Discover Sustainability

Research

Crop recommendation and forecasting system for Maharashtra using machine learning with LSTM: a novel expectation-maximization technique

Yashashree Mahale[1] · Nida Khan[1] · Kunal Kulkarni[1] · Shivali Amit Wagle[1] · Preksha Pareek[1] · Ketan Kotecha[1,2] · Tanupriya Choudhury[3,4] · Ashutosh Sharma[5,6]

Received: 5 January 2024 / Accepted: 22 May 2024
Published online: 28 June 2024

Abstract

Agriculture in Maharashtra has immense importance in India, acting as the back-bone of the economy and a primary livelihood source for a significant population. Being the third largest state in India, Maharashtra has a high scale crop production in the country which also has an important impact on the economy. Initially the study focus on developing predictive models that guide farmers in selecting suitable crops for the divisions in the state of Maharashtra. This study presents a Crop Recommendation System (CRS) designed to support Maharashtra's agricultural sector by utilizing a comprehensive dataset from 2001 to 2022 provided by the India Meteorological Department. This study helps in improvising technical efficiency and productivity of the farmers. Harvesting crops in optimal condition can help to produce efficient harvest hence the research concentrates on providing best crop recommendation system (CRS) with the help of Machine Learning and Deep Learning techniques. The data, enhanced for accuracy using expectation-maximization optimization, underpins predictive models that guide crop selection. EM contributes to a more robust and reliable dataset for subsequent analyses and modeling by iterative estimating and updating missing values based on probabilistic expectations. Key findings show that the Random Forest algorithm excels in predicting suitable crops with 92% accuracy. Further precision is achieved through a Long Short-Term Memory network forecasting weather patterns three months ahead, accommodating temporal data variations. Subsequently, the proposed system leverages these forecasts to recommend five ideal crops per division within Maharashtra, aiding farmers' decision-making and adapting to regional climatic conditions. A supplementary crop calendar offers monthly district-specific planting guidance. An intuitive Graphical User Interface delivers this information effectively, ensuring practical and informed agricultural choices across the state. In essence, the study provides an innovative tool for enhancing economic stability and sustenance in Maharashtra through technology-driven agriculture recommendations aligned with future weather expectations.

Keywords Crop recommendation · Expectation-maximization · LSTM · Weather forecast

✉ Ketan Kotecha, head@scaai.siu.edu.in; Yashashree Mahale, yashashree125@gmail.com; Nida Khan, nidak6478@gmail.com; Kunal Kulkarni, kulkarnikunal63@gmail.com; Shivali Amit Wagle, kulkarni_shivali@yahoo.co.in; shivali.wagle@sitpune.edu.in; Preksha Pareek, preksha.pareek@sitpune.edu.in; Tanupriya Choudhury, tanupriyachoudhury.cse@geu.ac.in; tanupriya1986@gmail.com; Ashutosh Sharma, Ashutosh@haust.edu.cn | [1]Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India. [2]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India. [3]Present Address: CSE Department, Graphic Era Deemed to be University, Dehradun, Uttarakhand 248002, India. [4]CSE Dept., Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India. [5]Business School, Henan University of Science and Technology, Luoyang, Henan 471023, China. [6]School of Computer Science, University of Petroleum and Energy Studies (UPES), Uttarakhand 248007 Dehradun, India.

1 Introduction

Agrarian activities in India date back to the Neolithic era. Indian Economic Survey 2018 indicated over half of the workforce in the nation is employed in agriculture, which also contributes 17–18 percent to its economy. Monsoons and Rainfall Patterns, Crop Varieties,Seeds, and agricultural practices highly affect the productivity and sustainability on farms of India. The factors mentioned are interconnected and have a major impact on the agricultural production of the country. Considering the limited agricultural land and unfavorable environment to produce food is a major aspect in growth of the country [1]. In India, the shift in climatic patterns has had a huge impact on crop yields. This has also affected soil, water, and pest prevalence in these regions by influencing the variety of crop cultivation in certain areas. It may affect irrigated agricultural production in agro-ecological zones. Considering the state of Maharashtra, it has a huge area under the cultivation of both cash crops and food crops. A farmer must always select the best crop considering the environment in mind [2]. Farmers mostly rely on historical weather data for crop selection and management. Maharashtra's monsoon season, which is crucial for agriculture, has shown some variations in onset, withdrawal, and distribution over time. Sowing and crop development can be impacted by delayed or erratic monsoons, resulting in lower yields. Moreover, many parts of Maharashtra have experienced temperature increases which has caused changes in the variety of crops to be cultivated. Droughts as well as water scarcity have further worsened the situation leading to crop failures in regions and affecting crops like sugarcane, pulses, and oilseeds. As a result, farmers have had to adapt by shifting towards drought shorter duration crops. However, this adjustment can potentially alter the landscape as a whole. This may not be economically sustainable, in the long term. Making educated decisions about crop selection, weather forecasts, and resource allocation requires consideration of the growing variability in climate patterns.

The field of machine learning enables computers to be taught without predetermined programming. These techniques solve agricultural frameworks that are either linearistic or nonlinear with exceptional forecasting ability [3]. Thus, to forecast or make decisions about agricultural processes, computer algorithms are trained based on labeled data with supervised machine learning approaches and statistical methods. Compared to conventional statistics, machine learning algorithms have shown comparatively greater potential [4]. With an ML-based crop recommendation system, it is possible to help farmers choose the appropriate crops for a particular location and desired weather condition thus mitigating the risk of crop failure and improving the overall management of farming operations. In this article, we propose a methodology along with the comparison of different supervised ML approaches for the weather-crop dataset of Maharashtra state alone that focuses on the recommendation of crops based on weather parameters like temperature and rainfall for different districts in the state. The Expectation-Maximization (EM) method's distinct probabilistic structure and iterative refining procedure set it apart from previous imputation preparation methods. Because EM can manage missing data in a probabilistic framework, it is very useful in situations like crop recommendation and forecasting. This study also demonstrates an LSTM model for the weather forecasting of the next 3 months from the available data. Based on these forecasts, the crop recommendation modules can make predictions of the crops. The study mainly focuses on the following aspects:

1. Application of an EM-based optimization technique to effectively clean the temporal dataset, enhancing the overall data quality and replacing null values and reliability.
2. Advances in forecasting of weather by employing Long Short-Term Memory (LSTM) networks on primary data for the 6 divisions of Maharashtra state thereby doing the crop recommendation for Agro-climatic zones.
3. The crop calendar designed operates on a month-by-month basis, offering a comprehensive framework to optimize agricultural practices and enhance precision in farming methodologies.

2 Related work

Several strategies for developing a crop recommendation system have been published in the literature. Some methods use a machine-learning approach while others use a deep learning approaches with Multisensory Machine Learning Approaches (MMLA).

In [5], authors proposed a smart way to manage crops and harvest them.A variety of machine learning models like KNN, Logistic Regression, Bagging, Naïve-Bayes, SVM, AdaBoost, Decision Tree, RF, Gradient-Boosting, XGB, IBGM were used for crop recommendation where Random Forest gave better accuracy of 97.18

An ensemble model named KKR (Korringa Kohn Rostoker) for proper crop cultivation in Bangladesh [6] uses distance-based KNN '&' second-order ensemble strategy which finds the combined predictors from KNN, RR, and RF to form an ensemble regressor predicting better. The proposed KKR performed better to investigate three major rice variations with potatoes and help authorities planning the food supply in the future. However, only the major crops are predicted. There were some factors that were omitted from the analysis, such as soil properties, production costs, and market prices.

Comparisons of finely tuned machine learning models explain the variability of wheat yields on Indian farmlands in the northwest Indo-Gangetic Plains has been made in [7]. The random forest model showed good fitting, accuracy and precision measures, which accounted for 25% of wheat yield variability. Future research is required to understand the relation between crop yield for prioritizing and ensuring sustainable farming.

Authors in [8] conducted an in-depth literature analysis for crop prediction with ML. Around 567 of relevant studies were retrieved, of which 50 were selected for further analysis. The authors conclude that temperature rainfall, and soil variety are the commonly used features and the most common algorithm used is CNN. In paper [9], the authors experimented with 5 different ML algorithms viz Decision Tree, Naive Bayes, SVM , Random Forest and Logistic Regression for West Bengali Agriculture focusing on 13 major crops for different districts in West Bengal. Rainfall, Temperature, Humidity, Sun Hours were considered and it was observed that RF and SVM helped to achieve superior results. Although the project is tested on 2 district datasets only, work needs to be done on updating the databases for error-free predictions.

In [10] the authors proposed a prediction model for rice cultivation with Support Vector Machine in China. PCA and fivefold cross-validation were used for dimensional reduction with three types of rice plantings and executed successfully. In addition to being convenient for data acquisition, the proposed open-crop model integrates multiscale factors well, is simple to parametrize and applies to a wide range of regions. The model can be further expanded to include variables like the cost of grain, fertilizer, seed, labour, location, traffic, support from the government, the true state of agriculture, and innovations in science and culture.

A deep learning model using CNN was studied in [11]. They identified infections in plants and studied them to prevent loss of crop prediction and help farmers grow healthy plants. This method increased the efficiency compared to traditional methods as it uses image processing. This technique can be modified by considering climatic factors and considering more diseases in plants. Although the system is currently only implemented in Karnataka, it can be expanded to cover the entire country.

In [12] Data Analytics techniques such as Linear Regression with neural networks has been used for prediction of prices for crops. Factors such as Area harvested, Area planted, etc. were considered and the authors concluded that XGBoost was the best technique for price prediction of various crops while in [13] authors predicted crop cultivation for farmers using machine learning techniques which suggests best conditions to plant, harvest, and water crops. They used pair plots, joint plots, heat maps, and Barh for decision tree prediction. Moreover, the models can be modified for fertilizer recommendation and app integration, so that farmers can have easy access to them. For fertilizer utilization and crop prediction, a machine learning regression algorithm based on naive Bayes classifiers was developed in [14]. A model was developed based on 4 crops and 7 parameters, including soil nutrients, in the Mysore district. For crops such as wheat, ragi, and paddy, the algorithm has a good prediction rate. Furthermore, it can also be modified through the use of farmer-friendly applications.

The study in [15] delves into the evolving nature of Indian Summer Monsoon (ISM) variability in Maharashtra, Western India, and its impact on agriculture and food security. Unlike previous coarse-scale analyses, this research focuses on a finer district-level examination, revealing significant Spatio-temporal heterogeneity. A monsoon variability index, incorporating six key parameters, identifies heightened vulnerability in Vidarbha and Marathwada districts. Structural equation modeling establishes connections between the index, average yield, and cropped area, aiding in the determination of optimal cropping patterns for vulnerable districts. The study proposes a district-level empirical model for monsoon variability, offering practical insights for regional climate action plans and guiding agricultural policies and climate adaptation measures. The below Table 1 gives detailed literature study including the details of the dataset used, comparative analysis of various models used with their performance.

Table 1 Summary of literature study

References	Year	Dataset used	Model used	Performance
[16]	2022	Data of United Nations Food and Agriculture Organization (FAO)	Predictions were done with Crop-Decision-Tree and KNN predict crop-yield values. Multivariate Logistic Regression was used to compare with other crop yield models	CDT- 94,65 CMLR-83,80
[17]	2021	Sentinel-1A SAR data from June to September each year.	Regression techniques that have been suggested include Gaussian Kernel Regression, Bayesian Formulation From GPR, and Bayesian Linear Regression	The Gaussian Kernel Regression gave highest prediction accuracy with r2 = 0.81
[18]	2023	Crop_recommendation Source: Kaggle	Authors proposed and compared modes like Random Forest, Gradient Boosting,AdaBoost, XGBoost, LightGBM, and CatBoost with their precision, recall and F1 scores respectively	LightBGM and XGBoost outperformed RF and AdaBoost with accuracy of 99.3%
[19]	2023	MODIS NDVI time series	Three statistical models—SVM, RFR, and MLR have been analyzed and contrasted by the authors. These models were verified with 70% of soil samples	RFR frequently offered the highest accuracy, significantly outperforming MLR
[20]	2023	Palm oil dataset	SVM NB RF ANN CNN MLR PCA Bayesian Networks	99.21% 85% 88.70% 94% 92.29% 92.5% MSE less than 0.01 75%
[21]	2022	Felin dataset	Feature selection: Boruta (RF) Random Feature Elimination Naive Bayes Decision Tree Support Vector Machines KNN RF Bagging	NB:70.64% DT: 73.22% SVM: 77.50% KNN:83.24% RF:87.43% Bagging:84%
[22]	2018	–	Kohonen Self Organizing Map (Kohenon's SOM) and BPN (Back Propagation Network)	The soil analysis module uses pH and location as input, analyzes the soil and classifies it using the soil crop comparison module
[23]	2023	Crop yield data with fertilizers	Using the Random Forest technique and the Support Vector Machine, classification of soil	Random Forest's accuracy is 86.35%, whereas Support Vector Machine's accuracy is 73.75%
[24]	2021	Palm oil yield data of Corn, Soybean and wheat	MLR Extremely randomized trees SVM XGBoost RF	92.5% 95% 93.16% 85.41% 95%
[25]	2022	–	SVM, RT, BG, BA, LR, KNN, ANN, SVR, RF, SVR, RF	

Table 1 (continued)

References	Year	Dataset used	Model used	Performance
[26]	2019	Data of the Corn Belts region of the Midwest(US)	Cropland Data Layer (CDL) time series is used as the crop planting prediction,with a prediction model of multi-layer artificial neural network	The machine-learned prediction's estimations of crop acreage were significantly correlated with R2 > 0.9
[27]	2020	IOT sensor data with crop database	Naive Bayes, Support Vector Machines, and Neural Networks (NB)	NB: 84% SVM:78% NN : 98.4%
[28]	2021	35,000 images of healthy and sick plant leaves	Deep Learning, TensorFlow, Keras, Max_pooling, Batch normalization	96.84 96.5 95 94 88.8 87

[29]	2020	Data is collected with the help of sensors and Iot	RSSI-based crop sensing
			technique; link quality indication
			(LQI), RSSI-based LAI
			model, Cross-validation
				R2 = 0.88
				nRMSE = 0.12

3 Proposed system

In this section, the entire flow of the crop recommendation system is presented as illustrated in Fig. 1.

3.1 Dataset

3.1.1 Temporal monthly weather data of Maharashtra state

The India Meteorological Department (IMD), Pune, an esteemed Indian government agency, provided the dataset used for this study. Including the state of Maharashtra, IMD runs a network of weather stations, observatories, and other equipment all throughout India. The dataset spans many stations throughout the state of Maharashtra and includes a comprehensive collection of weather-related metrics that are recorded monthly. It includes a comprehensive collection of meteorological metrics recorded across all weather stations throughout the state of Maharashtra and

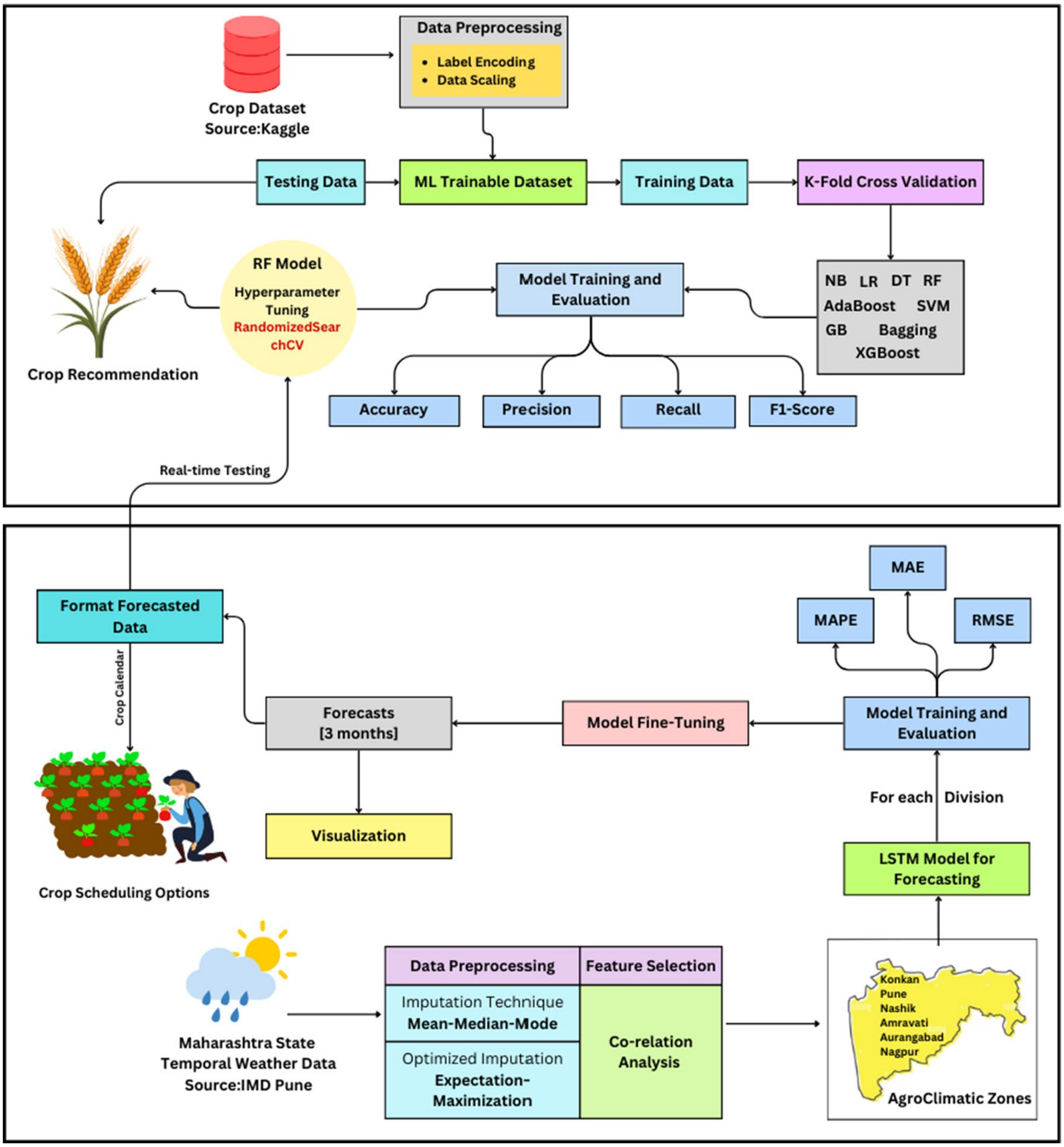


Fig. 1 Proposed system design

Table 2 Meteorological parameters in the dataset

Index	Index number of a station
MN	Month
DT	Date
OC	No. of occasions
RD	No. of rainy days in the month
HVYRF	Heaviest 24 h rainfall (mm)
MMAX	Mean maximum temp (°C)
MMIN	Mean minimum temp (°C)
HMAX	Highest maximum temperature (°C)
LMIN	Lowest minimum temperature (°C)
NO	No. of observations
TMRF	Total rainfall in the month (mm)
MWS	Mean wind speed (kmph)
MEVP	Mean evaporation (mm)
MSSH	Duration of sunshine (h)

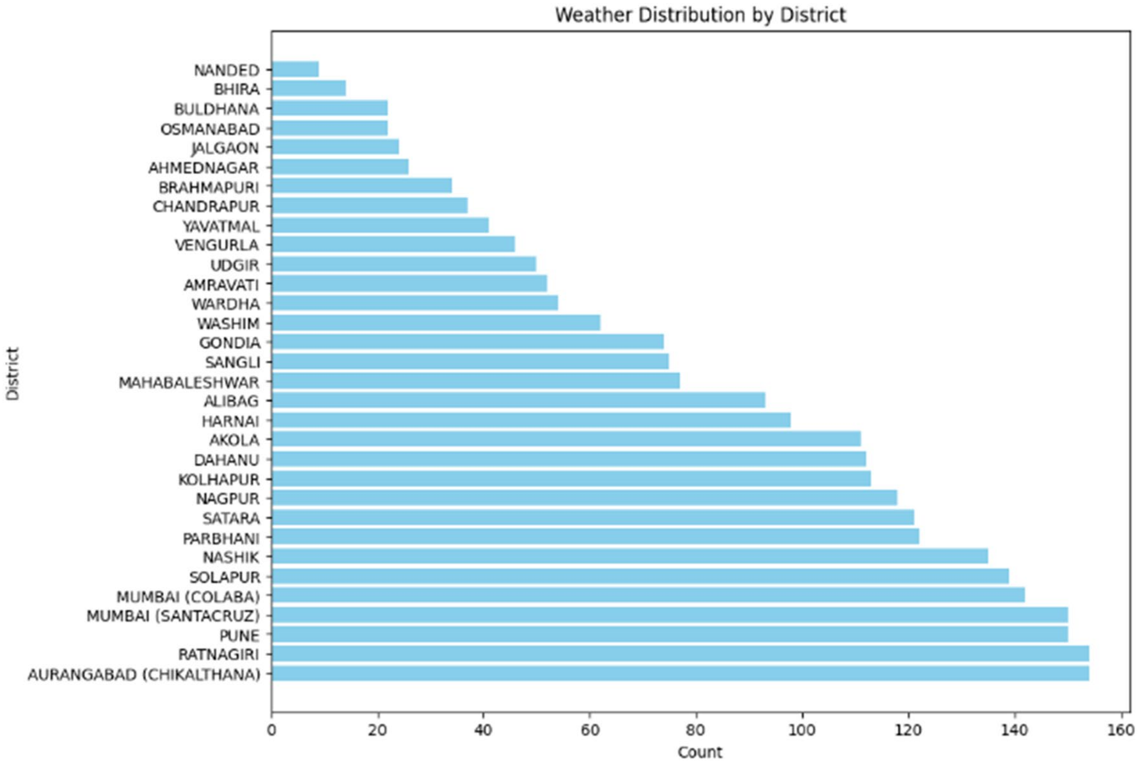


Fig. 2 Weather distribution by district

covers the years 2001 to 2021, with further data for 2022 and 2023 added later. Rainfall, temperature, wind speed, evaporation, sunlight hours, gull strength, dust accumulations, and storm frequency are some of these characteristics. Table 2 shows some of the parameters found within the dataset while Fig. 2 illustrates an analysis of the weather distribution by district.

The pair plot presented in Fig. 3 illustrates the temporal dynamics of key variables in Maharashtra state over the specified time period. The plot provides a comprehensive visual exploration of the relationships and trends among the variables, offering valuable insights into the temporal patterns within the region. Other valuable exploratory tools for uncovering temporal patterns and relationships within the dataset were also studied to do the thorough data analysis.

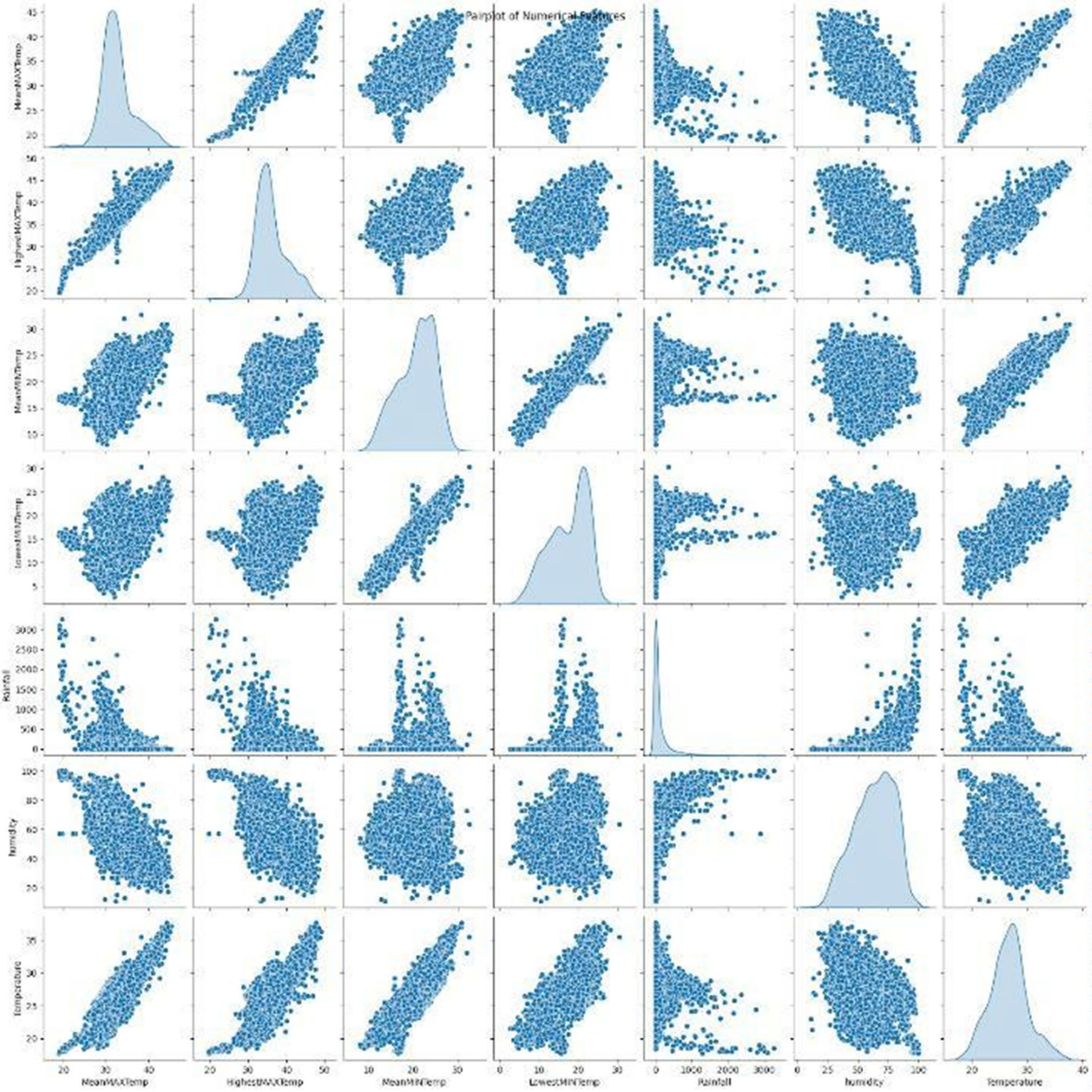


Fig. 3 Pair plot for temporal data

3.1.2 Crop recommendation dataset

Data for this study was sourced from Kaggle, a popular platform for sharing and hosting datasets. Agricultural data pertaining to crop recommendations are found in dataset for Crop Recommendation [30]. The dataset was formed with enhancing databases that included information on India's rainfall, climate, and fertilizer. This includes information on various crops and their recommended levels of the NPK factors (Nitrogen, Phosphorus,Potassium) along with the environmental parameters such as humidity, rainfall and temperature. The crop distribution is shown in Fig. 4.

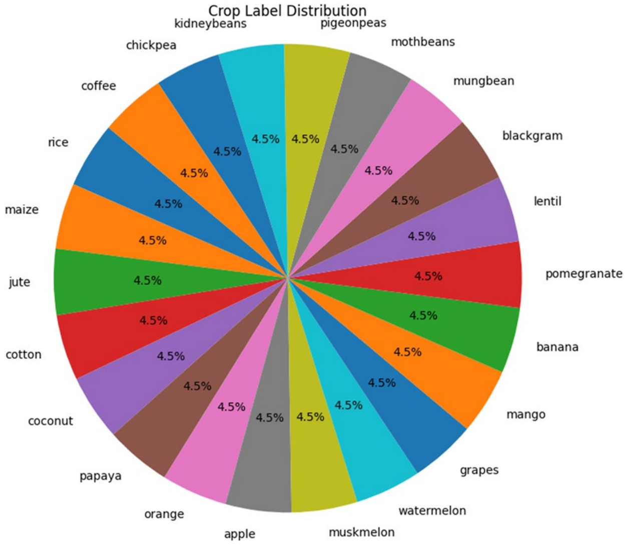


Fig. 4 Crop labels in crop-recommendation dataset

3.2 Preprocessing

Considering the temporal weather data of Maharashtra state, 49 objects had zero values to be treated. This is time series data, so optimization imputation is applied here.

3.3 Expectation maximization: an optimized imputation technique

Expectation maximization (EM) is a statistical algorithm commonly used to handle missing or null values for time series data. It is an iterative algorithm that alternates between two main phases: a expectation phase and a maximization phase [31]. EM treats missing data as latent variables and repeatedly estimates their values along with the model parameters, in contrast to some other approaches that might just fill in the missing values using means or medians. Due to its probabilistic character, EM is particularly helpful in agricultural environments where data may be noisy or lacking. It also captures uncertainty in the data. Additionally, EM has a number of benefits over conventional imputation techniques. Although imputation of the mean, median, mode, and regression is simple, they frequently produce estimates that are skewed and fail to adequately portray the underlying complexity of the data, particularly when absence of values is not entirely random. The assumption that similar instances have similar missing values is a major reliance of K-Nearest Neighbors (KNN) imputation. However, this assumption may not always hold true, especially in high-dimensional environments. Furthermore, these approaches don't offer a logical strategy to deal with the uncertainty brought on by missing data. Hence all things considered, EM is a better option for imputation preprocessing in crop recommendation and forecasting systems because of its probabilistic foundation, flexibility, and iterative refinement process, especially when working with complicated and partial datasets. The EM algorithm aims to find the maximum likelihood of estimating the parameters in a statistical model when there is missing data [32]. The steps used in the algorithm are listed below:

1. Initialization:
   Begin with initial parameter estimates (α_0). Initialize missing values.
2. Expectation step (E-step):
   Update missing values based on current parameter estimates using:

$$Q(\alpha \mid \alpha_t) = E_{\text{missing}}[\log L(\alpha;\text{observed, imputed}) \mid \text{observed},\alpha_t]. \tag{1}$$

Here, Q is the expected log-likelihood function, α is the parameter vector, and α_t are the current parameter estimates. Given the current parameter estimates and the observed data, the missing data is expected.

$$E(\text{missing} \mid \text{observed},\alpha_t). \tag{2}$$

3. Maximization step (M-step):
   Update parameters α_{t+1} by maximizing $Q(\alpha \mid \alpha_t)$.

$$\alpha_{t+1} = \arg\max_{\alpha} Q(\alpha \mid \alpha_t). \tag{3}$$

   The optimization problem is addressed in this step to determine the parameter values.
4. Convergence check:
   Verify convergence by evaluating the shift in parameter estimates.

$$\|\alpha_{t+1} - \alpha_t\| < \epsilon. \tag{4}$$

5. Iterative refinement:
   Iterate steps 2–4 until convergence.

3.4 Feature selection

The volume for feature space increases exponentially with various features. As result, the dimensionality curse may occur, making it difficult to build efficient and accurate models, especially with minimal number of samples [33]. In all 49 features has been identified in Maharashtra state dataset across all districts. By reducing feature numbers without affecting prediction performance, feature selection significantly improves computational efficiency. An analysis of the correlation matrix determined the interdependence between the features. For each pair of features, the correlation matrix provided a pairwise correlation coefficient, ranging from – 1 to 1. In general, values closer to 1 represent a strong positive correlation, values –1 represent a strong negative correlation, and values 0 represent no linear correlation. This coefficient quantifies the linear relationship between variables. 0.5 was set as the threshold for identifying highly correlated features. The coefficients of the features greater than this threshold are considered highly correlated. Using the correlation matrix of the selected features, only the features with significant correlations were retained. Table 3 gives the list of features obtained after feature selection and the conclusive correlation matrix is visually depicted in Fig. 5 through a heat map. Besides the selected columns, district names, latitudes, and longitudes are also added.

For crop-recommendation.csv, only weather parameters [temperature, rainfall and humidity] are selected aligning with the requirement of temporal weather data of Maharashtra state.

3.5 Modelling

Phase 1: crop recommendation model

(i) Data splitting

After preprocessing, meticulous data cleaning and preprocessing to ensure the integrity and quality of the dataset the ML trainable dataset is prepared and involved. The data was biforcated into two sets: a training set consisting with 80% and a testing set of 20%. This division facilitated the evaluation of machine learning models

Table 3 Meteorological parameters after feature selection

Index	Index number of a station
MMAX	Mean maximum temp (°C)
MMIN	Mean minimum temp (°C)
HMAX	Highest maximum temperature (°C)
LMIN	Lowest minimum temperature (°C)
NO	No. of observations
TMRF	Total rainfall in the month (mm)
MWS	Mean wind speed (kmph)
MEVP	Mean evaporation (mm)
MSSH	Duration of sunshine (h)
HVYRF	Heaviest 24 h rainfall (mm)
Humidity	Humidity

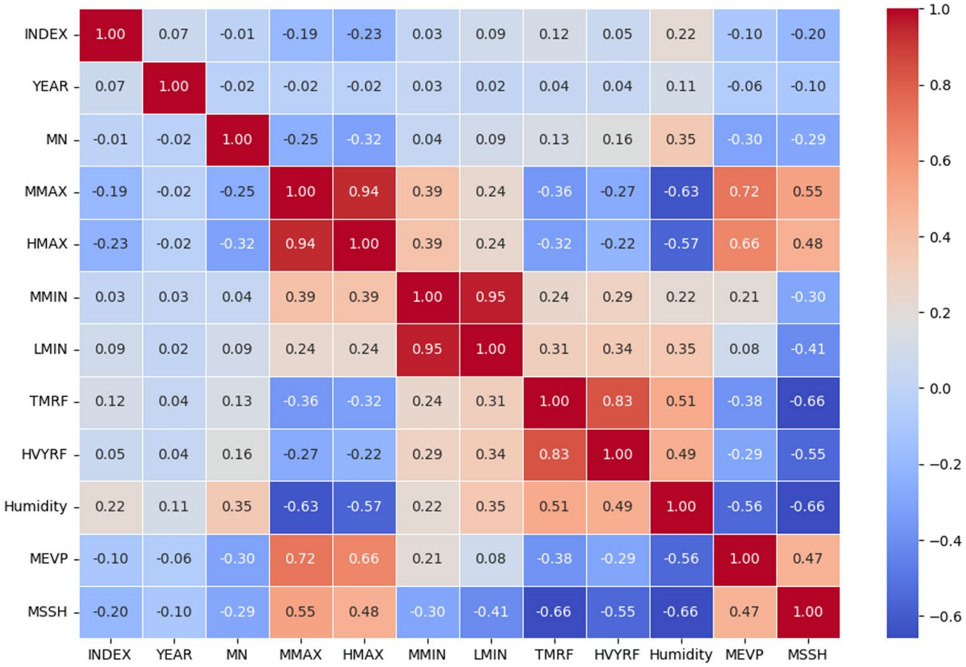


Fig. 5 Heatmap visualizing correlation matrix

on a comprehensive dataset, enabling a robust assessment of their performance on both familiar and unseen instances.

(ii) K-fold cross validation

Following this, k-fold cross-validation using the Stratified K-Fold technique with k = 5, shuffling the data for enhanced robustness and reproducibility, and setting the random seed to 77 for consistency in the results. This method ensures that each fold proportion of class labels is maintained, addressing potential imbalances in the dataset. The Stratified K-Fold approach partitions the dataset into five subsets, and in each iteration, four of these subsets are used for training and left were used as validation set. This procedure is executed five time, allowing each subset to act as a validation set exactly once. The shuffling of data prior to partitioning enhances the generalization of the model by mitigating the impact of any inherent order in the dataset.

(iii) Models used

A diverse set of machine learning algorithms to the dataset, including Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), AdaBoost, Bagging, XGBoost, and Gradient Boost as shown in Figure. The k-fold cross-validation framework was utilized to train and assess each algorithm, enabling a comparative analysis of their performance across various evaluation metrics such as precision, F1 Score, recall and accuracy.

(a) Naive Bayes

The working of Naive Bayes is based on the Bayes theorem, Naive Bayes is a probabilistic machine learning algorithm. It's a straightforward yet effective algorithm that works especially well on classification tasks. Naive Bayes is an algorithm that works well in practice, especially when handling high-dimensional data.

(b) Random forest

To create a class that is the mean prediction (regression) or the mode of the classes (classification) of the individual trees, the Random Forest ensemble learning algorithm builds a large number of decision trees during training. Random Forest is well known for its resilience and ability to handle complex data because it is robust and flexible.

(c) Support vector machine

The optimal hyperplane to split two classes of data is found using the discriminative classification algorithm Support Vector Machines (SVM). When the task has a clear distinction between these two classes, SVM performs exceptionally well. In addition, SVM has a reputation for performing well in tasks involving high-dimensional data.3.4 Logistic regression: A statistical model called logistic regression estimates the likelihood of a binary result. A straightforward and easily comprehensible model, logistic regression is frequently applied to tasks with only two possible outcomes.

(d) Decision tree

DT is a classification model that bases its predictions on a structure resembling a tree. Decision trees can be trained on a range of data types and are simple to comprehend and interpret.

(e) AdaBoost

This algorithm for ensemble learning creates a strong classifier by combining several weak classifiers. To increase the influence of the classifiers producing the most accurate predictions on the final prediction, AdaBoost works by boosting the predictions of the 4 weak classifiers.

(f) Bagging

This algorithm, too, uses ensemble learning to combine several examples of the same model to generate a more accurate prediction. Through the bootstrap aggregation of the training data, bagging ensures that every model instance is trained on a distinct subset of the data.

(g) XGBoost

This is an additional gradient-boosting-based ensemble learning algorithm. Gradient boosting is an algorithm that builds decision trees into a model one after the other in a sequential fashion, with each tree being trained to fix the mistakes of the one before it. XGBoost is a potent classifier that is frequently applied to tasks involving large amounts of data and/or features.

(h) Gradient boosting

An algorithm known as ensemble learning, or GB, combines several models to generate more accurate predictions. By gradually adding models to a model, each of which is trained to fix the mistakes of the preceding models, gradient boosting operates.

(iv) Best model identification and hyper-parameter tuning

In an extensive evaluation, Random Forest emerged to be the top-performing algorithm, demonstrating superior predictive capabilities. Encouraged by these results, we conducted further refinement through hyperparameter tuning. A Random Forest model configuration was optimized by systematically exploring the hyperparameter space via a randomized cross-validation search.

Randomized CV search

Unlike an exhaustive search over a specified grid of hyperparameter values, randomized search random samples a defined number of combinations from the hyperparameter space. This approach significantly reduces computation time while still exploring a diverse set of parameter configurations. By efficiently navigating the hyperparameter space, Randomized CV Search enables us to discover optimal settings that enhance the predictive power of our Random Forest model.

Following this, the final model of Random Forest is used for crop recommendation based on the weather parameters.

Phase 2: weather forecasting model

(i) Data division into agro-climatic zones

As a result of the Arabian Sea and Western Ghats, Maharashtra is divided into six regions Fig. 6: Konkan coastal region in the west; Marathwada (Aurangabad region) in the south-east; Vidarbha (Nagpur and Amravati division) in the east, Pune in the west-central region; and Nashik in the north-west; [35]. In the study, we partitioned the temporal weather dataset for Maharashtra into six distinct Agro-climatic divisions. The selected divisions—Nashik, Nagpur, Konkan, Pune, Amravati, and Aurangabad—represent areas with unique Agro-climatic conditions, encompassing variations in topography, soil types, and climate patterns. By segmenting the dataset based on these divisions, we acknowledge the heterogeneity of weather phenomena across the state. This stratification aimed to capture regional variations in weather patterns and optimize the forecasting model's performance by tailoring it to the unique characteristics of each division. This division facilitated a more localized and accurate prediction of weather conditions, acknowledging the diverse Agro-climatic zones within the state.

(ii) LSTM model for weather forecasting

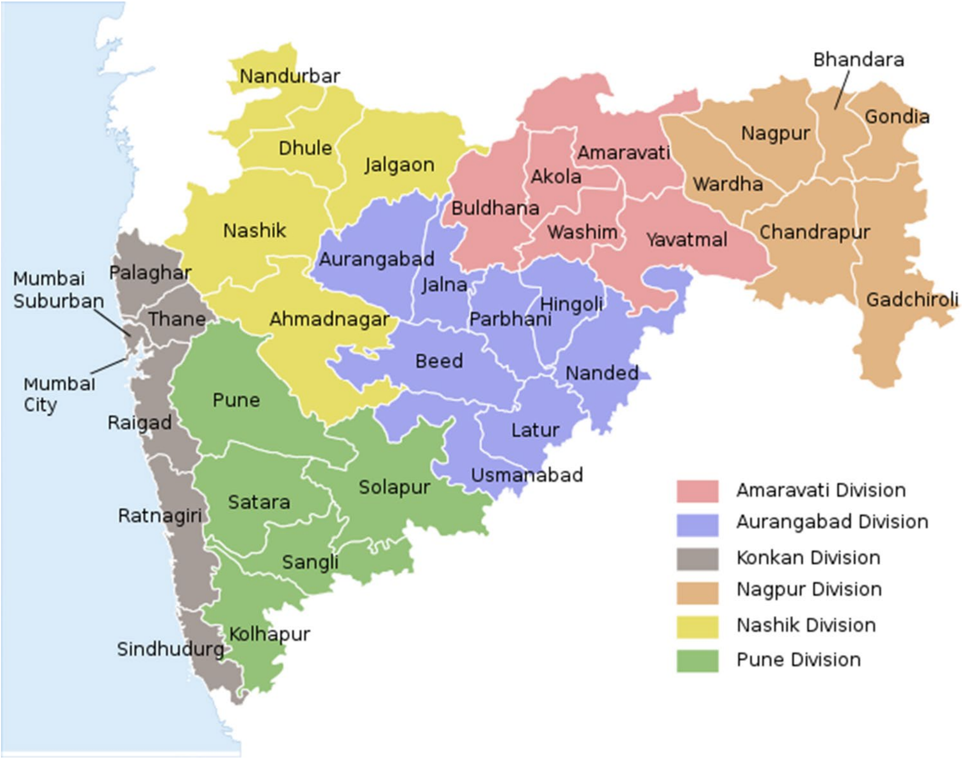


Fig. 6 Maharashtra state agro-climatic divisions source: wikipedia [34]

For weather forecasts, Long Short-Term Memory (LSTM) type is a good choice [36]. Since LSTMs are good at capturing temporal dependencies in time-series data, modeling the sequential nature of weather patterns is a particularly good use for them. Each of the interconnected layers that make up the LSTM architecture has a unique function designed to capture temporal patterns.

LSTM architecture

The linked layers that make up the LSTM architecture are depicted in Fig. 7. each have unique functions designed to capture temporal patterns. Here is a thorough analysis.

(a) Input layer

The input layer is set up to take historical weather data sequences. In this case, each sequence is presented as a time period, like a window which consists of things such as humidity, average sunshine, highest temperature,

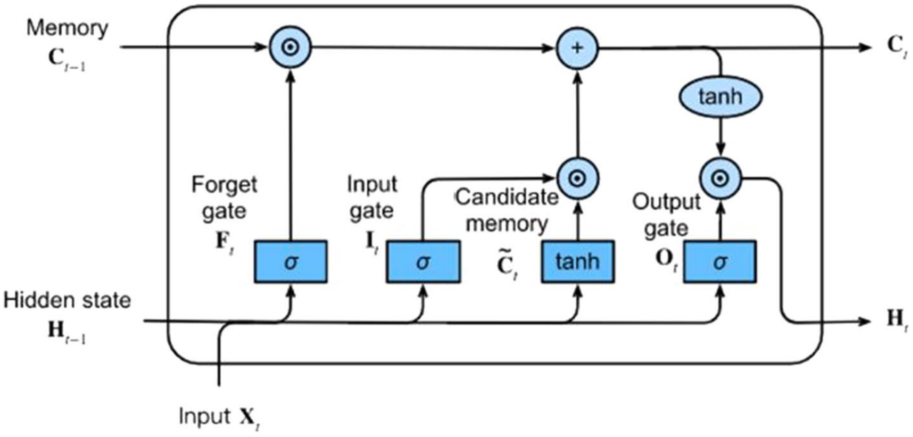


Fig. 7 LSTM architecture [37]

evaporation mean and lowest temperature. The input shape is defined by the sequence of number of features and time steps.

(b) LSTM layer

The LSTM layer comprises cells and gates for memory and is the main component of architecture.Memory cells help to retain the data for a long time.

The LSTM architecture basically consists of three gates: input gate, forget gate and output gate which helps in controlling the data flow in memory cells [38].These gates help in retaining sequential data for long term dependencies such as weather traits and LSTM manages to regain and discard information over time.

1. Input gate: Input gate provides the data to the memory gate, that is kept within the memory cell.It observes the characteristics such as precipitation,humidity,mean sunshine,mean evaporation,mean minimum temperature, and mean maximum temperature in context of weather data to which data is relevant for forecasting future weather status.
2. Forget gate: Forget gate tries to filter irrelevant retained information which would not be further needed for weather data prediction.It eventually decides and handles the data to be deleted from the memory cell.
3. Output gate: The data that is transferred from the memory cell to the subsequent time step is controlled by the output gate. It decides which sections of the data that is stored are relevant to the current prediction. In the case of weather forecasting, the output gate ensures that the LSTM model outputs relevant information about temperature, precipitation, and other features for the upcoming time step. The attributes used in the LSTM model are explained in Table 4.

(c) Dense layer

Following the LSTM layer is a dense layer responsible for delivering the final output. The number of similar features in the input data appears in the count of units in this layer. We apply this layer to map the learned representations from the LSTM layer to the desired output space.

(d) Forecasting

Once trained, the LSTM model is capable of forecasting weather conditions for the next 3 months. The forecasting process involves making sequential predictions, with each prediction influencing subsequent ones, allowing the model to generate realistic and coherent weather scenarios.

Integration with random forest for crop recommendations

Following the LSTM-based weather forecasting for the next three months in each Agro-climatic division, the obtained weather predictions play a pivotal role as essential input features for our crop recommendation system. The LSTM-generated weather forecasts encapsulate valuable information about mean maximum temperature, mean minimum temperature, humidity, precipitation, and mean evaporation, providing a comprehensive temporal context for understanding crop growth dynamics. A careful formatting process is applied to integrate this forecasted data into our crop recommendation system seamlessly. The LSTM-generated weather forecasts are transformed and structured to meet the specific requirements of the Random Forest model which are temperature, rainfall, and humidity. The architecture of the same is outlined in Fig. 8 where the average of Mean Min Temp and Mean Min Temp is taken as the parameter of temperature and the rest are humidity and rainfall. This formatted data, enriched with the temporal insights provided by

Table 4 Parameters of LSTM model

LSTM units	50
Activation function	ReLU
Dense layer units	Number of common features
Batch size	32
Epochs	50
Optimizer	Adam
Loss function	Mean squared error
Forecasting period	3 Months

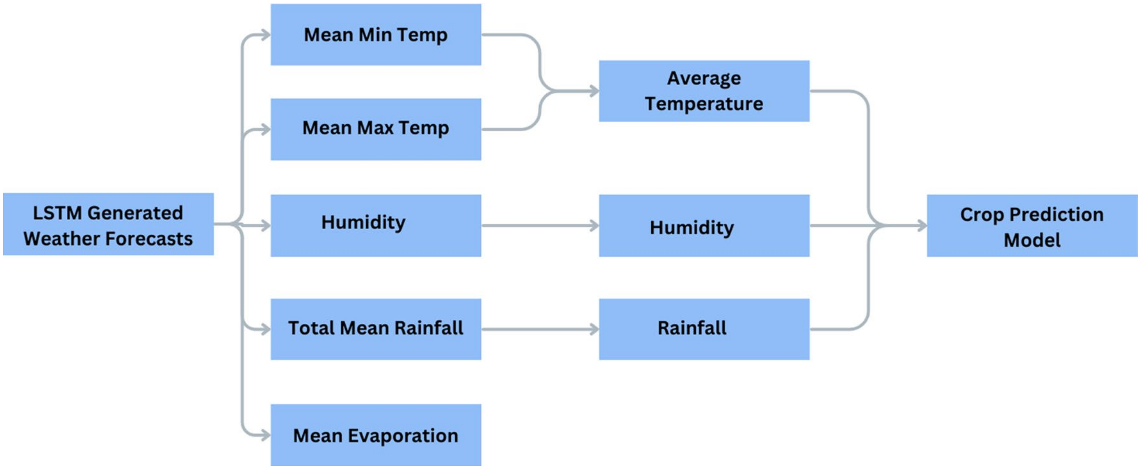


Fig. 8 Parameters for integration with crop prediction model

Table 5 Accuracy of models

Sr. No.	Models used	Mean accuracy
1.	Logistic regression	0.5716
2.	Naïve Bayes	0.9108
3.	Support vector machine	0.7221
4.	K-nearest neighbor	0.8420
5.	Decision tree	0.9102
6.	Bagging	0.9170
7.	AdaBoost	0.1432
8.	Gradient boosting	0.9051
9.	Random forest	0.9188

the LSTM, is then presented as input to the tuned Random Forest model. Leveraging the ensemble learning capabilities of Random Forest, the integrated model becomes adept at predicting optimal crops for cultivation in each Agro-climatic division.

4 Experimental results

The study in the first phase explored various machine learning models, including NB, RF, SVM, LR, DT, AdaBoost, Bagging, XGBoost, and Gradient Boost. Following this, the performance of each model was observed and the results were visualized using an accuracy metric and seen in the bar chart as shown in Table 5. One common metric used to assess a classification model's execution is accuracy. It shows the proportion of accurately predicted incidents to every instance in the dataset. The accuracy of a model can be easily evaluated by looking at its accuracy percentage, which indicates how accurate the model's predictions are overall. Hence it can be seen that random forest showed better accuracy of approximately 91.88% as compared to others. Figure 9 shows the bar chart of the accuracy of respective machine learning models.

$$\text{Accuracy} = \frac{\text{Tp} + \text{Tn}}{\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn}}. \tag{5}$$

From the bar chart, it is clear that the Random Forest performed better than other models., which achieved the highest accuracy of 91.59%. Therefore, we selected Random Forest as our primary model for further experimentation. After hyperparameter tuning. The optimized hyperparameters resulted in a remarkable improvement, achieving

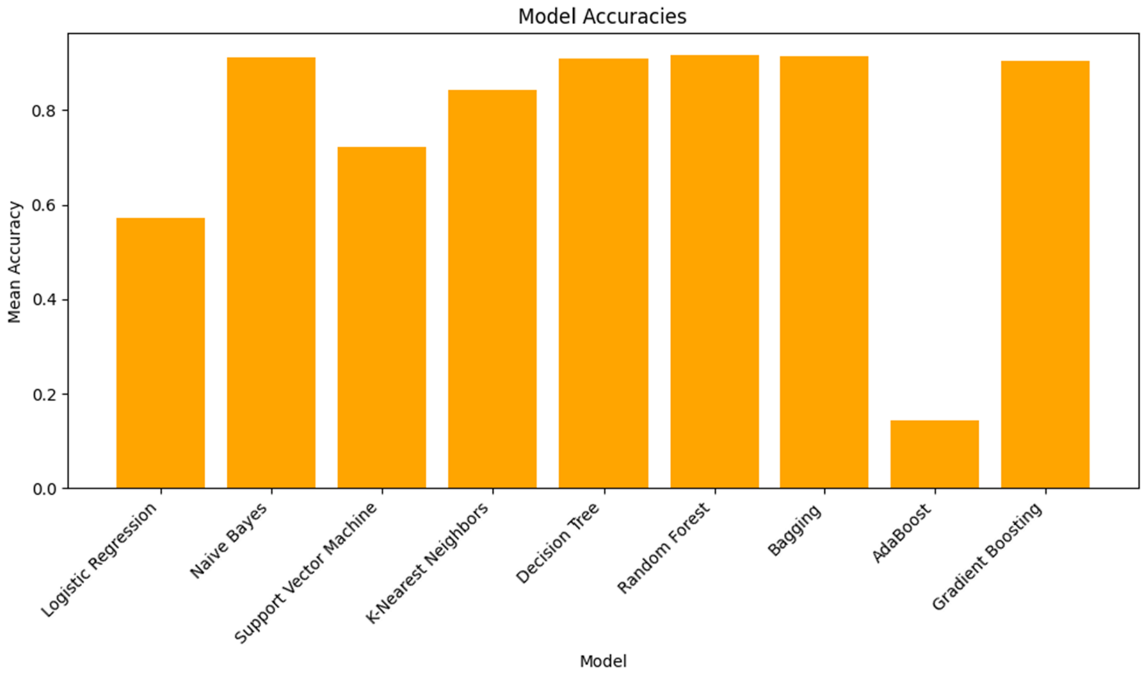
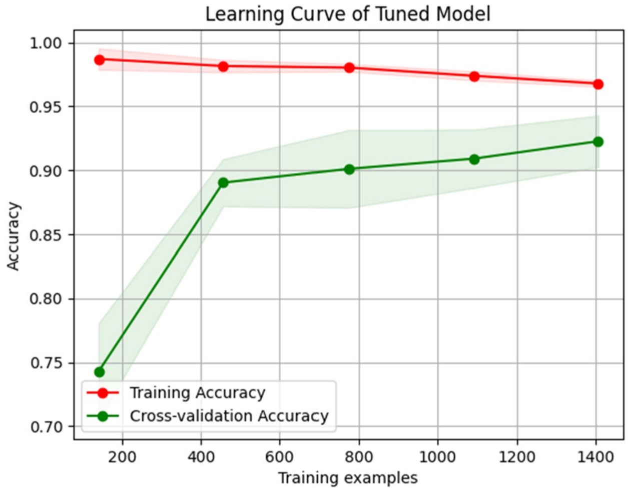


Fig. 9 Bar chart for various model accuracies

Fig. 10 Learning curve of RF

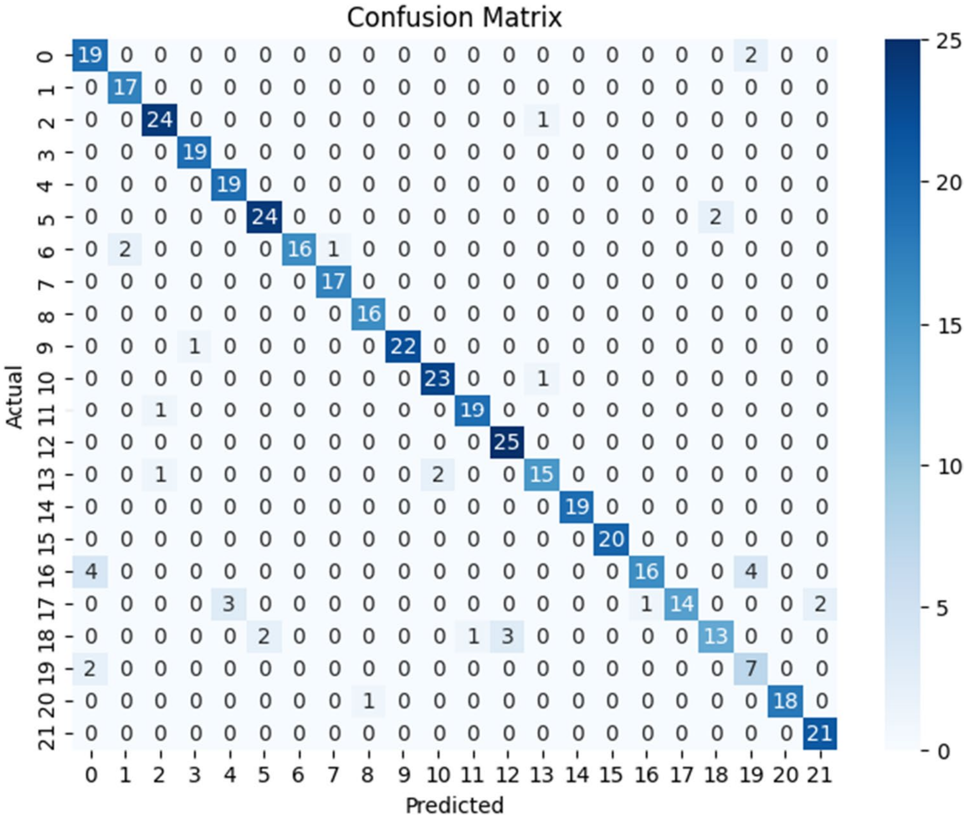


an accuracy of 96.53% on the training data and 92.73% on the test data. The average F1 Score, recall and precision are found to be 93% each. The learning curve plot Fig. 10 illustrates the performance of the models for both the training and validation sets during training iterations. It indicates that the model converges well without overfitting or underfitting.

To further evaluate the performance of our Random Forest model, a confusion matrix is generated as shown in Fig. 11. Confusion matrices summarize the performance of classification models in a tabular format. The matrix provides insights into the correctness of the model's crop classifications. The high values of the diagonal elements indicate accurate predictions, while off-diagonal elements represent some misclassifications.

- Positive true positives (TP): Number of positive predictions made correctly.
- True negatives (TN): Examples that were accurately forecasted as negative.
- False positives (FP): Cases in which the model predicts a positive outcome when the class is negative.
- False negatives (FN): The instances that are miscalculated to be negative (model predicts negative, but class is positive).

Fig. 11 Confusion matrix for RF

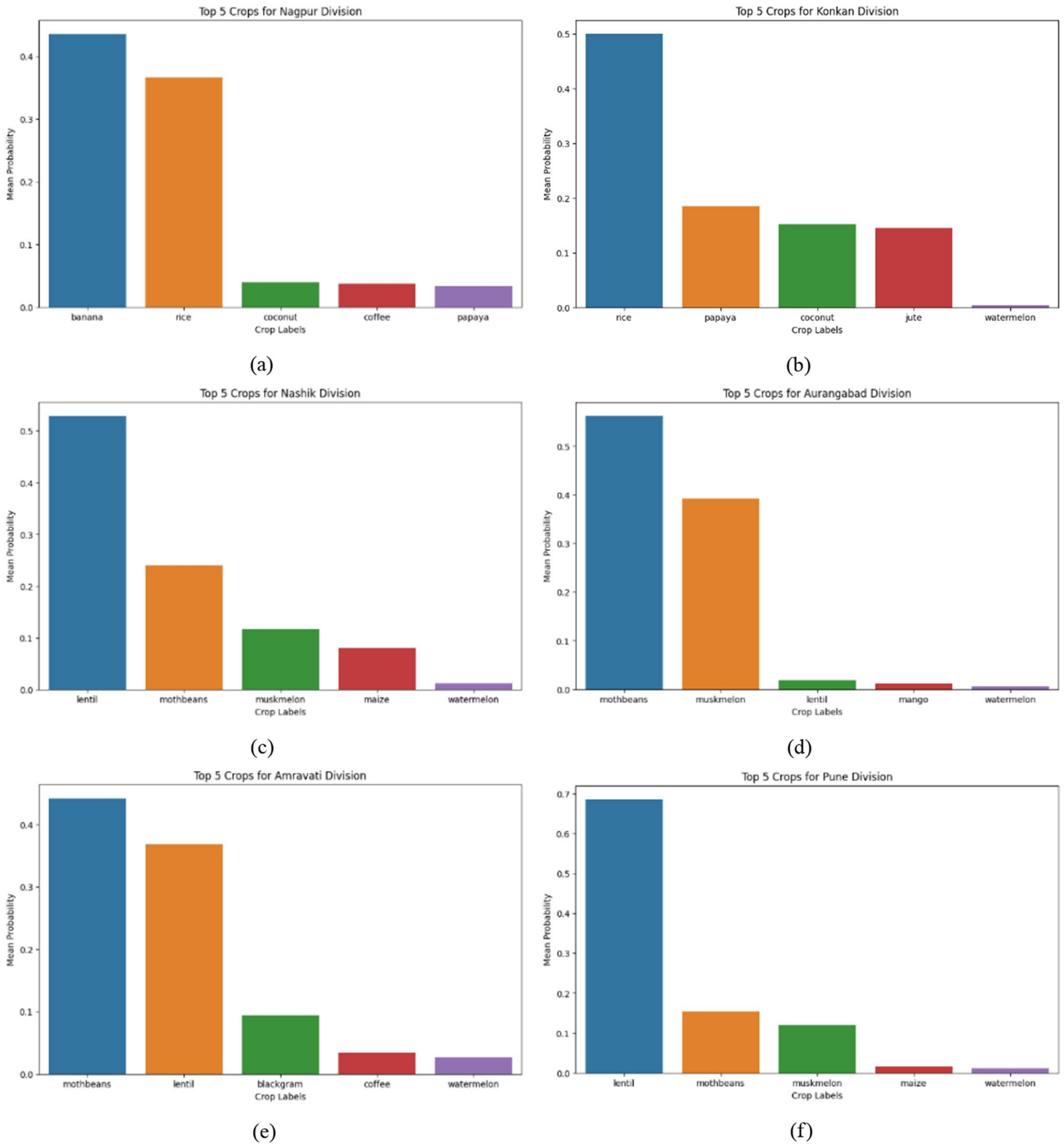


In the second phase, the formatted data generated from the LSTM networks for 6 divisions of Maharashtra state provide input to our best-performing model, Random Forest, for crop recommendation. The top 5 recommended crops based on classification probabilities are shown with the help of a bar chart. The predicted crops of the 6 divisions can be seen in Fig. 12. The graph is shown in Figure 12a for the Nagpur region for the top 5 crops as banana, rice, coconut, coffee, and papaya. Figure 12b shows the top 5 crops rice, papaya, coconut, jute, and watermelon for the Konkan region. Figure 12c presents top-5 crops, namely, lentils, mothbeans, muskmelon, maize, and watermelon for the Nashik region. Figure 12d presents the top 5 crops such as moth beans, muskmelon, lentil, mango, and watermelon for the Aurangabad region. Figure 12e shows the top 5 crops like moth bean, lentil, black gram, coffee, and watermelon for the Amravati region. Figure 12f presents the top 5 crops namely, lentil, moth bean, muskmelon, maize, and watermelon for the Pune region. The bar chart illustrates the top five recommended crops for each of the six divisions within the Maharashtra state, incorporating a three-month forecast and factoring in the probabilities derived from classification. This comprehensive analysis aims to provide strategic insights for agricultural planning and decision-making by highlighting crops that exhibit a higher likelihood of success based on the forecasted conditions. The inclusion of probabilities adds a layer of precision to the recommendations, allowing stakeholders to prioritize crops with a more favorable outlook.

Frontend

An interactive web app is created for the user to get the recommendations for a crop with streamlit application. Here, the user must specify the specific meteorological conditions, such as humidity, rainfall, and temperature. Figure 13 shows the interface of crop recommendation system and Fig. 14 shows how the user interface looks like and the predictions given by the proposed model.

With the given temporal dataset, the weather forecast is done for the next 3 months and displayed to the user. Based on these forecasts, the top 5 crop recommendations are shown for each row. The sample of one division is shown in Fig. 15. The user can do this for all 6 divisions by selecting the division name from the given drop-down menu. Users also get the option to upload the new data of any of the 6 divisions to get the forecast values and the required crop predictions. This is further supported by displaying a crop calendar as depicted in Fig. 16 for Maharashtra state enabling more crop



(a) (b) (c) (d) (e) (f)

Fig. 12 Predicted crops for 6 divisions

scheduling options. The comprehensive experimentation underscores the effectiveness of the Random Forest model across two pivotal aspects: traditional machine learning tasks and its integration with time series forecasting for crop recommendation. An improved view of the model's classification performance is provided using visual representation offering results for strategic crop planning.

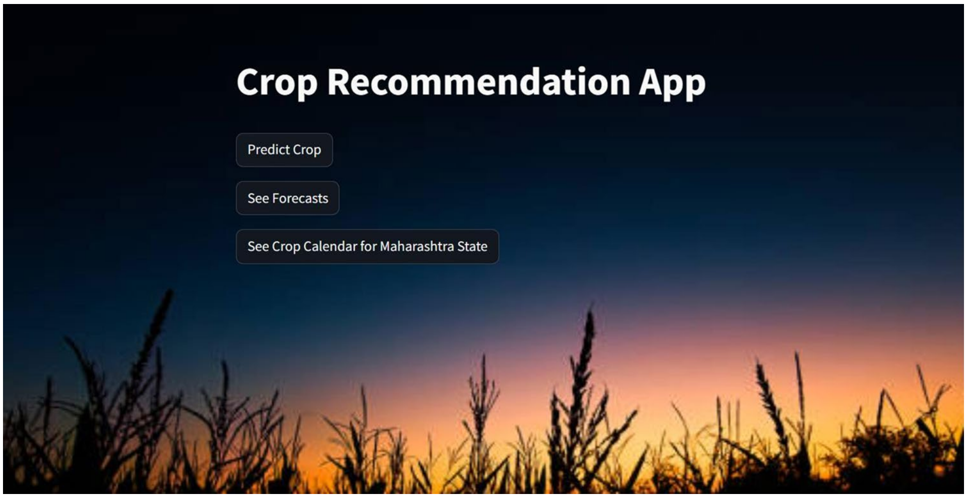
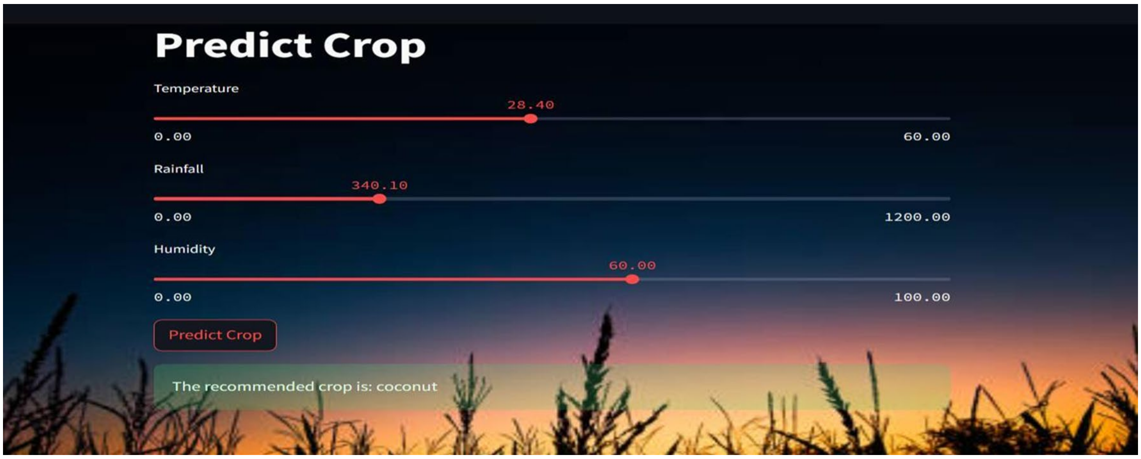


Fig. 13 Overview of the Web App



Temperature 28.40 (0.00 – 60.00)

Rainfall 340.10 (0.00 – 1200.00)

Humidity 60.00 (0.00 – 100.00)

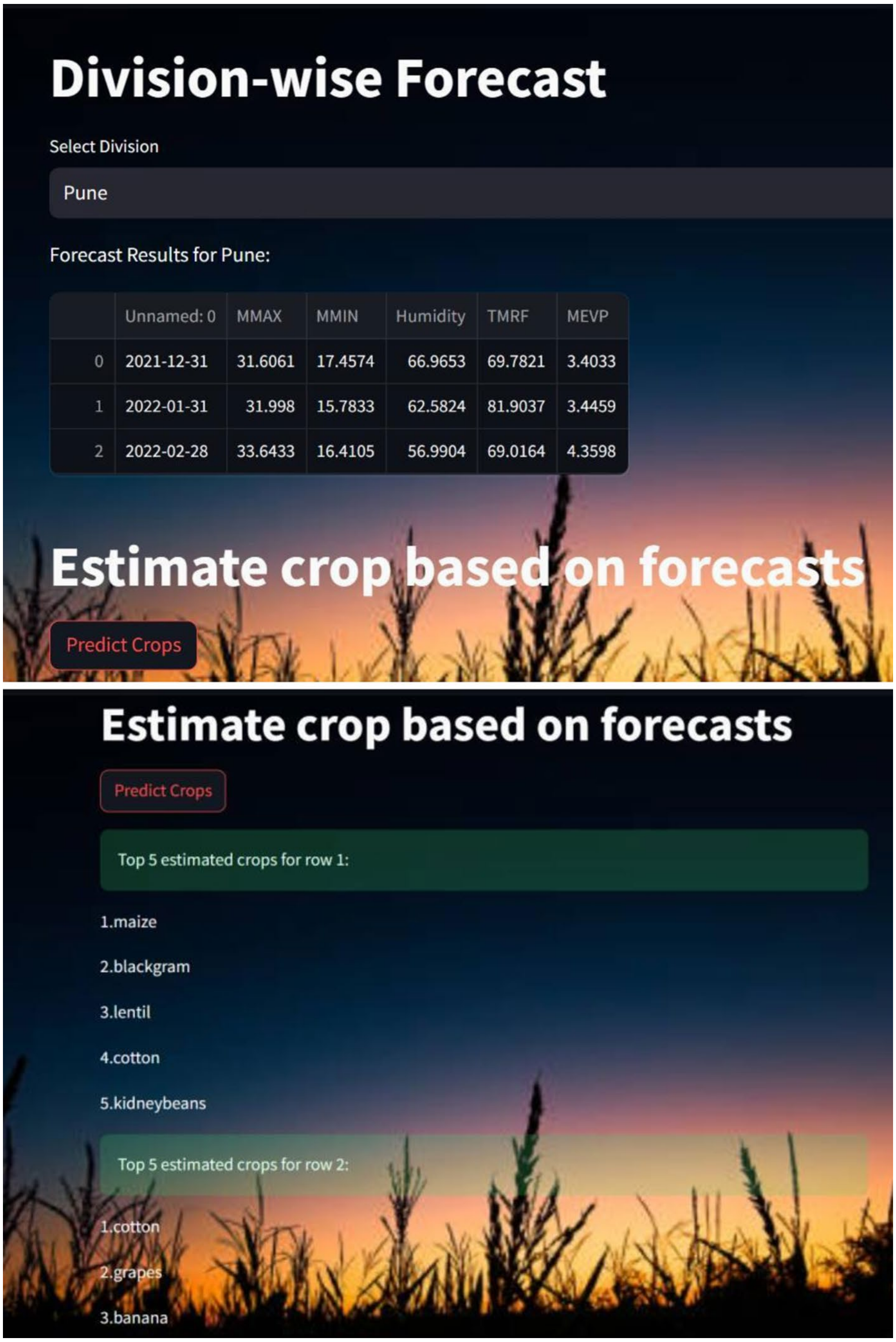
The recommended crop is: coconut

Fig. 14 Crop prediction based on user input

5 Conclusion

Research has culminated in the development of an advanced state of the art decision support system for agriculture, leveraging advanced machine learning and predictive analytics. Every country relies on agriculture. As a result, it must be monitored regularly. The Random Forest algorithm, known for its reliability in handling complex data, has been fine-tuned to enhance prediction accuracy significantly. Additionally, long short term memory interconnected networks have been integrated forecasting climate trends across Maharashtra's regions with high precision. This innovative combination provides farmers with actionable insights for crop management and resource optimization. A major breakthrough has been made in smart agricultural practices within Maharashtra by marrying intricate analysis techniques with localized weather forecasts. In addition, it sets the stage for future progress in creating sustainable, resilient farming methods customized to the climate of the region. In conclusion, this study offers critical contributions toward the development of smarter farming solutions that will improve the productivity of agriculture as well as the sustainability of the system.

Fig. 15 Division wise next 3-months forecasts and the estimated top5 crops



Division-wise Forecast

Select Division

Pune

Forecast Results for Pune:

	Unnamed: 0	MMAX	MMIN	Humidity	TMRF	MEVP
0	2021-12-31	31.6061	17.4574	66.9653	69.7821	3.4033
1	2022-01-31	31.998	15.7833	62.5824	81.9037	3.4459
2	2022-02-28	33.6433	16.4105	56.9904	69.0164	4.3598

Estimate crop based on forecasts

Estimate crop based on forecasts

Top 5 estimated crops for row 1:

1.maize

2.blackgram

3.lentil

4.cotton

5.kidneybeans

Top 5 estimated crops for row 2:

1.cotton

2.grapes

3.banana

6 Future work

Looking ahead, there are promising avenues for further refinement and expansion of the agricultural decision support system, with a continued focus on Maharashtra. An improved dataset with a larger number of attributes can be utilized. While the current model provides a single crop label recommendation, future enhancements could explore presenting

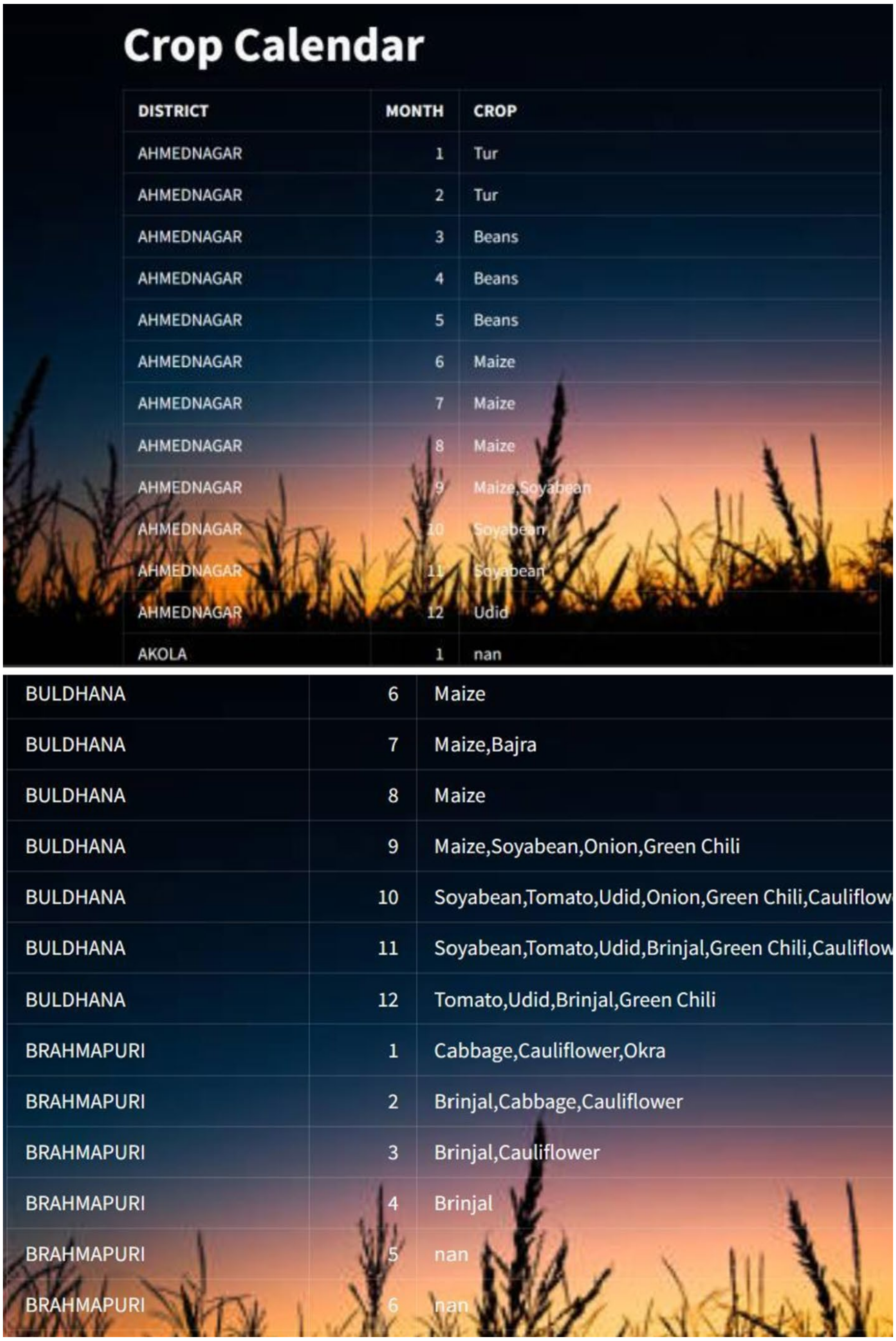


Fig. 16 Maharashtra state temporary crop calendar

Crop Calendar

DISTRICT	MONTH	CROP
AHMEDNAGAR	1	Tur
AHMEDNAGAR	2	Tur
AHMEDNAGAR	3	Beans
AHMEDNAGAR	4	Beans
AHMEDNAGAR	5	Beans
AHMEDNAGAR	6	Maize
AHMEDNAGAR	7	Maize
AHMEDNAGAR	8	Maize
AHMEDNAGAR	9	Maize,Soyabean
AHMEDNAGAR	10	Soyabean
AHMEDNAGAR	11	Soyabean
AHMEDNAGAR	12	Udid
AKOLA	1	nan

BULDHANA	6	Maize
BULDHANA	7	Maize,Bajra
BULDHANA	8	Maize
BULDHANA	9	Maize,Soyabean,Onion,Green Chili
BULDHANA	10	Soyabean,Tomato,Udid,Onion,Green Chili,Cauliflow
BULDHANA	11	Soyabean,Tomato,Udid,Brinjal,Green Chili,Cauliflow
BULDHANA	12	Tomato,Udid,Brinjal,Green Chili
BRAHMAPURI	1	Cabbage,Cauliflower,Okra
BRAHMAPURI	2	Brinjal,Cabbage,Cauliflower
BRAHMAPURI	3	Brinjal,Cauliflower
BRAHMAPURI	4	Brinjal
BRAHMAPURI	5	nan
BRAHMAPURI	6	nan

additional information. For instance, considering the probabilities associated with other potential crops in the top five predictions would offer farmers a more comprehensive view of viable alternatives based on the forecasted weather conditions. Future studies can concentrate on various enhancements and verification methods to boost the effectiveness and reliability of the system through historical validation and ground truth comparison to assess the crop recommendation component's performance. Additionally, more research has to be done to determine how useful the supplemental crop calendar is in providing monthly planting guidelines particular to each region. Farmers and agricultural experts opinions can be used to determine whether more regular and detailed instruction is needed. Examining the viability of offering district-specific planting guidelines every two weeks or more frequently, akin to current programs like CRIDA's Contingency Plans or IMD AgriMet endeavors, can augment the significance and use of the calendar. Efforts can be made to adapt the system for other states with similar agricultural landscapes, potentially creating a scalable model for precision agriculture in diverse regions of India.

Author contributions Yashashree Mahale: Conceptualization, Methodology, Formal Analysis, Visualization, Writing; original draft preparation. Nida Khan: Conceptualization, Methodology, Formal Analysis, Visualization, Writing; original draft preparation. Kunal Kulkarni: Conceptualization, Methodology, Formal Analysis, Visualization, Writing; original draft preparation. Shivali Amit Wagle : Methodology, Investigation, Supervision, Writing; original draft preparation. Preksha Pareek : Methodology, Investigation, Supervision, Writing; original draft preparation. Ketan Kotecha : Writing review and editing, Administration, Validation, Project Administration. Tanupriya Choudhury: Writing review; and editing, Administration, Validation, Project Administration. Ashutosh Sharma: Writing; review and editing, Administration, Validation.

Funding This research received no external funding.

Data availibility Sensitive reasons prevent the data supporting the study's conclusions from being publicly published, although they are available from the corresponding author upon justifiable request. The India Meteorological Department in Pune, India has controlled access data storage where the data is stored [39].

Declarations

Competing interests No conflict of interest is declared by the authors.

Declaration of generative AI in scientific writing The authors have not used AI tools in scientific paper writing. The authors have checked for plagiarism with Turnitin software.

References

1. Durai SKS, Shamili MD. Smart farming using machine learning and deep learning techniques. Decis Anal J. 2022;3: 100041.
2. Sharma P, Dadheech P, Senthil ASK. Ai-enabled crop recommendation system based on soil and weather patterns. In: Artificial Intelligence Tools and Technologies for Smart Farming and Agriculture Practices, pp. 184–199. IGI Global, 2023.
3. Elavarasan D, Vincent PD. Crop yield prediction using deep reinforcement learning model for sustainable Agrarian applications. IEEE Access. 2020;8:86886–901.
4. Cai Y, Guan K, Peng J, Wang S, Seifert C, Wardlow B, Li Z. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sens Environ. 2018;210:35–47.
5. Ed-daoudi R, Alaoui A, Ettaki B, Zerouaoui J. A predictive approach to improving agricultural productivity in morocco through crop recommendations. Int J Adv Comput Sci Appl 2023;14(3).
6. Moon MH, Marjan MA, Uddin MP, Ibn Afjal M, Kadry S, Ma S, Nam Y. Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. Front Plant Sci. 2023;14:1234555.
7. Nayak HS, Silva JV, Parihar CM, Krupnik TJ, Sena DR, Kakraliya SK, Jat HS, Sidhu HS, Sharma PC, Jat ML, et al. Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in northwest India. Field Crops Res. 2022;287: 108640.
8. Van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. Comput Electron Agric. 2020;177: 105709.
9. Banerjee S, Mondal AC. A region-wise weather data-based crop recommendation system using different machine learning algorithms. Int J Intell Syst Appl Eng. 2023;11(3):283–97.
10. Su Y-x, Xu H, Yan L-j. Support vector machine-based open crop model (sbocm): case of rice production in china. Saudi J Biol Sci. 2017;24(3):537–47.
11. Kedlaya A, Sana A, Bhat BA, Kumar S, Bhat N, et al. An efficient algorithm for predicting crop using historical data and pattern matching technique. Global Transit Proc. 2021;2(2):294–8.
12. Samuel P, Sahithi B, Saheli T, Ramanika D, Kumar NA. Crop price prediction system using machine learning algorithms. Quest J Softw Eng Simul. 2020.
13. Gupta T, Maggu S, Kapoor B. Crop prediction using machine learning. 2023.
14. Chandana C, Parthasarathy G. Efficient machine learning regression algorithm using naïve Bayes classifier for crop yield prediction and optimal utilization of fertilizer. Int J Performabil Eng. 2022;18(1).
15. Todmal RS. Future climate change scenario over Maharashtra, western India: implications of the regional climate model (remo-2009) for the understanding of agricultural vulnerability. Pure Appl Geophys. 2021;178(1):155–68.

16. Cedric LS, Adoni WYH, Aworka R, Zoueu JT, Mutombo FK, Krichen M, Kimpolo CLM. Crops yield prediction based on machine learning models: case of west African countries. Smart Agric Technol. 2022;2: 100049.
17. Alebele Y, Wang W, Yu W, Zhang X, Yao X, Tian Y, Zhu Y, Cao W, Cheng T. Estimation of crop yield from combined optical and sar imagery using Gaussian kernel regression. IEEE J Sel Topics Appl Earth Observ Remote Sens. 2021;14:10520–34.
18. Nti IK, Zaman A, Nyarko-Boateng O, Adekoya AF, Keyeremeh F. A predictive analytics model for crop suitability and productivity with tree-based ensemble learning. Decis Anal J. 2023;8: 100311.
19. Liu J, Yang K, Tariq A, Lu L, Soufan W, El Sabagh A. Interaction of climate, topography and soil properties with cropland and cropping pattern using remote sensing data and machine learning methods. Egypt J Remote Sens Space Sci. 2023;26(3):415–26.
20. Johnston DB, Pembleton KG, Huth NI, Deo RC. Comparison of machine learning methods emulating process driven crop models. Environ Modell Softw. 2023;162: 105634.
21. Raja S, Sawicka B, Stamenkovic Z, Mariammal G. Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. IEEE Access. 2022;10:23625–41.
22. Ghadge R, Kulkarni J, More P, Nene S, Priya R. Prediction of crop yield using machine learning. Int Res J Eng Technol (IRJET). 2018;5:2237–9.
23. Devan K, Swetha B, Sruthi PU, Varshini S. Crop yield prediction and fertilizer recommendation system using hybrid machine learning algorithms. In: 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), pp. 171–175, 2023. IEEE.
24. Rashid M, Bari BS, Yusup Y, Kamaruddin MA, Khan N. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. IEEE Access. 2021;9:63406–39.
25. Bali N, Singla A. Emerging trends in machine learning to predict crop yield and study its influential factors: a survey. Arch Comput Methods Eng. 2022;29(1):95–112.
26. Zhang C, Di L, Lin L, Guo L. Machine-learned prediction of annual crop planting in the us corn belt based on historical crop planting maps. Comput Electron Agric. 2019;166: 104989.
27. Shafi U, Mumtaz R, Iqbal N, Zaidi SMH, Zaidi SAR, Hussain I, Mahmood Z. A multi-modal approach for crop health mapping using low altitude remote sensing, internet of things (iot) and machine learning. IEEE Access. 2020;8:112708–24.
28. Aditya D, Manvitha R, Mouli CR. Detect-o-thon: identification of infected plants by using deep learning. Global Trans Proc. 2021;2(2):336–43.
29. Bauer J, Aschenbruck N. Towards a low-cost rssi-based crop monitoring. ACM Trans Internet Things. 2020;1(4):1–26.
30. Crop Recommendation Dataset. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset.
31. Cheng L, Chen J, Mao Y, Liao C, Zhu Q. Expectation maximization algorithm for gps positioning in multipath environments based on volterra series. Circ Syst Signal Process. 2023: 1–18.
32. Kulkarni K, Mahale Y, Khan N, Nandhini K, Gite S. Deep learning for anomaly detection in spatio-temporal Maharashtra weather data: a novel approach with integrated data cleaning techniques. Int J Intell Syst Appl Eng. 2024;12(12s):169–82.
33. Shaheen M, Naheed N, Ahsan A. Relevance-diversity algorithm for feature selection and modified Bayes for prediction. Alexandria Eng J. 2023;66:329–42.
34. Agro Climatic Zones. https://en.wikipedia.org/wiki/List_of_districts_of_Maharashtra.
35. Sali V, Nagrale D, Sushir M, Kadam D, Dighule S, Deshmukh D. Occurrence, diversity and characterization of effective soil yeast isolates from different agro climatic zones of Maharashtra. 2023.
36. Venkatachalam K, Trojovskỳ P, Pamucar D, Bacanin N, Simic V. Dwfh: an improved data-driven deep weather forecasting hybrid model using transductive long short term memory (t-lstm). Expert Syst Appl. 2023;213: 119270.
37. LSTM Architecture. https://d2l.ai/chapter_recurrent-modern/lstm.html.
38. Salman AG, Heryadi Y, Abdurahman E, Suparta W. Single layer & multi-layer long short-term memory (lstm) model with intermediate variables for weather forecasting. Proc Comput Sci. 2018;135:89–98.
39. Indian Meteorological Department data supply portal. https://dsp.imdpune.gov.in/. Accessed on 15 Jan 2024.