International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Classification and Estimation of Crop Yield Prediction in Karnataka using LSTM with Attention Mechanism

Nandini Geddlehally Renukaradya*[1], Kishore Gopala Rao[2], Anand Babu Jayachandra[3]

Submitted: 28/01/2024 **Revised:** 06/03/2024 **Accepted :** 14/03/2024

Abstract*:* Agriculture is an important occupation across the world with the dependency on the weather and rainfall. The objective of this paper is an early prediction of crop yield by using the climate, soil, and temperature factors. In this research, the classification-based crop yield prediction is proposed by using the Long Short-Term Memory (LSTM) with Attention Mechanism. The manual data is collected from the Economics and Statistics, Government of Karnataka department. This method utilized the dataset from the Department of Economics and Statistics of three crops named jowar, paddy, and ragi. The linear interpolation method is utilized for filling the missing and null values in the dataset. The feature selection process helps in the Correlation based Feature Selection Algorithm (CBFA) and Variance Inflation Factor Algorithm (VIF) for selecting and removing correlated feature sets. The model performance is evaluated by using Accuracy, R2, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The proposed LSTM model delivers the results through evaluation metrics such as accuracy, R2, MAE, MSE, and RMSE values about 98.23%, 0.43, 0.131, 0.054 and 0.232 respectively.

Keywords: *Attention mechanism, Correlation-based feature selection algorithm, Feature selection, Crop yield prediction, Long short-term memory, Variance inflation factor*

1. Introduction

Agriculture is the most important thing in the development of the Indian economy sector. Globally, many countries still facing the problem of huge food supply chain management demand due to the rapidly enhancing population. In this generation, the production of essential food crops has been integrated with agriculture [1]. Jowar, paddy and ragi are the most substantial food crops and second place in production. In the Indian economy, the jowar is the primary sector and it is cultivated in both rainy and after rainy seasons. The crop yield production helps the farmers to make better decisions about the appropriate time to cultivate crops based on environmental factors to produce efficient yield. The parameters such as climate, soil, temperature, biological, geographical, and other factors affect crop yield production [2, 3]. The crop yield prediction is the most arduous task in every stage for decision-makers of the farmers at the local and global levels. Prediction of crop yields is valuable to many stakeholders including agronomists, traders, farmers and policy makers [4]. Larger crop yield with a low field area makes it difficult to accomplish the objective. The farmers are analyzing a preferable yield production, depending on the collection of the agricultural data and developing the crop yield estimating strategies to enhance the rural insights and agronomy [5].

The crop yield prediction model helps the farmers to make better decisions about the appropriate time to cultivate the crops and what types of crops to cultivate based on environmental factors to produce better yield.

Precision agriculture is a recent approach compared to the method of traditional cultivation as well and it minimizes the farmer's time and economic cost [6]. Machine Learning (ML) can learn automatically from past experiences by continuously training and providing better prediction and classification results [7]. The modern data-based modeling approach has been applied in different agricultural fields to get a benefit over the last ten years for more accurate predictions, efficiency and relevant features [8]. ML image classifiers are utilized to classify the affected or diseased crops from the healthy crops. The predictive model is developed using various features such as model parameters determined using historical data in the training stage [9]. The ML algorithms such as Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF) are both parametric and non-parametric in nature as well as strongly dominating the crop yield prediction. In certain, Artificial Neural Network (ANN) is used for the identification and classification of crop yield prediction problems by seeing various factors such as CO2 fixation,

[1] *Department of Information Science and Engineering, Sri Siddhartha Institute of Technology, SSAHE, Tumakuru, and Visvesvaraya Technological University, Belagavi-590018, India*
[2] *Department of Information Science and Engineering, Jyothy Institute of Technology, Kanakapura, and Visvesvaraya Technological University, Belagavi-590018, India*
[3] *Department of Information Science and Engineering, Malnad College of Engineering, Hassan, and Visvesvaraya Technological University, Belagavi-590018, India*
** Corresponding Author Email: nandiniaradya@gmail.com*

solar radiation and water content [10]. In this study, Long Short-Term Memory (LSTM) is proposed for early prediction of crop yield. The major contributions of the proposed method are as follows:

- The linear interpolation method is used for data pre-processing because the dataset in the Karnataka region contains missing values.
- The correlation-based Feature Selection Algorithm (CBFA) is for selecting the most correlated feature set and it is used for filling in the missing values the variance Inflation Factor Algorithm (VIF) for removing whole multicollinearity independent features is used in the feature selection process.
- The Long Short-Term Memory (LSTM) with attention mechanism is utilized for the classification of better crop yield prediction compared with the three crops such as jowar, paddy and ragi and the performance of this model is evaluated using Accuracy, R2, MAE, MSE, RMSE and MAPE.

The rest of the paper is organized as follows: Section 2 provides a literature survey. Section 3 provides a detailed description of the proposed method. Section 4 provides experimental results and finally, Section 5 provides the conclusion.

2. Literature Survey

Gowda and Reddy [11] presented an ML approach for predicting the best crop yield prediction in particular agricultural regions. The suggested approach analyzed several climatic factors such as humidity, rainfall and temperature. The ML used the three main approaches Polynomial Regression, Decision Tree and Random Forest for performing the best yield production. The advantage of this proposed model was providing the best yield prediction, minimizing the farmers' loss face and increasing the economic capital. The limitation of this model was it required a longer time to train the three algorithms of the model.

Moraye [12] presented various machine-learning approaches for predicting the yield of the crop by using the web application of smart agriculture. The Random Forest (RF) approach with five climatic parameters was used to train the model to obtain higher accuracy. This method used the input as pest, chemical, and soil type factors. The advantage of this model is that it achieved the best accuracy and prediction result by using the technique of 10-fold cross-validation and increases the yield economy and marketing.

Kale and Patil [13] presented a prediction of the variety of crop yields by using Artificial Neural Network (ANN) regression modeling. The ANN uses the 3 neural network layers for a variety of crop yield predictions to improve the prediction accuracy by increasing neural network layers and the parameters. This model used the dataset from the website of the Indian Government with 2,40,000 records. This proposed model achieves the best accuracy result and time-consuming but, the model requires a greater number of parameters.

Jadhav and Monisha [14] presented a jowar and wheat yield prediction from the same season and the input parameters namely area, production, season, and crop. Ada-boost and random forest algorithms are used in ensemble techniques for developing and improving the yield prediction accuracy of jowar. The advantage of this model was solving the problems of both the classification and regression and using any model to improve the performance. The limitation of this model was the lower accuracy result due to the combination of both Ada-boost and random forest algorithms.

Israni [15] presented different machine-learning approaches for predicting the best prediction of crop yield. This model used different techniques such as Ridge Regression, LGBM Classifier, and XGB Regressor with hyperparameter tuning for the prediction of the crop to yield better accuracy results. The advantage of this model was XGB Regressor with the hyperparameter tuning gives the best accuracy result and optimal solution. The limitation of this model was causing the overfitting and starting modeling the noise.

Gopal and Bhargavi [16] presented a hybrid method of Multiple Linear Regression (MLR) as well as Artificial Neural Network (ANN) for efficient prediction of crop yield. The MLR coefficients and intercepts were included in the input layer of ANN weights and bias initialization to identify the lowest optimal error and improve the accuracy results. The advantage of the MLR-ANN model achieves the best accuracy result when compared to the other conventional models. The limitation of this model was a failure in the continuous prediction of outcomes and time-consuming.

Shidnal [17] presented an architecture of a multi-tier machine learning approach for prediction of crop yield. In the first level, the suggested approach was to identify the nutrient deficiency of the paddy crop by utilizing the neural network. In the next level, the k-means clustering approach was utilized to quantify the intensity of the corresponding yield value. The advantage of the multi-tier model achieves the best accuracy result by using tensor flow. The limitation of that model was time complexity, installation cost, and unscalable.

Gupta and Nahar [18] developed the hybrid ML approach based IoT for crop yield prediction. Correlation based Feature Selection (CBFS) and Variance Inflation factor were utilized for the feature selection process. The suggested approach utilized the two ML approaches, initially, the Adaptive K-nearest Centroid Neighbour Classifier was used and finally, the Extreme Learning Machine Approach (ELM) was used for the classification of various classes according to the input soil parameters.

The limitation of this model was causing the overfitting and starting modeling the noise.

The limitations found in the related works are time complexity, failure in continuous processing, requiring a large number of parameters, and overfitting. These limitations cause the LSTM model to deliver inappropriate results. To overcome this, a deep learning model – LSTM with an attention mechanism is proposed for crop yield prediction.

3. Proposed Method

In this proposed methodology, the dataset is collected from the yield production of three major crops such as jowar, paddy and ragi in the region of Karnataka. Then, classify the better yield prediction by comparing it with the three crops. This framework includes the major processes such as dataset, pre-processing, feature selection and LSTM. By using these methods, the classification-based better crop yield prediction is effectively performed with the attention mechanism. Fig. 1 shows the flowchart for the proposed crop yield prediction.

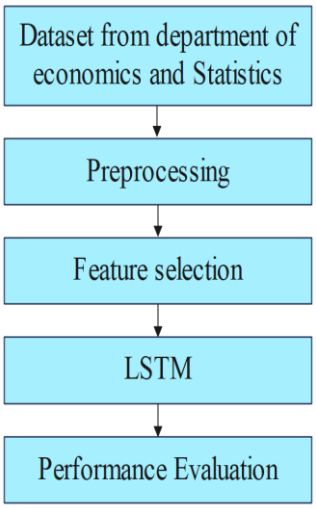

Fig. 1. Proposed Diagram for crop yield prediction

3.1. Dataset

Jowar, paddy and ragi are the most important cultivation crops in the Karnataka region. The dataset is collected from Economics and Statistics, Government of Karnataka department.

- **Jowar**

80% of the jowar productions come from the Karnataka districts such as Vijayapura, Koppal, Raichur, Belagavi, Kalaburagi, Belgaum, Bidar, Chitradurga, Bidar, Dharwar, Gulbarga from 2011 to 2021 with the around 18 million hectares and the average annual production of 8-10 million tonnes.

- **Paddy**

The major paddy-growing districts in Karnataka such as Raichur, Koppal, Ballari, Haveri, Uttar Kannada, Dharwad, Mysore, Hassan and Chitradurga in the period of 2020-2021 with around 7.86 lakh hectares and the average annual production of 23.73 lakh tonnes.

- **Ragi**

In the Karnataka region, the ragi yielded production of 13lakh tonnes from 2020 to 2021 and 50% of the total production comes from the market. Tumakuru district is the largest production, which is followed by Ramnagar, Bengaluru rural, Hassan, Mandya, Kolar, Chikballapur, Shivamogga, Chikkamagaluru, Chamarajnagar and Davanagere districts.

3.2. Pre-processing

After collecting the datasets, the pre-processing of data is an important stage whereas Deep Learning [19] does not manage noisy data such as outliers and errors. The pre-processing of data is done before classification, because some districts in Karnataka contain missing values, and null values, removing the unwanted data, and the appropriate range of data maintained in the production row [20]. In that row, the values can be replaced by the mean values and the dataset contains string values, which should be replaced by the numerical conversion for the process of splitting the train and testing the data. Linear Interpolation [21, 22] method is used for filling those missing values and null values. This method is an arithmetic process that determines the new data points with the existing data in a straight line in the same increasing order as the previous value. Whenever have time-series data, then to deal with missing values, Interpolation is a major method used for filling the missing values in data of time-series. The function of the linear interpolation Eq. (1) is as follows,

$$f(X) = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x - x_0) \qquad (1)$$

Where x is the independent variable, x_0 and x_1 are the independent variable known values and $f(x)$ is the dependent variable value for the independent variable value x.

3.3. Feature Selection

The feature selection is processed after the data pre-processing is done. The high-level feature selection plays a major role in obtaining the best accurate forecasting results, where datasets have a greater number of attributes. The feature selection is majorly used because of the ML approach to faster training, reduction of complexity of the model and it can make it easy to interpret the process. This can increase the model accuracy due to right subset selection as well as preventing overfitting. There are three feature selection methods utilized in attribute selection including wrapper, filter and embedded. According to these, the methods of wrapper and filter are initially used to select the best attributes. The wrapper method normally performs better than the filter method, however, the model

is highly expensive in computation. The embedded method contains the filter and wrapper method and uses the selection process of the attribute of their own. These can be used to obtain the best attribute from the original dataset. In this feature selection, the CBFA and VIF algorithm are utilized in filter-based feature selection method.

3.3.1. Correlation based Feature Selection Algorithm (CBFA)

The Correlation based Feature Selection Algorithm [23] (CBFA) is the most effective method, that selects a correlated feature set, that is majorly integrated with the crop yield prediction and is placed based on the evaluation function of the correlation heuristic. The evaluation of CBFA is close to the subset that has the features, which has the greater correlation against classes and uncorrelated together. The CBFA function aims to identify a subset of features that exhibit high correlations among themselves while remaining uncorrelated with each other.

The CBFS can be calculated in Eq. (2) as follows,

$$M = \frac{Nr\bar{c}}{\sqrt{N+N(N-1)r\bar{f}}} \quad (2)$$

Where, N – the sum of total features

$r\bar{c}$ – average correlation

$r\bar{f}$ – average feature.

There are different types of correlation used for the selection of features. In this method, the Pearson correlation approach was utilized to extract the most correlated features in the regression of similarity to the crop yield. It is the measure of two variables covariance division and the multiply of standard deviations. Pearsons's correlation is mathematically represented by Eq. (3) as follows;

$$\rho_{A,B} = \frac{con\ (A,B)}{\sigma_A\ \sigma_B} \quad (3)$$

Where, $\rho_{A,B}$ – coefficient of Pearson correlation between A and B features

$con\ (A,B)$ – covariance of A and B

$\sigma_A\ \sigma_B$ – A and B feature standard deviation.

In the CBFA method, the crop yield data is classified into two phases such as training and testing phases. In the pre-processing step, the CBFA selects the best feature set and which is mainly integrated with the crop yield.

3.3.2. Variance Inflation Factor Algorithm (VIF)

The variance Inflation Factor (VIF) algorithm verifies the multicollinearity in the independent features. Thus, the VIF can remove the whole multicollinearity-independent features. The VIF can evaluate the multicollinearity strength in the regression analysis of least squares. The VIF model can be utilized to eliminate the correlated features, which is quick and it accomplished the one-pass search to the predictor. This approach is arithmetically efficient in testing and eliminates overfitting problems. The model is accomplished due every independent variable regression. Permit Y on rest independent W and Z variables and confirming how many of Y is described by those parameters and it represented in Eq. (4) as;

$$Y = a_1x_1 + a_2\ x_2 + \ldots + a_ix_i + l \quad (4)$$

Where, $a_1, a_2 \ldots a_i$ - coefficient of regression and l – intercept. The predictors are included in linear relationships amidst, standard errors for various individual partial regression coefficients are extremely filled. The VIF formula can be calculated in Eq. (5) as follows,

$$V = \frac{1}{1-R^2} \quad (5)$$

Where, R^2 – variable X is collinear with the variables of Y and Z. The multicollinearity occurs when the value of VIF is more than 10.

3.4. Long Short-Term Memory (LSTM)

The hyperparameters applied during the training process were 11 hidden layers and 50 neurons in each layer.

Long Short-Term Memory (LSTM) is the most advanced model out there to forecast time series as well as classify the better crop yield prediction. The LSTM is utilized to classify which crop produces the better yield production by comparing it with the three crop's yield prediction. During the model training with the Deep Learning (DL) algorithms, there are a greater number of hyperparameters such as number of neurons, hidden layers, and the learning rate can be considered. Manually, the parameter setting is not always supportable. Hyperparameter optimization is the process of improving the model performance by selecting a correct combination of hyperparameters. In LSTM, hyperparameter optimization variables can be considered as the time steps, number of input and hidden layers as well as number of hidden neurons. The hyperparameter tuning may affect the model performance, so it can be done carefully. In the deep LSTM model, the layers will cause overfitting and slow convergence and the neurons of the hidden layer work similarly to the layer of LSTM. For forecasting the time-series data, past historical data are needed but the traditional neural network will consider only the present input data. RNN and LSTM can hold historical data, but compared with RNN LSTM models hold the previous for a long time.

LSTM contains two basic concepts, that are used to learning of temporal features from the data. The first thing is the memory concept, which introduced the cell state and the other is the cell concept, which can effectively train the fully connected layers. In LSTM, have different memory cells in the hidden layer of read, write, and delete operations, which are facilitated by three gates such as input, output, and forget gate and these three gates will determine the data, it is needs to be stored in memory. The cell state passes information from one layer to another. The

first stage is the forget gate which allows only the necessary data to pass through the cell state. The first stage in the input gate is the sigmoid layer that manages the value of output and the second stage is the Tanh layer, which develops the vectors of new feature values, both are stored in the cell state. Updated cell information is the output in the output gate. The LSTM deals with the previous historical data and current unknown patterns are analyzed by regulating fundamentally, to achieve patterns and this generates future predictions earlier. Fig. 2 shows a representation of LSTM functionality

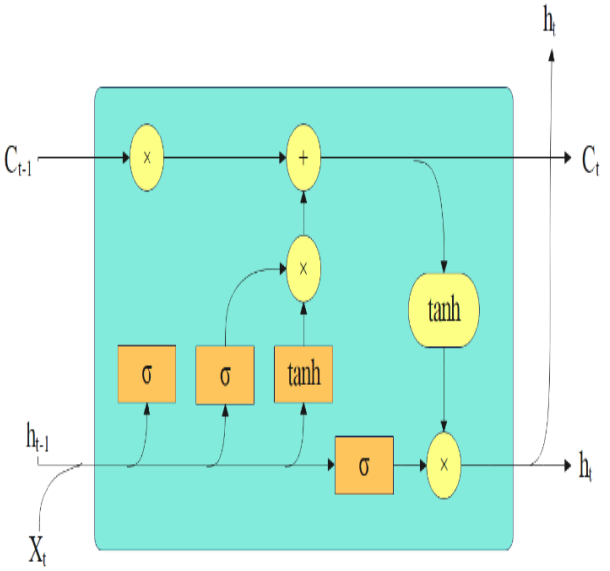

Fig. 2. Representation of LSTM functionality

h_{t-1} – previous memory output
C_t – current memory output.
LSTM cell is described in Eq. (6) as:

$$cg_t = Tanh(wt_{Cg} \times [hd_{cg-1}, x_{Cg}] + bs_{cg} \quad (6)$$

where, (cg_t) – current memory
(wt_{Cg}) – weight matrix
(bs_{Cg}) - bias

- The input gate controls the current memory input data update to the value of the memory cell and it is calculated in Eq. (7) as:

$$ig_t = \sigma(wt_{ig} \times [hd_{ig-1}, x_{ig}] + bs_{ig} \quad (7)$$

- The input gate controls the previous memory data update to the value of the memory cell and it is calculated in Eqs. (8) and (9) as:

$$fg_t = w_{tf} wt_{fg} \times [hd_{fg-1}, x_{fg}] + bs_{fg} \quad (8)$$

$$cu_t = fit \times lc_{t-1} + cg_t \quad (9)$$

Where, cu_t – current memory cell
lc_{t-1} – last LSTM cell value.

LSTM can also be worked in both the stacked and bidirectional form. In stacked, initially, LSTM operates on the input and subsequent LSTM, and then operate on the outputs of the temporal features, which is produced by the preceding models. Stacked LSTM used in higher level temporal learning features. The LSTM of bidirectional train the additional model compared to the unidirectional LSTM. A LSTM can read the input data from the sequence start to the end $(t_0 \rightarrow t_n)$ and at same time, the other read the input from end to start $(t_n \rightarrow t_0)$. Then these two models are combined to perform the temporal feature output. Bidirectional used to learn the model features from both sides of the input sequences. Further, the LSTM model includes the Attention Mechanism for improving the prediction accuracy. The LSTM works as an attention mechanism. The state is updated the decoder in every stage, it will examine all states of the encoder again. It is used in the training process; the input progression is encoded as the last time step hidden state. The Recurrent Neural Network (RNN) works as a long-term memory, then provides the attention mechanism and the output regression is updated more accurately. Therefore, the hidden state includes the input sequence complete information. This classification method provides powerful tools for accurately classifying the crop yield prediction and the performance and accuracy of the classified results are evaluated and tested in the following section.

4. Experimental Results

In this study, the proposed method is replicated using LSTM with the system requirements. The experimental result was evaluated by using various metrics such as Accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-Squared (R2). The mathematical expression for each metric is described in the Equation as follows.

Accuracy: Corrected prediction to the total number of predictions. This can be calculated by the below Eq. (10) as;

$$Accuracy = \frac{correct\ prediction}{Total\ number\ of\ prediction} \quad (10)$$

Mean Absolute Error: Mean of the absolute difference between actual values and the predicted values. This can be calculated by the below Eq. (11) as;

$$MAE = \sqrt{\frac{1}{m}} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (11)$$

Mean Squared Error: Measure average of squared error of the predicted value and the actual value. This is always a positive and a risk function. This can be calculated by the below Eq. (12) as;

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (12)$$

Root Mean Square Error: Difference between the predicted values using the estimator and observed values. This can be calculated by the below Eq. (13) as;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (13)$$

Mean Absolute Percentage Error: Absolute difference between a quantity observed value and the true value. This can be calculated by the below Eq. (14) as;

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (14)$$

R-Squared: Estimate the proposed model variance over the total variance and difference between observed and the predicted value. This can be calculated by the below Eq. (15) as;

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (15)$$

Where, n - number of points,
y_i - predicted value obtained from the neural network,
$\hat{y}_i$ - the real value and
$\bar{y}$- mean of the real value.

4.1. Quantitative and Qualitative Analysis

This section shows the quantitative and qualitative analysis of LSTM model in terms of achievable sum rate. Table 1 and Table 2 shows the experimental result of the LSTM model with various deep learning models.

Table 1. Accuracy results of the proposed method with existing methods

Methods	Accuracy (%)
CNN	83.78
DNN	86.39
RNN	89.56
GAN	91.67
LSTM with Attention Mechanism	98.23

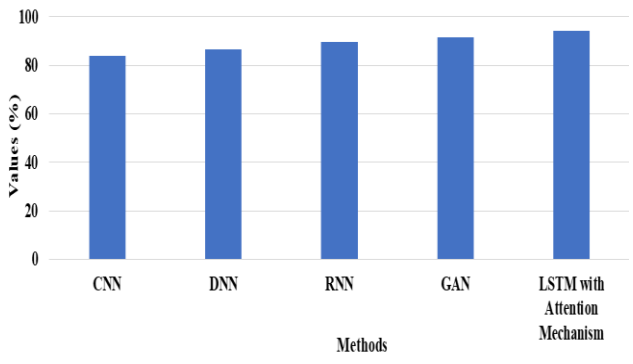

Fig. 3. Graphical representation of accuracy results of the proposed method with existing method

Table 1 and Fig. 3 show the accuracy results of the proposed method with existing methods. The existing methods such as Convolutional Neural Network (CNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Generative Adversarial Network (GAN) are compared with the proposed LSTM with an attention mechanism. The obtained accuracy results show that 83.78% of CNN, 86.39% of DNN, 89.56% of RNN, and 91.67% of GAN. The proposed LSTM with Attention mechanism achieved a better accuracy result of 98.23% compared with the other existing methods.

Table 2. Experimental results of the proposed method with existing methods

Methods	R2	MAE	MSE	RMSE
CNN	0.49	0.139	0.062	0.248
DNN	0.47	0.136	0.059	0.242
RNN	0.46	0.135	0.058	0.240
GAN	0.44	0.132	0.055	0.234
LSTM with Attention Mechanism	0.43	0.131	0.054	0.232

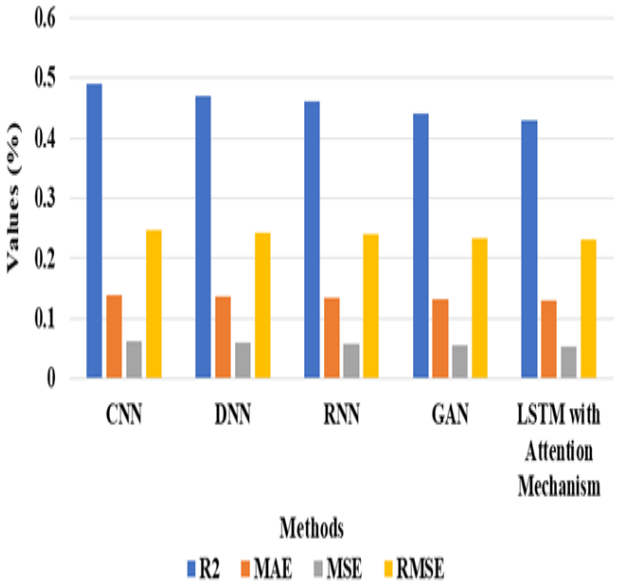

Fig. 4. Graphical representation of experimental results of proposed method with existing methods

Table 2 and Fig. 4 represent the experimental results of the proposed method with existing methods. Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Generative Adversarial Network (GAN) are measured and compared with the proposed LSTM with Attention mechanism. The obtained result shows that the proposed method achieves better results by using performance metrics like R2, MAE, MSE, and RMSE values of about 0.43, 0.131, 0.054 and 0.232.

4.2. Comparative Analysis

This section shows the comparative analysis of the LSTM with attention mechanism in terms of achievable sum rate. The existing research such as [16] and [18] is used for evaluating the efficiency of this model. The comparison results of the LSTM model and the existing model are represented in Table 3. The LSTM with attention mechanism achieved good performance results compared to existing comparative models.

Table 3. Comparative Analysis of proposed method with existing methods

Author	Method	Accuracy (%)	R2	MAE	RMSE
Gowda and Reddy [11]	RF	88	N/A	N/A	N/A
Jadhav and Monisha [14]	Ensemble method	74.3	N/A	N/A	N/A
Gopal and Bhargavi [16]	MLR-ANN	N/A	0.99	0.041	0.051
Gupta and Nahar [18]	aKCN-ELM-MOBA	N/A	0.81	0.064	0.301
Proposed LSTM with Attention Mechanism	LSTM with Attention Mechanism	98.23	0.43	0.131	0.232

4.3. Discussion

In this section, the advantages of the proposed method and the limitations of existing methods are discussed. The existing method has some limitations such as RF [11] requires larger time to train the three algorithms of the model. The ensemble method [14] obtained the less accurate result due to the combination of both Ada-boost and random forest algorithm. The MLR-ANN [16] had failed in the continuous prediction of outcomes and was time-consuming. The aKCN-ELM-MOBA [18] had caused the overfitting and started modeling the noise. The proposed LSTM with Attention Mechanism method outperforms these existing model limitations. The LSTM is much better at handling long-term dependencies.

5. Conclusion

The prediction of crop yield is an important role in the world food production. The effective prediction of the crop yield provides major support for the farmer and it increases economic value and food productivity. Due to climatic changes, crop yield production is affected due to factors such as rainfall, climate, soil, and temperature. This proposed methodology collects the yield production of three major crops in the region of Karnataka. Then, classify the better crop yield prediction by comparing it with the three crop's yield prediction. The proposed model uses the linear interpolation method to fill the missing values in the dataset. The feature selection process used the CBFA method to select the most correlated feature set and it was used for filling the missing values and the VIF method for removing the whole multicollinearity independent features. From the performance analysis, the proposed method achieves better results by using performance metrics like accuracy, R2, MAE, and RMSE values of about 98.23%, 0.43, 0.131 and 0.232 which are comparatively better than the existing methods. In future work, the proposed method will extend to utilize the number of soil parameters to improve the prediction results.

Author contributions

Nandini Geddlehally Renukaradya: Conceptualization, Methodology, Software, Field study, Writing-Original draft preparation, **Kishore Gopala Rao:** Data curation, Software, Validation, Writing-Reviewing and Editing, Field study **Anand Babu Jayachandra:** Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] A. Suruliandi, G. Mariammal, and S. P. Raja, "Crop prediction based on soil and environmental characteristics using feature selection techniques," *Math. Comput. Modell. Dyn. Syst.*, vol. 27, no. 1, pp. 117–140, Jan. 2021.

[2] D. Paudel, H. Boogaard, A. de Wit, M. van der Velde, M. Claverie, L. Nisini, S. Janssen, S. Osinga, and I. N. Athanasiadis, "Machine learning for regional crop yield forecasting in Europe," *Field Crops Res.*, vol. 276, p. 108377, Feb. 2022.

[3] S. Iniyan and R. Jebakumar, "Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (MSER)," *Wireless Pers. Commun.*, vol. 126, no. 3, pp. 1935–1964, Oct. 2022.

[4] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, "Machine learning for large-scale crop yield forecasting," *Agric. Syst.*, vol. 187, p. 103016, Feb. 2021.

[5] D. Elavarasan and P. M. D. R. Vincent, "Fuzzy deep learning-based crop yield prediction model for sustainable agronomical frameworks," *Neural Comput. Appl.*, vol. 33, no. 20, pp. 13205–13224, Oct. 2021.

[6] F. Abbas, H. Afzaal, A. A. Farooque, and S. Tang, "Crop yield prediction through proximal sensing and machine learning algorithms," *Agronomy*, vol. 10, no.

7, p. 1046, Jul. 2020.

[7] S. Fei, M. A. Hassan, Y. Xiao, X. Su, Z. Chen, Q. Cheng, F. Duan, R. Chen, and Y. Ma, "UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat," *Precis. Agric.*, vol. 24, no. 1, pp. 187–212, Feb. 2023.

[8] P. Nevavuori, N. Narra, P. Linna, and T. Lipping, "Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models," *Remote Sens.*, vol. 12, no. 23, p. 4000, Dec. 2020.

[9] M. P. S. Gopal and R. Bhargavi, "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms", *Applied Artificial Intelligence*, vol. 33, no. 7, pp. 621–642, Jun. 2019.

[10] U. Bhimavarapu, G. Battineni, and N. Chintalapudi, "Improved Optimization Algorithm in LSTM to Predict Crop Yield," *Computers*, vol. 12, no. 1, p. 10, Jan. 2023.

[11] S. Gowda and S. Reddy, "Design and implementation of crop yield prediction model in agriculture," *International Journal of Scientific & Technology Research*, vol. 8, no. 1, pp. 544–549, Jan. 2020.

[12] K. Moraye, A. Pavate, S. Nikam, and S. Thakkar, "Crop Yield Prediction Using Random Forest Algorithm for Major Cities in Maharashtra State," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, vol. 9, no. 2, pp. 40–44, Mar. 2021.

[13] S. S. Kale and P. S. Patil, "A Machine Learning Approach to Predict Crop Yield and Success Rate," in *2019 IEEE Pune Section International Conference (PuneCon)*, Pune, India, 2019, pp. 1–5.

[14] V. Jadhav and Monisha, "Wheat and Jowar Crop Yield Prediction Model using Ensemble Technique," in *Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021)*, 2021.

[15] D. Israni, K. Masalia, T. Khasgiwal, T. Khasgiwal, M. Tolani, and M. Edinburgh, "Crop-Yield Prediction and Crop Recommendation System," 2022. Available at SSRN 4111856.

[16] P. S. M. Gopal and R. Bhargavi, "A novel approach for efficient crop yield prediction," *Comput. Electron. Agric.*, vol. 165, p. 104968, Oct. 2019.

[17] S. Shidnal, M. V. Latte, and A. Kapoor, "Crop yield prediction: two-tiered machine learning model approach," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 1983–1991, Oct. 2021.

[18] A. Gupta and P. Nahar, "Classification and yield prediction in smart agriculture system using IoT," *J. Ambient Intell. Hum. Comput.*, vol. 14, no. 8, pp. 10235–10244, Aug. 2023.

[19] E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," *Environ. Dev. Sustainability*, vol. 22, no. 6, pp. 5687–5708, Aug. 2020.

[20] A. Kaneko, T. Kennedy, L. Mei, C. Sintek, M. Burke, S. Ermon, and D. Lobell, "Deep learning for crop yield prediction in Africa," in *International Conference on Machine Learning AI for Social Good Workshop*, Long Beach, United States, 2019.

[21] O. N. Oyelade, A. E. -S. Ezugwu, and H. Chiroma, "CovFrameNet: An Enhanced Deep Learning Framework for COVID-19 Detection," *IEEE Access*, vol. 9, pp. 77905–77919, 2021.

[22] A. Picornell, J. Oteros, R. Ruiz-Mata, M. Recio, M.M. Trigo, M. Martínez-Bracero, B. Lara, A. Serrano-García, C. Galán, H. García-Mozo, P. Alcázar, R. Pérez-Badia, B. Cabezudo, J. Romero-Morte, and J. Rojo, "Methods for interpolating missing data in aerobiological databases," *Environ. Res.*, vol. 200, p. 111391, Sep. 2021.

[23] F. Pereira, H. Lopes, P. Maia, B. Meyer, J. Nocon, P. Jouhten, D. Konstantinidis, E. Kafkia, M. Rocha, P. Kötter, and I. Rocha, "Model-guided development of an evolutionarily stable yeast chassis," *Mol. Syst. Biol.*, vol. 17, no. 7, p. e10253, Jul. 2021.