Machine Learning-Powered Forecasting of Climate Conditions in Smart Greenhouse Containing Agriculture

Ayesha Chowdhury[1], Farhana Jaman[2], Kayes Rasha Md Samaun[2],

*Ahmed Wasif Reza Sir[1]

Department of Computer Science and Engineering, East West University,
Dhaka, Bangladesh

Abstract. The effective control of greenhouse environments is essential for enhancing crop productivity and quality. This study focuses on forecasting key environmental factors temperature, relative humidity, CO2$_2$ levels in the greenhouse using ML models. Four models Random Forest, Xgbooster, CatBoost, Sarimax were developed and evaluated for their predictive performance. Among these, the CatBoost model showed better accuracy, achieving an R^2 value of up to 0.957 for temperature and humidity predictions. The study highlights the importance of data preprocessing, including lag feature engineering, to capture seasonal and temporal patterns effectively.

1 Introduction:

Agriculture and climate change are very dependent on each other. . As every crop have their different capabilities of adaptation in different weather conditions. Farmers also face a big loss because of certain climate changes. As climate conditions like excess heat, storm, cyclone or these alike situations can cause great harm to crop's growth. As day by day technologies have become more advanced so by the help of using different machine learning models now it has become possible to give accurate predictions on future climate conditions and their effect on crop production.

Despite having the benefits of machine learning models we have faced many challenges as some of them could not detect the right time series pattern of previous years to give a good accuracy based prediction for the future. Though having these challenges we have captured some of the good performing models for our research based purpose. Our research mainly focuses on

- Observing different machine learning model's performance of giving time series predictions on weather conditions
- Understanding the previous research findings and their observations
- Recovering previous research gaps by giving long term based time series predictions.
- Analyzing the impact of climate conditions on crop production.

In our research section 2 focuses on analyzing the previous research findings, section 3 focuses on the dataset description, their preprocessing and observing different machine learning model's predictions, section 4 focuses on analysis of the results of prediction and how our findings fulfill previous research gaps, section 5 focuses on our findings limitation and planning for future work.

2 Literature Review:

Researchers have explored diverse approaches to make time series predictions based on weather conditions containing agriculture. They have done this experiment using different sorts of deep learning models for finding the exact accuracy. Table 2.1 has information of some similar experiments.

2.1 Related Work:

The table shows key findings, objectives and limitations of research paper related to this topic

Table 1. Related Work Discussion

Ref	Objective	Highest Performing Model	Performance Metrics	Limitations	Research gaps
1	Develop a ML model to predict Greenhouse environment factors (Temperature, Humidity, CO2)	XGBoost	Temperature: R^2=0.9724,RMSE = 0.5471 Humidity: R^2=0.9656, RMSE=3.2789 CO_2R^2=0.9929,RMSE=9.59	Incomplete data, models may under or over predict due to this. Might face generalization problems	Ignoring the factor of predicting rainfall and the experiment is only done for one specific crop based
2	Develop a hybrid DF-RF-ANN ML model for Greenhouse to reduce dependency on IoT devices	DF-RF-ANN	Temperature: R^2=0.72,MAE=1.9C Humidity:R^2=0.68, MAE = 0.10 PAR:R^2=0.75, MA=76.39µmol/m^2/s $CO_{2:}$ R^2 = 0.59, MAE = 10.14 ppm	Inaccurate CO_2 prediction, external meteorological data used for prediction caused noise, more input data increased complexity which reduced accuracy	Prediction of good accuracy based are missing and this predictions are done for only short term based time series

3	Improve ML models internal temperature prediction by analyzing synthetic data produced by GANs	CNN+LSTM	Summer for 15 min interval: R^2 = 0.935, RMSE = 1.667° C MAE = 1.239°C Winter for 30 min interval,R^2=0.919, RMSE=0.932°C, MAE = 0.734°C	Impact of seasons and time intervals is high for result, risk of over fitting data due to GAN-generated data	Weather conditions like (temperature, humidity and rainfall) like more in detailed based predictions are missing and only have shown the weather based prediction not the crop production effect
4	Enhance prediction accuracy of environment factors for Greenhouse using BiGRU-Attention and LightGBM	PSO-BiGRU-Attention-LightGBM	Temperature: R^2=0.9902, RMSE = 0.5578,MAE=0.317 Humidity: R^2 = 0.9825,RMSE=2.2327,MAE=1.1981 PAR:R^2=0.8871, RMSE=39.1562, MAE = 18.4348	Poor PAR prediction due to imbalance dataset, accuracy reduction for long term prediction	Only have done the predictions by limited machine learning model based and the predictions are based on only short term based time series

3 Methodology:

The aim of the methodology is to provide sufficient detail for readers to understand the study. Figure 1 shows the steps of this experiment.

3.1 Dataset Description:

In our dataset there are two sets of district wise datasets of West Bengal. One is for Burdwan and another is for Hooghly. The Burdwan district dataset is almost like 355 and the Hooghly district dataset is almost like 246. There are 11 columns for each dataset. Both dataset are mainly about telling the weather condition effects on agriculture. For each year there are six seasons for each crop available showing the effects of weather conditions on the crop production. There are four categorical columns and the rest are numerical columns. Columns like state_name, district_name, crop_year, season , crop, area, production, temperature, rainfall, humidity and sun_hours.

As our predictions are many we are doing two types of preprocessing here.

3.2 Preparation of datasets for training and testing:

- Preprocessing dataset for time series weather forecasting:
  As forecasting for weather conditions we have mostly focused on weather related columns which really are much correlated for prediction of time series weather forecasts and have done the preprocessing following steps of the left side in Figure: 1

- Preprocessing dataset for crop production:
  As for prediction on crop production most of the columns have their own different effects on crop production so we have preprocessed the dataset for crop production prediction following the steps of the right side in Figure: 1

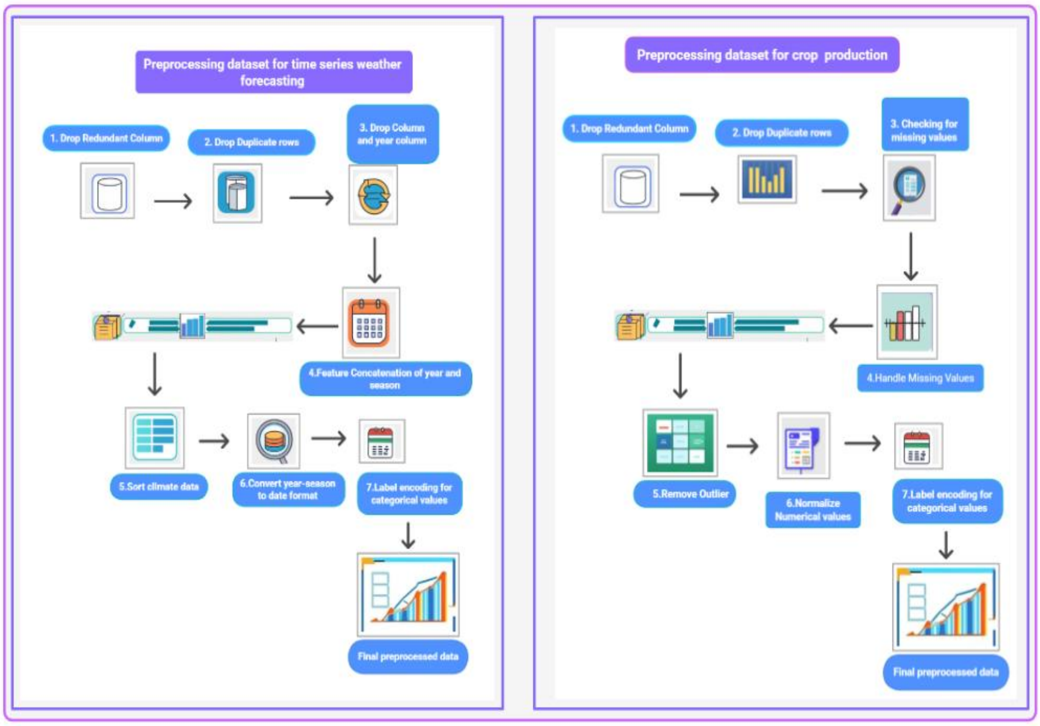
Preprocessing dataset for time series weather forecasting
1. Drop Redundant Column
2. Drop Duplicate rows
3. Drop Column and year column
4.Feature Concatenation of year and season
5.Sort climate data
6.Convert year-season to date format
7.Label encoding for categorical values
Final preprocessed data
Preprocessing dataset for crop production
1. Drop Redundant Column
2. Drop Duplicate rows
3. Checking for missing values
4.Handle Missing Values
5.Remove Outlier
6.Normalize Numerical values
7.Label encoding for categorical values
Final preprocessed data

Fig. 1. Preprocessing Dataset for time series weather forecasting and Crop production

3.3 Evaluation and discussion:

This study mainly focuses on predicting two sets of problems

- Time Series forecasting on weather conditions based on (Temperature, rainfall, humidity)
- Prediction for crop production

Time Series forecasting on weather conditions based on (Temperature, rainfall, humidity):

As in our datasets there are year and monthly based weather conditions. So we have tried to do a time series weather forecasting based on these monthly and yearly based data. After the weather data preprocessing we have created lag values for (temperature, rainfall, humidity) column data each . The lag values are basically past year and month values of temperature, rainfall and humidity dataset. The present values are treated as target values for prediction. Then from the lagged values 80% lag values are trained to understand the weather pattern for each month and year and 20% & 10% of the lagged data are being used for testing and prediction. These steps are done the same for temperature, rainfall and humidity each. We have used different parameters

R2, RMSE, MAE to evaluate how well it is predicting future values of weather conditions by using different machine learning models.

Machine Learning Models:.

- **Arima(Sarimax Model):** The Sarima model is the upgraded version of the Arima model. This is the upgraded version because it can forecast time series weather conditions within the seasonality pattern. We have used this model because we have the year and monthly based seasonality patterns on the weather conditions (temperature, humidity, rainfall). This model uses auto_arima function which generates the parameter values by itself instead of generating manual parameters like (p, d, q, P, D, Q, s).
  p means the number of auto regressive stages.
  d means the apart value for making series stationary
  q means moving average stages.
  (P, D, Q) means almost the same but the difference here is they are all used in seasonality patterns.

- **Random Forest Model:** For predicting continuous type results Random Forest model is used. It uses a “decision tree” which acts like a flowchart by spilling the data into small segments to make a final decision. These small segments all together behave like a team for generating the final result. Lag values are given as inputs for getting the time series forecasting.

- **Cat Boost Regressor:** Cat Boost regressor is also a machine learning model which is used for predicting continuous type results but here it performs more accurately than the other models (random Forest model). Here it creates a single decision tree at a time rather than creating all trees at a time. Each tree creates an update prediction by analyzing and correcting the previous prediction errors. This is how by correcting and updating each decision tree predictions it gives the final result. Here also lag values are used as input for time series prediction.

Machine Learning Model Performance and their Visualization Results:

As we have tried to forecast time series weather condition results for weather conditions like humidity, temperature and rainfall for Burdwan and Hooghly district of West Bengal. The results of each machine learning models and their visualization are given below:

- **For Burdwan District of West Bengal:**

Using Random Forest Regressor on the basis of 80% lag data trained and prediction on 20 % test data:

— Using Random Forest Regressor:

Table 2 . Forecasting future values of climate conditions using Random Forest Regressor:

Weather Condition	RMSE	MAE	R2
Temperature	2.4470	1.5003	0.819
Humidity	3.3659	2.3928	0.897
Rainfall	123.143	78.375	0.950

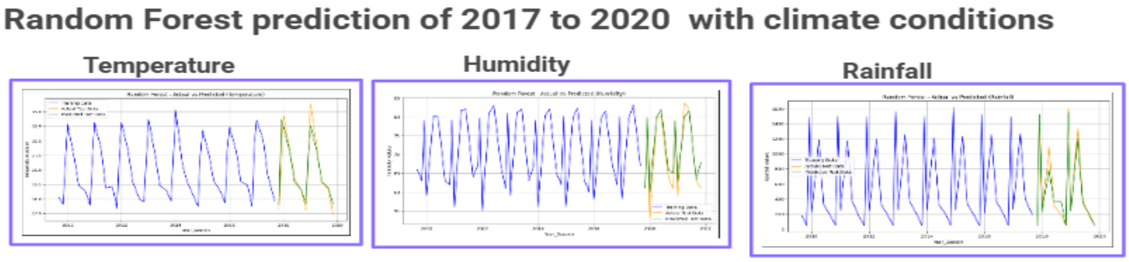

Fig. 2. Time Series predictions of future year's weather conditions using Random Forest

— Using Xgbooster:

Table 3. Forecasting future values of climate conditions using Xgboost Regressor:

Weather Condition	RMSE	MAE	R2
Temperature	2.7429	1.9568	0.773
Humidity	3.8183	2.7294	0.867
Rainfall	263.9598	116.0128	0.771

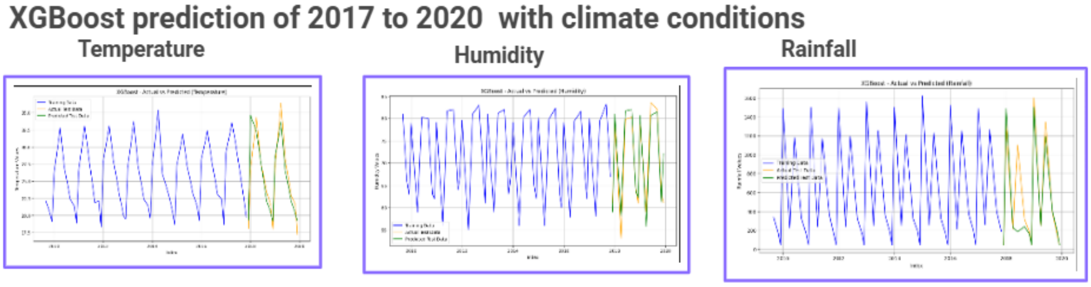

Fig. 3. Time Series predictions of future year’s weather conditions using XGBoost

— Using CatBoostRegressor:

Table 4. Forecasting future values of climate conditions using CatBoostRegressor:

Weather Condition	RMSE	MAE	R2
Temperature	1.505	1.108	0.931
Humidity	2.557	1.885	0.940
Rainfall	113.788	83.457	0.957

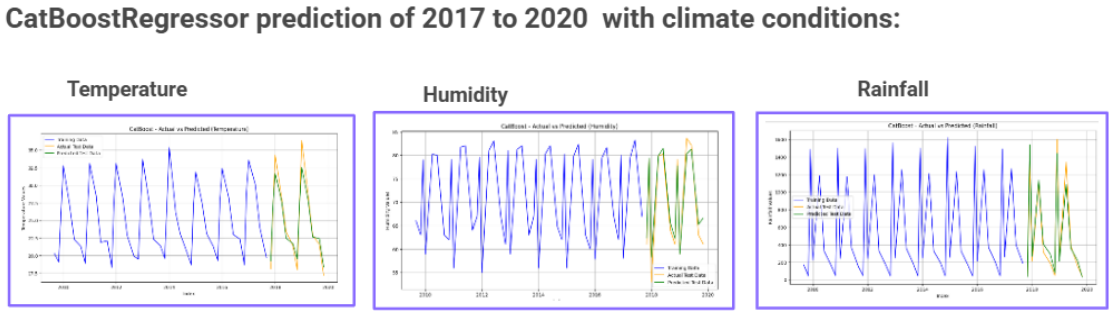

Fig. 4. Time Series predictions of future year’s weather conditions using CatBoostRegressor

— Using Sariamx (only for temperature):

Table 5. Forecasting future values of climate conditions using Sarimax

Weather Condition	RMSE	MAE	R2
Temperature	1.505	1.108	0.931

In figure 2, 3 and 4 there are different machine learning model's prediction based visualizations of Burdwan district where the blue lines analyzing the patterns about lagged values of previous year's weather conditions (temperature, humidity and rainfall), the green lines are the future year based prediction lines of (2017- 2020) by each models and the yellow lines are the actual values of each weather condition. By this we can easily understand how accurately they are predicting time series weather conditions. Among all of them CatBoostRegressor is giving the best predictions.

The confusion matrix and the learning curves of each weather conditions by CatBoostRegressor

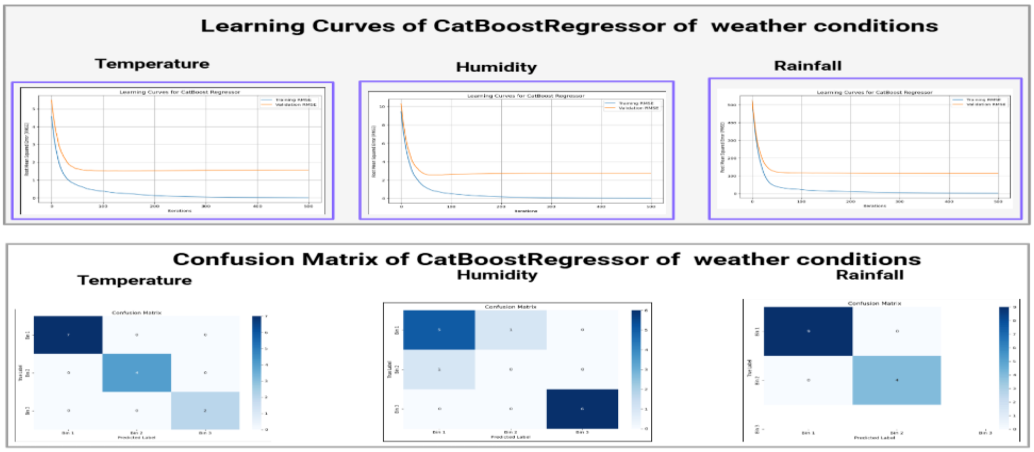

Fig. 5. The learning curves and confusion matrix plots of each weather conditions

In figure 5 the learning curves and confusion matrix is shown of weather conditions (temperature, humidity, rainfall). The learning curves are shown based on RMSE. The confusion matrix is for analyzing the accuracy of a model's prediction performance by binning them in different categories.
Visualization of the comparison between best two performing models using shap are given below:

In Figure 6 these graphs with shap values mainly show which lagged value has really impacted the most on each time series predictions of weather conditions.

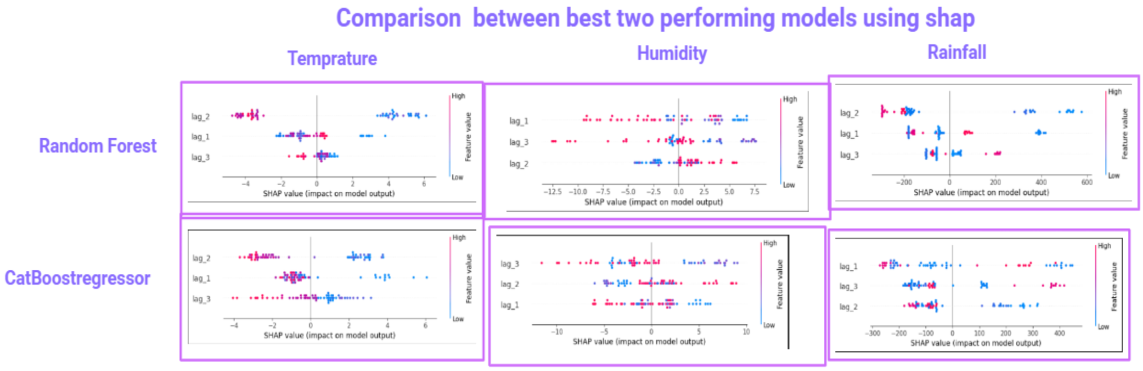

Fig. 6. Lag values effects and their visualization of different weather conditions using shap

- **For Hooghly District of West Bengal:**

Using Random Forest Regressor on the basis of 80% lag data trained and prediction on 20 % test data

— **Using Random Forest Regressor:**

Table 6. Forecasting future values of climate conditions using Random Forest Regressor:

Weather Condition	RMSE	MAE	R2
Temperature	0.833	0.72361	0.967
Humidity	3.846	2.7425	0.793
Rainfall	82.0286	60.4753	0.982

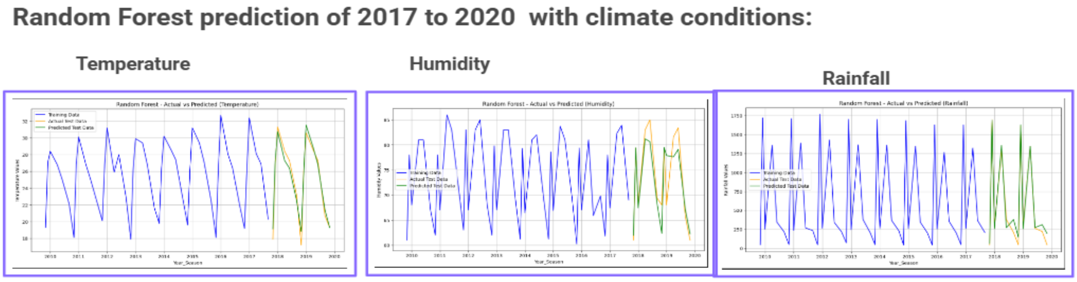

Fig. 7. Time Series predictions of future year’s weather conditions using Random Forest

— **Using Xgbooster:**

Table 7. Forecasting future values of climate conditions using Xgboost Regressor:

Weather Condition	RMSE	MAE	R2
Temperature	1.190	0.965	0.933
Humidity	5.569	3.957	0.567
Rainfall	309.505	107.940	0.744

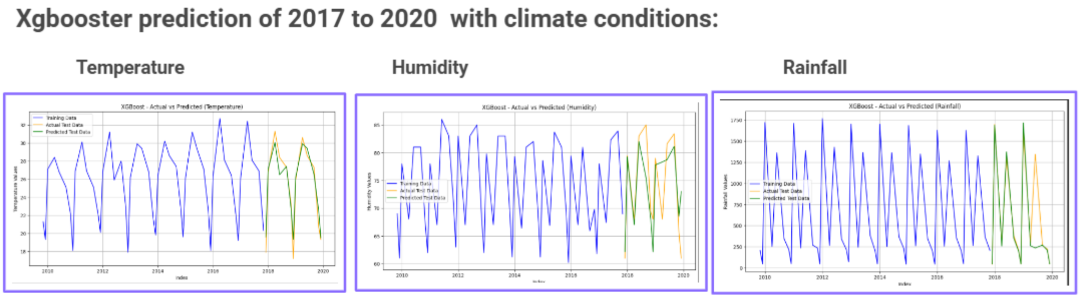

Fig. 8. Time Series predictions of future year’s weather conditions using XGBoost

— **Using CatBoostRegressor:**

Table 8. Forecasting future values of climate conditions using CatBoostRegressor

Weather Condition	RMSE	MAE	R2
Temperature	0.884	0.754	0.963
Humidity	4.12	3.101	0.762
Rainfall	145.790	95.650	0.943

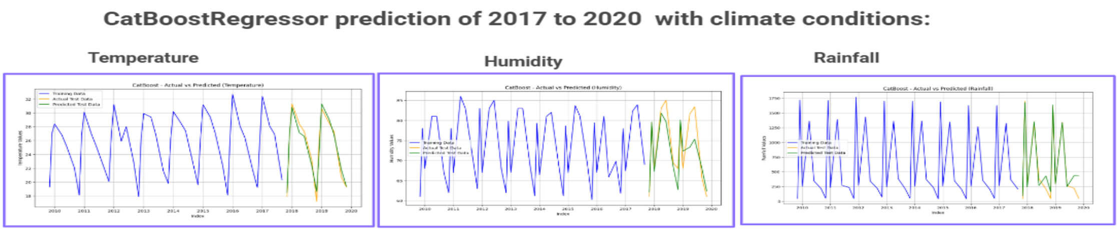

Fig. 9. Time Series predictions of future year’s weather conditions using CatBoostRegressor

— Using Sariamx (only for temperature):

Table 9. Forecasting future values of climate conditions using Sarimax:

Weather Condition	RMSE	MAE	R2
Temperature	2.709	2.131	0.725

In figure 7, 8 and 9 there are different machine learning model's prediction based visualizations of Hooghly District where the blue lines analyzing the patterns about lagged values of previous year's weather conditions (temperature, humidity and rainfall) , the green lines are the future year based prediction lines of (2017- 2020) by each models and the yellow lines are the actual values of each weather condition. By this we can easily understand how accurately they are predicting time series weather conditions. Among all of them CatBoostRegressor is giving the best predictions.

In figure 10 the learning curves and confusion matrix are shown of weather conditions (temperature, humidity, rainfall,) of CatBoost Regressor . The learning curves are shown based on RMSE. The confusion matrix is for analyzing the accuracy of a model's prediction performance by binning them in different categories.

In Figure 11 these graphs with shap values mainly show which lagged value has really impacted the most on each time series predictions of weather conditions.

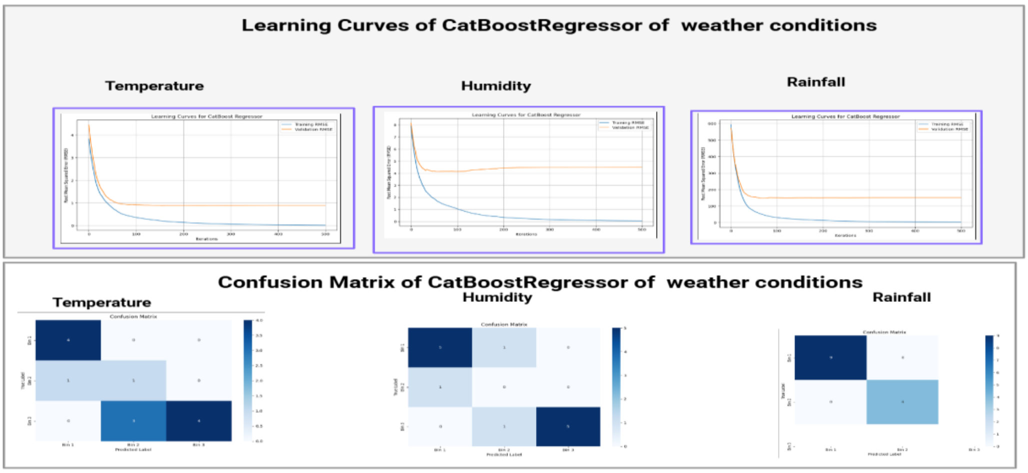

Fig. 10. The learning curves and confusion matrix plots of each weather conditions

Visualization of the comparison between best two performing models using shap

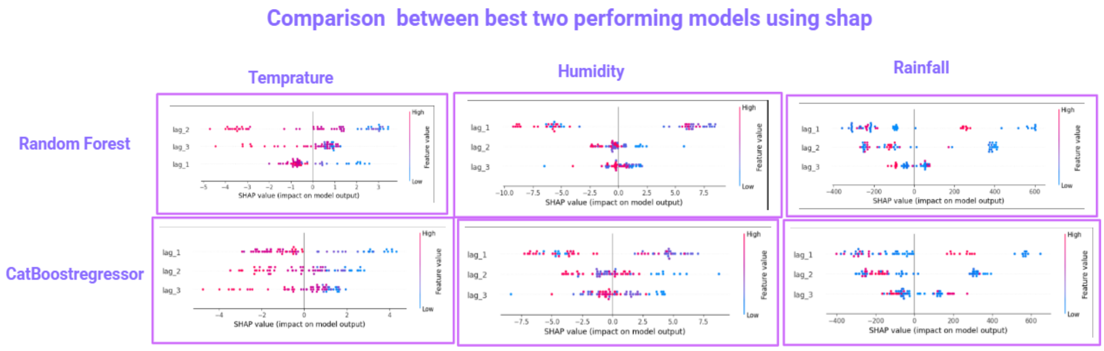

Fig. 11. Lag values effects and their visualization of different weather conditions using shap

Prediction for crop production:

As our project is based on forecasting weather conditions based on agriculture, how the weather conditions impact on each crop's total production is also a very important thing. As in different seasons there are different weather conditions based on which specific seasons are good for specific crop’s growth. If farmers can predict most accurately about specific year, season , weather conditions based crop production, if the production of that particular crop has less production then the farmers can choose other crops for farming and that will save them from a big loss. In our dataset we have tried different machine learning models for crop production based prediction for Burdwan and Hooghly district of West Bengal. The machine learning models are

- LinearRegression
- GradientBoostingRegressor
- Lasso Regression
- Ridge Regression

After giving inputs of weather conditions, year, crop, season ,area the upper given machine learning models have predicted the production of the crops. By using shap we can see here which feature has how much impact for this prediction.

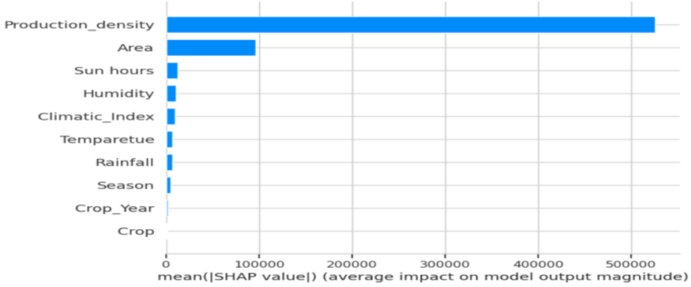

Fig. 12. Different features effects on the prediction of crop production by using shap

Some machine learning model's accuracy of the predictions and visualization are given below

Table 10. Different machine learning model's predictions based on crop production

Machine Learning Models	MSE	R2
GradientBoostingRegressor	128683.016	0.900
Linear regression	464628.99	0.703
Lasso Regression	464628.67	0.703

Some of their visualizations are given below

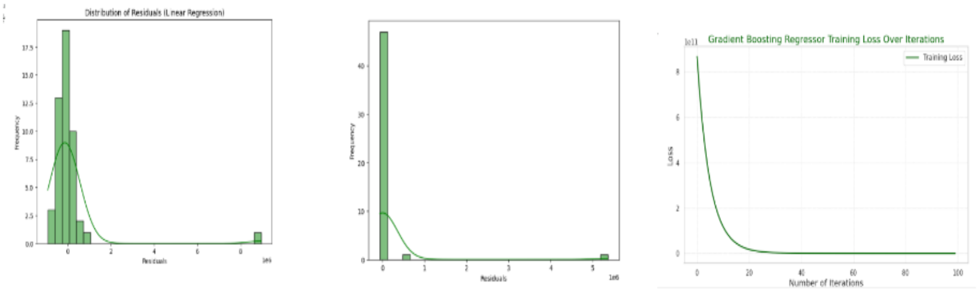

Fig. 13. Visualizations of different machine learning models of crop production based

These are some visualizations of crop production of the machine learning models. These observations play a vital role in green computing perspective because our goal is not only to predict the weather conditions but also to make sure that what are the impacts of the weather condition in each crop's growth. If we are able to most accurately predict these things these will be able to save crop yield and farmers will mostly be benefited.

4 Results and Discussion:

Here mainly two types of predictions are analyzed. One is for time series weather forecasting another is for understanding the impact of weather conditions in crop production. For time series prediction we have used 4 models which are Sarimax, Random Forest, Xgbooster and CatBoostRegressor for both Hooghly and Burdwan district. Here on average the CatBoostRegressor is giving the best performance on future prediction of weather conditions of temperature, humidity and humidity and Random Forest giving the 2nd highest. For crop production we have used Linear Regression, Lasso Regression and GradientBoostingRegressor. The GradientBoostingRegressor gives the best performance.

Table 11. Research gaps and their fulfillment in our paper

References	Research gaps	Fulfill the research gap
1	Ignoring the factor of predicting rainfall and the experiment is only done for one specific crop based	Along with other weather conditions rainfall based predictions are also done and our experiment is done for multiple crop based
2	Prediction of good accuracy based are missing and this predictions are done for only short term based time series	Here predictions with good accuracy are achieved and the time series predictions are long term based.
3	Weather conditions like (temperature, humidity and rainfall) like more in detailed based predictions are missing and only have shown the weather based prediction not the crop production effect	More detailed based weather condition predictions are achieved and here along with time series prediction the prediction for crop production based prediction is also done
4	Only have done the predictions by limited machine learning model based and the predictions are based on only short term based time series	Observed time series predictions with multiple machine learning models and the time series predictions are long term based.

5 Limitations and Future work:

There are many limitations in our proposed work. Our main limitation is that the dataset size is small and there is long time series data. So our first future goal is to make the dataset size more available as per need to make more accurate and real life based predictions and our second goal is to go to more depth so that our machine learnin model could also predict short time and longest time series based weather condition related predictions.

6 Conclusion:

This study demonstrates the potential of ML models in forecasting climate conditions within smart greenhouse. Among the evaluated models The XGBoost model outperformed the others, achieving the highest accuracy with an R2 value of 0.9929 and Residual Prediction Deviation (RPD) of 11.8464. The model's interpretability, as revealed through feature importance analysis, highlights its ability to understand complex interdependencies between microclimatic variables, including temperature, humidity and CO2 concentration. The findings of this study emphasize the practical benefits of ML in enabling percussive environmental control in greenhouses, leading to improved resource efficiency, higher crop yield and enhanced crop quality. However, the study has some limitations. The model's performance was validated on specific datasets, which may limit its generalizability to other greenhouse structure or crops types. Furthermore, real-time deployment and integration with existing greenhouse management systems require additional exploration to address potential challenges in sensor reliability, data acquisition and computational requirements.

References:

1. Y. J. Jeon, J. Y. Kim, K. S. Hwang, W. J. Cho, H. J. Kim, and D. H. Jung, "Machine Learning-Powered Forecasting of Climate Conditions in Smart Greenhouse Containing Netted Melons," Agronomy, vol. 14, no. 5, May 2024, doi: 10.3390/AGRONOMY14051070.

2. W. Sun and F. J. Chang, "Empowering Greenhouse Cultivation: Dynamic Factors and Machine Learning Unite for Advanced Microclimate Prediction," Water (Switzerland), vol. 15, no. 20, Oct. 2023, doi: 10.3390/W15203548.

3. W. H. Chen and F. You, "Semiclosed Greenhouse Climate Control Under Uncertainty via Machine Learning and Data-Driven Robust Model Predictive Control," IEEE Transactions on Control Systems Technology, vol. 30, no. 3, pp. 1186–1197, May 2022, doi: 10.1109/TCST.2021.3094999.

4. L. Wang, X. He, and D. Luo, "Deep reinforcement learning for greenhouse climate control," *Proceedings - 11th IEEE International Conference on Knowledge Graph, ICKG 2020*, pp. 474–480, Aug. 2020, doi: 10.1109/ICBK50248.2020.00073

5. X. Mao et al., "A variable weight combination prediction model for climate in a greenhouse based on BiGRU-Attention and LightGBM," Comput Electron Agric, vol. 219, Apr. 2024, doi: 10.1016/J.COMPAG.2024.108818.