MAUSAM, **77**, 3 (July 2026), 1059-1068

MAUSAM

DOI: https://doi.org/10.54302/mausam.v77i3.6782
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No.633.51:631.559:004.85(540.15)

Machine learning models to forecast cotton yield for Punjab

OPINDER KAUR[1*], MOHAMMED JAVED[1], CHETAN SINGLA[2] and GURJEET SINGH WALIA[3]

[1]*Department of Maths, Stats and Physics, Punjab Agricultural University, Ludhiana -141004*

[2]*Department of Soil and Water Engineering, Punjab Agricultural University, Ludhiana -141004*

[3]*Department of Statistics, Central University of Odisha Koraput, Odisha*

(*Received 01 August 2024, Accepted 15 April 2025*)

***Corresponding author's email: opinder664@gmail.com**

सार – कृषि में सटीक और समयपूर्व पूर्वानुमान सतत खेती और कृषि-क्षेत्र प्रबंधन को अनुकूलित करने के लिए अत्यंत आवश्यक हैं। फसल की पैदावार का पूर्वानुमान फसल के मौसम के दौरान फसल बीमा, भंडारण की मांग और अन्य महत्वपूर्ण कारकों के संबंध में किसानों के निर्णयों को महत्वपूर्ण रूप से प्रभावित करता है। फसल की पैदावार में अरेखीयता (Non-linearity) के कारण, आजकल पूर्वानुमान के उद्देश्यों के लिए अरेखीय मॉडलों का उपयोग लोकप्रिय हो गया है। इस शोध पत्र में, भारत के पंजाब क्षेत्र के लिए कपास की पैदावार का अनुमान लगाने हेतु मौसम के मापदंडों का उपयोग करके ANN और LSTM मॉडलों को प्रशिक्षित (Train) किया गया था। न्यूनतम त्रुटि के साथ पैदावार का पूर्वानुमान लगाना एक मुख्य चुनौती है। प्रच्छन्न परतों (Hidden layers) में ReLU सक्रियण फलन (Activation function) से युक्त ANN (10 10 10 1) मॉडल ने न्यूनतम MSE (0.0182) के साथ अन्य पूर्वानुमान मॉडलों की तुलना में बेहतर प्रदर्शन किया। न्यूरल नेटवर्क (NN) मॉडल के उपयोग द्वारा किए गए विश्लेषण से यह निष्कर्ष निकला कि मौसम के मापदंडों ने पौधों की वृद्धि को प्रभावित करने में महत्वपूर्ण भूमिका निभाई है। ये कारक पैदावार को उल्लेखनीय रूप से बढ़ा या घटा सकते हैं। संवेदनशीलता विश्लेषण (Sensitivity analysis) से पता चला कि सापेक्ष आर्द्रता (Relative humidity) सबसे महत्वपूर्ण मौसम मापदंड था, जिसके बाद वर्षा का स्थान था।

ABSTRACT. Accurate and early predictions in agriculture are essential for sustainable farming and optimizing field management. Crop yield prediction significantly impacts the farmer's decisions on crop insurance, storage demand and other important factors during the growing season. Due to non-linearity in crop yield, the use of non-linear models for forecasting purposes has become popular these days. In this paper, ANN and LSTM models were trained using weather parameters to forecast the cotton yield for Punjab, India. Predicting the yield with minimum error is a main challenge. ANN (10 10 10 1) model with ReLU activation function in hidden layers performed better than other forecasting models with a minimum MSE (0.0182). The analysis using the NN model concluded that the weather parameters played an important role in affecting the plant growth. These variables may enhance or reduce the yield significantly. Sensitivity analysis showed that relative humidity was the most important weather parameter followed by rainfall.

Key words – Crop yield, Forecasting, Weather variables, ANN, LSTM and MSE.

1. Introduction

Cotton, also termed as "White Gold", is one of the most essential commercial and widely grown cash crops that contribute significantly to employment, industry and the Indian economy. It provides direct livelihood to around 6 million farmers. Around 40-50 million people are indirectly engaged in cotton related activities (Meshram and Dange, 2022). Cotton significantly contributes to India's foreign exchange through the exports of raw cotton, yarn, fabric, garments and cottonseed oil. With 22.24% of the world's total cotton consumption, India is the second largest consumer of cotton. It is also the third largest exporter of textiles and apparel globally. Gujarat, Maharashtra and Telangana are the top three cotton producing states of India. Cotton yield in these states increases with time due to advancements in agriculture. Punjab used to be one of the top contributing

states in producing cotton. Now, Punjab is facing a continuous decline in the ranking and has become the ninth state in the production of cotton. There are several factors that are responsible for the slow growth in the production of cotton in Punjab. The cotton crop yield data shows that the yield follows a non-linear trend. Cotton cultivation reached its peak in 2006-07 and plummeted to its lowest point in 2015-16 covering 3,35,000 hectares and yielding an average of 196 kg per hectare (Dhillon and Pathak, 2020). The crop yield non-linearly depends upon different weather variables. These variables affect the yield positively or negatively according to the fluctuations in them. To predict the crop yield, there is a need to accurately map the yield using these weather variables. This mapping is done using ANN and LSTM modelling.

Researchers have continued to pay attention to use non-linear machine learning models in order to forecast the agricultural yield. Kaul, *et al*., 2005 analysed the efficiency of ANN models used to predict the corn and soybean yield in the Maryland region for typical climate conditions at the state, regional and local levels. The ANN models were compared with multiple linear regression models. The ANN models were considered a better method to predict the yield. Dahikar and Rode, 2014 developed an ANN architecture to forecast yield, in which the neural network was trained using the feed forward back propagation method. By using parameters related to soil and atmosphere, it was concluded that the ANN modelling gave better predictions.

Patel and Saha, 2020 made attempts using available weather data from 1901 to 2002 to forecast cotton yield using a time series model based on the moving average method and the neural network method. Then, using the weather parameters, forecasting of cotton yield is done. The results based on MAPE, MSE and RMSE revealed that the neural network method predicted the yield with minimum variation in actual and predicted values. Salehin, *et al*., 2020 used an artificially intelligent LSTM technique to develop monthly rainfall forecasting model for Bangladesh. The training was done on a detrended numerical dataset of different weather variables, whereas the output variable was categorical. An accuracy of 76% was achieved by the model when predicted rainfall data was compared with the actual data.

Sharma, *et al*., 2020 proposed a method to predict wheat yield at block level across several states in India. The proposed method involved deep CNN-LSTM network that worked directly on raw imagery data of satellites rather than any extracted hand-crafted features. It was found that additional information such as the location of farmlands and water bodies in the area helped in improving the yield estimates. Also, the proposed model outperformed over existing methods by 50%.

Pravallika, *et al*., 2021 studied different approaches like ANN, CNN, RNN and LSTM etc. to predict the yield of rice, millet and paddy. These models were trained on the values of different factors like genotype, rainfall, soil etc. The study concluded that for rice multilayer perceptron model outperformed with an accuracy of 97.5%. For the millet crop RF technique with an accuracy of 99.74% and for the paddy crop deep reinforcement learning technique with accuracy of 93.7% outperformed. Yildirim, *et al*., 2022 attempted to predict the cotton yield four months before harvest. The ANN model was trained using the limited data of 13 years related to four categories of input variables named meteorological data, drought indexes, vegetation indices and measured cotton yield for Menemen Plain, Turkey. It was concluded that the ANN model had the potential to predict the cotton yield with a MAPE value of less than 5%.

From different studies, it is found that cotton yield in Punjab follows a non- linear trend. Cotton yield prediction helps policy-makers, textile industry to make timely decisions about importing or exporting the crop. So, it is important to make such a model that predicts the yield accurately. The main objective of this research is to forecast the cotton yield with ANN and LSTM models using weather variables and the comparison between these models is also done on the basis of different performance statistics.

2. Data and methodology

2.1. *Study area and data collection*

The present study was conducted for Punjab, a state in northern India having coordinates 30.79ºN and 75.87ºE. The state covers an area of 50,362 square kilometres (19,445 square miles), which is 1.53% of India's total geographical area. The major cotton producing districts of Punjab are Bathinda, Fazilka, Muktsar, Abohar, Barnala, Sangrur and Faridkot. The cotton crop is affected by weather parameters like temperature, rainfall, windspeed and relative humidity etc. In the present study, the data from 1981-82 to 2020-21 was used. The secondary data of cotton yield (kg/ha) had been collected from the data published by the department of agriculture and farmers welfare (https://agri.punjab.gov.in). The rainfall IMD Grid data (in mm) was downloaded in CSV format from the India-Wris website (https://indiawris.gov.in/wris/#/). Nasa Power Data of maximum temperature (ºC) minimum temperature (ºC), relative humidity (%) and windspeed (m/s) was used to conduct this study (https://power.larc.nasa.gov/data-access-viewer/).

2.2. *Descriptive analysis*

The knowledge of weather parameters is needed as forecasting of cotton yield is aimed to done using weather parameters. Descriptive analysis of the data illustrates the properties of distribution of different data variables. Various descriptive measures like mean, median, mode and range etc. are calculated for cotton yield and the weather parameters of the study area.

To understand the linear relationship between the variables in the dataset, Pearson correlation coefficient (r) is used. The correlation value ranges from -1 to +1. The value nearest to +1 indicates a highly positive correlation between the variables and -1 indicates a high negative correlation. If its value is 0, then no correlation exists between them. If 'x' and 'y' are the variables for which the correlations are determined, the formula of the correlation coefficient is given as:

$$r = \frac{\sum_{i=1}^{n}(x_i-\overline{x}_i)(y_i-\overline{y}_i)}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x}_i)^2 \sum_{i=1}^{n}(y_i-\overline{y}_i)^2}} \quad (1)$$

There may exist a hidden non-linear pattern between the cotton yield and these weather variables. This non-linear pattern is used to forecast the cotton yield in this research using ANN and LSTM.

2.3. *Methodology*

An ANN is a computational model based on the structure and functioning of the human brain and biological neurons that simulates brain intelligence by mathematical equations, electronic circuits or software (Shao and Shen, 2022). The components of ANN are input layer, hidden layers, output layer, weights, biases and activation functions. All of these components work together to capture the important features of the data. The LSTM approach is an advanced version of Recurrent Neural Networks (RNN) that introduces long term memory components to effectively study and learn sequential data. In ANN, the output of each neuron is fed only to the next neuron. But, the RNN is a type of neural network where the output from the previous step is fed as input to the current step. It allows the network to maintain an internal state and simulate memory. In standard RNNs, there is a problem of vanishing and exploding gradients. Both problems limit the learning capacity of the model. To solve these problems efficiently, LSTM approach with a more complex model of units and a gated architecture is a very suitable choice.

Both ANN and LSTM approaches use supervised machine learning algorithms for forecasting. In a supervised machine learning algorithm, the model anticipates output based on well-labelled training input data with the appropriate output. The cotton training data is supplied to the algorithms as a supervisor who trains them to accurately predict the output. This method seeks to identify a mapping function between the input variable (x) and the output variable (y).

This neural network modelling is done using the Python programming language. The packages used in the study are Pandas, NumPy, Scikit-learn and TensorFlow. Due to cross-platform processing capability, open-source code, and extensive support libraries, Python is chosen for the work environment. The models based on the ANN and LSTM approach are built having different layers like input layer, hidden layers and output layers. The collected dataset is rescaled to avoid convergence problems and extremely small weighing factors using the given formula (Tayfur and Singh, 2006):

$$x_{\mathrm{ni}} = \frac{x_i\text{-}x_{\min}}{x_{\max}\text{-}x_{\min}} \quad (2)$$

where 'x_{ni}' is a normalized dimensionless variable, 'x_i' is the observed value of variable; 'x_{max}', 'x_{min}' are the maximum and minimum value of the observed variable. Whole data is divided into training and testing datasets in a ratio 80:20. Then the input layer of the model receives the pre-processed training data and the output layer produces the predicted crop yield. The next step is to train the models using the training data. During the training process, the model adjusts its weights to minimize the error between the predicted crop yield and the actual crop yield using optimization algorithms. Adam as an optimizer is used in the present research. After training, the performance of the models using the testing data is evaluated. Then, trained models are used to predict the future crop yield.

Different ANN and LSTM models with one to five hidden layers having different activation functions have developed. The optimal number of nodes is found using minimum MSE values for the testing data. Then these one to five layered ANN and LSTM models with an optimal number of nodes are compared on the basis of MSE of training & testing data and one model is selected for each activation function. The results of these selected models are analysed thoroughly. On the basis of different statistics, the model that gives better results is selected and used for forecasting the cotton yield.

2.4. *Statistical analysis*

The performance of the models is evaluated using mean absolute percentage error (MAPE), root mean

square percentage error (RMSPE), normalized root mean square percentage error (NRMSE), mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), adjusted R^2 ($\bar{R}^2$) and correlation between observed and predicted values (ρ). Smaller values of MAPE, RMSPE, NRMSE and MAE indicate higher predictive accuracy. The performance index NRMSE avoids bias towards underestimating and overestimating. It should be less than 0.3 for the better performance of the model. When 'y' is the actual yield and '$\hat{y}$' is the predicted yield, 'n' is the number of observations and 'k' is the number of independent variables, then:

$$\text{MAPE}=\frac{1}{n}\left(\sum\left|\frac{\text{y-}\hat{\text{y}}}{y}\right|\right)*100 \quad (3)$$

$$\text{RMSPE}=\sqrt{\frac{1}{n}\sum\left(\left|\frac{\text{y-}\hat{\text{y}}}{y}\right|*100\right)^2} \quad (4)$$

$$\text{NRMSE}=\frac{1}{(\bar{y})}\sqrt{\frac{\sum(\text{y-}\hat{\text{y}})^2}{n}} \quad (5)$$

$$\text{MAE}=\frac{\sum|\text{y-}\hat{\text{y}}|}{\text{n}} \quad (6)$$

$$\text{RMSE}=\sqrt{\frac{\sum\left(\text{y-}\hat{\text{y}}\right)^2}{\text{n}}} \quad (7)$$

$$\text{R}^2=1-\frac{\sum\left(\text{y-}\hat{\text{y}}\right)^2}{\sum\left(\text{y-}\bar{\text{y}}\right)^2} \quad (8)$$

$$\bar{\text{R}}^2=1-\left[\left(1-\text{R}^2\right)\frac{\text{n-1}}{\text{n-k-1}}\right] \quad (9)$$

$$\rho=\frac{\sum\left(\text{y-}\bar{\text{y}}\right)\left(\hat{\text{y}}-\bar{\hat{\text{y}}}\right)}{\sqrt{\sum\left(\text{y-}\bar{\text{y}}\right)^2\sum\left(\hat{\text{y}}-\bar{\hat{\text{y}}}\right)^2}} \quad (10)$$

To check the significance of R^2, F-test is used.

$$\text{F}=\left(\frac{R^2}{1-\text{R}^2}\right)*\left(\frac{\text{n-k-1}}{k}\right) \quad (11)$$

If $F > F_{k,n-k-1}(\alpha)$, then R^2 is significant at $(1-\alpha)*100\%$ level of significance.

2.5. *Sensitivity analysis*

To identify the most important input parameter, sensitivity analysis is performed. First, the MSE of the model containing all 5 variables is recorded. Then, each input variable is eliminated one by one and the network error is noted again after re-applying the training process. A new $MSE_{(i)}$ is calculated for all the 5 models. The ratio of new error to original error is defined as:

$$\text{W}=\frac{\text{MSE}_{(i)}}{\text{MSE}} \quad (12)$$

The larger the value of 'W' for the eliminated variable, the more sensitive the network is to the lack of that variable and this indicates that the eliminated variable is affecting the cotton yield significantly.

3. Results and discussion

3.1. *Descriptive analysis*

A complete description of dependent variable (cotton yield) and independent variables (rainfall, temperature, windspeed and relative humidity) is shown in Table 1. In the study period, high variation in the total yearly rainfall was observed that ranged from 413.52 mm to 1198.17 mm. Rainfall followed the distribution with a mean value 657.84 mm with a standard deviation of 163.89 mm. Similarly, relative humidity values of the study period followed the distribution with an average value 37.27% having a standard deviation 4.97%. The average yearly maximum windspeed data used to develop models ranged from 4.74 m/s to 5.91 m/s with an average value of 5.29 m/s and a standard deviation 0.32 m/s.

Table 1 shows that cotton yield values of the studied period followed a distribution whose values deviated from the average value (520.55 kg/ha) with the standard deviation 175.7 kg/ha having a high range of 647 kg/ha. The mean was more than the median and the kurtosis had a negative value.

The correlation analysis between cotton yield and other variables is given special attention. The observed correlation between the variables is presented in Fig. 1. Cotton yield showed a non-significant negative linear correlation with maximum windspeed, maximum temperature and rainfall whereas it showed a low positive correlation with relative humidity and minimum temperature indicated that crop yield might be non-linearly correlated with these weather variables.

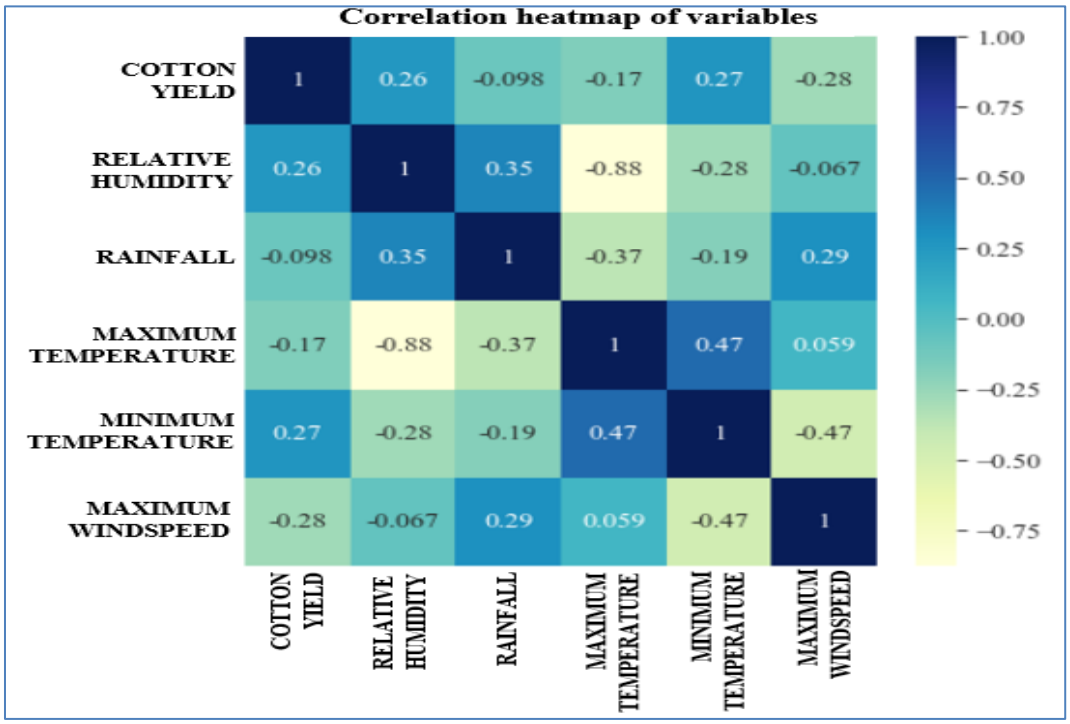

Fig. 1. Correlation heatmap of variables

TABLE 1

Different measures of variables used in the study

Parameter	Cotton (kg/ha)	Max. Temp. (^{0}C)	Min. Temp. (^{0}C)	Rainfall (mm)	Relative humidity (%)	Max. Windspeed (m/s)
Mean	520.55	37.45	14.38	657.84	37.27	5.29
Median	520.00	37.46	14.22	641.52	37.65	5.39
Maximum value	827.00	39.14	15.92	1198.17	50.61	5.91
Minimum value	180.00	35.92	13.06	413.52	28.05	4.74
Range	647.00	3.22	2.86	784.65	22.56	1.17
Standard deviation	175.70	0.85	0.68	163.89	4.97	0.32
Skewness	-0.22	0.17	0.28	1.12	0.21	-0.06
Kurtosis	-0.71	-0.75	-0.36	1.57	0.12	0.92

3.2. *ANN models for Punjab*

ANN models with one to five hidden layers were considered. The optimal number of nodes was found using the least MSE values of testing data. Activation functions, namely ReLU and tanh, were used to develop different models.

For the ReLU activation function in hidden layers, the performances of different ANN models having different numbers of hidden layers with an optimal number of nodes were observed. ANN (**10 10 10 1**) performed better with MSE values 0.0123 and 0.0418 for training and testing data of cotton yield. For the whole study period, MSE value was 0.0182. When performance

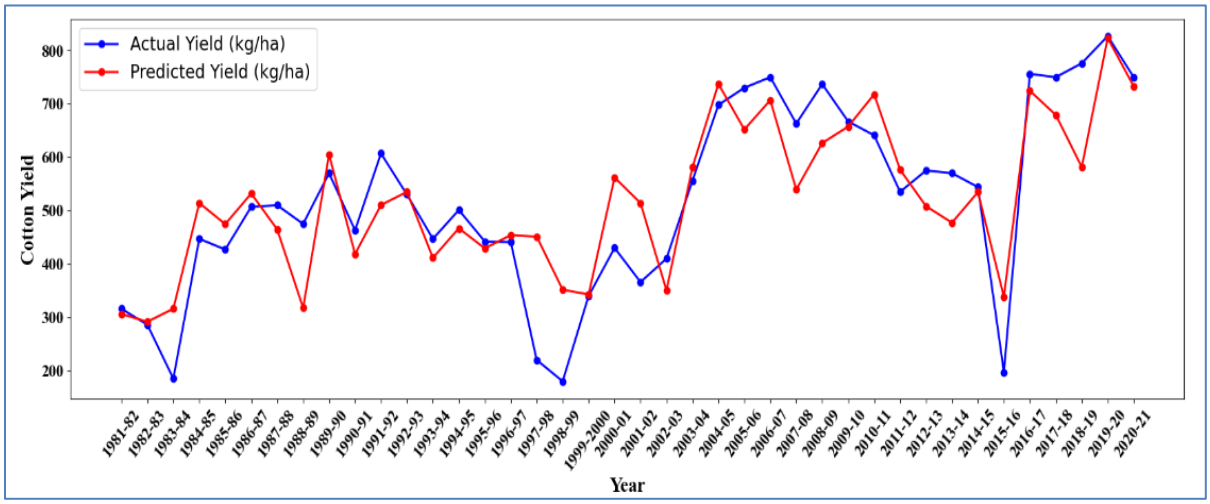

Fig. 2. Cotton yield forecasted by ANN (10 10 10 1) model

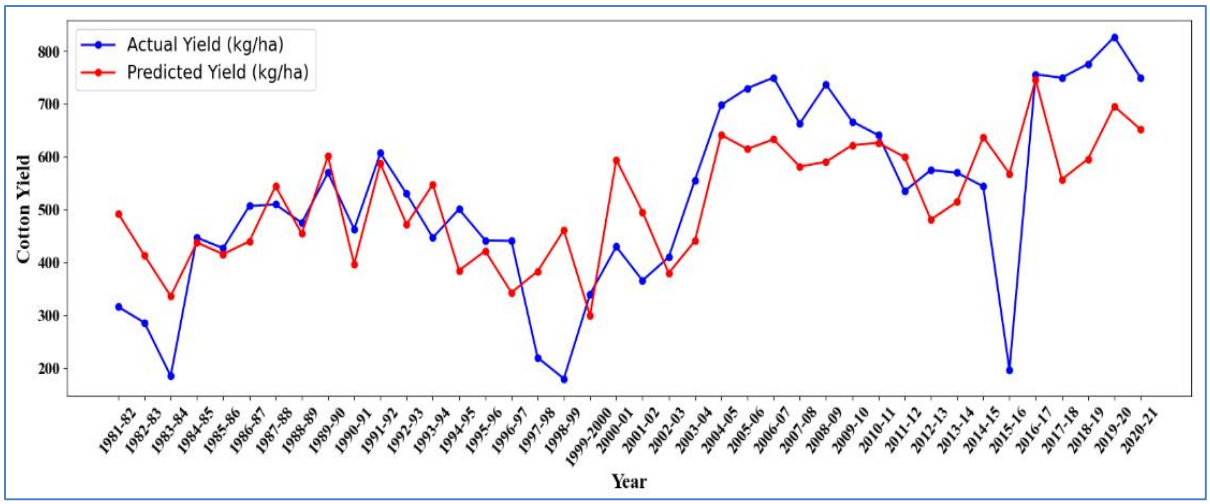

Fig. 3. Cotton yield forecasted by ANN (10 9 9 1) model

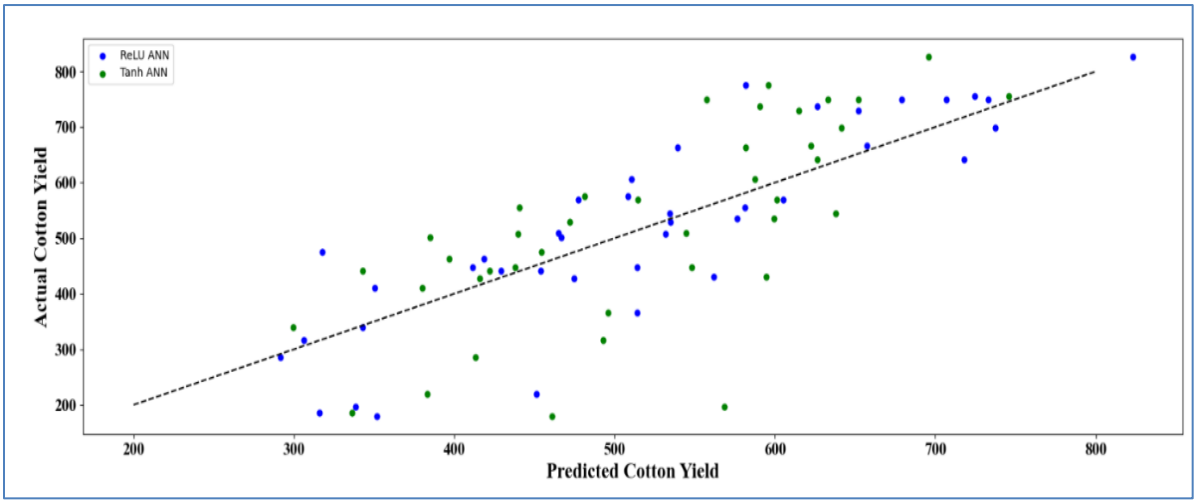

Fig. 4. Scatter plot for observed and predicted values of cotton yield for ANN models

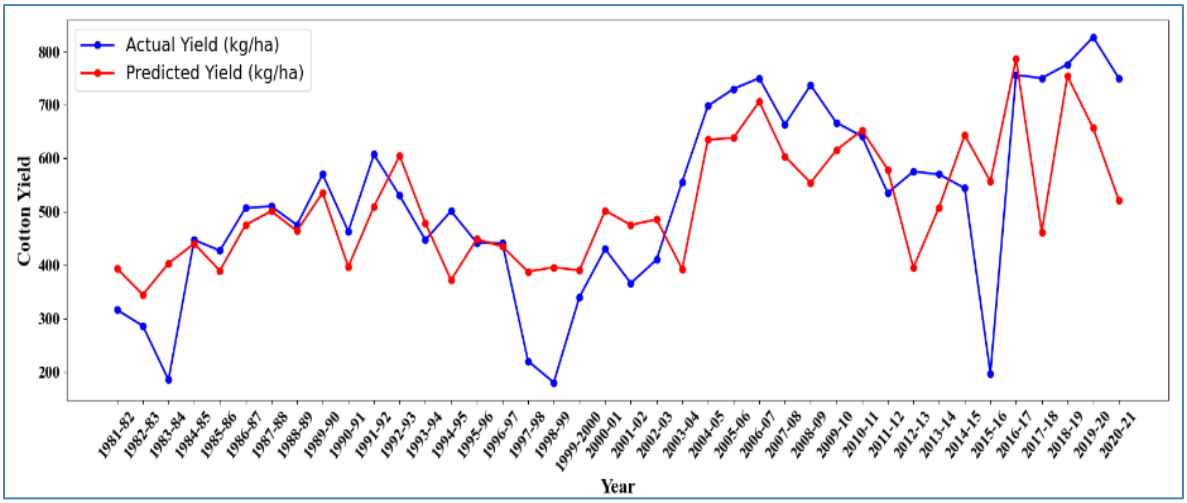

Fig. 5. Cotton yield forecasted by LSTM (10 4 4 4 4 4 1) model

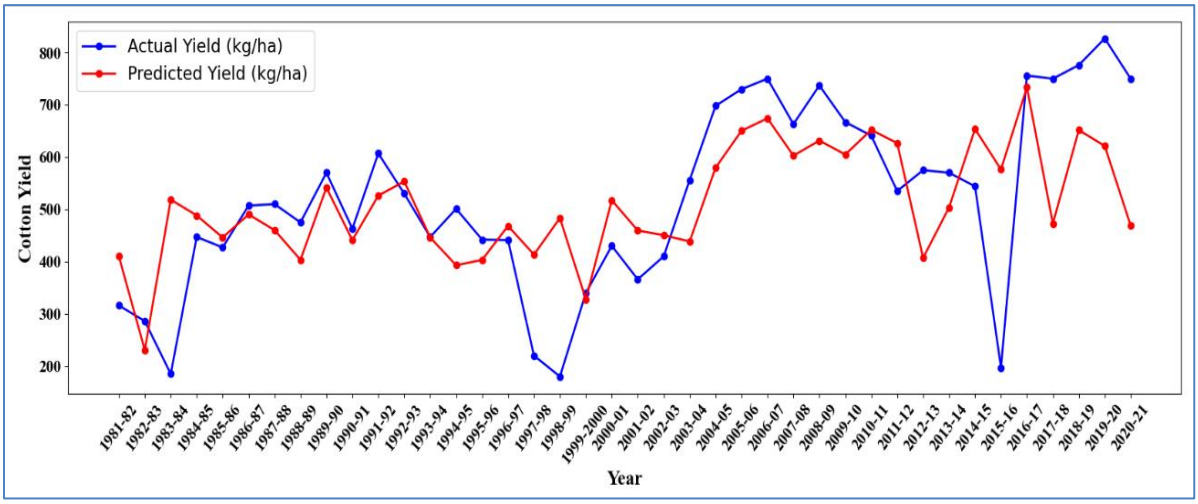

Fig. 6. Cotton yield forecasted by LSTM (10 3 3 3 1) model

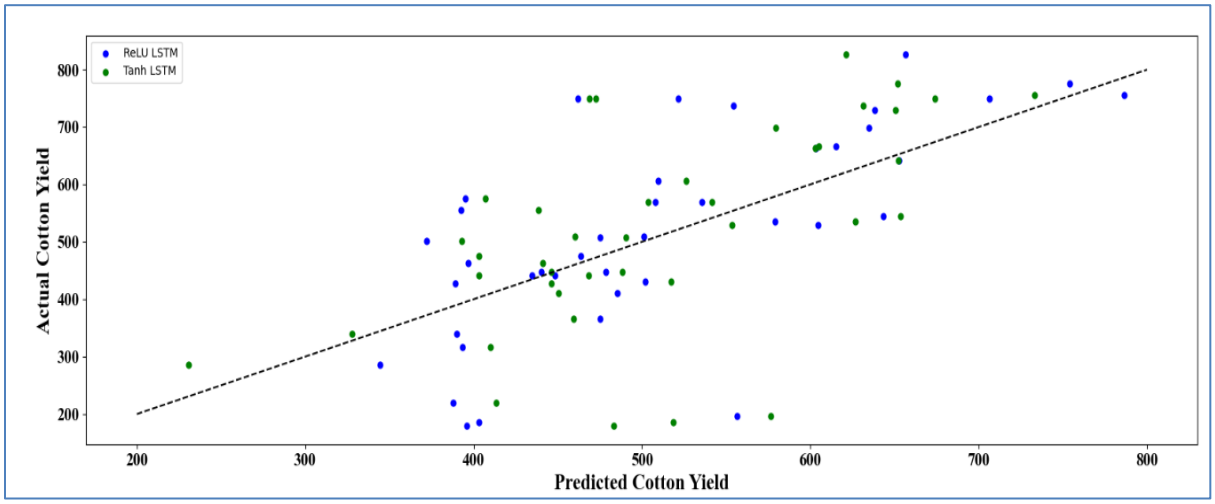

Fig. 7. Scatter plot for observed and predicted values of cotton yield for LSTM models

TABLE 2

Different statistics for ANN and LSTM models

Statistics	ANN		LSTM	
	ReLU in hidden layer (10 10 10 1)	**Tanh in hidden layer (10 9 9 1)**	**ReLU in hidden layer (10 4 4 4 4 4 1)**	**Tanh in hidden layer (10 3 3 3 1)**
MAPE in (%)	17.83	26.63	24.83	28.34
RMSPE in (%)	30.52	46.46	44.14	53.86
NRMSE	0.17	0.23	0.24	0.28
RMSE	0.13	0.19	0.19	0.21
MAE	0.10	0.15	0.14	0.16
R^2	0.74**	0.50**	0.48**	0.29**
$\bar{R}^2$	0.70	0.42	0.41	0.19
ρ	0.86**	0.71**	0.70**	0.60**

**represents significant at 1% significance level (p-value < 0.01)

TABLE 3

Mse and w values of sensitivity analysis

Variable removed	MSE	W
NONE	0.0182	1.0000
Relative Humidity	0.0466	2.5615
Rainfall	0.0356	1.9560
Maximum Temperature	0.0334	1.8352
Minimum Temperature	0.0236	1.2967
Maximum Windspeed	0.0200	1.0989

of different ANN models with tanh function in hidden layers was observed using MSE, ANN model (**10 9 9 1**) had lower MSE values than other ANN models. MSE values of training & testing data of yield were 0.0351 and 0.0401 respectively.

The observed and predicted cotton yield for ANN model when ReLU activation function is used in hidden layers is presented in Fig. 2. For the ANN (**10 9 9 1**) model with tanh activation function in hidden layers, the observed and predicted cotton yield is shown in Fig. 3. Blue lines exhibit the recorded crop yield. Red lines represent the forecasted values predicted by ANN models.

The scatter plot of recorded and forecasted cotton yield by ANN models is shown in Fig. 4. The predicted cotton yield values are uniformly distributed about the 1:1 line, indicating a reasonably accurate prediction. The predicted values are in close approximation with the corresponding observed values. Blue dots exhibit the values of ANN (**10 10 10 1**) model and green dots represent the values of ANN (**10 9 9 1**) model.

3.3. *LSTM models for Punjab*

Like ANN, LSTM models with one to five hidden layers were considered. The optimal number of nodes was found using the least values of MSE. In hidden layers, activation functions, namely ReLU and tanh, were used to develop different models.

For the ReLU activation function in hidden layers, LSTM model with 10 nodes in the input layer, 4 nodes in 5 hidden layers performed better with MSE values 0.0316 and 0.0592 for training and testing data of cotton yield. For whole crop data, MSE value was 0.04615. When performance of different LSTM models with tanh function in hidden layers is observed, LSTM (10 3 3 3 1) model had lower MSE values than other models. MSE values of training & testing data are 0.03925 and 0.07374 respectively.

The observed and predicted cotton yield for LSTM (10 4 4 4 4 4 1) model when ReLU activation function is used in hidden layers is presented in Fig. 5. For LSTM (10 3 3 3 1) model with tanh activation function in hidden layers, the observed and predicted cotton yield is shown in Fig. 6. Blue lines exhibit the recorded crop yield. Red lines represent the forecasted values predicted by LSTM models. The scatter plot of observed and predicted cotton yield by LSTM models is shown in Fig. 7. Blue dots exhibit the values of LSTM (10 4 4 4 4 4 1) model and green dots represent the values of LSTM (10 3 3 3 1) model.

3.4. *Comparison of the performance of ANN and LSTM models*

Comparison between the selected ANN and LSTM models was done on the basis of statistics as MAPE, RMSPE, NRMSE, MAE, RMSE, R^2, $\bar{R}^2$ and correlation between predicted and observed cotton yield. Table 2 represents the comparison between NN models. A test of significance for goodness of fit and correlation coefficient was also done. ANN model (**10 10 10 1**) with ReLU activation function in hidden layers outperformed with MSE (0.0182) and MAE (0.10). The model significantly explained approximately 70% variability in the model due to weather variables. The predicted values of this model had a highly significant correlation (0.86) with actual cotton yield. MAPE and RMSPE values were 17.83% and 30.52% respectively.

3.5. *Sensitivity analysis*

Sensitivity analysis of ANN for cotton yield was exhibited as discussed in Section 2.5. After eliminating each parameter one by one and running ANN with four parameters, a new MSE was determined for each eliminated variable. Then ratios (W) of new MSE and original MSE were calculated as given in Table 3. The cotton yield in Punjab was mostly affected by relative humidity followed by rainfall. Maximum windspeed had the lowest effect on the output. In general, the importance of input weather parameters of the model can be differentiated with the use of sensitivity analysis for the outperformed ANN (10 10 10 1) model.

3.6. *Discussion*

Predicting the cotton yield is one of the major concerns for farmers and policymakers in Punjab, India. The literature survey showed that the cotton yield data from 1981-82 to 2020-21 followed a non-linear trend (Dhillon and Pathak, 2020). The machine learning models based on ANN and LSTM approaches have gained significant importance and are used world-wide for non-linear forecasting purposes. In the present study, average yearly maximum temperature, minimum temperature, maximum windspeed and relative humidity from 1981-82 to 2020-21 were used for forecasting the cotton yield. Total rainfall values were also taken as input. ANN and LSTM models having one to five hidden layers with ReLU and tanh activation were considered. The optimal number of nodes in these layers were found using the minimum value of MSE. These models were compared and ANN (10 10 10 1) model with ReLU activation function in hidden layers performed better than other ANN and LSTM model. Many studies have resulted that LSTM models outperformed ANN models when models were trained on larger dataset. But some studies have revealed that ANN models were better to forecast agricultural predictions (Ju, *et al*., 2019). The outperforming model in the present study predicted the cotton yield with MAPE (17.83%), RMSPE (30.52%) and MAE (0.10). This model significantly explained approximately 70% variability in the model due to weather variables. The correlation between the observed and predicted values was 0.86. Sensitivity analysis using the outperformed model showed that relative humidity was an important factor that significantly impacted the cotton yield followed by rainfall. Maximum windspeed had the lowest effect on cotton yield. All the weather parameters used in the study affected the cotton yield. These results aligned with the reviewed literature survey. Cetin and Basbag, 2010 claimed that air temperature and relative humidity significantly impacted the productivity of cotton. Cotton productivity was also affected due to rainfall (Dhir, *et al*., 2024).

4. Conclusions

The present study was done to develop a prediction model using different techniques to forecast cotton yield for Punjab. Climatic factors like rainfall, relative humidity, temperature and rainfall were considered as input. ANN and LSTM techniques were used to develop the forecasting model. ANN, having the ReLU activation function in 2 hidden layers with 10 nodes, performed better with MAPE (17.83%), RMSPE (30.52%) and MAE

(0.10). The model significantly explained approximately 70% variability in the model due to weather variables. Furthermore, sensitivity analysis revealed that relative humidity was the most sensitive parameter that significantly affected the cotton yield followed by rainfall. Maximum windspeed had the lowest effect on cotton yield.

Acknowledgements

The authors would like to thank department of Maths, Stats and Physics, Punjab Agricultural University, Ludhiana for providing the required support and facilities. The authors are highly grateful to the department of agriculture and farmers welfare, India-Wris website and data access viewer.

Data Availability

The sources of the obtained data are mentioned in the Section 2, subsection 2.1 (Study area and data collection).

Authors' contributions

Opinder Kaur: Manuscript Writing, model formation and data analysis. (*email: Opinder664@gmail.com*).
Mohammed Javed: Supervision and advisory support. (*email: mjaved@pau.edu*).
Chetan Singla: Co-supervision, technical guidance, methodology support and manuscript review. (*email: chetan_singla@pau.edu*).
Gurjeet Singh Walia: methodology support and guidance. (*email: gswalia@cuo.ac.in*).

Disclaimer: The contents and views presented in this research article/paper are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Cetin, O. and Basbag, S., 2010 "Effects of climatic factors on cotton production in semi-arid regions - A review", *Research on Crops*, **11**, 3 785-91. https://www.researchgate.net/publication/259713780.

Dahikar, S.S. and Rode, S.V., 2014, "Agricultural crop yield prediction using artificial neural network approach" International Journal of Innovation Research in Electrical, Electronics, *Instrumentational and Control Engineering*, **2**, 1, 683-86. https://api.semanticscholar.org/CorpusID:16167655.

Dhillon, B.S. and Pathak, D., 2020 "An analysis of cotton cultivation in Punjab and role of Bt cotton", *Agricultural Research Journal*, **57**, 6, 965-967. http://dx.doi.org/10.5958/2395-146X.2020.00142.8.

Dhir, A., Pal, R.K., Kingra, P.K. and Kaur, R. 2024 "Climate-smart cotton (Gossypium herbaceum) crop production in Punjab: A comprehensive review of sustainable management practices", *Indian Journal of Agricultural Sciences*, **94**, 2, l. 119-128. https://doi.org/10.56093/ijas.v94i2.141408.

Ju, S., Lim, H. and Heo, J., 2019, "Machine learning approaches for crop yield prediction with MODIS and weather data." *The 40th Asian Conference on Remote Sensing (ACRS 2019)*, Daejeon Convention Center (DCC), Daejeon, Korea.

Kaul, M., Hill, L.R. and Walthall, 2005, "Artificial neural networks for corn and soybean yield prediction." *Agricultural Systems*, **85**, 1 . 1-18. https://doi.org/10.1016/j.agsy.2004.07.009.

Meshram, P.S. and Dange, R.K., 2022, "Geographical analysis of cotton and its proportion under total crops and cash crops (2000 to 2020)", *International Journal of Food and Nutritional Sciences*, 11, 13, 343-346.

Patel, R. and Saha, G., 2020, "Regression of weather parameters for cotton in Gujarat using neural network for data (1901-2002)", https://www.researchgate.net/publication/347342770.

Pravallika, K., Karuna, G., Anuradha, K. and Srilakshmi, V., 2021. "Deep neural network model for proficient crop yield prediction", *3rd International Conference on Design and Manufacturing Aspects for Sustainable Energy*, 309 (01031) https://doi.org/10.1051/e3sconf/202130901031.

Salehin, I., Talha, I.M., Hasan, M., Dip, S. T., Saifuzzaman, M. and Moon, N.N., 2020, "An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network", *International Women in Engineering (WIE) Conference on Electrical and ComputerEngineering*.http://dx.doi.org/10.1109/WIECONECE52138.2020.9398022.

Shao, F. and Shen, Z., 2022, "How can artificial neural networks approximate the brain?", *Frontiers in Psychology*. **13**: 970214. https://doi.org/10.3389/fpsyg.2022.970214.

Sharma, S., Rai, S. and Krishnan, N.C., 2020, "Wheat crop yield prediction using deep LSTM model", https://doi.org/10.48550/arXiv.2011.01498.

Tayfur, G. and Singh, V.P., 2006, "ANN and Fuzzy Logic Models for Simulating Event-Based Rainfall-Runoff"' *Journal of Hydraulic Engineering*, **132**, 12, 1321–1330. https://doi.org/10.1061/(ASCE)0733-9429(2006)132:12(1321).

Yildirim, T., Moriasi, D.N., Starks, P.J. and Chakraborty, D., 2022, "Using artificial neural network (ANN) for short-range prediction of cotton yield in data-scarce regions", *Agronomy*, 12, 828. https://doi.org/10.3390/agronomy12040828.