Received: February 14, 2023. Revised: April 5, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Optimized Farming: Crop Recommendation System Using Predictive Analytics

Meeradevi[1]* **Monica R. Mundada[2]**

[1]*Department of Artificial Intelligence & Machine Learning, M S Ramaiah Institute of Technology, India*
[2]*Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, India*
* Corresponding author's Email: meera_ak@msrit.edu

Abstract: Agriculture is been majorly contributing to the country's GDP. So, it is foremost important to improve the production quality and quantity to improve the agricultural economy. Using modern technologies like IoT, big data, blockchain, robotics, 5G technologies, etc., farming can be improved by replacing humans in agricultural operations. The proposed work focuses on forecasting of yield and prices of agro-products which will be useful for farmers to increase their productivity and which in-turn increase their economy. Data science techniques can be used to perform this by predicting which crop can be grown for given environmental features and what is the outcome before harvesting period. The proposed model uses computational data-driven approach for crop yield prediction and price forecasting of agro-products. The study uses hybrid approach for crop yield prediction which integrates LSTM and genetic algorithm based evolutionary algorithm known as enhanced long short-term memory (ELSTM). The weight assignment plays a major role in learning the behaviour of the network. The gradient value with respect to weight is calculated for each epoch, and weight is updated to compute the new weight of the network to minimize error. The model accurately predicts the crop yield for the dataset which comprises of features like soil data, rainfall, history of production, fertilizers etc., the accuracy of the model is measured using metric root mean square error (RMSE) and mean absolute error (MAE). In the proposed study, the computation model is used to enhance the knowledge about agricultural yield before the crop sowing period. The proposed model will help farmers and government agencies to improve production. The proposed ELSTM model is compared with other machine learning models such as naïve bayes, decision tree, which shows accuracy between 75% to 80% and the experimental result show the proposed ELSTM model is performing better with 85% accuracy. The price forecasting model is necessary to identify the commodity prices well in advance. The agro-product price forecasting is made using auto-regressive integrated moving average (ARIMA) which uses dataset collected from various agricultural produce market committee (APMC). Auto correlation function (ACF) and partial auto correlation function (PACF) time series plots have been used to assess the proposed model's performance in predicting prices of crops.

Keywords: Forecasting, Farming, Agro-products, RMSE, MSE, LSTM, ARIMA.

1. Introduction

The survival of global human population is based on food produced from agricultural land. In current situation with increasing population, the productivity needs to be increased. Farmers are using traditional way of farming which cannot meet modern agriculture requirements that leads to reduced production rate. Using data-driven models crop yield can be predicted that can help farmers decide which plant to grow and when to grow. The proposed research develops prediction model using machine learning technique. Farmers are lagging in usage of modern tools and technologies, due to which they are not getting expected yield. Due to lack of yield mapping solutions in agriculture development our farmers face huge number of problems. There are the models for predicting yield based on soil parameters and vegetation indices. It is important to understand yield limiting factors such as soil characteristic, weed, plant disease etc., [1, 2] which reduces the production. Monitoring of environmental parameters and collecting the real time data using sensors will reduce the cost and time and also increases the efficiency. Using the modern technologies like IoT, artificial intelligence, big data, 5G communication technologies can guide the agriculture industry to improve the production [3, 4] by automated disease detection, automated weeding etc., The existing research uses only few soil parameters to develop yield prediction model, environmental conditions have not been considered such as rainfall, humidity, soil ph [5] etc., which plays major role in yield prediction. The proposed work uses the combination of multiple features of soil, history of crop production pattern and environmental parameter to develop a yield prediction model. The new production method, known as unmanned farming, eliminates the need for labour-intensive manual crop monitoring. This strategy employs cutting-edge technologies for production via autonomous system [6]. The data collected is transferred to cloud through reliable communication technology and cloud platform analyzes the data and process the data using ML technologies and makes the intelligent decision to improve the productivity [7]. The proposed model use hybrid approach ELSTM for yield prediction and ARIMA for agro-products price prediction. Prediction of crop production requires huge data related to production, environment etc., yield mapping and estimation is done based on historical data. The machine learning (ML) techniques are used for analysis of future using the previous data [6, 8].

Detection of crops and classifying features of crop quality can improve the cost of the crop. The study uses data from Gandhi Krishi Vigyana Kendra (GKVK), University of agricultural sciences, Bangalore and official web site IndiaStat.com. The need for decision making system furnishing strategies for effective management of input, such as dosage of fertilizer input, crop pattern and irrigation pattern is managed efficiently [4]. This greatly influences the process related to plant growth and agricultural production science. With the modern computational power large amount of data is collected. Data is pre-processed to extract relevant data from large dataset and new correlation patterns are obtained. Data analytics is considered to be one of the main key tools in supporting the innovation in ML techniques. ML is playing major role in society and has much social impact with respect to improving social and environmental sustainability [9]. This paper adopts machine learning algorithms in the field of agriculture in order to manage potential risks in yield and price of the crop and help them to achieve socio economic sustainability [10]. The dataset comprises of rainfall, temperature, humidity, production, soil nutrients like nitrogen, phosphorus, potassium (NPK) and soil type such as red, black, sandy, alkaline. The data also includes state name, district name. The dataset also comprises of price data such as arrival quantity, market, commodity, min price, max price for the day. The collected data is pre-processed and given as input to the model. The model is trained using scikit python library. Proposed work is conveyed with the following objectives.

- Predicting crop yield and price of commodity by implementing hybrid model
- Analyzing the effectiveness of models that have been used.
- Comparing the outcomes of the proposed model with the existing model.

The main aim of the proposed work is to develop an efficient yield prediction and price prediction model in order to help farmers in decision making on when to grow which crop and when to sell the crop.

Further the study is divided into 4 sections with introduction being first section and second section with materials and methods comprises of literature survey, dataset and proposed methodology for crop yield prediction with results, third section being crop price prediction model with implementation model and fourth section for concluding remarks.

2. Materials and methods

2.1 Literature survey

The country like India is more dependent on agriculture output. With increasing population estimating the production of crops is also difficult. Incorporating new technologies in farming such as deep learning, big data, IoT, robotics, 5G etc., will help farmers to transit from conventional farming to precision farming. ML is one of the decision support tools for agriculture yield prediction. Accuracy is the metric to measure the performance of the model. The yield prediction can be made for specific districts based on the dataset availability. Various ML models are used for predicting crop yield like regression, decision tree, random forest with approximately 67% accuracy [11, 12]. The neural network model showed better accuracy with approximately 80% accuracy. The short-term crop price forecasting models are developed using artificial neural network. A feed forward neural network and ARIMA model is developed for forecasting of prices for various crops. The dataset includes weekly and monthly wholesale price from 1996-2010. The study shows ANN as an efficient tool for price forecasting one day before or one week, which outperforms times series ARIMA model. The feed forward neural network model shows the relative error of 5.0% [13]. The time series image data is collected and proposed a spiking neural network model for prediction of crop yield for China country before six-weeks of harvesting with average error rate of 0.236 t/ha with highest accuracy of 95% and correlation coefficient of 0.801. Spiking Neural Network (SNN) used NDVI time series data for computation. SNN outperformed other traditional approaches [14, 15].

The authors made an extensive analysis on soil [5]. Soil is one of the important factors which plays major role in food production. For the good yield the soil is provided with organic and inorganic fertilizers which are consumed by plants. Soil quality mainly depends on water holding capacity, amount of organic matter and type of clay. The author has done study on physico-chemical properties of Bagalkot district soil. The study shows seasonal effect on soil fertility, which will in turn change the status of soil properties. The author collected 18 soil samples from six villages and analysed pysico-chemical properties of soil in Kharif, Rabi and summer season [5, 9]. Advanced machine learning techniques have been used for monitoring environmental parameters to predict maize crop yield. The prediction is done using NDVI time series dataset [16]. The effectiveness of the model is evaluated in comparison to the traditional regression techniques as well as random forest redresser (RFR), boosted regression tree (BRT), and Gaussian process regression (GPR). The proposed method outperforms when compared with other methods with 85% accuracy. The experiments were conducted for Maize crop. Soil already consists of some number of macronutrients in terms of animal and plant waste. The analysis is done for uptake of NPK and farm yard manure (FYM) which is required for increased crop production [6, 7]. The market price prediction is done for Taiwan market. The paper does the price forecasting on S.A.M.P. SAMP is smart argi-management platform this model retrieves the previous dataset and trains the model using time series analysis ARIMA model and ANN [7, 17]. The nitrogen (N) is used in much application for prediction without harming nature just by using the outcome of previous years. Recent papers show DNN algorithms are quite similar to ANN algorithms only varying in number of hidden layers [18]. CNN models showed improved accuracy in classification and LSTM neural network is most widely used model which shows great accuracy for regression analysis [19]. Time series data is analyzed using LSTM and gated recurrent unit (GRU). For the analysis agricultural dataset is used. The dataset consists of climate data which helps in predicting yield. Experimental results show the use of bidirectional GRU (BGRU) which performs better than LSTM on validation dataset [20]. MSE values are used to measure the model performance which ranges between 0.017 to 0.039 [8, 21]. Various application is listed in precision farming like crop selection based on soil, fertilizer application, pest management, monitoring of yield, irrigation pattern, crop diseases. Soil condition and type of soil is one of the foremost important parameters for improved production. Soil testing is required to analyse the properties of soil. The irrigation pattern is also equally important, over watering and under watering leads to huge loss in production. Fertilizers like NPK provide nutrients to soil which is required for healthy crops. The deficiency in nutrient leads to crop loss [22]. So, right amount of nutrients should be supplied for improved production [23]. Crop disease is one of the major threats which should be prevented by applying right amount of pesticides in right time. Over dose of pesticides will harm crops and human health. Lastly, in order to maximize the yield, the monitoring of yield is required using advanced technologies. [24, 25] Adaption of IoT and data analytics helps farmers to grow more crops. The proposed work uses neural network-based approach which consist of multiple layers and at each layer neurons receive signals and generate the output and send it to the next layer. The ReLu activation function used in neural network which send only the positive values to the next layer. The error value defines how efficient the model is. The metric used to measure the loss is RMSE and MAE. If the loss is high the model uses back propagation trough time to transfer the error values from last layer to first layer and then the computation takes for ‘n’ epochs until the optimal results are obtained. This paper used various machine learning models to predict crop yield. The model accuracies are as listed linear regression with 81% accuracy, decision tree with 79.2%, lasso regression with 80% and LSTM with 86.3% accuracy. The proposed ELSTM shows 96.7 accuracy [26]. [27] This study focuses on maximizing crop yield and automation of irrigation management. The paper also predicts the crop type suitable for particular soil using soil parameters and weather condition. The model uses XGboost algorithm which yield 92% which is less than proposed model. [28] In this paper authors have done prediction of soybean and corn yield. Linear regression model is used for prediction after robust feature selection and interaction selection. The accuracy of the model is measured using RMSE which ranges between 6% to 7%.

2.2 Dataset

The soil dataset is collected online from

Karnataka soil health. Crop production data is taken from GKVK. The dataset comprises of data from various districts of Karnataka. The rainfall dataset has historical values of precipitation data collected from various districts from 1998 to 2014. The crop production dataset consists of the season wise (summer, rabi, and kharif) production data. The area of farmland is measured in a hectare, and production is measured in tonnes per hectare. Fertilizer such as NPK concentration applied in kg/ha like urea and DAP, temperature, humidity, soil type (Black, Red), Soil pH. In order to perform data pre-processing, libraries have been used like NumPy, pyplot, matplotlib, pandas. The Pandas library is used for importing and managing datasets. Next, import the collected data, extract dependent and independent variables, and scikit library is used to handle missing values in the dataset. Some of the data were categorical and some were continuous numeric which causes problems and does not provide optimal solution. Hence, standard feature scaling was performed on all of the data to normalize and bring them into a common scale. Additionally, feature selection was done to reduce over fitting issues. The prices for the crops were collected from different APMC's of Karnataka. The data collected are the market prices of different commodities sold on that day the dataset was collected from 2016-2018. The dataset attributes are date, commodity, variety, grade, market, arrival unit (Quintal), minimum price, maximum price, and modal price. The data is collected once a week and stored in CSV file format. The attribute market is the district/taluk name, the commodity is the type of crop arrived at market and variety is the crop (local, hybrid, broken rice, Katta sambar a rice variety, etc.,), arrival unit is the quantity of crop arrived at particular market measured in quintal. The dataset is split into training data as 80% and testing data as 20%. ARIMA model is used for training and test data is used on the trained to evaluate performance. Price of the crop is predicted for two upcoming years.

2.3 Proposed methodology for crop yield prediction using ELSTM

The proposed model uses ELSTM for crop yield prediction. With growing population predicting the crop becomes foremost important. The proposed hybrid model ELSTM uses dataset that includes environmental data and crop data. The features used for crop yield prediction are history of production, soil type, pH, rainfall, NPK, seasons (kharif, summer, rabi) from year 1998 to 2014. As shown in Fig. 1 the dataset is divided into training and testing data with 80 percent and 20 percent. As the dataset is with different magnitude and units, the data pre-processing is done, the data is normalized to scale down the values to 0 and 1. There were few columns which had a small portion of null values in the production column. This did not affect much to the prediction model, mean of the particular column was taken and replaced the null values. Feature extraction is performed on the dataset because some of the data were categorical and some were continuous numeric, this causes problem to the model during training and which may not provide optimal solution [29]. Hence, standard feature scaling is performed on all of the data to normalize and bring them into a common scale. Additionally, feature selection was done to reduce over fitting issues. Next, train data is used as initial population to generate the optimized LSTM parameters such as window size and number of layers. A set of random population is selected and used to generate new individual using evolutionary algorithm. The main task is identifying LSTM time window size and number of layers which is evaluated using evolutionary algorithm in this work. The LSTM model is trained based on the parameters selected during evolutionary process. The genetic algorithm based evolutionary approach uses fitness function to computer fitness value of dataset. the proposed model use RMSE values as a fitness value. Lower the RMSE value satisfies the termination criteria. If the termination criteria are not satisfied than the process of selecting new chromosome starts over with genetic operator’s selection, crossover and mutation. Crossover rate is 0.8 and mutation rate is 0.15 for the population size 80. Optimized parameters (time window size and number of layer) are obtained and tested with test data.

2.3.1. ELSTM algorithm for crop yield prediction

Crop yield predictions are made using the proposed ELSTM algorithm. Firstly, the dataset is pre-processed and normalized using MinMaxScaler() and reshape(), which restructures the data based on the training window size 'tw' that is examined using EA.

Training window consist of a collection of patterns used to forecast. The input vector X={x1, x2, x3 …xtw}. The module takes the input vector, i.e., training window (tw), and previous information (dataset). Replacing the not a number (NAN) values and missing values with the mean of attributes. The dataset is split using test train split function into training and testing sets in order to implement LSTM.

The detailed sequence of steps carried out to implement the proposed algorithm ELSTM is

Received: February 14, 2023. Revised: April 5, 2023.

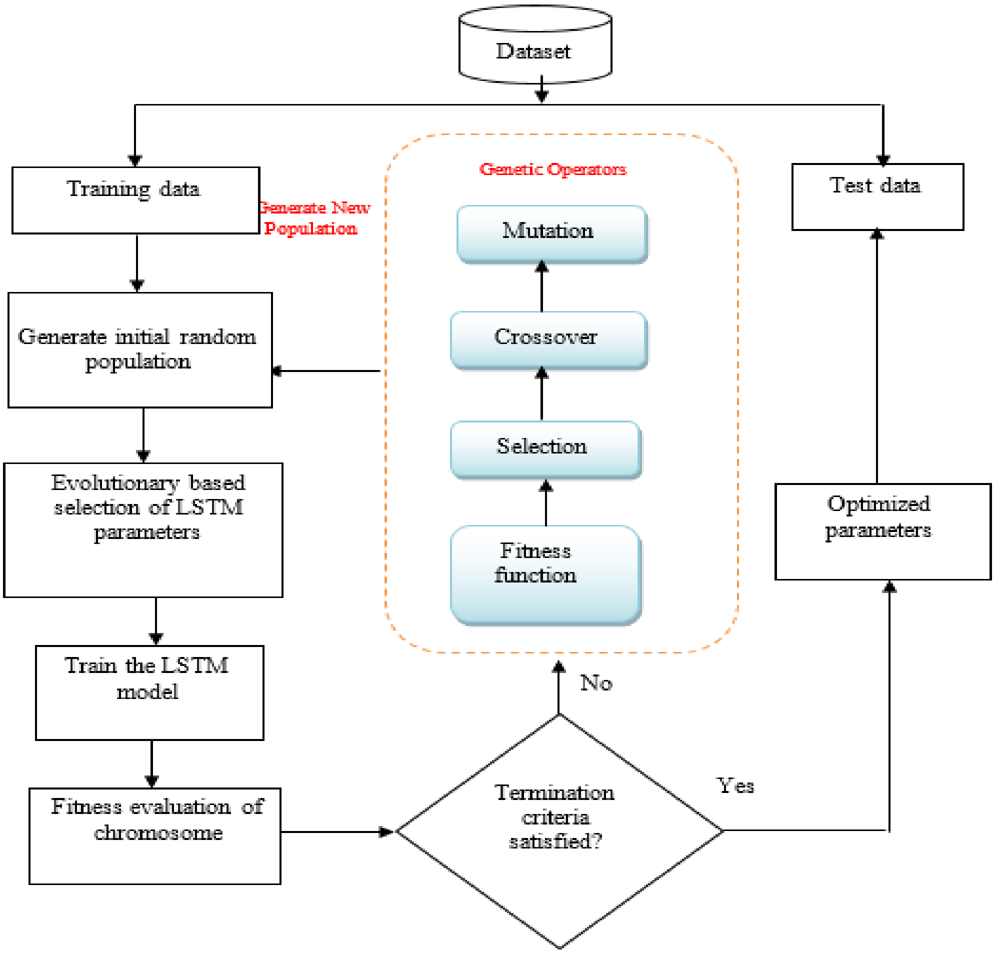

Figure. 1 Workflow of proposed model for crop yield prediction

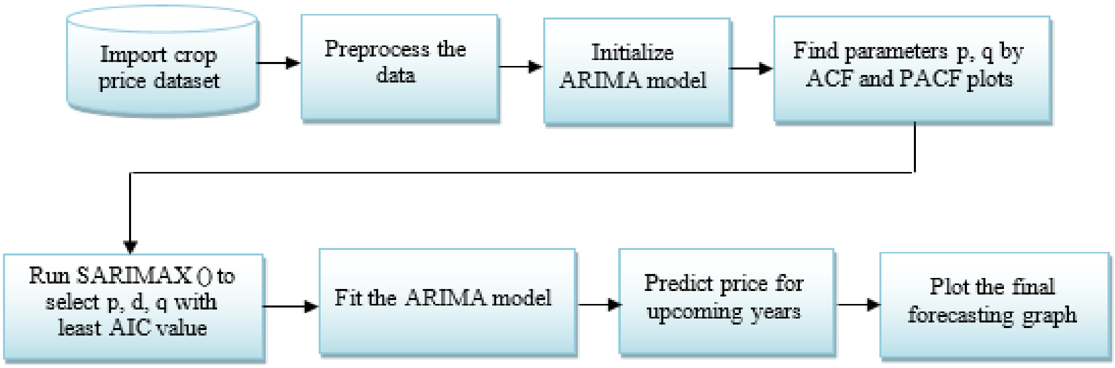
Import crop price dataset
Preprocess the data
Initialize ARIMA model
Find parameters p, q by ACF and PACF plots
Run SARIMAX () to select p, d, q with least AIC value
Fit the ARIMA model
Predict price for upcoming years
Plot the final forecasting graph

Figure. 2 Workflow of the proposed ARIMA model for crop price prediction

illustrated as shown in the algorithm.

The model uses many-to-one LSTM network which is efficient with two hidden layers when compared with four hidden layers. The increase in number of hidden layers leads to overfit in the model. Proposed model is trained using the training dataset by varying the batch size to (4, 16, 32 and 64), the model showed improved performance with 32 batch size and the obtained model is cross-validated with the test data and accuracy is measured using RMSE and MAE. The algorithm utilizes the principles of biological evolution to obtain the elite value of the defined crop parameters. The implementation of the agricultural yield optimization problem has been manifested using the genetic algorithm. The proposed ELSTM model provides better performance with 85% accuracy and metric precision, recall, and f-measure values are used to assess the model's performance.

```
Algorithm: Crop yield prediction using hybrid model ELSTM
Input: Pre-processed data with suitable attributes
Output: ELSTM model predicting crop yield with graphical analysis
Step 1: Consider the pre-processed dataset with appropriate attributes (crop type, production, year, season, soil type, NPK, soil pH) required for implementation.
Step 2: Divide the dataset into a 20% test set and an 80% training set.
Step 3: Normalize the data in the range 0 to 1 using below Eq. (1).

    $Xnorm = \frac{x - min(x)}{max(x) - min(x)}$    (1)

Step 4: Feed the normalized data to the evolutionary algorithm for parameter selection
Step 5: Using an evolutionary strategy, choose the training window size (tw), the number of LSTM units, and arrange the dataset properly.
Step 6: for 'n' epoch and batch 'b' do
            Train the model (L)
        end for
Step 7: Evaluate the proposed hybrid model and compute fitness.
Step 8: If the reproduction process' output meets the termination requirements (a lower RMSE value), derive the near-optimal solution for selected parameters.
        else
Repeat step 5, i.e., the process of evolutionary search (selection, crossover, mutation) for new parameter values until optimal parameter with lower RMSE is obtained
Step 9: Plot the crop yield graph for season Kharif, Rabi, and summer (year Vs. yield)
```

3. ARIMA model for crop price forecasting

The proposed ARIMA model's process is depicted in Fig. 2. Initialize the ARIMA model using a predetermined set of parameters by importing the dataset, dataset is pre-processed to remove null or missing values. In the autoregressive model parameters (p,d,q) is used, p stands for a variety of time lags. Moving average is used to check the stationarity, or error terms, in the data before constructing the ARIMA model, where d stands for difference, or the number of times past data values are removed. Then, in order to configure ARIMA to determine the ideal parameters, ACF and PACF plots are obtained.

```
Algorithm: ARIMA model
Input: Dataset with appropriate attributes (price, variety, commodity, market, grade, arrival unit, date)
Output: Commodity price prediction
Step 1: Data preparation and processing.
In this step, the data is read from the dataset, and time is taken as an index for the data frame. The mean of the data is calculated month-wise for the forecasting.
Step 2: Optimal set of parameters selection using grid search for (p, d, q, m). m is the seasonal parameter which is 12 in this study. Whereas p is autoregressive and d is for difference and q is used for moving average
Step 3: Fitting the ARIMA model.
Using the optimal parameter, the data is fitted using sm.tsa.state space. SARIMAX() imported from the statsmodels library. The fitted result is stored in the result variable for each market price, the ARIMA model is applied to all the data points as shown in Eq. (2).

    $y_t = \mu_t + \sum_{i=0}^{p} \Phi_p y_{t-p} + \varepsilon_t + \sum_{i=0}^{q} \theta_q \varepsilon_{t-q}$    (2)

Step 4: Validate the model and forecast the result for the next 24 months, and the mean of the upper and lower bound of the prediction is considered as the forecasted values. The data is visualized with the actual data to analyze how the prices may vary.
```

3.1 Implementation of price prediction model for agro-products

This section describes the prediction model using ARIMA. The three basic components of ARIMA are the AR term, I term, and MA term. Autoregressive, which models historical data, referred to as AR. I stand for Integrated, which is a phrase that contrasts the data to make the process steady. The moving average that regulates historical data is referred as MA. Grid search is used to generate MA based on past values for the parameter selections p, d, and q. The moving average component is represented by q, the autoregressive component by p, and the differencing term by d. Error terms are calculated for autoregression.

The Eqs. below show how moving-average and autoregression polynomials combine to form a complex polynomial. For each market price, the ARIMA model is applied to all the available data points. Eq. (3) can be used to express the AR(p), which is an autoregressive model of y_t series.

AR(p) term,

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \Phi_1 y_{t-2} + .. + \Phi_p y_t - p + \varepsilon t \quad (3)$$

y_t: the actual data over time t,

y_{t-1}, y_{t-2} …… $y_{t-\mathrm{i}}$ is data at a different time with lags, εt is the error factor

Φ_0, Φ_1,… Φ_p are coefficients of AR term

p: p is the number of previous observations necessary to predict the value of the series at the present time.

Similarly, MA (q) term which is the moving average model expressed in Eq. (4)

MA(q) term,

$$y_t = \mu_t + \varepsilon_t + \theta_t\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q} \quad (4)$$

μ_t: the mean value of the time-series data at time t,

θ 1 …. θ q moving average coefficients (MA)

q: the number of cut-off lags of the moving average process

Combine both AR and MA term will give a general form of the ARIMA model, which is expressed as in Eq. (5)

$$y_t = \mu_t + \sum_{i=0}^{p} \Phi_p y_{t-p} + \varepsilon_t + \sum_{i=0}^{q} \theta_q \varepsilon_{t-q} \quad (5)$$

Φ being an AR term, θ is an MA term, ε noise of the time series data, q: lagged forecast errors in prediction

d represents the order of differentiation which is calculated using Eq. (6)

$$d\ term, y`t = yt - yt - 1 \quad (6)$$

To consider the seasonal component, the SARIMA model is proposed as shown in Eq. (7)

$$Xt = SARIMA\,(p, d, q)\,(P, D, Q)\,m, \quad (7)$$

The model performance is estimated using AIC. The lower the AIC, the better the model. AIC is given by Eq. (8)

$$AIC = -2(\hat{L}_t) + 2k \quad (8)$$

k is the number of parameters estimated in the model $\hat{L}_t$is the log-likelihood function for the model with maximum value.

ACF and PACF can be used to verify the value obtained for AR (p) and MA (q) is optimal. This function helps in estimating parameters that can be used for forecasting using the ARIMA model.

ACF and PACF can be used to verify the value obtained for AR (p) and MA (q) is optimal. This function helps in estimating parameters that can be used for forecasting.

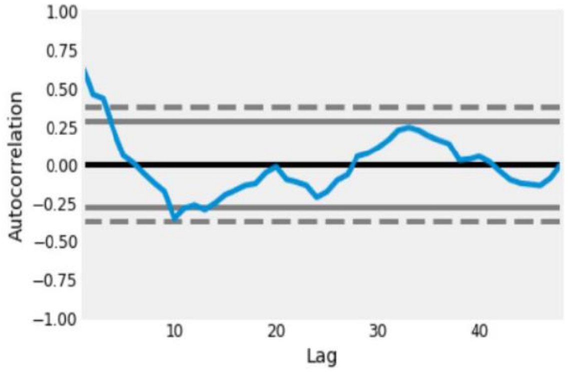

Figure. 3 Autocorrelation plot (ACF)

3.2 Model diagnostics

Figs. 3 and 4 provide a methodical technique for plotting the lag values that is used to assess the goodness of fit of the proposed model. The fitted ARIMA (0, 1, 1) time series plots of ACF and PACF showed a significant pattern; the proposed model was regarded as significant for forecasting.

Fig. 3 shows the ACF plot. From the graph, it can be analyzed that 95% of the spike is within the confidence zone, i.e., within +0.25 to -0.25. The graph shows a spike at lag 0 while the other spikes are within the zone. Hence, the model performs well for the order of p, d, q (0, 1, 1), and best fits most of the dataset because they obtained lower AIC values of 266.149.

Fig. 4 plots PACF for order of p, d, q as (0,1,1). The plot shows the spike at lag 1, 10, and 20, remaining spikes are more significant i.e., within the blue range. Hence 95% of the spikes are within zone therefore the model performs well. ACF and PACF correlation function plot are used to show correlation between present values and past values where X-axis shows lag values, and the y-axis shows autocorrelation. 95% of our data is within the given confidence zone, i.e., -0.2 to +0.2 as per the standard rule; if 95% of the distribution is within the zone, we conclude the model is efficient with normal data distribution in the model without much randomness. Hence, the proposed model accurately predicts the price.

Partial autocorrelation finds the correlation with the residuals. Autocorrelation shows the degree of correlation between time series data and lag data over successive intervals. In Figs. 3 and 4 shows ACF and PACF graphs which identify the configuration parameters (p, q) of the ARIMA model. PACF values are between +0.2 and -0.2, only few of the spike values are deviating from standard values; hence the model is accurate in predicting. The ARIMA model

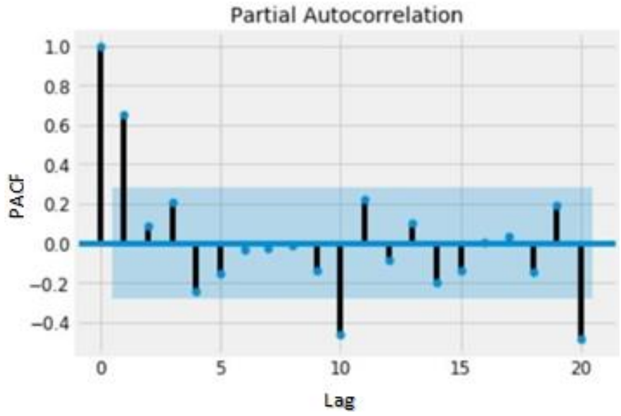

Figure. 4 Partial autocorrelation plot (PACF)

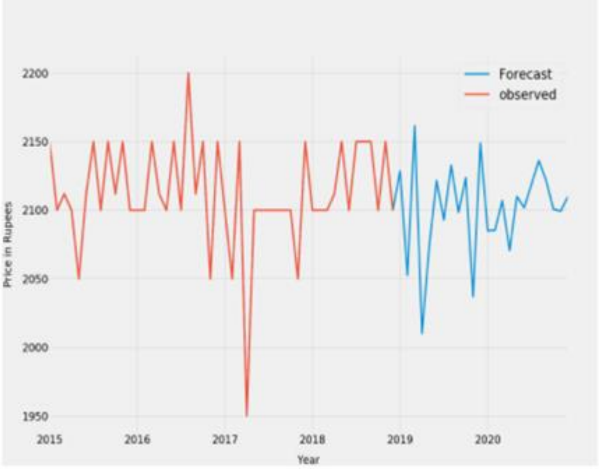

Figure. 5 Commodity price forecast for Kalaburgi district rice crop

is set up and the dataset corresponding to crop price prediction is pre-processed, and ACF and PACF plots are obtained.

Fig. 5 shows the forecast for commodity rice with broken rice variety for Kalaburgi market Karnataka state. The price is decreasing for the years 2019 and 2020 when compared to previously observed original data. So, the model recommends framers of Kalaburgi district that the yield and price of rice crop are less in near future.

The model also aims at developing a web interface called Raita Mitra, which has two features, i.e., crop price and yield prediction, which are required most of the time for farmers. The website design is kept simple without making it complex with only the yield and price tab. Based on a selection of district and type of crop and variety, the model will display the price and yield of the given crop for the future.

The plot_diagnostics in Fig. 6 object permits to create model diagnostics and research for any unusual conduct. In this framework, the diagnostics infer that the model residuals are normally distributed based on the following observations:

- As seen in the top right plot, the N(0,1) normal distribution standard notation is closely followed by the orange KDE, which indicates that the errors are normally distributed.
- As seen in the bottom right correlogram or autocorrelation plot, the time series residuals have a lower correlation with lagged versions of themselves. These perceptions drove us to infer that our model creates a satisfactory fit that could assist us with understanding our time-series data and estimate future price and yield estimates. Standardized residual on the top right corner shows the residual for every two months from 2017 May to 2018 Nov. Since the residuals lies between +2 and -2 the model is normally distributed. In bottom left corner the data points are close to hyperplane hence the model is normally distributed.

ARIMA model in the proposed study removes the linear tendency present in the dataset. The residuals of ARIMA contain the non-linear component. ARIMA is good at handling linear data, but ARIMA does not handle non-linear data. A normal distribution graph is plotted to analyse the data points. Fig. 6 shows the data point are normally distributed in standardized residual and correlogram with 95% of spikes within confidence zone. Fig. 6 on bottom left Normal Q-Q regression line shows the data points on the slope which is normally distributed and in the histogram on top right in the graph shows the bell-shaped curve without much skewness on left or right. Hence the model is significant with normal distribution.

The auto correlation ACF and PACF graphs are used to calculate the SARIMA parameters. SARIMA with the parameters (0,1,1) (0,1,1,12) is chosen for the ARIMA model since it has the lowest AIC value.

SARIMAX () is configured in Table 1 with the parameters p,d,q as (0,1,1) and P,D,Q,S as (0,1,1,12) and obtained log likelihood value as -130.075 which estimates AIC value of 266.149 that corresponds to least AIC value. Table 2 shows the final estimates of price parameters. The coefficient of AR and MA and p-value work together, which shows the significance of the proposed model. The coefficient shows the mathematical relationship between the dependent and independent variable, and the lower p-value shows whether these coefficients are significant. The usual standard significant level is considered in the proposed ARIMA model as 0.05. The coefficient is not statistically significant if the p-value is higher

Received: February 14, 2023. Revised: March 17, 2023.

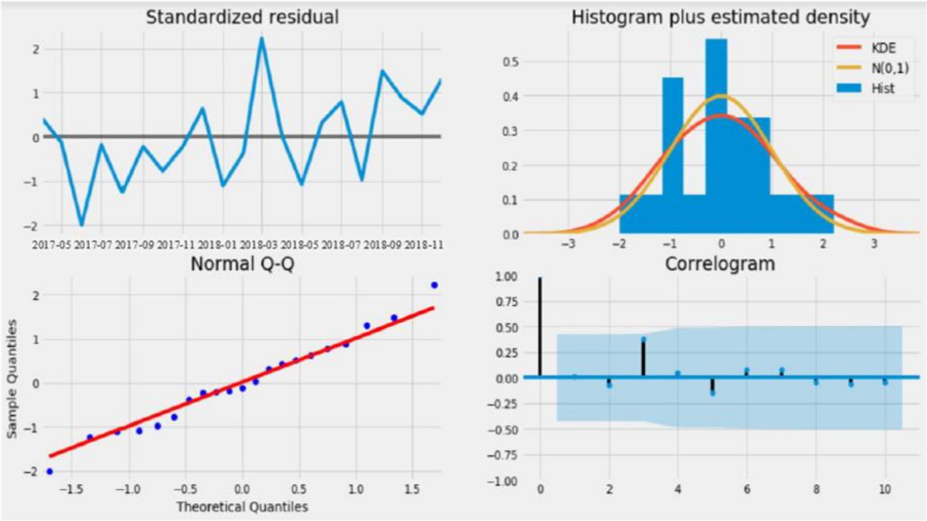

Figure. 6 Plot diagnostics

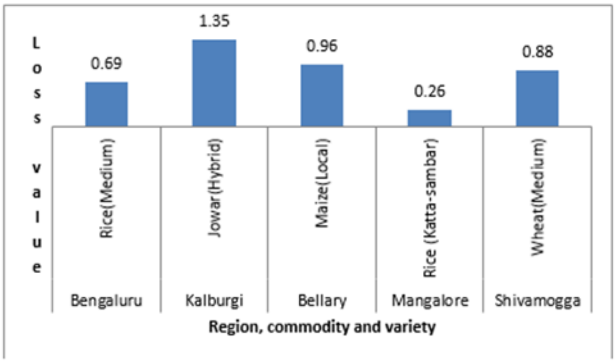

Figure. 7 Loss values with respect to crop, variety, and district

Table 1. Configuration summary

Dependent Variable	Modal
Model	SARIMAX (0, 1, 1)x(0, 1, 1, 12)
No. of Observations	56
Log-Likelihood	-130.075
AIC	266.150
BIC	269.283
HQ	266.830

Table 2. Final estimates of price parameters

Type	Co-efficient	.err	P
AR lag1	-0.3782	0.314	0.028
AR lag12	-0.4323	0.181	0.017
MA lag1	-0.0965	0.396	0.808
MA lag12	-1.000	0.359	0.005

than the significance level i.e., 0.05; the relationship shows the coefficient is significant for prediction. Table 2 shows the p-value AR at lag 1 and MA at lag 1 and MA at lag 12 where the predictor value (P) variable is less than 0.05, whereas only with MA at lag 1 P-value is greater than the significance level. Hence the model is 95% significant for prediction. The negative coefficient in Table 2 shows that the dependent variable tends to drop as the independent variable rises. The value of standard deviation is much less; hence the proposed model is normally distributed with a significant confidence level. The standard error shows the estimate of AR and MA parameters and estimates the standard deviation with its sampling distribution, the value of std. err is less than 1, so the predictions made are accurate.

The model performance is evaluated using RMSE. The RMSE value for the ARIMA model is 111.63, and RMSE for the LSTM model is 135.08 with a train loss value of 0.0157. ARIMA model is very well suited for stationary data. The residuals of ARIMA are fed as an input to the LSTM model which handles non-linear data.

Fig. 7 shows the loss value for crops which are majorly grown in various regions for given variety. In the graph, it can be seen the loss measured based on features like crop type, variety, and districts, etc., The prediction accuracy shows a reduced train loss value. Hence, the proposed model is efficient in making an accurate prediction. This model will help to analyze the dataset and helps in decision-making.

Table 3 shows the loss value considering environmental seasons and crop types. Therefore, the loss value is significantly less for all crop types. The proposed model accurately predicts the crop price without overfitting as loss values are very less. Thus, the proposed model helps the government and farmers analyse the pattern and decide which crop to be grown for a particular season and the price during the harvesting season.

The primary objective of SARIMA is to predict

Table 3. Performance analysis using loss value

Crop	Season	Loss Value
Jowar	Summer	0.17
Garlic	Whole Year	0.33
Onion	Kharif	0.17
Maize	Summer	0.12

Table 4. RMASE values using ARIMA model

ARIMA Model	RMSE
ARIMA(1,1,1)(1,1,0,12)	0.314
ARIMA(1,1,1)(1,0,1,12)	0.393
ARIMA(1,1,1)(1,1,1,12)	0.213
ARIMA(1,1,1)(1,1,0,12)	0.181

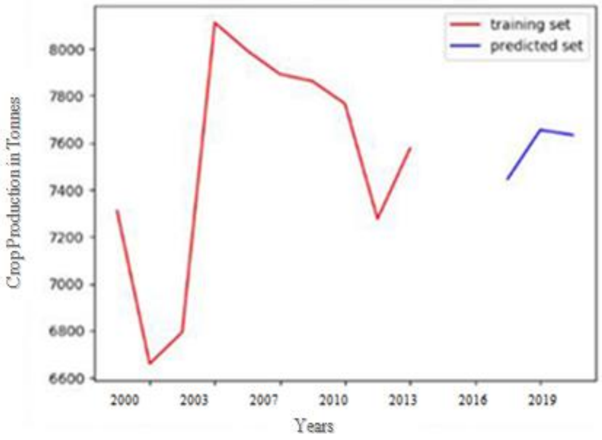

Figure. 8 Crop yield prediction for bagalkot district wheat crop

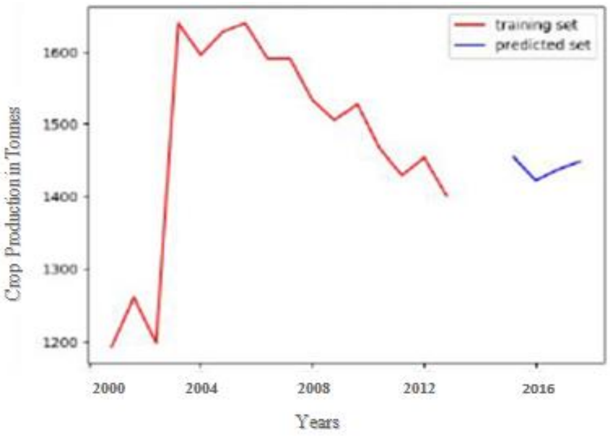

Figure. 9 Crop yield for bagalkot district rice crop

accurately the crop price. The proposed SARIMAX model is found to outperform the traditional models. Results from SARIMAX () reveal p, d, and q values (0,1,1) which gives less AIC values. Less the AIC values higher the accuracy of the model. The model is run using grid search for the range 0 to 2 with seasonality value 'm' being 12. Hence the order of (p, d, q) and (p, d, q, m) being (0,1,1) (0,1,1,12) for 12 for seasonality which is presented in Table 4.

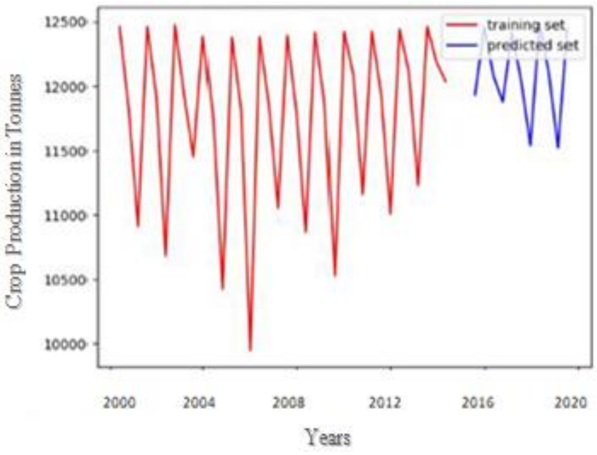

Figure. 10 Crop yield prediction for bagalkot district maize crop

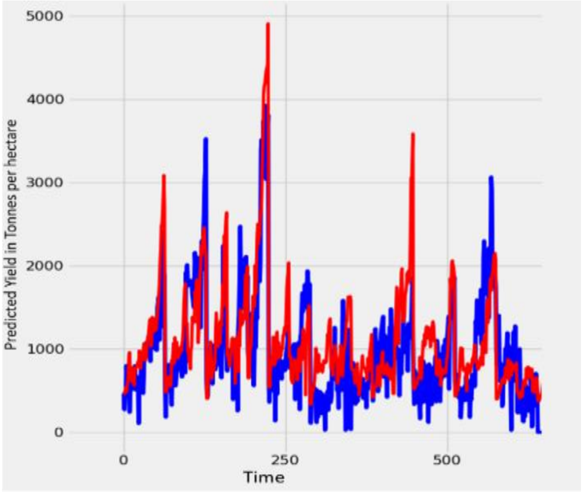

Figure. 11 Comparison between the actual and predicted yield values

In Fig. 8, the graph shows the observed (training set) and predicted value for future growth of the Wheat crop. As per proposed prediction model from the year 2017 to 2019, the Wheat crop is grown very less in Bagalkot district compared to the previous year. The yield estimation for wheat is around 7633 tonnes for Bagalkot agricultural land. This model will help the farmer to make a proper decision on which crop should be grown. Therefore, the prediction accuracy considering only wheat crop is 98.78% for black soil type, kharif season with NPK concentration 4:2:1 applied in three different growing stages of plant.

Fig. 9 shows rice crop production for the bagalkot region, which is less compared to observed data. The predicted yield production for rice crop is around 1449.07 tonnes for bagalkot region with black soil, NPK applied in three different times of growth stage of plant with concentration 4:2:1. The prediction accuracy is 98.06% for rice data. The model recommends farmers make a profitable decision

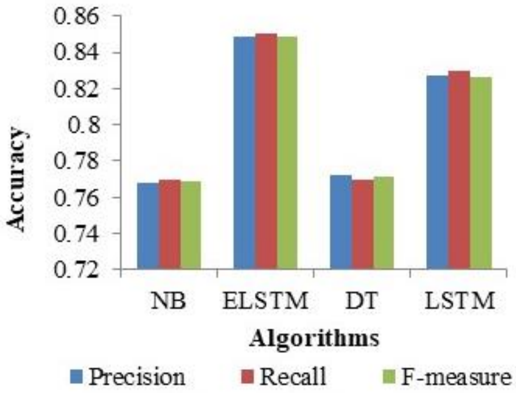

Figure. 12 Comparison of proposed ELSTM model accuracy with other algorithms

based on observed data.

The maize crop will be grown more than the wheat crop in the Bagalkot region. Fig. 10 shows bagalkot region maize crop production which more than previous years. So, it can be recommended that farmers can grow maize instead of the wheat crop. The prediction model achieves good accuracy with 98%. The yield estimation for Maize is around 12447.38 tonnes for the Bagalkot region in the near future for black soil type, with NPK concentration 4:2:1 applied three times with varying growth stages of plant.

The normalized prediction output of the proposed ELSTM model is shown in Fig. 11. The graph depicts model performance with respect to recorded data and predicted data. It is observed from the graph that the prediction curve can trace the recorded curve reasonably. The recorded and forecast curves show little variation. Thus, the model is capable of making a precise prediction. The effectiveness of the model is assessed using appropriate hyperparameters such activation function, an ideal number of neurons, and an appropriate LSTM window size. In terms of prediction error rate, the model's performance with permuted tuples of hyperparameters can be described. The metrics used such as RMSE and MAE as loss function and accuracy of the model is measured using metric precision, recall and f-measure.

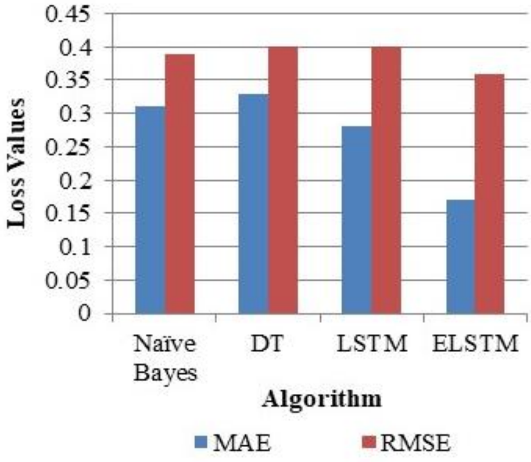

Figure. 13 Comparison of proposed model loss with other algorithms

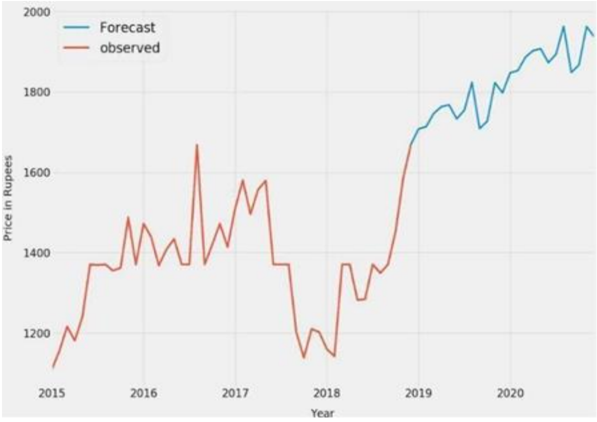

Figure. 14 Price forecast for jowar crop (year 2020 and 2021)

3.2.1. Comparative study

Fig. 12 shows the accuracy of various ML models. The Fig. 12 shows a comparative study. The proposed model ELSTM is compared with conventional algorithms naïve bayes (NB) and decision tree (DT). The model is also compared with LSTM, which shows reduced accuracy when compared with ELSTM. NB models takes all features as independent variable so it cannot learn the relationship between features. Hence shows reduced accuracy. The problem with DT model is as the tree grows the computation go more complex compared another model. Support vector machine (SVM) models are not stable for large dataset and requires large training time. Logistic model tree (LMT) cannot handle large number of variables efficiently which in turn leads to reduced accuracy as the proposed model has more features. J48 fails in showing branching and looping in the tree. The proposed model ELSTM predicts accurate yield with increased accuracy of 85%.

Accuracy of ELSTM model is shown in Fig. 12, which depicts model performance. The performance parameters for ELSTM precision, recall, and F-measure values are more than traditional models such as NB and DT and LSTM models. Recall is the total percentage of relevant results that are properly classified by the proposed ELSTM algorithm, and precision is the total proportion of relevant results. The recall value for ELSTM is high, which concludes that the proposed algorithm correctly classifies the relevant results with positive prediction.

Performance metric (error measure) plays a vital

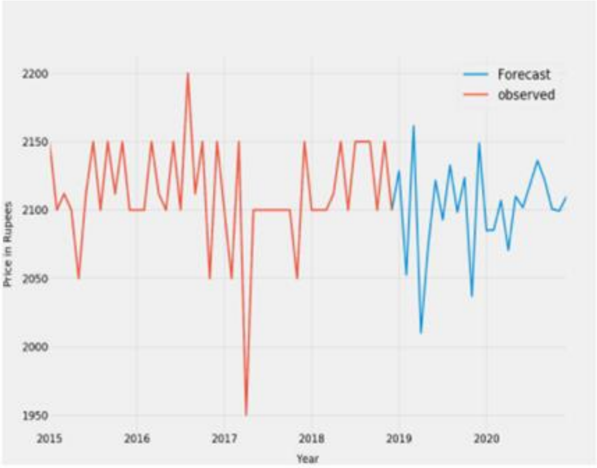

Figure. 15 Price forecast for rice crop (year 2020 and 2021)

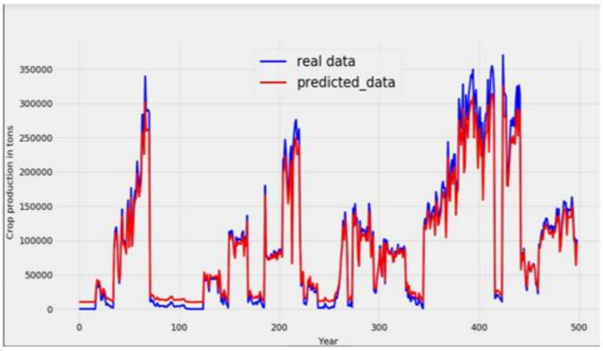

Figure. 16 Real vs predicted data (yield)

role in the evaluation framework. Fig. 13 shows the RMSE and MAE values for ELSTM are less when compared to NB, DT and LSTM models. Hence the proposed algorithm predicts the crop yield more accurately.

Fig. 14 shows the forecast for commodity Jowar for hybrid variety in Bellary market. The price is increasing for year 2020 and 2021 when compared to observed data. So, the proposed model gives an idea to farmer which crop has to be grown in the field.

Fig. 15 shows the forecast for commodity rice for broken rice variety in Kalaburgi market. The price is decreasing for year 2020 and 2021 when compared to observed data. The proposed model gives the recommendations to farmers which crop gives better yield.

Fig. 16 shows the real and predicted values for maize crop. It is observed from graph that both real and predicted values are same there is not much deviation. Hence, the error in the proposed model is less. Therefore, the proposed model is best suited for yield prediction with optimal values.

The comparative study is made with traditional algorithms which is shown in Table 5. The accuracy of the proposed model ELSTM is more optimal than other algorithms.

Table 5. Comparison of accuracy with other algorithms

Algorithm	Precision	F1-score	Recall
LMT	0.919	0.958	1.000
Naïve Bayes	0.966	0.932	0.901
J48	0.920	0.791	0.694
SVM	0.778	0.811	0.846
ELSTM	0.9691	0.9648	0.9607

LMT, NB, J48, SVM shows low accuracy as they work well for smaller dataset. With the increased features and high dimensional dataset these ML model will not give optimal results. ELSTM being a neural network model works well for huge dataset size.

4. Conclusion

The study investigates the prediction of yield using computation model ELSTM and ARIMA model is used to develop a price prediction model. The model ARIMA model focuses on the algorithm highlighting price along with crop yield prediction, since it performs joint prediction. The proposed model is said to be hybrid model. The architecture adds a computational component to the analysis of the aforementioned to improve understanding of the yield prior to crop sowing. The proposed model gives 85% prediction accuracy. The proposed ELSTM model is a better algorithm when compared other algorithms and ARIMA model is best suited for price prediction which shows lower AIC value of 266 and RMSE is <1 and ACF and PACF shows the acceptable probability which is below 0.5. The proposed study will be helpful for farmers and government agencies for yield and price forecasting. In the proposed work crop production prediction models provide improved accuracy with lower losses, while price prediction models foresee crop prices well in advance, assisting farmers in making decisions.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

For this research work all authors' have equally contributed “conceptualization, Author 1 and 2 and; methodology, Author 1; software, Author 1; validation, Author 1 and 2; formal analysis, Author 1; investigation, Author 1; resources, Author 1 and 2; data curation, Author 1; writing—Author 1 and 2, writing—Author 1 and 2; visualization, Author 1 and 2; supervision, Author 1;

References

[1] A. C. Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture", *Environmental Research Letters*, Vol. 13, No. 11, p. 114003, 2018.

[2] A. F. Colaço, R. G. Trevisan, F. H. S. Karp, and J. P. Molina, "Yield mapping methods for manually harvested crops", *Computers and Electronics in Agriculture*, Vol. 177, p. 105693, 2020.

[3] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction", *IEEE Access*, Vol. 9, pp. 63406-63439, 2021.

[4] J. Shook, T. Gangopadhyay, L. Wu, B. Ganapathysubramanian, S. Sarkar, and A. K. Singh, "Crop yield prediction integrating genotype and weather variables using deep learning", *PLOS ONE*, Vol. 16, No. 6, p. e0252402, 2021.

[5] D. Paudel, H. Boogaard, A. D. Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, "Machine learning for large-scale crop yield forecasting", *Agricultural Systems*, Vol. 187, p. 103016, 2021.

[6] T. Wang, X. Xu, C. Wang, Z. Li, and D. Li, "From Smart Farming towards Unmanned Farms: A New Mode of Agricultural Production", *Agriculture*, Vol. 11, No. 2, p. 145, 2021.

[7] B. M. Kalshetty, R. R. Chavan, S. R. Kaladagi, and M. B. Kalashetti, "A Study on Nutrient Status of Soil in Bagalkot District of Karnataka State, And Fertilizer Recommendation", *IOSR Journal of Environmental Science, Toxicology and Food Technology (IOSR-JESTFT)*, Vol. 9, No. 2, pp. 30-39, 2015.

[8] H. Chung and K. Shin, "Genetic Algorithm-Optimized Long Short-Term Memory Network for Stock Market Prediction", *Sustainability*, Vol. 10, No. 10, p. 3765, 2018.

[9] D. Paudel, H. Boogaard, A. D. Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, "Machine learning for large-scale crop yield forecasting", *Agricultural Systems*, Vol. 187, p. 103016, 2021.

[10] A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada, and S. Bhardwaj, "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture", In: *Proc. of 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, pp. 1-7, 2015.

[11] A. T. Nugraha, G. Prayitno, A. W. Hasyim, and F. Roziqin, "Social Capital, Collective Action, and the Development of Agritourism for Sustainable Agriculture in Rural Indonesia", *Evergreen Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*, Vol. 8, No. 1, pp. 01-12, 2021.

[12] D. Elavarasan and P. M. D. Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications", *IEEE Access*, Vol. 8, pp. 86886-86901, 2020.

[13] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks", *Frontiers in Plant Science*, Vol. 10, p. 621, 2019.

[14] S. Wolferteral, L. Ge, C. Verdouw, and M. Bogaardt, "Big Data in Smart Farming – A review", *Agricultural Systems*, Vol. 153, pp. 69-80, 2017.

[15] T. V. Klompenburga, A. Kassahuna, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review", *Computers and Electronics in Agriculture*, Vol. 177, p. 105709, 2020.

[16] T. Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review", *Computers and Electronics in Agriculture*, Vol. 177, p. 105709, 2020.

[17] P. Kamath, P. Patil, S. Shrilatha, Sushma, and S. Sowmya, "Crop Yield Forecasting using Data Mining", *Global Transitions Proceedings*, Vol. 2, No. 2, pp. 402-407, 2021.

[18] R. Suresh and R. Santhi, "Soil Test Crop Response Based Integrated Plant Nutrition System for Maize on Vertisol", *International Journal of Current Microbiology and Applied Sciences*, Vol. 7, No. 8, 2018.

[19] J. S. Bhanua, S. D. Bigul, and A. Prakash, "Agricultural internet of things using machine learning", *AIP Conference Proceedings*, Vol. 2358, No. 1, p. 080010, 2021.

[20] D. Paudel, H. Boogaard, A. D. Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, "Machine learning for large-scale crop yield forecasting", *Agricultural Systems*, Vol. 187, p. 103016, 2021.

[21] H. Aghighi, M. Azadbakht, D. Ashourloo, H. S. Shahrabi, and S. Radiom, "Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI", *IEEE Journal of Selected*

Topics in Applied Earth Observations and Remote Sensing, Vol. 11, No. 12, pp. 4563-4577, 2018.

[22] Y. H. Peng, C. S. Hsu, and P. C. Huang, “Developing crop price forecasting service using open data from Taiwan markets”, In: *Proc. of 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, Tainan, Taiwan, pp. 172-175, 2015.

[23] A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, and A. Sharma, “IoT and Machine Learning Approaches for Automation of Farm Irrigation System”, *Procedia Computer Science*, Vol. 167, pp. 1250-1257, 2020.

[24] H. Pariaman, G. M. Luciana, M. K. Wisyaldin, and M. Hisjam, “Anomaly Detection Using LSTM-Autoencoder to Predict Coal Pulverizer Condition on Coal-Fired Power Plant”, *Evergreen Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*, Vol. 8, No. 1, pp. 89-97, 2021.

[25] S. Iniyan, V. A. Varma, and C. T. Naidu, “Crop yield prediction using machine learning techniques”, *Advances in Engineering Software*, Vol. 175, p. 103326, 2023.

[26] R. Kumar and V. Singhal, “IoT enabled crop prediction and irrigation automation system using machine learning”, *Recent Advances in Computer Science and Communications*, Vol. 15, No. 1, pp. 88-97, 2022.

[27] J. Ansarifar, L. Wang, and S. V. Archontoulis, “An interaction regression model for crop yield prediction”, *Scientifc Reports*, Vol. 11, p. 17754, 2021.

[28] A. E. García, G. M. S. Zarazúa, M. T. Ayala, E. R. Araiza, and A. G. Barrios, “Applications of Artificial Neural Networks in Greenhouse Technology and Overview for Smart Agriculture Development”, *Applied Sciences*, Vol. 10, No. 11, p. 3835, 2020.