc. of 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pp. 316-321, 2020.
[21] https://www.kaggle.com/nickwong64/daily-wheat-price
[22] https://www.kaggle.com/mahirkukreja/delhi-weather-data
[23] Box, George EP, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

Preetha K G has completed her Ph. D in Mobile Ad hoc Networks from Cochin University of Science and Technology in 2018. She has completed her M Tech and B Tech Degree in Computer Science and Engineering. She has been associated with Rajagiri School of Engineering & Technology since 2004 and is now working as an Associate Professor in the Department of Computer Science & Engineering. She has around 20 years of academic experience. Her research interests include Mobile Computing, Wireless Networks, Ad-hoc Networks, Data Analytics etc. She has around 25 national and international conference papers (IEEE, Springer, ACM) and international journal papers (SCI/Scopus Indexed). She is a member of ISTE and CSI. She is the recipient of the Research Excellence award in 2018. She is also a reviewer of international conferences and journals and a registered research guide in APJ Abdul Kalam Technological University. She has also a National Patent published to her credit.

K R Remesh Babu holds BSc. degree in Mathematics, B.Tech in Information Technology and ME in Computer Science and Engineering. He received Ph.D. degree in cloud computing from Cochin University of Science and Technology (CUSAT). Currently he is working as Professor in the department of Information Technology, Government Engineering College, Palakkad, India. He has published several research papers in International Journals and Conferences. His research interests include cloud computing, Internet of Things, Machine learning, Wireless Sensor Networks, and Big Data Analytics.



Sangeetha U received B Tech degree in Information Technology from the Cochin University of Science and Technology (CUSAT), Kochi, India and MTech degree in Computer Science and Engineering from NIT Trichy. She received PhD in from NIT Kozhikode in 2022. Currently she is working as Associate Professor and Head in the department of information technology at Government Engineering College, Palakkad, India. She has published more than ten research articles in international journals and conferences. Her research interest includes internet of things, wireless communication, computer networks and cloud computing.

Rinta Susan Thomas currently a B Tech final year Computer Science student at Rajagiri School of Engineering & Technology under APJ Abdul Kalam Technological University. She has completed her higher secondary education in computer science from Marian Senior Secondary School in the year 2017. Her research interests are Machine Learning and Web Development.

Saigopika R is currently a B Tech final year Computer Science & Engineering student at Rajagiri School of Engineering & Technology under APJ Abdul Kalam Technological University. She has completed her higher secondary education in computer Science at St. Aloysius HSS in 2017. Her research interests are Web development, Programming, and Machine Learning.

Shalon Walter is currently a Bachelor of Technology in Computer Science & Engineering final year student at Rajagiri School of Engineering &Technology affiliated under APJ Abdul Kalam Technological University and will be completing her course by the year 2021. She has completed her higher secondary education in computer science at Good Shepherd Higher Secondary School in 2017. Her research interests are Machine learning, soft computing, Computer vision, Web development etc.

Swapna Thomas currently a B Tech final year Computer Science & Engineering student at Rajagiri School of Engineering &Technology affiliated under APJ Abdul Kalam Technological University and she has completed higher secondary education in Computer Science from Placid Vidya Vihar in 2017. Her research interests are Web development, Block Chain and Machine Learning.