International Journal of Multidisciplinary Research and Growth Evaluation.

Machine learning prediction of agricultural produces for Indian Farmers using LSTM

Dixika Grewal [1*], Moayed D Daneshyari [2]
[1-2] Department of Computer Science, California State University, East Bay, Hayward, California, United States

* Corresponding Author: **Dixika Grewal**

Article Info

ISSN (online): 2582-7138
Volume: 03
Issue: 05
September-October 2022
Received: 09-07-2022;
Accepted: 07-08-2022
Page No: 113-119
DOI:
https://doi.org/10.54660/anfo.2022.3.5.5

Abstract
The agriculture sector is India's primary source of national income and occupation. Today India is considered a global agricultural powerhouse, but the agriculture sector suffers from serious maladies. As a result, Indian agriculturists and farmers are poor. As per the experts, one of the major factors contributing to farmers' poverty is poor access to reliable and timely market information and the unavailability of commodity price forecasts. In this study, our main goal is to build a web application for Indian farmers with a simple user interface to provide them easy access to current and historical prices of their commodities as well as predict daily, weekly and monthly commodity prices using a long-short term memory (LSTM) machine learning model. For the price forecast, we evaluated different methods based on input data type and Mean Square Error (MSE). Upon evaluation, we used the LSTM model for its efficacy and precision on time-series data and lowest MSE among contemporaries.

Keywords: Long Short-Term Model, Recurrent Neural Network, Price Prediction, Time Series, Machine Learning Forecasting, Indian Farmers, Agriculture Forecasting

1. Introduction
The agriculture sector is the main source of national income and occupation in India. The agricultural sector contributes almost 18% to the Indian GDP [3] and 58% of India's population relies on farming as a source of livelihood. Currently, India is a global agricultural powerhouse, but the sector suffers from serious indispositions, which are keeping Indian agriculturists and farmers impoverished. The major factors contributing to the poor status of Indian farmers are inadequate access to reliable and timely market information for the farmers and the absence of supply and demand forecasting. The Indian Agriculture department provides an abundance of raw data on the government website (https://enam.gov.in/web/dashboard/trade-data), but most of the farmers in India are uneducated and non-tech savvy. This data is of no use to them. We want to contribute our work to solve these problems with the help of machine learning models. The goal of this study is to provide an agriculture visual interface to show the farmers the historical prices of their commodities as well as predict the daily, weekly, and monthly commodity prices.

2 Problem Statement
Livelihood of the farmers in India depends directly on the prices of the cultivated crops in the local agriculture marketplaces, commonly known as mandi. For making any income from the produce, it is necessary for them to be able to sell when the total selling price is more than their cost. These costs include the cost of storage facilities in case the farmer plans to wait to sell at a speculated better price in future. As such, it is important for the farmer to know at what price her produce may sell at some time in future.

In this study, we want to provide them with mandi price information including, current prices, historical prices, and most importantly, short-term, and long-term predictions of commodity prices based on selected mandi and specific crop. The agriculture commodity price prediction will not only be helpful for farmers, but also for Indian policymakers and administrative offices.

For instance, if a farmer can check the future prices of any commodity, then they will be able to decide what commodity crop they should harvest in the future. Apart from farmers, price prediction functionality can also help government officers and policymakers to design and develop government subsidy schemes and import/export schemes.

3 Literature Review

The agricultural sector lacks technological advancements, but the computer science field can provide tools, such as big data and machine learning, to improve the situation. Even though the application of machine learning models nowadays largely focuses on predicting stock price and cryptocurrency predictions. There are some scientists who are researching different techniques and computational models for efficient price predictions of agricultural commodities around the world. We worked to understand their work by reading journal articles and research papers to acknowledge the pros and cons of each technique. The related works are:
We weighed them considering their methodologies, Mean Absolute Error (MAE) / Mean Square Error (MSE), and the provided results of each of the proposed techniques [23]. The comparative MAE and MSE of techniques are given in Figure 1.

Techniques	MAE	MSE
Arfima	0.016034	0.026589
Arima	0.016043	0.024775
Neural Networks	0.024259	0.026
LSTM	0.0155	0.00502

Fig 1: Mean Absolute Error (MAE) / Mean Square Error (MSE) of each technique

Apart from LSTM which is a special neural network, the other two good models are Autoregressive Integrated Moving Average (ARIMA) and Neural Network. As per research mentioned in [17], Neural Network is the most feasible and efficient and has better results in forecasting agricultural commodity prices. ARIMA stands for Auto-Regressive Integrated Moving Average. It is a variation of Box Jenkins models used to predict stationary or non-stationary time-series data.
The authors also compared the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Square Error (MSE) among others to evaluate the results for efficacy and precision of the techniques.
In Figure 2, the results show a progressive inclination toward the use of neural networks among all the prediction models from 2011 to 2020 [24].

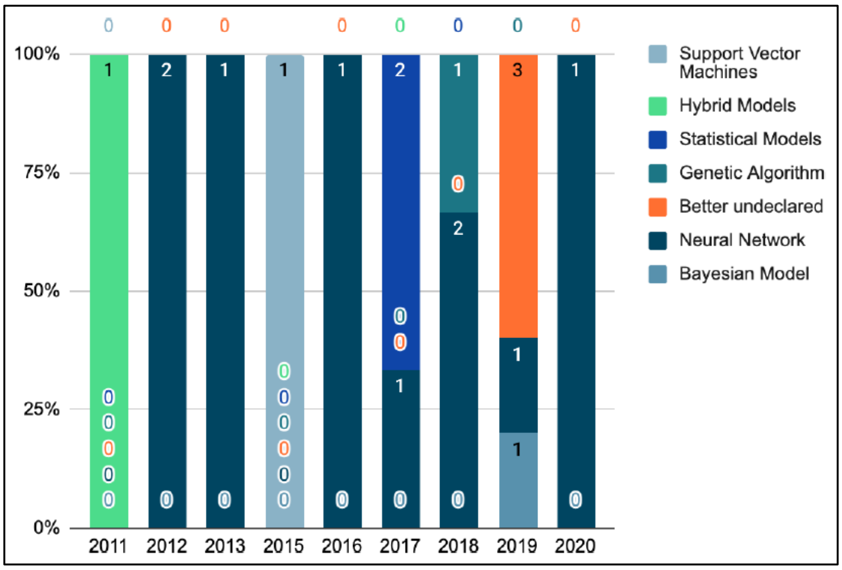

Fig 2: Prediction Models from 2011 to 2020

In the literature review [20], These study results concluded that the Neural Network is a better model for forecasting agricultural commodity prices than the Autoregressive integrated moving average (ARIMA) model. The review states that the Neural Network model is suitable for nonlinear time series data and the ARIMA model is suitable for linear time series data.
As per the paper [15], the neural network presents a good alternative for "short term" forecasting, while the Box–Jenkins method performs better for very short-term forecasting. IT also states that neural network even without a hidden layer works in similar ways as Box–Jenkins method.
In [17], the authors presented the Neural Network approach for multivariate time series data. They used the dataset of flour prices from three cities, and from training and testing results, they concluded that the Neural Network model can be used for forecasting.
In [18] as well, the Neural Network approach was used for

electrical load forecasting. They used the characteristics of the Neural Network to learn from the relationship between the historical data, current, and future temperatures. Based on the testing data, the result of the Neural Network is very satisfactory.

4 Methodology

We developed the web app in React for the front-end and will use the Django framework for the server-side interaction. For price prediction, we implement the LSTM - Recurrent Neural Network Model to train and test the historical data. The historical data is non-linear computational data which is ideal for time-series prediction techniques. There are two prediction models in the time-series technique (a) classical Box–Jenkins Models and (b) machine learning models (RNN). The classical Box–Jenkins Models only work with linear data, whereas the machine learning models (RNN) work well on a wide range of data including nonlinear data. Furthermore, as per the research regarding the prediction of agriculture commodities prices in developing countries like India [13-19], Recurrent Neural Network is specified best for nonlinear, adaptable, and strongly mapped data. Therefore, we will be using the LSTM (special recurrent neural network) model, over other models like the ARIMA model, for predicting the agriculture commodity price.

To start with, we load the commodity price historical data from the National Agriculture Market website. This data is not readily available, so we use internal APIs to get this data for preprocessing. We will use "pandas" data frames to normalize and preprocess the data into datasets. Based on these datasets, we will build the Long Short Term Memory (LSTM) special Recurrent Neural Networks to train 80% of the time-series dataset. LSTM provides the ability to memorize and update the new information or delete it if irrelevant. We build three different training models monthly, weekly, and daily and save them in 3 different files. With these trained models, we start making predictions and evaluate these predictions to the test data (20%). We tuned the model for finer results and minimal Mean Absolute Error (MAE). The predicted results will be provided by a graph using Matlab Plot.

4.1 Frameworks and Technologies

Frameworks and Languages

- Python-To develop and train LSTM model
- React-used to build front end
- Django - used to build backend and server Libraries
- Keras-Keras is an open-source software library that provides a Python interface for artificial neural networks. It is a high-level neural network library that runs on top of TensorFlow. We used it to build and load sequential LSTM models.
- Tensor flow-Tensor Flow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on the training and inference of deep neural networks.
- Pandas-pandas is a software library written for the Python programming language for data manipulation and analysis. It is used to store real-time data as datasets and cleaning and preprocessing.
- Scikit-Learn - Scikit-learn is a free software machine learning library for the Python programming language. It is used to normalize the columns of pandas data frames for the machine learning model.
- Numpy-Numpy is Python's library used as a powerful and interpolable tool for numerical computing.
- Matlab plot-Matplotlib is a comprehensive Python library used for creating static, animated, and interactive visualizations in Python.

4.2 Curated Dataset

The dataset was parsed from the Indian National Agriculture Market website https://enam.gov.in/. The dataset contains a total of 5 features. The details for them are as follows:

1. Min Price-It is the market minimum price for the commodity for that particular day.
2. Max Price-It is the market maximum price for the commodity for that particular day.
3. Modal Price-It is the average market price for the commodity over a two-month period time frame.
4. Commodity Arrivals-It is the quantity of the commodity arrived at the selected market located within the selected timeframe.
5. Commodity Traded-It is the quantity of the commodity traded at the selected market located between the selected timeframe.

This data is then preprocessed and normalized by removing the noise, dealing with the missing values, and transforming the input value before feeding it to the training and testing model. We removed all irrelevant columns. The top 5 rows of the dataset are given below in Figure 3.

Date	Model Price	Min Price	Max Price
2020-11-24	2751	2200	2811
2020-11-26	2751	1625	2801
2020-11-27	2781	1696	2851
2020-11-28	2800	1635	2800
2020-11-29	2800	1821	2827

Fig 3: Columns of the dataset, which will be divided into training and testing datasets

4.3 LSTM Model

The LSTM model was trained using the parameters listed in Figure 4. Figure 5 lays out the procedure to build the model using Keras library in Python.

Parameters	Inputs
neurons	100
activation function	linear
dropout	20%
loss function	MSE
optimizer	adam
window length	5
training epoch	20
batch size	32

Fig 4: LSTM model parameters

```
procedure build_lstm_model(input_data, output_size, neurons=100, activ_func='linear',
            dropout=0.2, loss='mse', optimizer='adam'):
    model = Sequential()
    model.add(LSTM(neurons, input_shape=(
            input_data.shape[1], input_data.shape[2])))
    model.add(Dropout(dropout))
    model.add(Dense(units=output_size))
    model.add(Activation(activ_func))
    model.compile(loss=loss, optimizer=optimizer)
    return model
```

Fig 5: LSTM model algorithm

4.4 System Architecture and Data Flow

Final app consists of a backend web server and front end web app. The server is developed in Python using the Django framework. The server contains different views for prediction and for historical data. For the prediction, we trained models for the predictive interval [a day, week, month] and were served on the disk of the server and accessed by the Django views. For historical data, Django returns the historical time series data from web API and then plots on the Front-end. For the front end, the web app connects API for each of the views and displays the results on an easy to understand user interface. Material UI and Chart JS frameworks used to develop a clean and easy to use user interface. The overall architecture is shown in Figure 6.

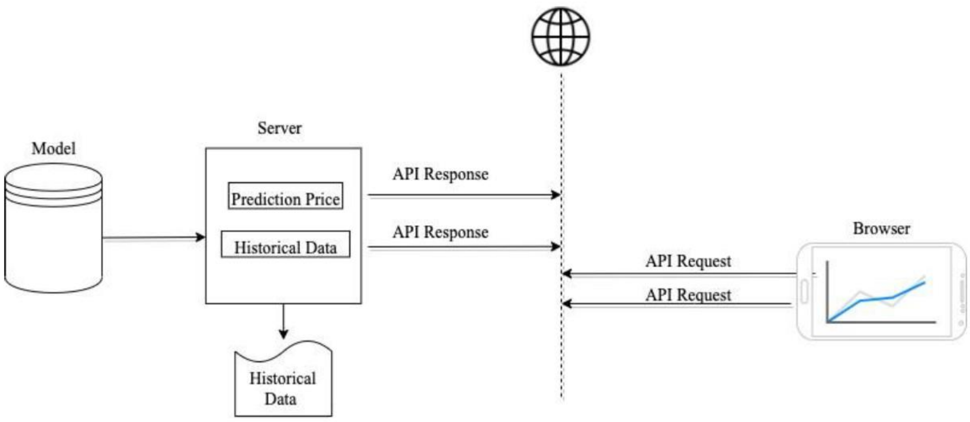

Fig 6: Overall architecture

5 Result

As laid out in the system architecture, a backend server and a frontend app was developed and deployed for generating and visualizing the historical and predicted prices of the commodities, namely cotton, paddy and wheat for four *mandis* in Hisar district of Haryana, India. The prediction for next day prices for paddy in Adampur mandi, Hisar was least uncertain and had an MSE of 0.06 as shown in Figure 8. The dataset for paddy in Adampur mandi, Hisar divided into 80:20 shown in Figure 7. The graph showing actual and predicted values in Figure 9.

Figures 10, 11, 12 shows webapp screenshots of American Cotton in *Mandi* Adampur, Hisar, Haryana for daily, weekly and monthly intervals.

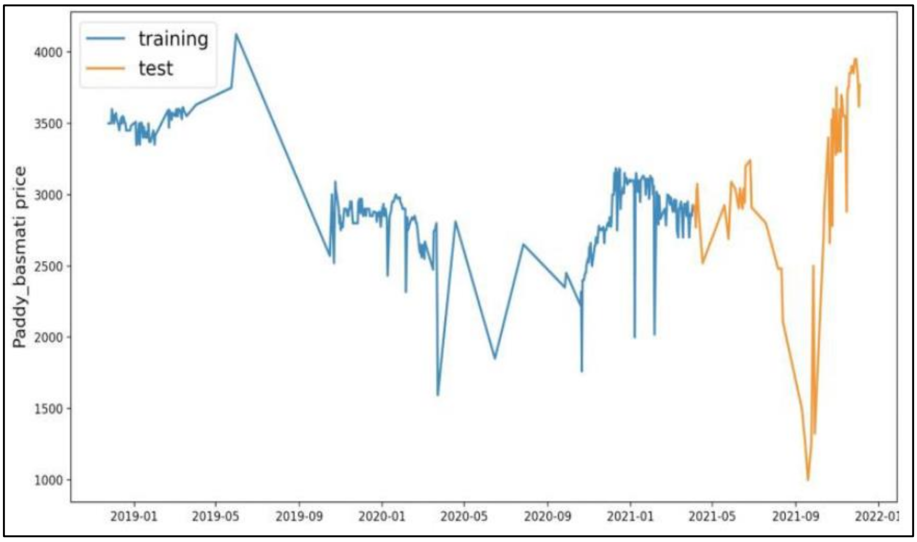

Fig 7: Graph showing data split into two subset 80% train and 20% test

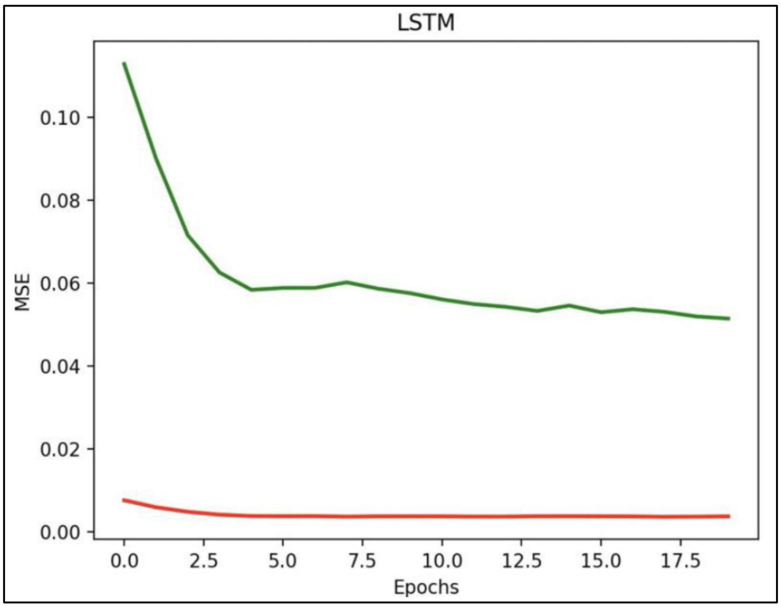

Fig 8: MSE = 0.06 for paddy in Adampur, Hisar; no over fitting and under fitting

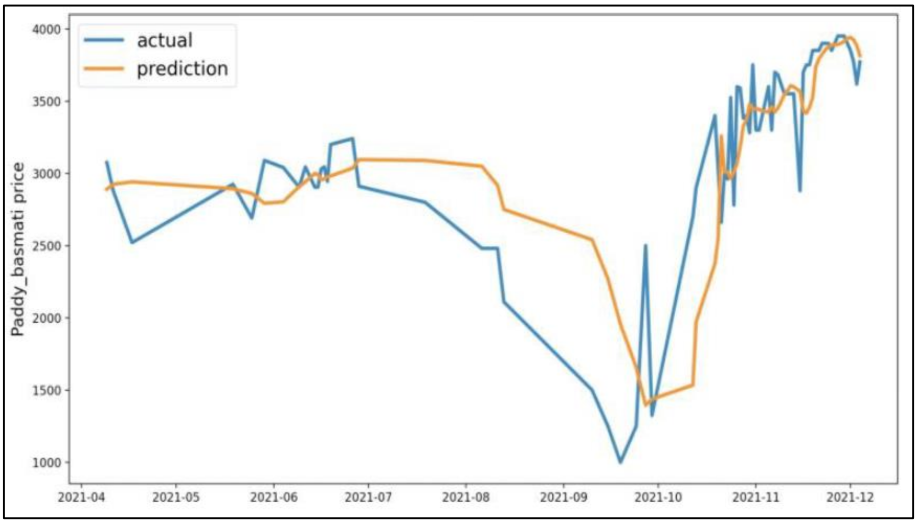

Fig 9: Prediction and actual prices of the Paddy in Adampur

Fig 10: Interactive graph and predicted value of Adampur Mandi, American cotton

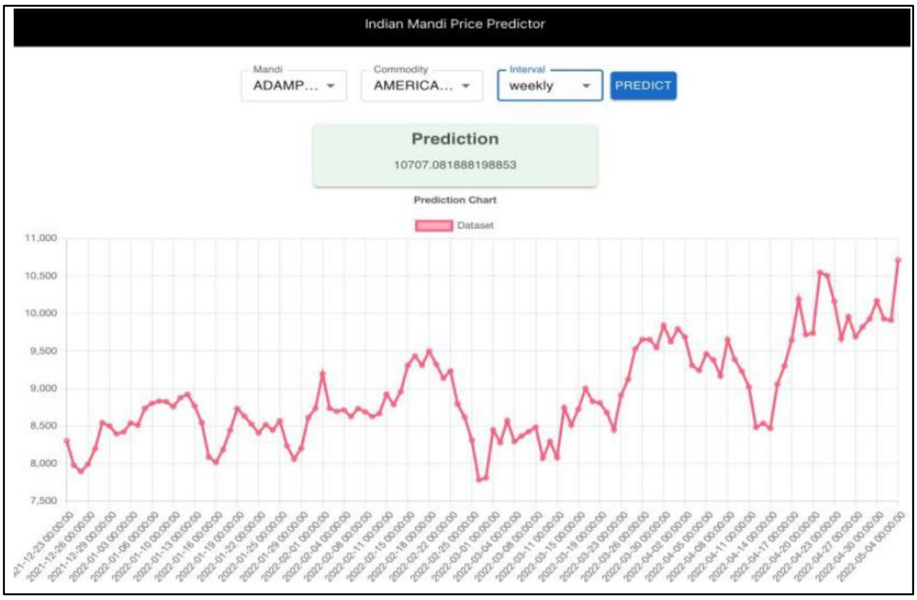

Fig 11: Historical prices and weekly prediction of Adampur Mandi, American cotton

Fig 12: Monthly predicted price and real-time price graph Adampur Mandi, American cotton

6 Conclusion

The commodity price prediction is a much needed advancement in the field of Indian agriculture. The long-short-term model prediction to forecast commodity future prices is solely based on historical time series data. It definitely can be useful to the farmers. It is able to predict and give a good ballpark most of the time, but not always. Agriculture commodity prices are dependent on many other factors apart from historical data, like climate changes, weather forecasts, import/export policies, and the global market. Henceforth, any single prediction model should not be taken for granted. Personally, we don't think any of the commodity prediction models taken for granted can be blindly relied on.

7 Future Work

The future work should focus on finely tuning weekly and monthly models. Alongside, the applied LSTM model can be integrated with one or more models like RNN and ARIMA model for better long-term predictions. As the price of Agriculture commodities are dependent on several factors including weather forecast, and import and export policies, a possible extension of the prediction system would augment it with a weather forecast and news feed analysis from social media platforms such as Twitter, where emotions are gauged from the articles. This sentiment analysis can be linked with the LSTM to better train weights and further improve accuracy.

8 References

1. A Das, M Senapati, J John. Impact of agricultural credit on agriculture production: an empirical analysis in India, Reserve Bank of India Occasional Papers.
2. J Ahmad, D Alam, MS Haseen. Impact of climate change on agriculture and food security in India, International Journal of Agriculture, Environment and Biotechnology.

3. KM Sabu, TKM Kumar. Predictive analytics in agriculture: forecasting prices of areca nuts in Kerala, Procedia Computer Science.
4. D Hellerstein, M Scott. The Influence of Rising Commodity Prices on the Conservation Reserve Program, Economic Research Service, Washington, DC, USA, 2011.
5. AC Harvey, PHJ Todd. Forecasting economic time series with structural and Box-Jenkins models: A case study, Journal of Business & Economic Statistics, 1.
6. FX Diebold. The past, present, and future of macroeconomic forecasting, Journal of Economic Perspectives, 5.
7. JE Hanke, AG Reitsch, DW Wichern. Business Forecasting, Prentice-Hall, Hoboken, NJ, USA, 2001, 9.
8. R Gupta, A Kabundi, SM Miller. Forecasting the US real house price index: structural and non-structural models with and without fundamentals, Economic Modeling.
9. C Deb, F Zhang, J Yang, SE Lee, KW Shah. A review on time series forecasting techniques for building energy consumption, Renewable and Sustainable Energy Reviews.
10. A Tealab, H Hefny, A Badr. Forecasting of nonlinear time series using ANN, Future Computing and Informatics Journal.
11. JH Stock, MW Watson. A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series, National Bureau of Economic Research, Cambridge, MA, USA, 1998.
12. G Zhang, B Eddy Patuwo, MY Hu. Forecasting with artificial neural networks: the state of the art, International Journal of Forecasting.
13. B Sharma, S Kumar, P Tiwari, *et al*. ANN-based short-term traffic flow forecasting in undivided two-lane highway, Journal.
14. GK Jha, K Sinha. Agricultural price forecasting using neural network model: an innovative information delivery system, Agricultural Economics Research Review.
15. Z Tang, C De Almeida, PA Fishwick. Time series forecasting using neural networks vs. Box-Jenkins methodology, Simulation.
16. KJ Kim. Financial time series forecasting using support vector machines, Neurocomputing, 55.
17. K Chakraborty, K Mehrotra, CK Mohan, S Ranka. Forecasting the behavior of multivariate time series using neural networks, Neural Networks, 5, 6.
18. DC Park, MA El-Sharkawi, RJ Marks, LE Atlas, MJ Damborg. Electric load forecasting using an artificial neural network, IEEE Transactions on Power Systems, 6, 2.
19. T Šestanović. Jordan neural network for inflation forecasting," Croatian Operational Research Review, 10.
20. Anil Kumar Mahto, M Afshar Alam, Ranjit Biswas, Jawed Ahmed, Shah Imran Alam. Short-Term Forecasting of Agriculture Commodities in Context of Indian Market for Sustainable Agriculture by Using the Artificial Neural Network, Journal of Food Quality, vol. 2021, Article ID 9939906, 13 pages, 2021.
21. Masters Dominic, Luschi Carlo. Revisiting Small Batch Training for Deep Neural Networks, 2018.
22. Practical recommendations for gradient-based training of deep architectures, 2012.
23. Sabu Kiran M, TK Manoj Kumar. Predictive Analytics in Agriculture: Forecasting Prices of Areca Nuts in Kerala. Procedia Computer Science. 2020; 171:699-708, 10.1016/j.procs.2020.04.076. Accessed 6 May 2022.
24. (PDF) Machine Learning for Price Prediction for Agricultural Products. Research Gate, www.researchgate.net/publication/352260358_Machine_Learning_for_Price_Prediction_for_Agricultural_Products, 10.37394/23207.2021.18.92.
25. Colah.github.io. Understanding LSTM Networks -- colah's blog, 2019. [online] Available at: https://colah.github.io/posts/2015-08-UnderstandingLSTMs/ [Accessed 6 Aug. 2019].
26. Sepp Hochreiter, *et al*. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 1998; 06(02):107-116