Applied Intelligence (2023) 53:28291–28309
https://doi.org/10.1007/s10489-023-05005-5

Optimizing LSTM and Bi-LSTM models for crop yield prediction and comparison of their performance with traditional machine learning techniques

V. Kiran Kumar[1] · K. V. Ramesh[1] · V. Rakesh[1]

Accepted: 5 September 2023 / Published online: 29 September 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Advance prediction of crop yield is very critical in the context of ensuring food security as the region specific challenges in social and environmental conditions often infringe plan of policy makers. This study presents a generic methodology to configure and fine tune the state-of-the-art Long Short-Term Memory (LSTM) based Deep Learning (DL) model through hyperparameter optimization for prediction of yield (annual crop production) in Wheat, Groundnut and Barely over India based on multiple independent input variables identified using multicollinearity test. The Monte Carlo cross-validation method is used to validate the optimized LSTM models. Results from the LSTM model tuning showed that among the 4 optimizers tested, Adam was found to perform better irrespective of the crop and Bi-LSTM outperformed sLSTM in terms of prediction accuracy. The percentage reduction in error with Bi-LSTM compared to sLSTM in predicting wheat and groundnut crop yield was 39% and 13% respectively while in case of barley crop, error reduction was marginal (0.34%). The performance of optimized Bi-LSTM model is compared with the performance of traditional machine learning (ML) models such as support vector regression (SVR) and SVR polynomial {2nd and 3rd order}, Auto Regressive Integrated Moving Average (ARIMA) and ARIMAX (ARIMA with exogenous variables) and Vector Auto-regression (VAR). The Bi-LSTM model is found to be superior to ML models; the percentage reduction in mean absolute scaled error with the Bi-LSTM compared to the best performing ML model was 94%, 72%, and 71% in predicting wheat, groundnut and barley yield respectively. This study showed that by choosing proper explanatory (independent) variable and hyperparameter optimization, a simple (single layer) structure of deep neural network (LSTM) outperformed traditional ML models in terms of accuracy for crop yield prediction application.

Keywords Machine leaning · Deep learning · Long Short-Term Memory (LSTM) · Support Vector Regression · Auto Regressive Integrated Moving Average (ARIMA) · Vector Auto-regression (VAR)

1 Introduction

Time series forecasting models are capable in providing advance guidance for timely decision making in various practical applications [1]. The type of data along with the underlying characteristics is the dominant factor determining the performance and accuracy of time series data analysis and forecasting techniques [2, 3]. Conventional forecasting models often utilize the linear regressions for model fitting, and then a moving average for prediction purposes. Most of these models like Auto-Regressive Integrated Moving Average (ARIMA), ARIMAX (ARIMA with exogeneous variable), Vector Autoregression (VAR) and Support Vector Regression (SVR) are well studied and found effective for many simpler forecasting problems. ARIMA model is widely used for various applications in agricultural sector [4–7]. ARIMA model is deployed successfully for forecasting annual cocoa production in Nigeria [4] while ARIMAX model is found to be quite effective in forecasting sugarcane yield in Northern Agro-climatic zone of Haryana, India [6]. Sapankevych and Sankar [8] discussed the applicability of

✉ K. V. Ramesh
kvram55@gmail.com

V. Rakesh
rakeshv82@gmail.com

[1] Earth and Engineering Sciences Division, Council of Scientific and Industrial Research-Fourth Paradigm Institute (CSIR-4PI), Bengaluru 560037, Karnataka, India

support vector machines (SVMs) for time series forecasting in various applications while Kok et al. [9] reviewed the effectiveness of SVM in precision agriculture applications. Raj et al. [10] used VAR model in predicting tea yield over South India and showed that model could capture the nonlinear short-term fluctuations as well as the long-term variations. A fuzzy based SVR modeling framework was found to be suitable to predict rice crop production with high accuracy [11]. However, such traditional methods have limitations like failing to focus on complete (missing) data, less effective for multivariate problems and less precise in forecasts having long-time horizon (multi-time steps) [12, 13].

Machine Learning (ML) and more importantly Deep Learning (DL) based algorithms are emerging techniques in Artificial Intelligence (AI) based data analytics. Primary difference between conventional ML techniques and DL is that the former tries to establish relationships between variables while the latter is designed for making predictions as accurate as possible from available past data [14]. DL based methods show robustness to noisy inputs and has the capability to approximate nonlinear functions [15]. It is also reported that DL based algorithms have distinct advantage in many applications in agricultural sector compared to traditional ML models [16–18]. Current generation Artificial Neural Network (ANN) algorithms have the capability to learn noisy and non-linear relationships, with arbitrarily defined but fixed number of inputs and outputs, making them suitable for multivariate and multi-step forecasting [19]. Many studies in the past highlighted the potential of neural network algorithms in prediction problems of various sectors including agriculture [20–22]. Srivastava et al. [23] used a Convolutional Neural Network (CNN) model to predict wheat yield in Germany from environmental and phenological data and concluded that CNN model outperformed traditional linear models. Another study reported the effectiveness of back propagation neural networks in predicting rice yield over mountainous region in Fujian province of China based on weather parameters and discussed the advantage of using these algorithms compared to multiple linear regressions [24]. O'Neal et al. [25] demonstrated the effectiveness of neural network algorithm for maize yield prediction in USA from weather parameters. Wolanin et al. [26] used a deep neural network based algorithm to predict wheat yield in India and showed that the performance of this model was superior to regression based models.

DL methods open new avenues in time series forecasting such as it is both data driven and non-linear, having no requirement for an explicit underlying model (nonparametric), and it has automatic learning skill with longer dependency in data without making strong assumptions [3, 27, 28]. However, traditional Recurrent Neural Networks (RNNs) have limitations in learning time-series data that obstruct their training [29]. DL based RNNs algorithms such as Long Short-Term Memory (LSTM) has gained lot of attention in recent years for their applications in time series prediction [30]. The LSTM is known for its ability in learning long term dependencies and extract the useful information from the historical record and then predicting the future [30, 31]. LSTM model captures important features from the input and preserves this information over a long period. The feature of LSTM model memory cell to selectively remember key information makes it very suitable for prediction in case of the random non-stationary data [31]. LSTM models are suitable for multivariate time series prediction due to its ability in preserving and training the features of given data for a longer period [32].

In recent years; LSTM has been widely used in the agricultural applications [20, 21, 33]. The LSTM model is successfully deployed for predicting country level maize yields in United States driven by primarily openly available weather data [33]. A hybrid modeling system combining multiple linear regression and deep neural network (LSTM) is found to be effective in rice crop yield forecasting over Tamilnadu, India [34]. Crisóstomo et al. [35] used Bidirectional LSTM (Bi-LSTM) for rice crop detection from Sentinel-1 time series data and showed that Bi-LSTM performance was superior to other ML based algorithms. Ramesh et al. [36] showed that Bi-LSTM model reduced the error in wheat yield prediction over India to the order of 50% compared to conventional statistical models. Similarly, Tian et al. [37] deployed Bi-LSTM for improving wheat yield estimates in the Guanzhong Plain, China. Bi-LSTM is successfully implemented for prediction of evapotranspiration, a crucial parameter in agricultural management [38] while Nishu and Anshu [39] demonstrated application of Bi-LSTM in wheat yield prediction of Punjab region in north India.

The objective of this study is configuration and validation of LSTM model for prediction of crop yield in wheat, groundnut and barely over India and comparison of its performance with that of traditional ML models. We have selected these three crops because the trend in annual crop production of these crops showed distinct pattern making them suitable candidate to test the prediction skill of LSTM. The predictive explanatory (independent) variables for multivariate LSTM is identified by conducting multicollinearity test. The LSTM model is configured by hyperparameter optimization by investigating the influence of the number of hidden nodes, activation function, optimizers, number of training times (epochs) and batch size on model performance. The remainder of the study is organized as follows. Section 2 of the paper describes data processing and various model used. Details of model configuration and validations methodology are given in Section 3. Results of the study are discussed in Section 4 followed by major conclusion of the study in Section 5.

2 Data used and evaluation methodology

Datasets used in this study are sourced from openly available records. Annual crop production data from 1961 to 2018 were obtained for India from the FAOstat (www.fao.org/faostat/en/#data/QC). Weather data for the same period is obtained from India Meteorological Department (IMD) (mausam.imd.gov.in). There are many factors which affect the crop yield [9] and among them, those affect the crop production significantly were considered as input variable in this study. The input variables chosen were crop area (the total harvested area of the crops for the year in hectares), number of wet days (the number of rainy days during the crop season; DWET), temperature (^{0}C), rainfall (in mm) and population. Among the input parameters mentioned above, crop related parameters are obtained from FAOstat while weather input parameters are taken from IMD.

2.1 Data pre-processing

Data pre-processing in DL is a pivotal step that improves the quality of data which helps in extraction of useful information and hence improves model learning ability. During pre-processing, we removed inaccurate, inconsistent and noisy data as well as checked multicollinearity (very high inter-associations among the independent variables) [40]. In the presence of inadequacies and high multicollinearity among independent variables, the statistical inferences made using the dataset may not be reliable. Variance Inflation factor (VIF) analysis of our dataset showed high multicollinearity between the variables average rainfall and average temperature which can hamper the interpretability of model results. Therefore, instead of quantitative rainfall, rainfall anomaly and interannual variability in rainfall were derived as follows

$$R_a = \left[\frac{R_i - \overline{R_i}}{\overline{R_i}}\right] \times 100, \text{ Where, } R_a \text{ is rainfall anomaly, } R_i \text{ is annual rainfall, } i = \{1, 2 \ldots \ldots 58\} \text{ and } \overline{R_i} \text{ is the average rainfall.} \quad (1)$$

$$R_{YoY} = \left[\frac{R_i - R_{i-1}}{R_{i-1}}\right] \times 100, \text{ Where, } R_{YoY} \text{ is the interannual variability in rainfall and } R_{i-1} \text{ is the rainfall for the previous year.} \quad (2)$$

Similarly, based on VIF analysis, seasonal minimum and maximum temperature is considered instead of average temperature. Then, multicollinearity is analyzed by calculating VIF as shown in Table 1. Here, we have considered VIF < 5.85 (p = 5) to be the cutoff for no collinearity, similar to the previous study [41]. In order to compare the variables of different scales, the data were subjected to normalization, which rescales the input data from original range so that all values are within the new range of 0 and 1. Min–Max scalar transformation is performed to normalize the original data using the formula y = (x–min) / (max–min). Where, the minimum and maximum pertain to the value of x being normalized [42]. Eventually, area cultivation of respective crop, population and weather data comprising maximum temperature, minimum temperature, number of WET days (DWET), rainfall anomaly and interannual variability in rainfall are merged as exogenous inputs to predict the crop production.

Table 1 Identification of significant explanatory variables for formulating multivariate prediction model

Multicollinearity (Variance Inflation Factor)	
Parameters	VIF Value
Area Cultivation	2.39
Annual rainfall anomaly	2.90
Year over Year rainfall	1.43
WET days	1.29
Maximum Temperature	5.78
Minimum Temperature	5.79
Population	1.29

2.2 Training and validation data sets

The performance of DL based model is very much dependent on how the entire data sets are split into training and validation data [42]. In this process input dataset is divided into two subsets. The first subset is used to train and fit the model and is referred to as training dataset. The second subset is validation dataset, which is not part of the training dataset. The validation data is provided as input to the trained model to evaluate the actual performance. The important configuration parameter in this procedure is the size of the train and validation data sets.

The model is trained based on 80% (of total 58 years) of data which account to 46 years of data for the period 1961 to 2006. Then the actual performance of the model is assessed using the validation data sets. The validation is carried out with remaining 20% of data which accounts to 12 years of data (2007 to 2018) which is not part of training dataset. In this study, for all the models (DL based (Bi-LSTM and S-LSTM) and ML based models) the training data sets is the 80% of total data (ie data for the period 1961–2006). Then the remaining 20%

of data (ie for the period 2007–2018), is used for validation all the model (Algorithm 1).

2.3 Evaluation metrics

The error metrics or evaluation metrics are the measures used for assessing the accuracy of prediction by the models. Multiple metrics were used in the past to measure the performance of time series prediction models [43]. We have used a number of evaluation parameters (scale dependent and scale independent) to assess the performance of LSTM models and other traditional ML models such as Fractional Bias (FB), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Relative Absolute Error (MRAE), Root Mean Square Percentage Error (RMSPE), Mean Absolute Scaled Error (MASE), and Relative Improvement Parameter. The initial shortlisting of LSTM model is done based on MASE such that only those models having MASE < 1 is only retained for further analysis. The MASE > 1 implies that the actual forecast does worse out of sample than a naive forecast did in sample, in terms of mean absolute error. The relative Improvement Parameter is computed as the percentage reduction in error with Bi-LSTM compared to the sLSTM and other statistical models are computed using the formula

$$R_{dn}(\%) = \left(1 - \frac{B_L}{O_M}\right) x100 \tag{3}$$

Where, R_{dn} is the percentage reduction in error,B_L is the error of Bi-LSTM and O_M is the error of other models.

3 Model configuration and fine tuning

In this section, we discuss the how the LSTM model and other statistical models are configured for conducting experiments in this study. A brief discussion on the mathematical formulation and setting up parameters for the ML methods used is given. The ML methods chosen in this study to compare the performance of LSTM model is based on their popularity in similar applications presented in this study. The LSTM model configuration and hyperparameter tuning procedure is also discussed.

3.1 Description of ML models used

We compared the performance of LSTM variants with traditional ML models like SVR, SVR polynomial (2nd and 3rd order), ARIMA, ARIMAX and VAR models. These were chosen to compare the LSTM models based on earlier studies that demonstrated their capability in prediction of crop production from past data [3, 7, 9]. The details of ML models used are shown in Table 2 with their mathematical formulations. SVR, a supervised learning algorithm, is one of the most widely used techniques in time series data prediction [44]. SVR Polynomial is derivative of SVR in which the order is higher than 1 and it trains using a symmetrical loss function that equally penalizes high and low miss estimates [44]. SVR polynomial with order 2 and 3 were used in this study. Another popular model used for time series forecasting is ARIMA [45]. ARIMA model with endogenous variable is called ARIMAX and this model is also used in many

Table 2 Summary of traditional Statistical Models used in this study: General equations and notations

Models	General Equation	Notations
Support Vector Regression	$f(x) = \omega\phi(x) + b$	$\omega and barecoefficients$ $\phi(x)$ high dimensional feature space
Support Vector Regression Polynomial	$f(x, w) = \sum_{i=1}^{M} w_i x^i$	M is the order of the polynomial w is vector magnitude
ARIMA	$\phi_p(B)\Delta^d Y_t = c\prime + \theta_q(B)a_t$	Y = dependent variable B = lag operator a = error term $\phi_p(B)$ = non-seasonal AR $\Delta^d = (1-B)^d$ = non-seasonal difference $\theta_q(B)$ = non-seasonal MA t = time
ARIMAX	$\left(1 - \sum_{x=1}^{p} a_s L^s\right)\Delta y_t = \mu * \sum_{x=1}^{q} \beta_s L^s x_t + \left(1 + \sum_{x=1}^{d} \gamma_s L^s\right) e_t$	$\Delta y_t = y_t - y_{t-1}$ $e_t = errors$
VAR	$y_t = c + A_1 y_{t-1} + A_2 y_{t-1} + \cdots + A_p y_{1-p} + e_t$	$y_{i,t}$ = Variable vector at time t A_i = time variant e_t = error term

prediction applications [46]. The VAR model is an alternative to traditional models to study the interactions among the inter-related time series data [47]. The hyperparameters optimization in ARIMA and ARIMAX was considered by defining range for autoregressive (p), integration (d) and moving average (q) (p = 1, 2, d = 1, q = 1, 2, 3) in the model. Then, the best performing model is selected using minimum Bayesian Information Criteria (BIC) and Akaike Information Criteria (AIC) value. In VAR model, the max lag order was set to 12 which is identified through iterative method such that the AIC value was the least.

3.2 Configuration of LSTM models

3.2.1 Standard LSTM (sLSTM) and Bidirectional LSTM (Bi-LSTM)

Crop yield prediction uses a sequence of crop production values with *n* historical time (year wise) steps as the input data, which can be represented by a vector,

$$X_T = \left[x_{T-n}, x_{T-(n-1)}, \ldots, x_{T-2}, x_{T-1}\right] \tag{4}$$

The crop production is mainly influenced by the crop area, number of wet days, temperature, rainfall and population as described in the Section 2.1. To reflect the temporal attributes of the crop production and simplify the expressions of the equations in the following subsections, the speed matrix is represented by a vector, $X_T^V = \left[x_{T-1}, x_{T-(n-1)}, \ldots, x_{T-2}, x_{T-1}\right]$, in which each element is a vector of the V values. In the deep neural network calculation, at each iteration time, *t*, the hidden layer maintains a hidden state, h_t, and updates it based on the layer input, x_t, and previous hidden state, h_{t-1}, using the following equation:

$$h_t = \sigma_h\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right) \tag{5}$$

where W_{xh} is the weight matrix from the input layer to the hidden layer, W_{hh} is the weight matrix between two consecutive hidden states (h_{t-1} and h_t), b_h is the bias vector of the hidden layer and σ_h is the activation function to generate the hidden state. The hidden layer of the standard LSTM is named as LSTM cell as shown in Fig. 1a. During each iteration time, *t*, the LSTM cell has the layer input, x_t, and the layer output, h_t. The complicated cell also takes the cell input state, x_t, the cell output state, C_t, and the previous cell output state, C_{t-1}, into account while training and updating parameters (Fig. 1a). The description of basic LSTM structure and formulations are available in [48].

The final output of a sLSTM layer should be a vector of all the outputs, represented as.

$\boldsymbol{Y}_T = [h_{T-n}, \ldots, h_{T-1}]$. Here, when taking the crop production prediction problem as an example, only the last element of the output vector, h_{T-1}, is the value need to be predicted. Thus, the predicted crop production value for the next iteration time, *T*, is h_T. sLSTM processes the information in one direction (forward/backward) only. In order to beat this shortcoming, Schuster and Paliwal [49] introduced the concept of bidirectional RNN, and then, Graves and Schmidhuber [50] introduced a full gradient version of the Bi-LSTM algorithm by merging desirable features of bidirectional RNN with those of sLSTM. The Bi-LSTM architecture can be trained in both the time directions (forward and backward) simultaneously to preserve the future and the past information, by splitting sLSTM into forward layers and backward layers.

The structure of a Bi-LSTM layer, containing a forward LSTM layer (L_F) and a backward LSTM layer (L_B), is shown in Fig. 1b. The forward layer output sequence, $\overrightarrow{h}$, is iteratively calculated using inputs in a positive sequence from time *T-n* to time *T*-1, while the backward layer output sequence, $\overleftarrow{h}$, is calculated using the reversed inputs from time *T-n* to *T*-1. Similar to the sLSTM layer, the final

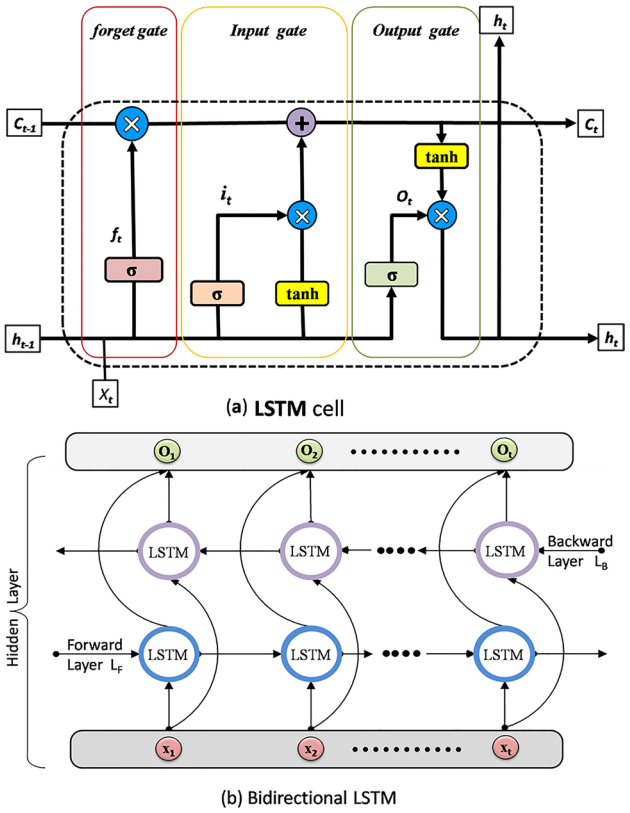

Fig. 1 Schematic of LSTM variants (**a**) sLSTM cell with sigmoid (σ) and tanh function having forget gate (f_t), with multiplication as point operation, Input gate (i_t) addition and multiplication as point operation and Output gate (O_t) with multiplication point operation; where x_{t0} is input, h_{t-1} output of previous cell, C_{t-1}memory of previous cell, h and C_t are output and memory of the cell (**b**) Bi-LSTM with forward and backward layers in hidden layer, $x_1 \ldots x_t$ is the input vector in input layer and $O_1 \ldots O_t$ is the output vector

output of a Bi-LSTM layer can be represented by a vector, $\boldsymbol{Y}_T = [y_{T-n}, \ldots, y_{T-1}]$, in which the last element, y_{T-1}, is the predicted speed for the next time iteration when taking crop production as an example.

3.2.2 LSTM model configuration

In this section, we describe how the LSTM model is configured for prediction of crop yield for selected crops. We have selected wheat as primary crop to apply LSTM for prediction of annual crop yield. We have selected two more crops to test LSTM such that their annual production data were distinctly differing from wheat production time series data. To identify these crops, we have computed correlation of wheat production data with other prominent crops in the country as shown in Table 3a. We have identified barley and groundnut as their production data showed stronger dissociation with wheat production when compared to other crops (Table 3a). The crop production data for barley showed significant strong negative correlation with wheat production while groundnut production also showed weaker correlation with wheat production compared to others as evident from the slopes and d-bar value (Table 3b). The long-term trend in annual production for wheat, barley and groundnut clearly indicated the distinct behavior of these crops (Fig. 2). Wheat production showed strong increasing trend, barley production showed decreasing trend and groundnut production did not show any significant trend. Table 4 and Algorithm 1 explain the experimental setup and algorithm design of different experiments. Here, wheat, groundnut and barley are the response variables which are considered individually and variables like crop area, rainfall anomaly, interannual variability in rainfall, minimum temperature, maximum temperature, DWET and population are the explanatory variables. The same sets of variables are considered for both deep learning and multivariate statistical models.

Table 3 Identification of distinct crops for testing LSTM model for prediction of crop yield, (a) Correlation of wheat production data with production of prominent crops in India (b) Slope and d-bar value of the distinct crops Wheat, Barley and Groundnut as highlighted as bold

(a)			
Correlation table			
Time series	r	Time series	R
Paddy	0.9787	Papaya	0.8781
Sugarcane	0.9662	Orange	0.8467
Potato	0.953	Onion	0.8391
Coconut	0.9526	Cotton	0.8384
Eggplant	0.9434	Pigeon pea	0.7862
Rapeseed	0.9406	Beans	0.7753
Soybean	0.9227	Chickpea	0.6424
Tomato	0.9194	Cassava	0.5785
Banana	0.9174	Millet	0.5238
Jute	0.9006	Groundnut	**0.5229**
Maize	0.8997	Sorghum	-0.5955
Mango	0.8927	Barley	**-0.7491**
b			
Time series		Slope	d-bar value (*** 99% significance level)
Wheat vs Barley		-36.018	13.572 ***
Wheat vs Groundnut		9.725	12.841 ***
Barley vs Groundnut		-0.192	19.373 ***

3.2.3 Hyperparameter tuning

Although the performance of LSTM algorithms in time series forecasting is promising, accuracy of the algorithm depend heavily on how the model parameters are tuned for specific applications. One important aspect in optimizing LSTM model for specific application is the choice of hyperparameters that is crucial in algorithm training and overall model performance [51]. The random selection of hyperparameters might cause either overfitting or underfitting of the data which affects the model performance. Hyperparameter

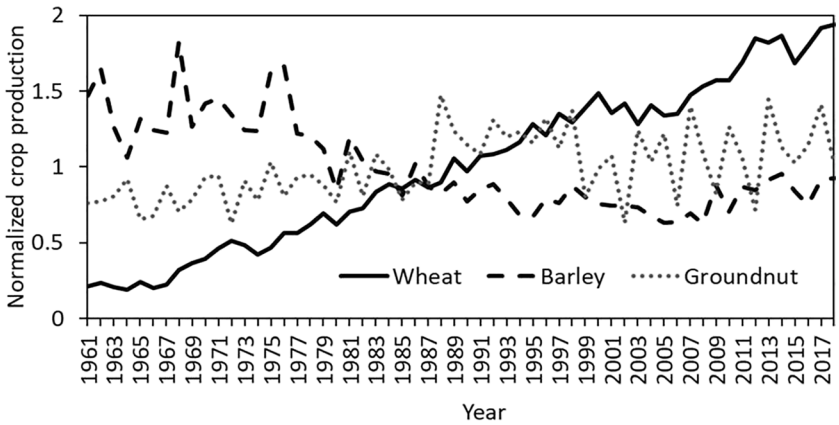

Fig. 2 Yearly production in wheat, barley and groundnut in India normalized with their respective average production

Table 4 Experimental setup of LSTM: List of model parameters used to tune the LSTM model

Variables/ Hyperparameters	LSTM Variants
Response Variables	3 (wheat, groundnut and barley)
Explanatory Variables	7 (crop area, anomaly rainfall, year-over-year rainfall, minimum temperature, maximum temperature, number of WET day (DWET) and population)
Nodes (Neurons)	2, 3, 4, 5 ……, 100, 101
Batch size	4, 8, 16, 32
Dropout Layer	0.20, 0.25, 0.30, 0.35, 0.4, 0.45,0.50
Optimizers	RMSProp, AdaDelta, Adam, Nadam, Adagrad and Adamax
Activation	ReLU, tanh
Epochs	100, 150, 200, 250, 300
Total Simulations per crop per LSTM variant (Combinations)	168,000

optimization or tuning is interpreted as a problem of choosing a set of optimal parameters to find the right combination of their values which helps to find the minimum loss or maximizes the model performance and yields an optimal model [52]. Here, an approach for hyperparameter optimization is proposed and demonstrated its effectiveness in algorithm training for predicting crop yield from past data. The focus is on hyperparameter optimization which includes choice of number of hidden neurons, dropout percentage, activation function, optimizer, number of epochs and batch size for two variants of LSTM, Unidirectional (sLSTM) and Bidirectional (Bi-LSTM).

The methodology of selection of hyperparameters and proposed prediction strategies for two LSTM variants and other statistical models is depicted in Fig. 3. The DL models are subjected to supervised learning with the training window of 2 points for 1 prediction point. The models were tuned for multiple set of hyperparameters as shown in the Table 4. Algorithm 1 details the procedure to train and validate LSTM and ML based models for crop yield

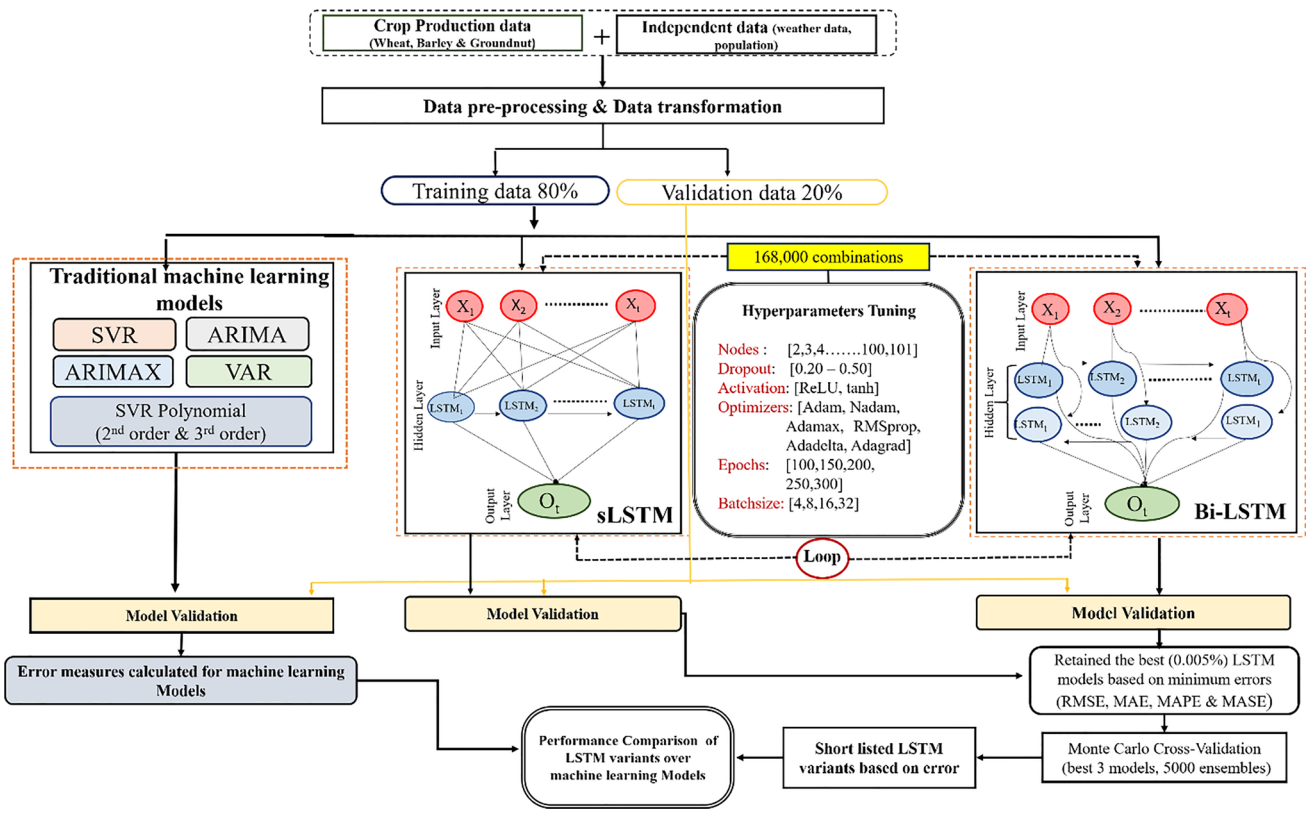

Fig. 3 Schematic of methodology to identify the best prediction models for crop production time series data prediction

Input crop area, rainfall anomaly, year-over-year rainfall, minimum temperature, maximum temperature, number of WET days and population

Models SVR, SVR (2^{nd} & 3^{rd}), ARIMA, ARIMAX, VAR, LSTM, Bidirectional LSTM

Output Crop yield (wheat, groundnut, barley)

Procedure

1 Step 1 LOAD dataset X:
SPLIT *dataset into train samples (X_{train}) and validation samples (X_{test})*

2 Step 2 EXTRACT X_{train}
TRAIN *traditional ML models with* X_{train}
EXTRACT X_{test}
$\boldsymbol{Y}_T$= PREDICT (X_{test})
GENERATE *error matrix*

3 Step 3
```
for no. of nodes do
  for no. of dropout do
    for no. of optimizer do
      for no. of epoch do
        for no. of batch do
          LOAD X_train
          TRAIN sLSTM & Bi-LSTM with X_train
          LOAD X_test
          Y_T= LSTM PREDICT (X_test)
          GENERATE error matrix
          end for
        end for
      end for
    end for
  end for
```
SELECT best 3 models based on error matrix
PERFORM Monte-carlo cross validation

4 Step 4 PERFORMANCE COMPARISON *LSTM variants and Traditional ML model error metric*

5 REPORT results

End procedure

Algorithm 1: To compare crop yield prediction skill of LSTM models with ML based models

prediction. From these simulation models, best performing (0.005% of total) models were retained based on computed errors during training such as RMSE, MAE and MAPE (Fig. 3). From these combinations, best 3 models for each LSTM variant were selected based on the minimum MASE and the performance of these tuned LSTM models was compared with performance of ML models by computing relative error metrics like fractional bias (FB), mean relative absolute error (MRAE), root mean square percentage error (RMSPE) (Fig. 3).

Details of LSTM configuration and model parameters selected are as below.

(1) Nodes (or hidden neurons): As per literature, there is no clear definite thumb rule on selection of number of nodes (or hidden neurons). Very often a trial-and-error approach only gives the best result for different applica-

tions. Sheela and Deepa [53], reviewed the methods to fix the neurons in neural networks and their study considered 101 variations (2 to 101) of neurons and compared their results with other methods. Here, we followed similar procedure (Fig. 3) for determining number of neurons [53].

(2) Batch size: Batch size denotes the subset size of the training sample used to train the network during its learning process. Each batch trains network in successive order considering the updated weights coming from the previous batch. It also controls how often one should update the weights of the network. Kandel and Castelli [54] studied the effect of batch size on the generalizability of neural network and recommended batch sizes to be selected as a power of two and not to exceed the number of samples. Accordingly, the batch sizes used here are powers of 2 such as 4, 8, 16, and 32 (Fig. 3).

(3) Dropout: Dropout introduces an extra hyperparameter (the probability of retaining a unit). It controls the intensity of dropout. This layer will help to prevent overfitting by ignoring randomly selected neurons during training, and hence reduces the sensitivity to the specific weights of individual neurons. Simple way to prevent overfitting of neural networks is by considering typical values of dropout rate for hidden units in the range 0.20 to 0.50 [55]. The dropout rates used in this study were 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, and 0.50 (Fig. 3).
(4) Optimizer: Optimizers are methods or algorithms used to update the weights of the LSTM network in order to reduce the loss function. Optimizers play a very crucial role in increasing the accuracy of the model [56]. Optimization algorithms considered in this study are RMSProp, AdaDelta, Adam, Adagrad, Adamax, and Nadam (Fig. 3).
(5) Activation: The activation function is to help the network in learning complex patterns in the data. The activation function decides what is to be fired to the next neuron during the learning process. It adds the ability to model non-linear processes to the neural network. The activations functions used in this study are tanh and Rectified Linear Activation function (ReLU) [57].
(6) Epoch: Epoch decides the number of times that the LSTM learning algorithm will iterate through the entire training dataset. It determines how many samples in the training datasets have opportunity to get updated by the internal model parameters. The epochs used in this study were 100, 150, 200, 250, and 300 (Fig. 3).

The Monte-Carlo Cross-Validation (MCCV) method is applied to ascertain diversity in the training dataset and to obtain an ensemble of the prediction. This also helps to estimate the prediction errors while considering the uncertainty arise from random splitting patterns (random sub sampling) as well as the random initialization of the algorithm weights [58]. MCCV was applied to create an ensemble of predictions to get the average performance of the models. The entire original input–output patterns were randomly sub sampled with suitable replacement to generate the training dataset for tuning the model parameters and generate testing dataset to evaluate the model performance. The weights of the individual neurons were initialized with a set of random values to produce variability in the outputs. In this study neural networks were repeatedly trained using 5000 sets of randomly sub sampled data and 5000 sets of initial weight vectors with the chosen numbers of input variables and hyperparameters (Fig. 3). In each time, datasets were divided into training and test datasets with random replacement. 5,000 simulations were carried out with selected models for each crop using randomly assigned training data to construct prediction ensembles (Algorithm 1).

3.3 Computational setup and computing time

The experiments were carried using python 3.8 and experiments are conducted on a computer with Intel Xenon Gold 6254 CPU at 3.10 GHz and 512 GB of memory using x86_64 Linux system. For each crop, the total number of simulations for each LSTM variant on account of all combinations of model hyperparameters is 1,68,000 (100 (nodes) × 4 (batch size) × 7 (drop out) × 6 (optimizers) × 2 (activation function) × 5 (epochs)). The computation time per simulation for LSTM models and different ML models for different crops is shown in Table 5. The computation time per simulation for LSTM variants is much higher compared to statistical models. Bi-LSTM is slightly computationally expensive compared to sLSTM. As expected, the computational time per simulation for different models among the crops is similar.

Table 5 Computational time (seconds) per simulation for LSTM models and different statistical models for different crops

Model	Computational time (seconds) per simulation		
	Wheat	Barley	Groundnut
Standard-LSTM	14.28	14.19	14.94
Bidirectional-LSTM	14.63	14.61	14.87
SVR	0.06	0.04	0.04
SVR Polynomial (2)	0.04	0.05	0.04
ARIMA	0.18	0.14	0.23
ARIMAX	0.60	0.62	0.63
VAR	0.57	0.65	0.51

4 Results and discussion

4.1 Results

In this section, we describe the results from the experiments carried out for identifying the optimum hyperparameters for the LSTM variants in predicting time series of three distinct (which are independent to each other) crop (wheat, barley and groundnut) production data. We also present the results of how the LSTM models are optimized and a comparison of the performance of two LSTM variants (sLSTM and Bi-LSTM).

The prediction error of both sLSTM and Bi-LSTM in predicting yield for the three crops is shown as 3D scatter plot; RMSE in x-axis, MAE in y-axis and MAPE in z-axis (Fig. 4). It is clear that there are many model combinations for both the LSTM variants which showed minimum errors in predicting crop production in wheat (Fig. 4(a), (b)). Prediction errors for both the LSTM variants are slightly higher for barely and groundnut compared to wheat. In general, the prediction error for the Bi-LSTM models was lower than that of the sLSTM models particularly for wheat yield prediction (Fig. 4). From the error distribution for different combinations of LSTM variants, best performing model

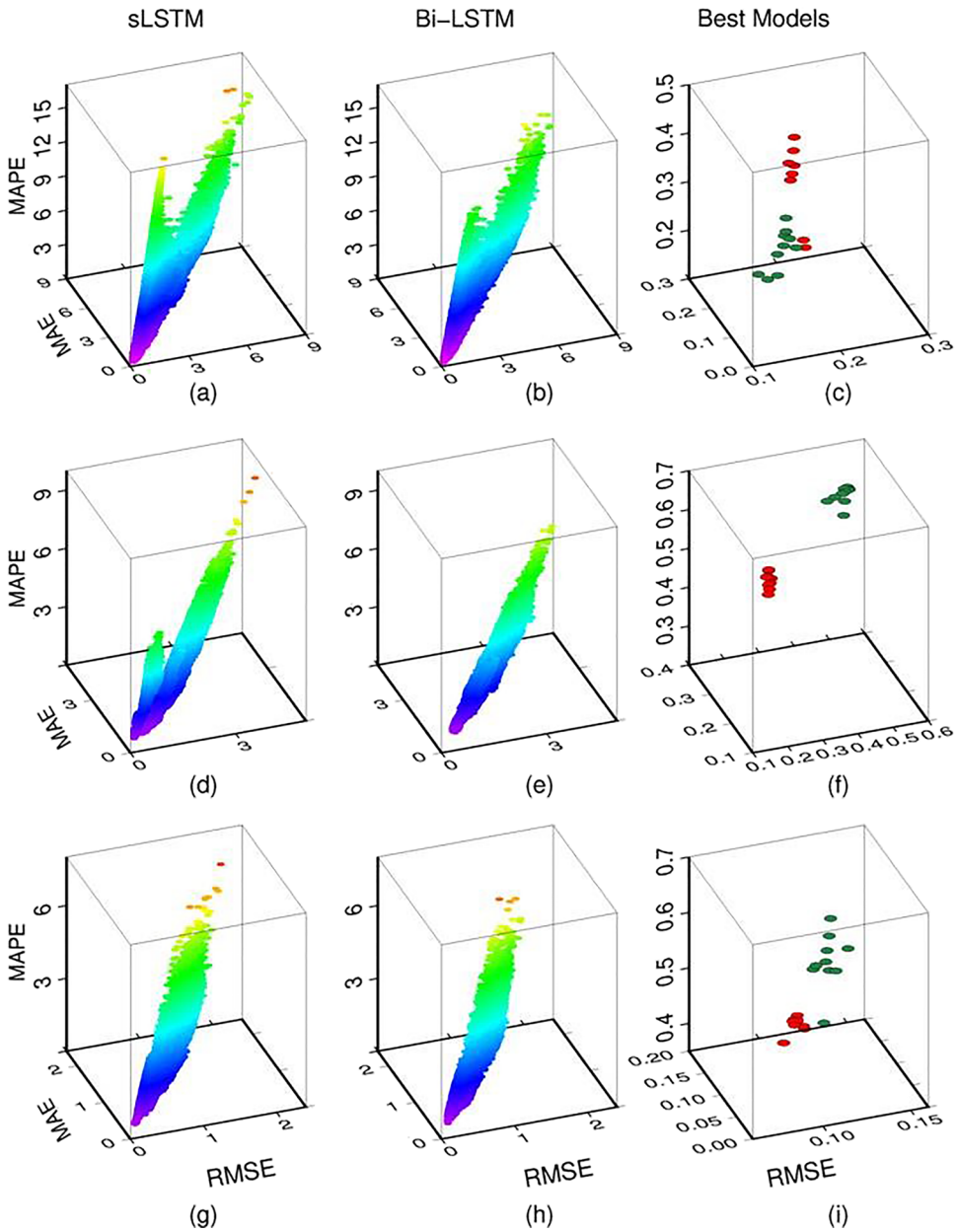

Fig. 4 Primary error metrics (RMSE on X-axis, MAPE on Y-axis & MAE on Z-axis) for different combinations of sLSTM (left panel; **a**,**d**, **g**), and Bi-LSTM (middle panel; **b**, **e**, **h**) model simulations in predicting crop production for wheat (top panel; **a**, **b**, **c**), barley (middle panel; **d**, **e**, **f**) and groundnut (bottom panel; **g**, **h**, **i**). The best LSTM models (red dots for sLSTM and green dots for Bi-LSTM) identified for each crop in terms of minimum error in prediction is also shown (right panel; **c**, **f**, **i**)

combinations (top 0.005% of the total combinations in terms of minimum error) were identified. Best models shortlisted for wheat, barley and groundnut for sLSTM (shown as red dots) and Bi-LSTM (shown as green dots) is shown in Fig. 4(c), (f) and (i) respectively. It was found that the LSTM variants performed differently for different crop data. These shortlisted models have same number of nodes, activation and to some extent similar batch size for respective crops. The major varying hyperparameters were optimizers and dropout percentage. From these top performing models, best three models were identified based on the least value of composite error (MASE) for each crop and LSTM variant. Then, MCCV is carried out for these identified models. The finalized (best three from the shortlisted models) models for both the LSTM variants with optimal hyperparameters for the three crops were tabulated in Table 6.

MCCV is used to evaluate the models described in the Table 6 and compare the prediction ability of different models. The average prediction uncertainty for 1% of total simulations is compared with that of 0.1% in terms of MASE and found that there is not much significant difference between them. However, the average uncertainty of all ensembles showed large error compared to 1% in terms of MASE. The optimized LSTM models after the MCCV are shown in Table 6. The average error metrics of the optimized models for sLSTM and Bi-LSTM are shown in the Table 7. Best performing sLSTM for wheat crop yield prediction from MCCV had the following characteristics; 89 neurons, tanh as an activation function, two optimizers Nadam and Adam with batch size of 32, epoch of 300, 0.25 and 0.45 as dropout percentage (Table 6) and average error statistics such as MAPE of 6.8, MASE of 0.03 and R-square of 98% with a negligible fractional bias of 0.0008 (Table 7). Similarly, the best performing Bi-LSTM model for wheat had specifications such as 91 neurons, ReLU as an activation function, RMSprop and Adam as optimizers with two batch sizes of 4 and 8, 100, 150 and 250 as epochs and 0.40, 0.50 as dropout percentages (Table 6) and error statistics like MAPE of 4.14, MASE of 0.04, R-square of 99% and fractional bias of 0.0015 (Table 7).

Table 6 Best three performing DL models with optimal hyperparameters shortlisted based on the proposed approach

Optimized models with hyperparameters			Nodes	Dropout	Activation	Optimizer	Epoch	Batch size
Wheat	sLSTM	Model 1	89	0.45	Tanh	Nadam	300	32
		Model 2	89	0.25	Tanh	Nadam	300	32
		Model 3	89	0.25	Tanh	Adam	300	32
	Bidirectional LSTM	Model 1	91	0.4	ReLU	RMSprop	100	4
		Model 2	91	0.4	ReLU	RMSprop	250	8
		Model 3	91	0.5	ReLU	Adam	150	4
Groundnut	sLSTM	Model 1	68	0.5	ReLU	RMSprop	300	4
		Model 2	68	0.35	ReLU	RMSprop	300	4
		Model 3	68	0.45	ReLU	RMSprop	300	4
	Bidirectional LSTM	Model 1	89	0.5	ReLU	Nadam	300	4
		Model 2	89	0.2	ReLU	Adam	250	8
		Model 3	89	0.35	ReLU	Adam	250	8
Barley	sLSTM	Model 1	57	0.5	ReLU	RMSprop	200	4
		Model 2	57	0.4	ReLU	RMSprop	300	4
		Model 3	57	0.5	ReLU	RMSprop	300	4
	Bidirectional LSTM	Model 1	8	0.4	ReLU	Adamax	200	32
		Model 2	8	0.4	ReLU	RMSprop	100	32
		Model 3	8	0.3	ReLU	Adam	100	16

Table 7 Relative error measures of the best LSTM variant from Monte Carlo cross validation

Time series	LSTM Variants	Fractional Bias (FB)	MRAE	RMSPE	MASE	RMSE	MAE	MAPE	R^2 (%)
Wheat	sLSTM	0.0008	0.230	9.174	0.038	4,163,645	3,043,640	6.83	98
	Bi-LSTM	0.0015	0.210	6.227	0.040	1,780,106	1,400,039	4.14	99
Groundnut	sLSTM	-0.0124	1.524	9.510	0.204	539,184	408,054	6.77	68
	Bi-LSTM	-0.0007	1.062	8.010	0.192	495,479	373,553	5.86	74
Barley	sLSTM	-0.0048	1.277	8.313	0.141	134,266	97,839	5.81	91
	Bi-LSTM	-0.0097	1.087	7.723	0.135	129,761	97,655	5.83	85

The optimized sLSTM model for groundnut had the characteristics of 68 neurons, ReLU as activation function with RMSprop as optimizer, batch size of 4, epoch of 300 and 0.35, 0.45 and 0.50 as dropout percentages (Table 6) with average MAPE of 6.77, MASE of 0.20, R-square of 68% and fractional bias of -0.0124 (Table 7). Similar configuration for best Bi-LSTM model for groundnut had 89 neurons, ReLU as activation function, Adam and Nadam as optimizer with two batch sizes of 4 and 8, epochs of 250 and 300 with 0.20, 0.35 and 0.50 as dropout percentages (Table 6) and average MAPE of 5.86, MASE of 0.19, R-square of 74% and fractional bias of -0.0007 (Table 7). The optimized sLSTM configuration for barley was found to have 57 neurons, ReLU as activation function with RMSprop as optimizer, batch size of 4, epochs of 200 and 300, dropout percentages of 0.40 and 0.50 (Table 6), with average errors such as MAPE of 5.81, MASE of 0.14, and R-square of 91% with negligible negative fractional bias of -0.0048 (Table 7).

Similar configuration for the best Bi-LSTM model for barley had 8 neurons, ReLU as activation function, Adam, Adamax and RMSprop as optimizers with two batch sizes of 16 and 32, epochs of 100 and 200, and 0.30, 0.40 as dropout percentages (Table 6) with average errors like MAPE of 5.83, MASE of 0.13, and R-square of 85% with negligible fractional bias of -0.0097 (Table 7). The overall analyses from Tables 6 and 7 showed that the performance of LSTM variants varied for different crops. It is noted from Table 8 that MAPE and relative errors were less in Bi-LSTM than sLSTM. Prediction skill for the winter crops (wheat and barley, which has trends even after normalizing) seems to be much better compared to groundnut (which is summer and Kharif crop) for both the LSTMs. Bi-LSTM performed better than sLSTM in predicting crop yield in most cases.

In order to find the best LSTM configuration for each crop from the optimized models, additional analyses were carried out with different sets of hyperparameters and dropout layers. Experiments were conducted by varying dropout percent from 0.2 to 0.5 with an interval of 0.05 with 4 different optimizers (Adam, Adamax, Nadam and RMSprop) for each crop time series data. Figure 5 describes the sensitivity of selection of optimizers and dropout on the performance of the shortlisted models. For wheat crop, sLSTM model simulations were carried out with 89 hidden layers, tanh activation function, 32 batch size and 300 epochs. Results (Fig. 5(a)) showed that sLSTM model performed better with Adam and Nadam as optimizers in which Adam performed consistently irrespective of dropout percentage while Nadam performed

Table 8 Performance comparison of DL models with ML models. Model with lowest error for each crop and for each error measure is highlighted as bold. The results of Wilcoxon rank test for different models is also shown (p-value)

Models	Time series	Error Measures					
		FB	RMSPE	MRAE	MAPE	MASE	*p*-value*
SVR	Wheat	-0.047	7.41	0.8	5.6	1.39	0.0043
	Groundnut	0.016	18.29	1.34	13.52	0.78	0.0084
	Barley	-0.069	12.61	**0.85**	9.71	0.45	0.0084
SVR polynomial 2nd order	Wheat	0.977	65.78	9.41	63.85	17.53	0.0021
	Groundnut	0.311	39.12	6.03	34.44	2.28	0.0707
	Barley	-0.446	69.25	6.08	61.27	2.97	0.0002
SVR polynomial 3rd order	Wheat	0.075	17.43	2.1	14.93	4.07	0.1701
	Groundnut	0.014	24.06	2.53	19.23	1.08	0.0222
	Barley	0.56	49.71	3.76	33.03	1.75	0.0063
ARIMA	Wheat	0.095	8.08	2.81	9.32	0.76	0.1780
	Groundnut	0.075	**4.69**	2.3	20.46	0.69	0.0034
	Barley	0.413	32.66	3.89	32.66	0.83	0.0654
ARIMAX	Wheat	-0.06	8.29	0.91	6.5	1.62	0.0056
	Groundnut	0.076	18.4	1.93	15.31	0.94	0.0453
	Barley	-0.188	23.8	2.07	21.22	1.05	0.0845
VAR	Wheat	-0.01	**4.14**	0.68	**3.06**	0.81	0.0033
	Groundnut	-0.011	22.27	1.95	16.87	0.89	0.0969
	Barley	0.313	30.51	5.11	27.96	1.79	0.0024
sLSTM	Wheat	**0.0008**	9.17	0.23	6.83	**0.03**	0.0005
	Groundnut	-0.012	9.51	1.52	6.77	0.2	0.0530
	Barley	**-0.0048**	8.31	1.27	**5.81**	0.14	0.0029
Bi—LSTM	Wheat	**0.001**	6.22	**0.21**	4.14	0.04	0.0003
	Groundnut	**-0.0007**	8.01	**1.06**	**5.86**	**0.19**	0.0421
	Barley	-0.0094	**7.72**	1.08	5.83	**0.13**	0.0026

lowest error for each crop and for each error measure is already highlighted as bold

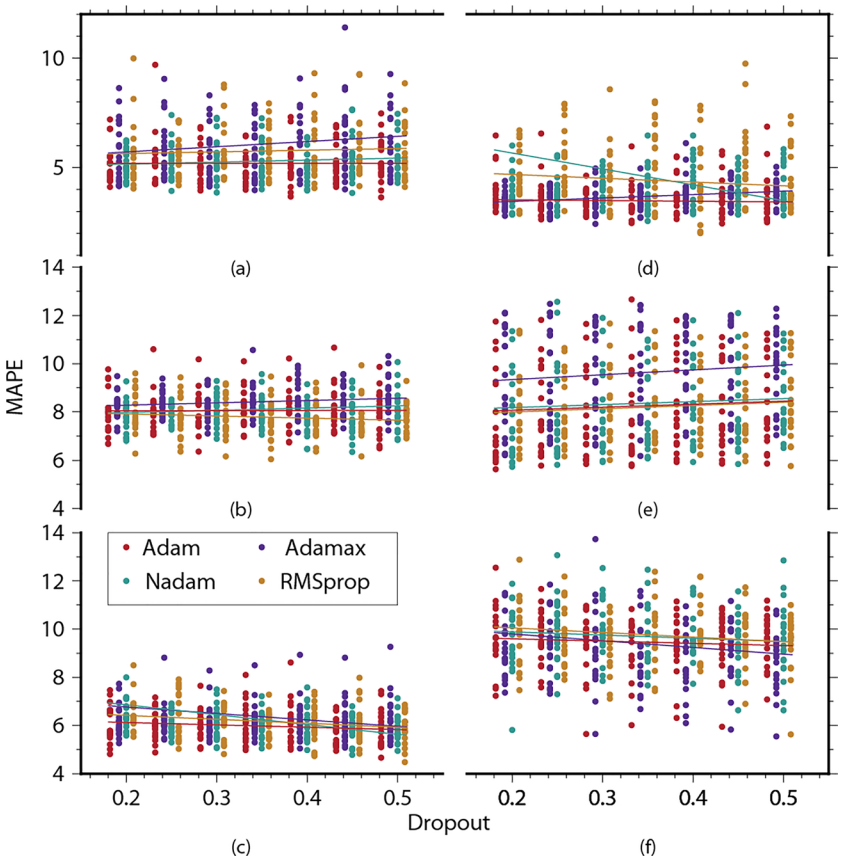

Fig. 5 Hyperparameters sensitivity of sLSTM (left panel) and Bi-LSTM (right panel) for different crop time series data prediction; (**a**) and (**d**) for wheat, (**b**) and (**e**) for groundnut, (**c**) and (**f**) for barely. MAPE values were plotted against dropout for different optimizers

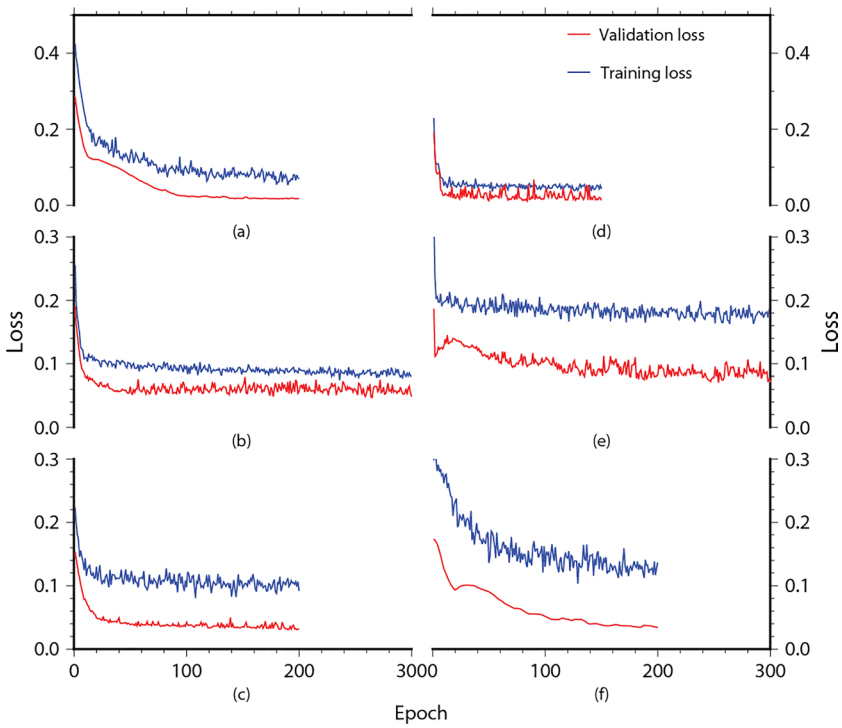

Fig. 6 Training and validation loss for the optimized sLSTM (**a**, **b**, **c**; left panel) and Bi-LSTM (**d**, **e**, **f**; right panel) model for in predicting wheat (top panel), barley (middle panel) and groundnut (bottom panel) production

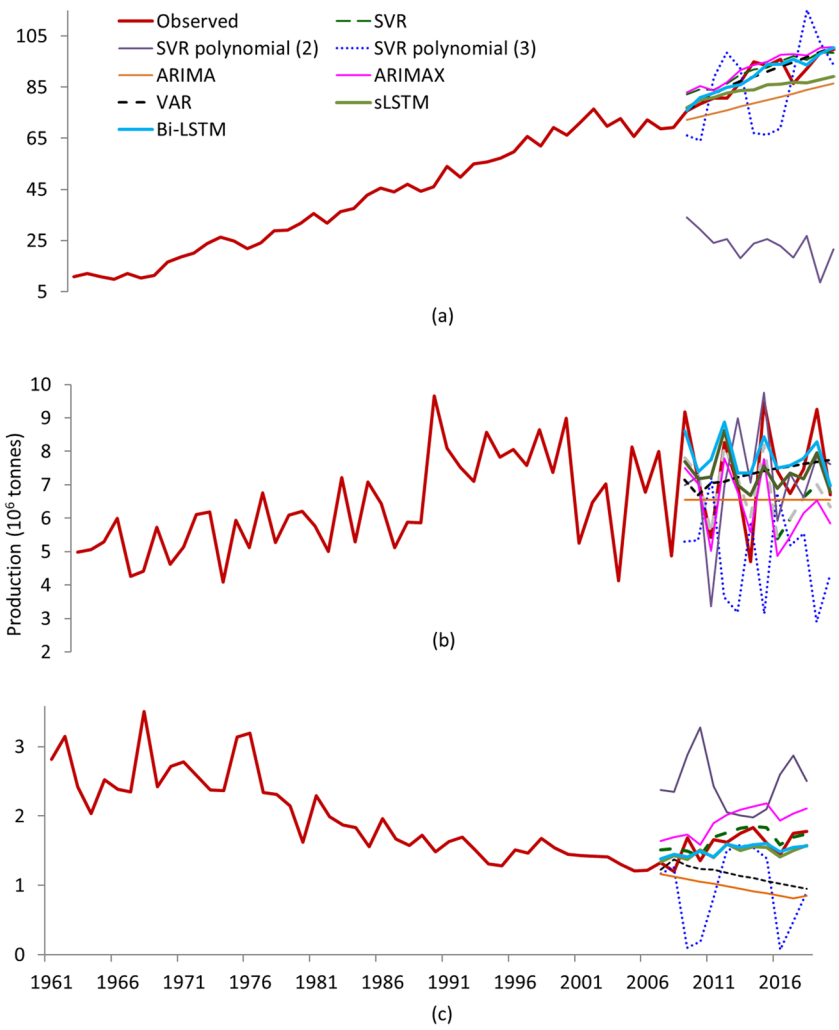

Fig. 7 Comparison of predicted annual crop production in India using DL and ML models with the observation; **a** observed and model predicted values for wheat, **b** observed and predicted values for groundnut and **c** observed and predicted values for barley

better only at higher dropout. Similarly, Bi-LSTM model simulations were carried out for wheat crop with 91 hidden layers, ReLU as activation function, batch size of 8 and with 250 epochs. Model error distribution (Fig. 5(d)) against dropout showed that though RMSprop performed better at high dropouts for selected models, Adam performed consistently better for a wide range of dropouts. Figure 5(b) describes the performance of sLSTM for groundnut crop with 68 hidden units, ReLU as activation function, and batch size of 4 with 300 epochs. As seen in Table 6, RMSprop optimizer is found to perform better consistently for different dropouts fallowed by Adam. Bi-LSTM simulation for groundnut with 89 hidden units, ReLU as activation function and batch size of 8 with 300 epochs showed that Adam and Nadam optimizers performed better at lower and higher dropout percentage respectively (Fig. 5(e)). Figure 5(c) describes the sLSTM performance for barley crop with 57 hidden layers, ReLU as activation function, and batch size of 4 with 250 epochs. RMSprop optimizer was found to be performing better at higher dropout (0.40 – 0.50) while Adam showed consistent performance for all dropout levels (Fig. 5(c)). Bi-LSTM simulation for barley with 8 hidden layers, ReLU as activation function, batch size of 32 with 200 epochs showed that Adamax optimizer performed better at higher dropout percentages and similar to sLSTM, Adam was found to perform consistently for different dropout percentages (Fig. 5(f)). In general, among the 4 optimizers, Adam was found to perform consistently for both the LSTM variants irrespective of the crops.

Figure 6 depicts fit of the optimized version of sLSTM and Bi-LSTM for training and validation datasets. For a well-trained model, performance should be comparable for both the training and validation datasets. Figure 6(a) and (d) explain the training and validation loss of sLSTM and Bi-LSaTM for wheat crop. Relatively lower validation loss compared to training loss and narrowing gap between training and validation loss with epoch indicated good model fit. Figure 6(b) and (e) describe fit of sLSTM and Bi-LSTM models for the barley crop yield prediction; higher gap between training and validation loss curve for Bi-LSTM indicated potential of further scope in model improvement.

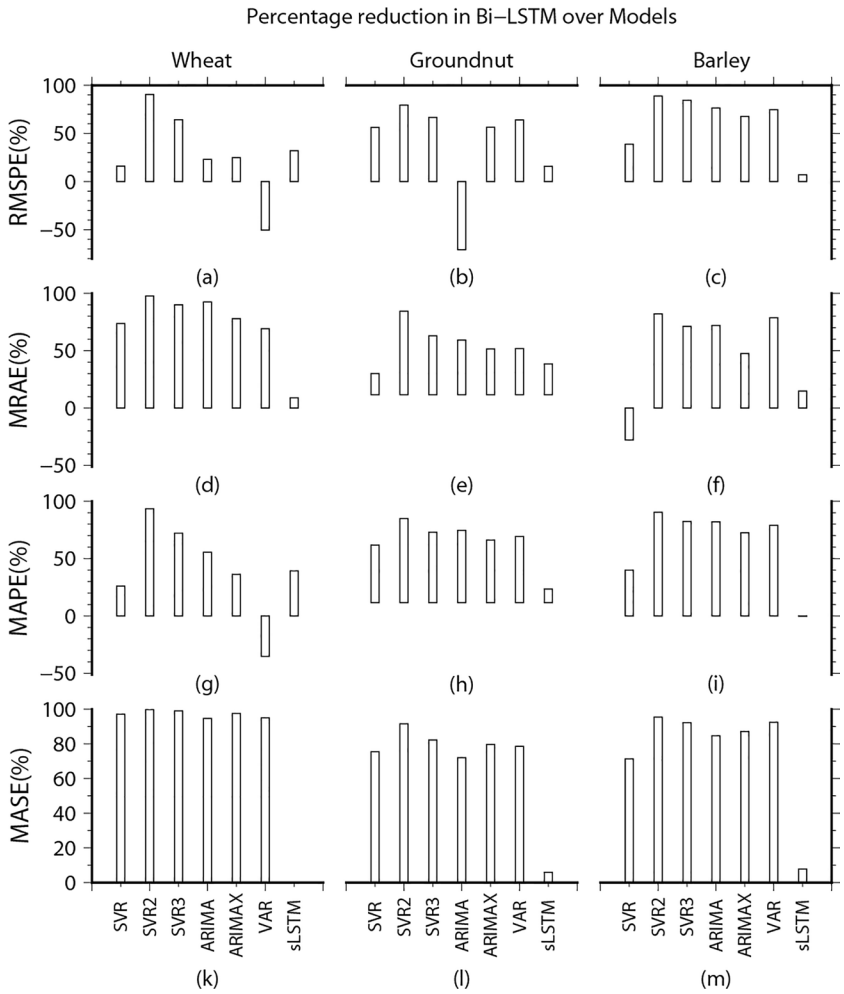

Fig. 8 Percentage reduction in average errors with Bi-LSTM compared to sLSTM and other ML models for predicting annual production in wheat (left panel; **a**, **d**, **g**), groundnut (middle panel; **b**, **e**, **h**) and barely (right panel; **c**, **f**, **i**)

Model fit for groundnut crop prediction (Fig. 6(c) and (f)) showed that validation loss is much lower for Bi-LSTM and it is decreased with epoch. In general, validation losses of models were much smaller and it followed training loss pattern indicating that neural network was not overtrained and provided a good fit to the distinct time series data.

4.2 Discussion

In this section, we compare the performance of best performed LSTM model in crop yield prediction with the state of the art ML based models. For this purpose simulations are carried out with ML based model with the same data sets used for training and validation of LSTM model. For brevity, the best performed LSTM model errors are only compared with that from traditional ML based models and percentage reduction in error with the use of LSTM model compared to ML based is computed.

Comparison of the performance of DL and ML models in predicting annual crop yield for the three distinct crops are shown in Fig. 7. Interannual variability in yield was observed highest in case of groundnut followed by barley and wheat. Observed (actual) and predicted data from different models for the validation period (2007–2018) clearly showed that LSTM models have clear edge over conventional statistical models in predicting annual crop yield for all the three crops (Fig. 7). It is also noted that Bi-LSTM model performance was superior to sLSTM for wheat and groundnut crops (Fig. 7(a) and (b)), while their performances were comparable for barley crop (Fig. 7(c)). In the case of groundnut and barley where interannual variability in yield is very high, ML based models particularly ARIMAX and SVR completely failed in simulating observed variability in yield (Figs. 7(b) and (c)). These results indicate limitations of ML based models in prediction when nonlinearity in time series data is very high.

Comparison of average errors of different models in predicting annual crop production is shown in Table 8. Bias in predicted values from the LSTM variants were near to zero (-0.0007 to 0.001) and much lower when compared to other

statistical models (Table 8) indicating that predicted values were closer to the true values. SVR ML model showed large RMSPE in yield prediction irrespective of crops. The MRAE was lowest (0.21 and 1.06) in Bi-LSTM model for predicting wheat and groundnut production respectively; however, the lowest MRAE in prediction for barley crop yield was shown by SVR model. MAPE is computed for intercomparison of the performance of different models and found that MAPE was least in DL models except for wheat crop. All the ML based models showed very high MAPE in predicting yield for barely and groundnut. It should be noted here that the interannual variability was much less for wheat crop compared to groundnut and barely. This clearly shows that ML based models has serious limitation in time series data prediction where inherent variability is very high and to a larger extent DL based models overcome this limitation. In order to intercompare errors from different model, MASE in prediction is computed and results showed that MASE was lowest in LSTM variants for all the three crops. Between the LSTM variants, Bi-LSTM showed less error compared to s-LSTM.

We have also conducted Wilcoxon test for comparing the performance of different models deployed in this study and to know how they ranked among themselves in predicting the yield compared to observed yield (Table 8). Wilcoxon rank test also confirmed that LSTM models showed much superior agreement with observation compared to ML based models; a feature clear from lower p values in LSTM models compared to ML based models. Figure 8 shows percentage reduction in various error measures with Bi-LSTM compared to sLSTM and other statistical models. Here, we computed how much the standard error is reduced with the use of BI-LSTM when compared to its LSTM counterpart s-LSTM and other ML based statistical models. Bi-SLTM performed better compared to s-LSTM; slight reduction in all the error measures is noted with Bi-LSTM compared to s-LTM for all the three crops (Fig. 8). When compared to ML based models, LSTM significantly reduced the error in yield prediction in terms of all the error measures (Fig. 8). On average, Bi-LSTM outperformed ML based models by reducing the model error measures such as MASE, MAPE, MRAE and RMSPE by 88.08%, 51.20%, 72.68%, 43.04% respectively for wheat crop (Fig. 8). Similarly, percentage reduction in MASE, MAPE, MRAE and RMSPE with Bi-LSTM for groundnut crop was 75.88%, 63.85%, 56.35% and 62.55% respectively while for barely crop error reduction was 69.37%, 60.14%, 48.04% and 58.50% respectively.

It should be noted that the computational time of deep learning models was many folds higher than that of statistical model and computational cost was proportional to complexity of the model. This generally puts a limitation in deploying complex deep neural networks with multiple layers in practical applications. This study demonstrated that by choosing proper explanatory (independent) variable and hyperparameter optimization, a simple (single layer) structure of deep neural network (LSTM) outperforms traditional ML models in terms of accuracy in crop yield prediction. However, as future studies, authors also have plan to investigate the performance of other variants of LSTM like (Stacked LSTM and Stacked Bi-LSTM) with complex neural network structure (such as multilayer) in similar prediction problems and compare their performance with the models used in this study. This study will be extended for many other crucial prediction problems to demonstrate applicability of the LSTM configuration used in this study.

5 Conclusion

This paper demonstrated application of DL techniques in prediction of crop yield for three crops (having distinct characteristics in their annual crop production) in India. This study presented an objective way of optimizing various hyperparameters in the DL based LSTM models. Two variants of LSTM were optimized for predicting crop yield namely, sLSTM and Bi-LSTM. The performance of trained LSTM models was evaluated against independent validation datasets which was not part of training data set. The performance of LSTM models were compared with performance of traditional ML models such as SVR, SVR polynomial, ARIMA, ARIMAX and VAR. The main conclusions derived from the present study are summarized below.

1. This study highlighted that pre-defined hyperparameters in DL models cannot be applied for prediction in different domain and a generic methodology was proposed to fine tune them for crop yield prediction.
2. The model accuracy is significantly improved with hyperparameter optimization and Monte-Carlo Cross-Validation
3. LSTM models performance was superior to ML models in crop yield prediction; Bi-LSTM algorithm improved the prediction by 88% on average compared to traditional ML models.
4. Between the LSTM variants, Bi-LSTM outperformed sLSTM with improvement in prediction accuracy of the order of 4–5%.

Acknowledgements The authors would like to thank National Mission on Himalayan studies for funding "Integrated system dynamical model to design and testing alternative intervention strategies for effective remediation and sustainable water management for two selected river basins of Indian Himalayas" and "Enhancement of the quality of livelihood opportunities and resilience for the people in the Indian

Himalayas, through design of intervention strategies aimed at maximizing resource potential and minimizing risks in urban-rural ecosystem (NMHS-2017/MG-04/480 and NMHS-2017-18/MG-02/478).

Author's contribution Kiran Kumar contributed in analyzing data and generating figures. Ramesh and Rakesh contributed in conceptualize, design, data analysis and drafting the manuscript.

Funding The research was not supported any funding other than institutional support from CSIR, INDIA.

Data availability Openly available.

Code availability Codes will be provided on reasonable request.

Declarations

Ethics approval Complied with Ethical Standards of Applied Intelligence Journal.

Consent to participate All listed authors have approved the manuscript before submission.

Consent for publication All authors agreed with the content and gave explicit consent to submit and that they obtained consent from the responsible authorities at the institute/organization where the work has been carried out, before the work is submitted.

Conflicts of interest None.

References

1. Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. John Wiley & Sons, USA
2. Armstrong JS (ed) (2001) Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic, Boston, MA, p 30
3. Hajirahimi Z, Khashei M (2019) Hybrid structures in time series modeling and forecasting: a review. Eng Appl Artif Intell 86:83–106
4. Wessel M, Quist-Wessel PF (2015) Auto-regressive integrated moving average (ARIMA) modeling of cocoa production in Nigeria: 1900–2025. J Crop Improv 33(4):445–455. https://doi.org/10.1080/15427528.2019.1610534
5. Wen Q, Wang Y, Zhang H, Li Z (2019) Application of ARIMA and SVM mixed model in agricultural management under the background of intellectual agriculture. Clust Comput 22(6):14349–14358
6. Verma U (2022) ARIMA and ARIMAX models for sugarcane yield forecasting in northern agro-climatic zone of Haryana. J Agrometeorol 24(2):200–202. https://doi.org/10.54386/jam.v24i2.1086
7. Mgaya JF (2019) Application of ARIMA models in forecasting livestock products consumption in Tanzania. Cogent Food Agric 5(1):1607430. https://doi.org/10.1080/23311932.2019.1607430
8. Sapankevych NI, Sankar R (2009) Time series prediction using support vector machines: a survey. IEEE Comput Intell Mag 4(2):24–38
9. Kok ZH, Shariff ARM, Alfatni MSM, Khairunniza-Bejo S (2021) Support vector machine in precision agriculture: a review. Comput Electron Agric 191:106546
10. Raj EE, Ramesh KV, Rajkumar R (2019) Modelling the impact of agrometeorological variables on regional tea yield variability in South Indian tea-growing regions: 1981–2015. Cogent Food Agric 5(1):1581457. https://doi.org/10.1080/23311932.2019.1581457
11. Umoh U, Asuquo D, Eyoh I, Abayomi A, Nyoho E, Vincent H (2022) A fuzzy-based support vector regression framework for crop yield prediction. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2020, Volume 1 (pp. 173–185). Springer Singapore
12. Parmezan ARS, Souza VM, Batista GE (2019) Evaluation of statistical and machine learning models for time series prediction: identifying the state-of-the-art and the best conditions for the use of each model. Inf Sci 484:302–337
13. Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and machine learning forecasting methods: concerns and ways forward. PLoS ONE 13(3):e0194889
14. Schmidt J, Marques MRG, Botti S et al (2019) Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater 5:83. https://doi.org/10.1038/s41524-019-0221-0
15. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V (2020) Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina 56(9):455
16. Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, Bochtis D (2021) Machine learning in agriculture: a comprehensive updated review. Sensors 21(11):3758
17. Thai TH, Omari RA, Barkusky D, Bellingrath-Kimura SD (2020) Statistical analysis versus the M5P machine learning algorithm to analyze the yield of winter wheat in a long-term fertilizer experiment. Agronomy 10(11):1779
18. Meshram V, Patil K, Meshram V, Hanchate D, Ramkteke SD (2021) Machine learning in agriculture domain: a state-of-art survey. Artif Intell Life Sci 1:100010
19. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H (2018) State-of-the-art in artificial neural network applications: a survey. Heliyon 4(11):e00938
20. Khan T, Jiangtao Q, Muhammad AAQ, Muhammad SI, Rashid M, Waqar H (2020) Agricultural fruit prediction using deep neural networks. Procedia Computer Science 174:72–78
21. Van Klompenburg T, Kassahun A, Catal C (2020) Crop yield prediction using machine learning: a systematic literature review. Comput Electron Agric 177:105709
22. Akbar A, Kuanar A, Patnaik J, Mishra A, Nayak S (2018) Application of artificial neural network modeling for optimization and prediction of essential oil yield in turmeric (Curcuma longa L.). Comput Electron Agric 148:160–178
23. Srivastava AK, Safaei N, Khaki S et al (2022) Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. Sci Rep 12:3215. https://doi.org/10.1038/s41598-022-06249-w
24. Ji B, Sun Y, Yang S, Wan J (2007) Artificial neural networks for rice yield prediction in mountainous regions. J Agric Sci 145(3):249–261
25. O'Neal MR, Engel BA, Ess DR, Frankenberger JR (2002) AE—Automation and emerging technologies: neural network prediction of maize yield using alternative data coding algorithms. Biosys Eng 83(1):31–45
26. Wolanin A, Mateo-García G, Camps-Valls G, Gómez-Chova L, Meroni M, Duveiller G, Guanter L (2020) Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. Environ Res Lett 15(2):024019
27. Yan W (2012) Toward automatic time-series forecasting using neural networks. IEEE Tran Neural Netw Learn Syst 23(7):1028–1039
28. Zheng C, Wang S, Liu Y, Liu C, Xie W, Fang C, Liu S (2019) A novel equivalent model of active distribution networks based on LSTM. IEEE Trans Neural Netw Learn Syst 30(9):2611–2624

29. Ergen T, Kozat SS (2017) Efficient online learning algorithms based on LSTM neural networks. IEEE Trans Neural Netw Learn Syst 29(8):3772–3783
30. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
31. Huawei Technologies Co., Ltd.. (2023) Overview of Deep Learning. In: Artificial Intelligence Technology. Springer, Singapore. https://doi.org/10.1007/978-981-19-2879-6_3
32. Wu Y, Yuan M, Dong S, Lin L, Liu Y (2018) Remaining useful life estimation of engineered systems using vanilla LSTM neural networks. Neurocomputing 275:167–179
33. Jiang Z, Liu C, Ganapathysubramanian B, Hayes DJ, Sarkar S (2020) Predicting county-scale maize yields with publicly available data. Sci Rep 10(1):1–12
34. Sathya P, Gnanasekaran P (2023) Paddy yield prediction in Tamilnadu Delta Region using MLR-LSTM model. Appl Artif Intell 37(1)
35. Crisóstomo de Castro Filho H, Abílio de Carvalho Júnior O, Ferreira de Carvalho OL, Pozzobon de Bem P, dos Santos de Moura R, Olino de Albuquerque A, Trancoso Gomes RA (2020) Rice crop detection using LSTM, Bi-LSTM, and machine learning models from sentinel-1 time series. Remote Sensing 12(16):2655
36. Ramesh KV, Rakesh V, Rao EVS (2020) Application of big data analytics and artificial intelligence in agronomic research. Indian J Agron 65(4):383–395
37. Tian H, Wang P, Tansey K, Zhang J, Zhang S, Li H (2021) An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong plain, PR China. Agric Forest Meteorol 310:108629
38. Yin J, Deng Z, Ines AV, Wu J, Rasu E (2020) Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (bi-LSTM). Agric Water Manag 242:106386
39. Nishu B, Anshu S (2021) Deep learning based wheat crop yield prediction model in Punjab Region of North India. Appl Artif Intell 35(15):1304–1328. https://doi.org/10.1080/08839514.2021.1976091
40. Maharana K, Mondal S, Nemade B (2022) A review: data preprocessing and data augmentation techniques. Global Transit Proc 3(1):91–99. https://doi.org/10.1016/j.gltp.2022.04.020
41. Salmerón R, García CB, García J (2018) Variance inflation factor and condition number in multiple linear regression. J Stat Comput Simul 88(12):2365–2384
42. Van Houdt G, Mosquera C, Nápoles G (2020) A review on the long short-term memory model. Artif Intell Rev 53(8):5929–5955
43. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688
44. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
45. Annamalai N, Johnson A (2023) Analysis and forecasting of area under cultivation of Rice in India: univariate time series approach. SN Comput Sci 4:193. https://doi.org/10.1007/s42979-022-01604-0
46. Anggraeni W, Andri KB, Mahananto F (2017) The performance of ARIMAX model and vector autoregressive (VAR) model in forecasting strategic commodity price in Indonesia. Procedia Comput Sci 124:189–196
47. Holtz-Eakin D, Newey W, Rosen HS (1988) Estimating vector autoregressions with panel data. Econometrica J Econ Soci 56(6):1371–1395
48. Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J (2017) LSTM: A search space odyssey. IEEE Trans Neural Netw Learn Syst 28(10):2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924
49. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Proces 45:2673–268142
50. Graves A, Schmidhuber J (2005) Frame wise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5–6):602–610
51. Yadav A, Jha CK, Sharan A (2020) Optimizing LSTM for time series prediction in Indian stock market. Procedia Comput Sci 167:2091–2100
52. Ghimire S, Yaseen ZM, Farooque AA et al (2021) Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. Sci Rep 11:17497. https://doi.org/10.1038/s41598-021-96751-4
53. Sheela KG, Deepa SN (2013) Review on methods to fix number of hidden neurons in neural networks. Math Probl Eng 2013:425740. https://doi.org/10.1155/2013/425740
54. Kandel I, Castelli M (2020) The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT Express 6(4):312–315
55. Baldi P, Sadowski P (2014) The dropout learning algorithm. Artif Intell 210:78–122
56. Verma P, Tripathi V, Pant B (2021) Comparison of different optimizers implemented on the deep learning architectures for COVID-19 classification. Mater Today Proc 46:11098–11102
57. Farzad A, Mashayekhi H, Hassanpour H (2019) A comparative performance analysis of different activation functions in LSTM networks for classification. Neural Comput Appl 31(7):2507–2521
58. Xu QS, Liang YZ (2001) Monte Carlo cross validation. Chemom Intell Lab Syst 56(1):1–11

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

V. Kiran Kumar received the B.Sc. degree in Agricultural Marketing and Co-operation from the College of Agriculture under University of Agricultural Sciences Bengaluru, Karnataka, India, in 2016, and M.Sc. degree in Agricultural statistics from College of Agriculture under University of Agricultural Sciences Dharwad, Karnataka, India, in 2018. He was working as Senior Research Fellow with Council of Scientific and Industrial Research – Fourth Paradigm Institute, Bengaluru, Karnataka, India. His research interests include time series analysis, machine learning and deep learning.

K. V. Ramesh received his B.Sc. (1995) in Physics in Madurai Kamaraj University, Madurai, India, M.Sc. (1997) in Physics and Post graduate diploma in computer application (1996) from Bharathidasan University, Trichy, India, MTech (1999) in Andhra University, India, and Ph.D. degree in Physics from the University of Pune, India, in 2005. He is a Senior Principal Scientist at Council of Scientific and Industrial Research-Fourth Paradigm Institute (CSIR-4PI), Bengaluru, India, since 2004 and Professor in Academy of Scientific and Innovative Research (AcSIR), India. His research interests are primarily system dynamical modelling, Mathematical modelling, machine and deep learning in different sectors like agriculture, water resources, climate and weather informatics, renewable energy, environmental impact assessment, remote sensing-based image analytics, and crop modelling. He has published more than 30 peer-reviewed research papers. He was a recipient of INSA Young Scientist medal and CSIR Young Scientist Award 2009.

V. Rakesh received his graduation from Kerala University, Kerala, India in Physics in the year 2002, master's from Cochin University of Science and Technology, Kerala, India in Meteorology in the year 2004, and Ph.D. degree in science from the Gujarat University, India, in 2009 for the research work carried out at Space Applications Center, Indian Space Research Organization, Ahmedabad, India. He is a Senior Principal Scientist at Council of Scientific and Industrial Research Fourth Paradigm Institute (Erstwhile CSIR Centre for Mathematical Modelling and Computer Simulation (C-MMACS)), India since 2009 and Professor in Academy of Scientific and Innovative Research (AcSIR), India. His research interests are primarily in atmospheric modeling, data assimilation, machine and deep learning techniques in different sectors like agriculture, water resources, climate and weather informatics, remote sensing-based image analytics, and crop modelling. He has published more than 40 peer-reviewed research papers. He is a recipient of CSIR Young Scientist Award 2016.