UDC 004.8:631.559:633.18
IRSTI 68.35.31

https://doi.org/10.55452/1998-6688-2026-23-2-290-311

[1]*Kulyal M.,
ORCID ID: 0000-0002-2367-0897,
*e-mail: malika_21@rocketmail.com
[2]Dr. Umang,
ORCID ID: 0000-0002-9458-5817,
e-mail: anilumang@yahoo.co.in
[1]Dr. Saxena P.,
ORCID ID: 0009-0007-0544-593X,
e-mail: parul_saxena@yahoo.com
[3]Dr. Pant J.,
ORCID ID: 0000-0003-2279-5556,
e-mail: geujay2020@gmail.com

[1]Department of Computer Science, Soban Singh Jeena University, Almora-263601, Uttarakhand, India
[2]Department of Computer Applications, Kumaun University, Nainital, Uttarakhand, India
[3]Department of Computer Science and Engineering, Graphic Era Hill University, Bhimtal Campus, Uttarakhand, India

MACHINE LEARNING-DRIVEN PADDY YIELD PREDICTION: COMPARATIVE EVALUATION OF BASELINE VS. ENSEMBLE MODELS IN UDHAM SINGH NAGAR, UTTARAKHAND

Abstract

Rice is a cornerstone of food security in India, supporting millions of livelihoods and the national economy. However, erratic climate patterns are making paddy yields increasingly unpredictable. This study develops a machine learning framework for rice yield prediction in Udham Singh Nagar district, Uttarakhand, by integrating weather, soil, and crop data. Among baseline classifiers, CatBoost performed best with 80.85% accuracy and a ROC-AUC of 0.90. To further enhance performance, Optuna-tuned CatBoost, XGBoost, and LightGBM models were combined into hybrid ensembles. The Weighted Hard Voting classifier, giving higher weight to CatBoost ([3,1,1]), achieved the highest accuracy of 97.37%, followed by Stacking (95.6%) and Soft Voting ensembles (up to 96%). These results were supported by strong ROC-AUC scores. Overall, the study shows that carefully optimized ensemble models can significantly improve yield prediction accuracy, offering a practical tool for more precise and sustainable rice farming in climate-sensitive regions of India.

Key words: CatBoost, Gradient Boosting, Hybrid Ensemble, Machine learning, Random Forest, XG Boost, Yield.

Received April 13, 2026; revised May 27, 2026; accepted May 30, 2026.

Introduction

As one of the most important crops in the world, rice is very important in maintaining food security in the world, especially in India which is also among the largest rice-producing countries. Paddy farming is the main source of rural livelihoods in Uttarakhand at Udham Singh Nagar district. According to the reports on Agriculture Department Uttarakhand, the issues of climate variability, soil erosion, and unreliable agricultural practices are the main factors that influence the paddy yield and require the implementation of predictive instruments that are reliable and can assist farmers in

their endeavours to enhance agricultural production (Agriculture Department Uttarakhand, 2025). According to Sharma et al. (2023), one of the agribusiness regions of Uttarakhand State is Udham Singh Nagar District. It is famous by its agriculture and irrigation which are through the process of historical evolution and is known to have a productivity of paddy crops all over Uttarakhand, to gain a status of Chawal ki Nagari, thus enhancing its impact in the groundwater resource of the district. Almost 64 percent of the total number of personnel is engaged in agriculture on the incredibly fertile Tarai development. The advancement in machine learning has now reconstituted predictive analytics with new solutions advancing to optimize solutions to difficult agronomy challenges. All the environmental factors such as soil properties, climatic conditions, and crop properties are instilled in machine learning models to deliver correct AI-driven analytics on yield forecasting. The tools increase the decision-making and promote sustainable use of resources to develop sustainable agriculture and food security further.

It has been proposed that Precision Agriculture, through the use of recent advances in Machine learning, can be used to be transformational in responding to these challenges. The use of these Machine learning algorithms can help in drawing significant patterns through complex data sets of climatic variables, soil traits, and crop attributes which can be utilized in formulating good predictions of yield. Specifically, the use of analytic approaches may lead to improved resolutions on the issue of farming and its sustainability in such locations as Udham Singh Nagar, a region of the Uttarakhand state, which is among the largest rice-producing regions in the country.

According to Tan et al. (2021), ensemble learning methods, which is a synthesis of the merit of over one base learner, have already proven to have essential steps in prediction accuracy and model resilience. It is common with other methods such as XGBoost, LightGBM, and CatBoost due to its ability to work with a large dataset using a complex dataset and minimum preprocessing. Recently, predictive tools such as stacked generalisation, voting (hard, soft, and weighted), and boosting ensembles have been proven to be effective in agricultural predictions (Tan et al., 2021; You et al., 2020). With hyperparameter optimisation algorithms, such as Optuna, those models provide better results even on limited and heterogeneous data, which is the common case in practice in real-world agriculture. These innovations can be seen as one of the recent trends in the same domain: combining data science with agronomy to enhance management of crops, to make the most out of resources, and to ensure livelihoods in case of environmental uncertainty.This paper affirms the applicability of the ensemble learning approach to agricultural production forecasting and provides a scalable potent instrument of making data-driven decisions in precision farming. This paper is founded on this assertion and the hybrid ensemble model that has been proposed and evaluated is applicable in the prediction of rice in Udham Singh Nagar. The study will focus on high-quality forecasts, sustainable agricultural activities, and policy formulation within similar agro-climatic areas by incorporating weather, soil, and crop parameters into a single machine learning pipeline, and using the optimal models of Optuna-tuned XGBoost, LightGBM, and CatBoost.

In the study by Chandrakumar et al. (2023), regression-based machine learning models that included Gradient Boosting, Random Forest and Support Vector Machines were used to predict rice production in the Tamil Nadu region. The environmental factors that were integrated by the authors include rainfall, soil nutrients and temperature. Gradient Boosting model had the best predictive accuracy when compared to the models.

Renju et al. (2022) examine the climate risk factors by analysing the benefits that are brought about by combining machine learning methods with agricultural databases. The study explores the significance of Gradient Boosting and Random Forest algorithms, together with other models, for climate issue management and increased predictive accuracy. It demonstrates how different machine learning systems help computer networks and decision trees to identify crop outcomes by examining multiple data types, including environmental, soil, and satellite measurements. These procedures enhance yield forecasting outcomes and improve resource utilization.

(Joshua et al., 2021) this study tested multiple modern learning tools to forecast crop yields for all farming types particularly rice. It demonstrates that the weaknesses of data are required, as well as the choice of the significant features, to produce improved models. Gradient Boosting with the addition of the Random Forest was effective in dealing with agricultural data. Application of machine learning (ML) methods in agriculture has transformed the process of predicting crop yields especially in rice which is one of the main food commodities in India. Recent research has shown that different ML models are effective at predicting yields of rice through the analysis of complicated datasets containing climatic, soil, and crop variables.

In another study, De Clercq et al. (2024) modelled Kharif season rice yield in 247 Indian districts with models like CatBoost and LightGBM, with an out of sample R 2 of up to 0.82. On the same note, Yewle et al. (2025) have come up with RicEns-Net a deep ensemble model, which used a combination of synthetic aperture radar together with optical remote sensing data as well as meteorological variables, which led to a mean absolute error of 341 kg/ha.

In the study by Kamilaris and Prenafeta-Boldu (2018), machine learning was applied to analyze remote sensing data in which the authors trained XGBoost and CatBoost models to estimate crop yield. As vegetation indices and climatic data were incorporated into ensemble techniques, the techniques produced better results than the conventional regression techniques. Ensemble learning techniques have become notable due to the enhancement in the degree of predictive precision through the synthesis of numerous models.

To predict crop yields, Manjunath and Palayyan (2023) created a hybrid machine learning model that included Decision Tree, XGBoost, and Random Forest that had a high predictive accuracy of 98.6. Similarly, Chandraprabha and Rajesh Kumar Dhanraj (2023) introduced an ensemble based deep learning methodology of stacking and using Deep Neural Networks, Deep Belief Networks as well as Support Vector Machines to predict rice yield depending on the level of soil nutrition, which they reported with a 89.5% accuracy.

Precision farming practices have led to the use of technology that has been used to enhance optimal utilization of available resources and achieve better crop yields. TNN (2025) the partnership of Punjab Agricultural University and BITS-Pilani is focused on the implementation of AI, IoT, and geospatial technologies in the field of agriculture and ensuring the use of data to make decisions. In addition, novel methods of farming such as Seeding of Rice on Beds (SRB) has been implemented to deal with the problem of water scarcity by consuming as much as 80 percent less water than that of the traditional methods.

IGARSS 2019 exhibited the process of predicting the number of rice harvests with the help of integrated meteorological and soil data. Guruprasad et al. (2019) trained machine learning algorithms, Neural Networks and Random Forest, using district-level data and distributed their results to taluk levels to create better maps. The research showed that better results in yield measurement could be obtained thanks to spatial resolution data, which produced only a 6% error for average measurements.

Materials and methods

The dependent variable used in the analysis was the district level yield of rice in the current year, and the independent variables were the selected weather, soil and crop parameters which were standardized before the model execution. The models were trained using data that covered the past 10 years ensuring that the training pipeline is efficient and reliable. Consequently, training models have been evaluated based on developed methods of training.

The model training pipeline included extensive data preprocessing, involving treatment of the missing values, encoding the categorical variables, and the normalization of the numerical ones. The dataset was then stratified into 80:20 train and test sets in order to maintain the proportion of classes, Figure 1.

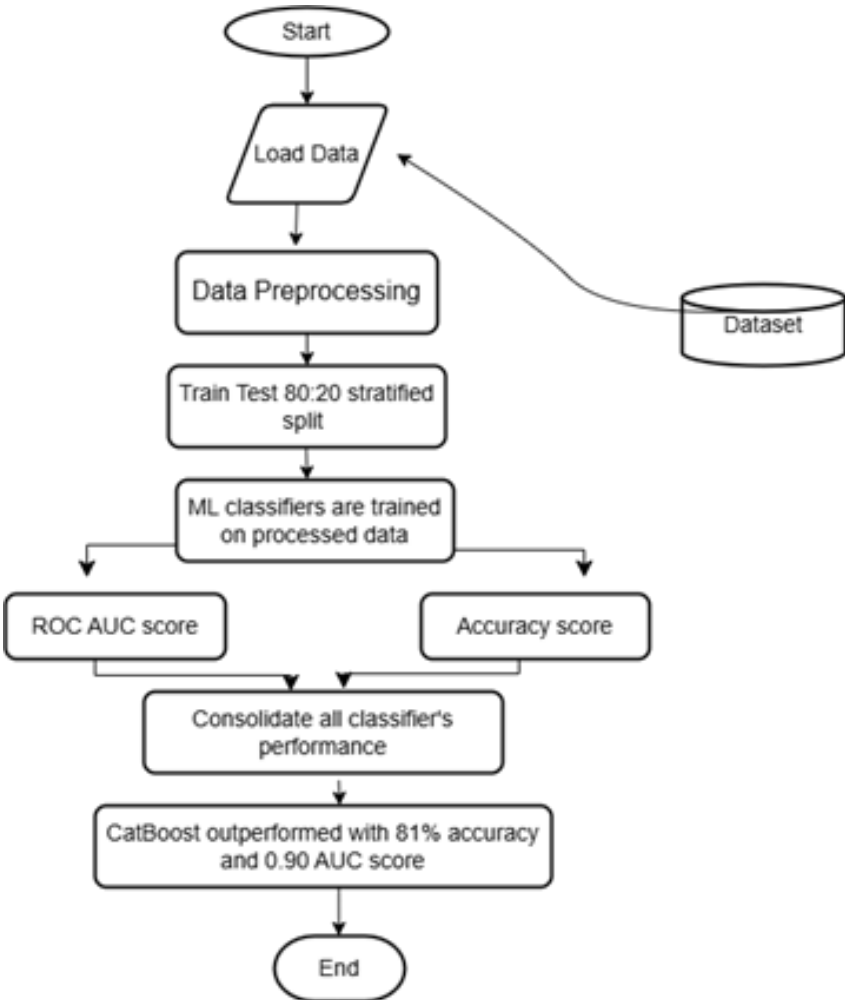


Figure 1 – Workflow Diagram of traditional base classifiers

The target variable (Yield Category) is a binary (0 = Low Yield 1 = High Yield) variable that is calculated according to whether each yield value falls below or above the median yield of its zone. This methodology guarantees the equal distribution of the classes and also puts into consideration the varying yield potentials in the rainfall areas.

We have developed our model in a systematic process, starting with an initial and simple classification models with more complex ensemble architectures. A brief summary of each algorithm administered in this evaluation is provided in following sections.

According to Yasaswy et al. (2022), the Gradient Boosting type of classification model has been a very useful ensemble ML model applied in the prediction of the outcome of agricultural yield, such as paddy yield. It builds multiple feeble decision trees sequentially where each new decision tree improves the model by rectifying the errors in the earlier decision trees. The objective of this iterative process is to minimize a loss criterion that has been preprogrammed which enables the Gradient Boosting process to discover complex dependencies between input qualities, such as climatic cycles, soil properties and sustainable agricultural methods.

According to Badshah et al. (2024), the Random Forest classification model is a form of machine learning model that is widely applied in predictive activities, including forecasting agricultural yields. In the training of the model, it constructs many decision trees and combines integrated results

in order to minimize discrepancies. All trees are computed based on random sub setting of the data, and a probabilistic method of attribute representation is evaluated at every split, such that there is fluctuation across trees. This randomness renders Random Forest applicable to noisy inputs, and in a position to discover more complex correlations among the input variables.

According to Chen and Guestrin (2016), the XGBoost classification model, which is also referred to as the Extreme Gradient Boosting model is among the sophisticated machine learning algorithms and is very effective when it comes to predictive tasks like predicting the yield of paddies. XGBoost is a specialized version of the Gradient Boosting framework designed to provide high-performance and scalable solutions. Features such as parallel processing, a sparsity-aware algorithm, and parameter regularization enhance performance and help prevent overfitting.

Yandex (2019) report that CatBoost (also known as Categorical Boosting) is a gradient boosting algorithm developed to specifically handle categorical features with minimal preprocessing, i.e., no one-hot encoding is required. CatBoost is superior to other gradient boosting algorithms in that it employs ordered boosting to minimize overfitting. It is regarded to be very reliable, correct and easy to use especially with tabular datasets. CatBoost uses a technique known as target-based statistics that uses categorical data directly, limiting the likelihood of overfitting. It also has little hyperparameter tuning, and it can also be used in conjunction with most data science packages.

According to Yamparla et al. (2022), one of the effective statistical methods that are commonly used in classification including classification of rice output is the Logistic Regression. In Agri-tech applications, Logistic Regression has the ability of classifying the level of yield, either high or low, in relation to climatic conditions, soil characteristics, and farming practices which can add valuable information to the decision-making process of precise agriculture. It uses the correlation of a categorical dependent variable with another factor or a combination of multiple factors to compute the likelihood of the outcome occurring within a given label of a class. The algorithm makes use of the logistic function, which is frequently modeled as an S-shaped sigmoid curve, to map the values of features into constrained probability distributions between 0 and 1.

Hyperparameter tuning for CatBoost, XGBoost, and LightGBM was conducted using Optuna, an automated optimization framework. Optimal configurations were selected based on cross-validated performance on the training data, enabling improved generalization and robustness relative to default parameter settings.

First, hyperparameters tuning of the ensemble classifiers- CatBoost, XGBoost, and LightGBM was used with Optuna. The best settings derived out of this tuning were then applied in the implementation of the above hybrid models. The CatBoost Classifier was optimally set and some of the parameters that were set included depth of 8, learning rate of about 0.15, and the hyperparameters of the regularization and sampling including l2leafreg, border count, and bagging temperature. These values were discovered by optuna by trial and error to maximize the results on the training data. Equally, XGBoost and LightGBM hyperparameters were also optimised that contained the following hyperparameters; iterations=1049, depth=5, learningrate=0.105 and other boosting-specific hyperparameters such as l2leafreg, random strength and grow_policy= SymmetricTree. These optimized settings are likely to enhance model accuracy and generalization much more than default settings, which constitute a solid basis of a high-performance ensemble model.

To enhance the classification accuracy, a set of ensemble learning methods, such as Hard Voting, Soft Voting (weighted and weightless), and Stacking, were used to combine three fine-tuned gradient boosting classifiers: CatBoost, XGBoost, and LightGBM.

Hard Voting Classifier

An implementation of a Hard VotingClassifier was done whereby the predictions are done depending on the majority class vote of the base learners. The weight configuration [3, 1, 1] was used to provide CatBoost with more power. The accuracy of this ensemble was high and reached 97.37, which supports the effectiveness of majority rule in the case of well-calibrated and complementary base learners.

Mathematically this can be put as:

$$\hat{y} = \underset{(c \in C)}{ar\ g\ m\ ax} \sum_{m=1}^{M} 1_{[h_m(x)=c]} \cdot \omega_m$$

Where:

- $\hat{y}$: Final predicted class
- $c \in C$: Class labels (e.g., low / high yield)
- M: Number of base classifiers
- $h_m(x)$: Prediction of the *m-th* base classifier
- $\mathbf{1}_{[h_m(x)=c]}$: Indicator function
- w_m: Weight assigned to the *m-th* classifier

The hard voting classifier was adopted in this research as a hybrid ensemble model, and it consists of CatBoost, XGBoost, and LightGBM.. The assignments were made by hand, to give a better score to CatBooust (e.g. wCatBoost=3, wXGB=1, w{XGB} = 1, wLGBM=1, wLGBM=1), as it has a better standalone performance. Based on this strategy, the accuracy rate was 97 percent, which was higher than all base models, which showed that ensemble voting, when properly weighted, is a robust method.

Soft Voting Classifier

In the Soft Voting arrangement, the prediction in the classifier was done by the probability averaging. There were three configurations that were investigated:

◆ Weighted Soft Voting [3, 1, 1] with weights weighted favoring CatBoost was the best with an accuracy that was 95.6%.

◆ Weighted Soft Voting with [2, 2, 3] also reached 95.6%, showing consistent performance across similar configurations.

Unweighted Soft Voting (equal weight) resulted in a marginally greater accuracy of 96 showing that even when the model was not weighted, the ensemble was still very effective as the individual strengths of parts made it.

Ahmed (2023) explains that in a soft voting classifier, the final predicted class is calculated based on the mean of the predicted probabilities of each of several classifiers, and the class with the largest combined probability is selected. The equation can be expressed as:

$$\arg\max_{c \varepsilon C} \sum_{m=1}^{M} w_m \cdot P_m(c|x)$$

Where:

- $\hat{y}$: Final predicted yield class
- $c \in C$: Set of yield classes
- M: Number of base classifiers
- w_m: Weight of the m-th classifier
- $P_m(c \mid x)$: Class probability from classifier m

The soft voting classifier plays a crucial role in this study, as it facilitates a probability-based hybrid ensemble that can intelligently fuse the predictions of CatBoost, XGBoost, and LightGBM. In comparison to hard voting, which uses majority class designation, soft voting takes into account the level of confidence in each model's prediction; thus, it is especially efficient when the capabilities of each model are complementary.

Stacking Classifier

Two Stacking Classifiers had been built:

One of them used a Random Forest as the meta-learner which was trained on predictions of the out-of-fold predictions of the base learners. The accuracy of this setup with cross-validation of cv=5 was 95.6%.

Lin et al. (2023) report that the other CatBoost model was used as the meta-learner. When passthrough=True, both base model outputs and original features were forwarded to the meta-learner, which improved learning. This approach also achieved 95.6 percent accuracy, demonstrating the versatility and power of stacking with alternative meta-models.

Equation:

$$Z = \begin{bmatrix} B_1^{oof}(X) \\ B_2^{oof}(X) \\ \vdots \\ B_k^{oof}(X) \end{bmatrix}^T \in \mathbb{R}^{n \times k}$$

$$\hat{y} = \Phi_{RF}(\mathrm{Z})$$

Where:

The meta-learner (CatBoost) is trained on both:

- $B_j^{\mathrm{oof}}(X)$: Out-of-fold predictions generated by the *j-th* base learner.
- Φ_{RF}: Random Forest classifier trained on dataset Z.
- $\hat{y}$: Final predicted class labels.

- Out-of-fold predictions Z from base learners.
- Original input features X.

Equation:

$$Z = \begin{bmatrix} B_1^{oof}(X) \\ B_2^{oof}(X) \\ \vdots \\ B_k^{oof}(X) \end{bmatrix}^T \in \mathbb{R}^{n \times k}$$

$$\tilde{Z} = [X \,\|Z] \in \mathbb{R}^{n \times (d+k)}$$

$$\hat{y} = \phi_{CatBoost}\left(\tilde{Z}\right)$$

Where:

$\tilde{Z}$: Concatenation of the original feature set and the outputs of the base learners.

Φ_{CatBoost} : CatBoost-based meta-learner.

This strategy enriches the meta -model with additional information, there by enhancing its learning capacity and predictive performance.

Both stacking approaches are:

$$\hat{\bar{y}} = \phi([Base\ Model\ Outputs] or\ [X \| Base\ Model\ Outputs])$$

The key difference lies in the input to the meta-learner:

Random Forest: Φ(Z)

CatBoost (passthrough=True): Φ([X ‖ Z])

Both achieved 95.6% accuracy, illustrating the power and adaptability of the stacking framework with different meta-learners and feature strategies.

The workflow diagram of Hybrid Ensemble models shown in Figure. 2, hybrid models are always more effective accurate compared to base classifiers. Although the base classifiers are about 53-81 percent, the hybrid models reach greater than 95 percent accuracy indicating that ensemble methods are effective in improving the predictive ability.

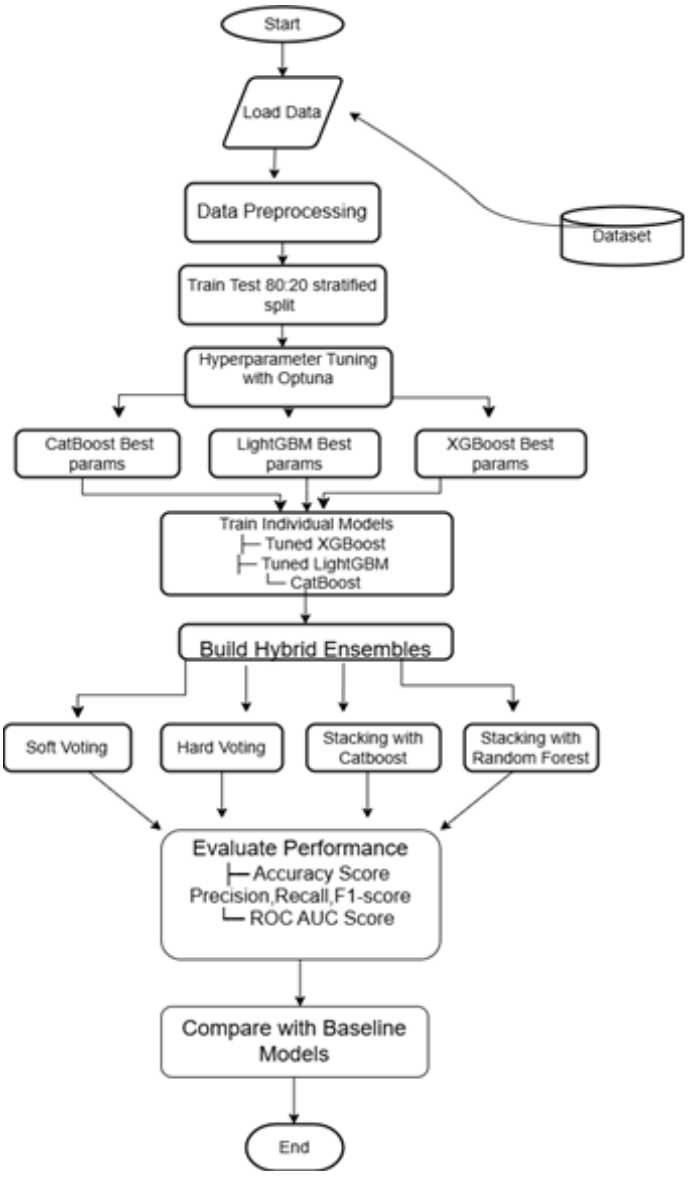


Figure 2 – Workflow of ensemble models

Study area

Udham Singh Nagar district is situated in the Tarai region of Kumaon Division. The district spans a total area of 3,055 km², standing 9th in terms of geographical size within Uttarakhand. It's situated between latitudes 28° 53' N and 29° 23' N, and longitudinally expands over 78° 45' E to 80° 08' E, (Kumar et al., 2021) Figure 3.

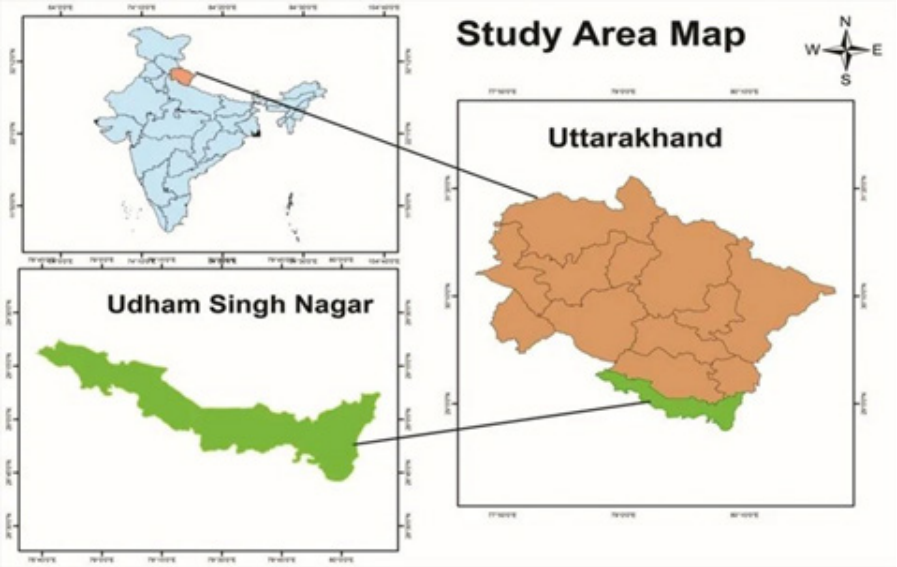


Figure 3 – Location map of Udham Singh Nagar district

The data set, Table 1, used in this study is holistically divided into three broad categories, which are weather data, soil data and crop data, which makes it easy to evaluate thoroughly the factors influencing paddy production in the region.

Table 1 – Dataset with independent features and dependent variable (Yield)

Yield (quintal/hect	temp	temp_min	temp_max	dew_point	pressure	humidity	wind_speed	wind_deg	rain_1h	clouds_all	Area (Ha)	Soil_pH	EC (mmos/cm^2)	Soil_O/C (%)	Available phosphorus (kg/ha)	Available Potash (kg/ha)	Soil_Zinc (ppm)	Mineralizable Nitrogen (kg/ha)
4.70257949	28.82276	28.2660345	29.248972	24.91544024	1004.192	79.15957	1.780747277	143.7580174	0.351858235	50.45183845	15800	7.2	0.81	0.58	17.9	174.58	1.64	289.76
6.32073643	26.53995	25.371644	27.418904	18.4175	1006.803	55.88043	2.035472826	148.6209239	0.057960372	25.76992754	1002	7.1	0.65	0.75	18.24	210.26	1.09	395.87
6.46853408	28.71452	25.6357107	27.623859	17.61774004	1007.244	54.65808	1.944322466	147.7284493	0.030057065	24.30543367	1226	6.8	0.79	0.73	17.85	205.36	1.19	385.97
6.73783033	25.86867	24.8075091	26.711712	17.16966938	1007.848	56.37047	1.952445652	166.2626852	0.126013408	27.29528985	1200	7.3	0.81	0.62	16.89	198.9	1.43	378.58
7.01345016	26.49332	25.5945199	27.249468	16.30553004	1007.727	51.91712	1.997042572	194.8467391	0.025897464	24.91394628	1247	6.8	0.9	0.71	18.23	211.51	1.51	397.89
7.07434268	24.66786	23.7320482	25.372115	17.83665026	1009.078	62.30208	1.912699842	161.495471	0.07214221	33.63924858	1101	7.1	0.82	0.74	19.58	209.89	1.28	391.85
7.28244259	27.39877	26.4412911	28.225371	19.07423007	1005.725	57.77808	1.875611413	146.5285326	0.042567029	18.98867754	1167	7.2	0.9	0.64	16.54	168.54	0.43	298.75
7.07250925	25.48213	24.5360398	26.224728	15.65445199	1008.989	55.22056	1.971485878	175.9244588	0.035095109	33.50490588	1274	6.9	0.91	0.71	19.01	210.09	0.77	395.46
7.43789737	27.53265	26.4857482	28.402842	17.82918949	1006.747	52.98505	2.003063394	170.1571358	0.062982246	38.26322246	1056	6.9	0.8	0.75	19.5	208.78	1.23	399.98
7.67174373	26.47958	25.5972283	27.245819	14.80522358	1008.107	48.77038	2.08973279	210.1132246	0.010958291	27.12327899	712	7.2	0.79	0.65	20.1	173.54	1.09	319.58
8.03196773	27.28491	26.3176223	28.122585	16.26352809	1006.184	49.05435	1.9579407	184.6512681	0.011621577	24.12994565	724	7.1	0.86	0.65	22.31	178.92	2.6	387.42
8.92125499	27.13549	26.6225453	27.734728	16.5689538	1006.87	55.10967	2.340953804	172.8446558	0.442792298	20.75867754	745	7.2	0.79	0.72	20.21	217.12	1.45	388.95
8.69433018	25.67581	24.8490582	26.352151	17.84142099	1008.577	57.84311	1.889535301	172.428442	0.079981413	32.3125929	815	6.9	0.76	0.61	21.3	184.28	1.21	287.58
8.70199522	26.85269	26.1229801	27.525978	15.41306159	1007.945	50.6975	2.311264203	198.3129529	0.34491039	24.48503725	837	6.9	0.81	0.64	17.89	188.1	1.76	384.65
8.8109329	27.87532	27.3096575	28.535716	18.5459317	1009.32	59.98254	2.121222826	160.5815217	0.532918667	34.68621577	884	7.2	0.76	0.53	15.56	160.45	0.45	298.67
9.40580748	25.33144	24.6952264	26.004087	16.12578804	1008.934	60.15987	2.128902174	168.2201449	0.435314367	32.89039853	1026	6.8	0.72	0.52	16.9	178.21	1.02	294.35
9.43442869	25.30179	24.8540353	25.826013	16.71133905	1008.856	62.21875	2.23125	174.572358	0.486858301	32.33695652	927	6.8	0.79	0.65	18.66	187.92	1.26	384.25
9.51023936	26.42467	25.8834401	27.04572	16.01936225	1007.603	56.35978	2.221795428	173.3586957	0.830143879	25.17630696	819	6.9	0.81	0.55	15.98	166.3	1.67	289.3
9.70673913	25.33144	24.6952264	26.004087	16.12578804	1008.934	60.15987	2.128902174	168.2201449	0.435314367	32.89039853	920	6.8	0.79	0.68	17.88	199.82	0.63	367.54
10.0506957	27.539	27.0000942	28.196532	16.27882246	1006.422	53.36957	2.371721014	175.6431159	0.567028945	23.66712956	785	6.9	0.99	0.74	20.98	213.32	1.72	387.28
10.83640358	25.85708	25.0941304	26.557722	15.50498094	1008.425	57.07065	2.150265217	233.9506341	0.030502717	27.69330768	503	6.9	0.88	0.59	15.54	171.59	1.5	336.22
11.2909787	27.49505	26.8261775	28.164105	15.28539855	1005.929	51.99955	2.57464759	228.6854259	0.016548913	22.45172101	515	6.8	0.8	0.56	16.22	168.21	0.45	318.6

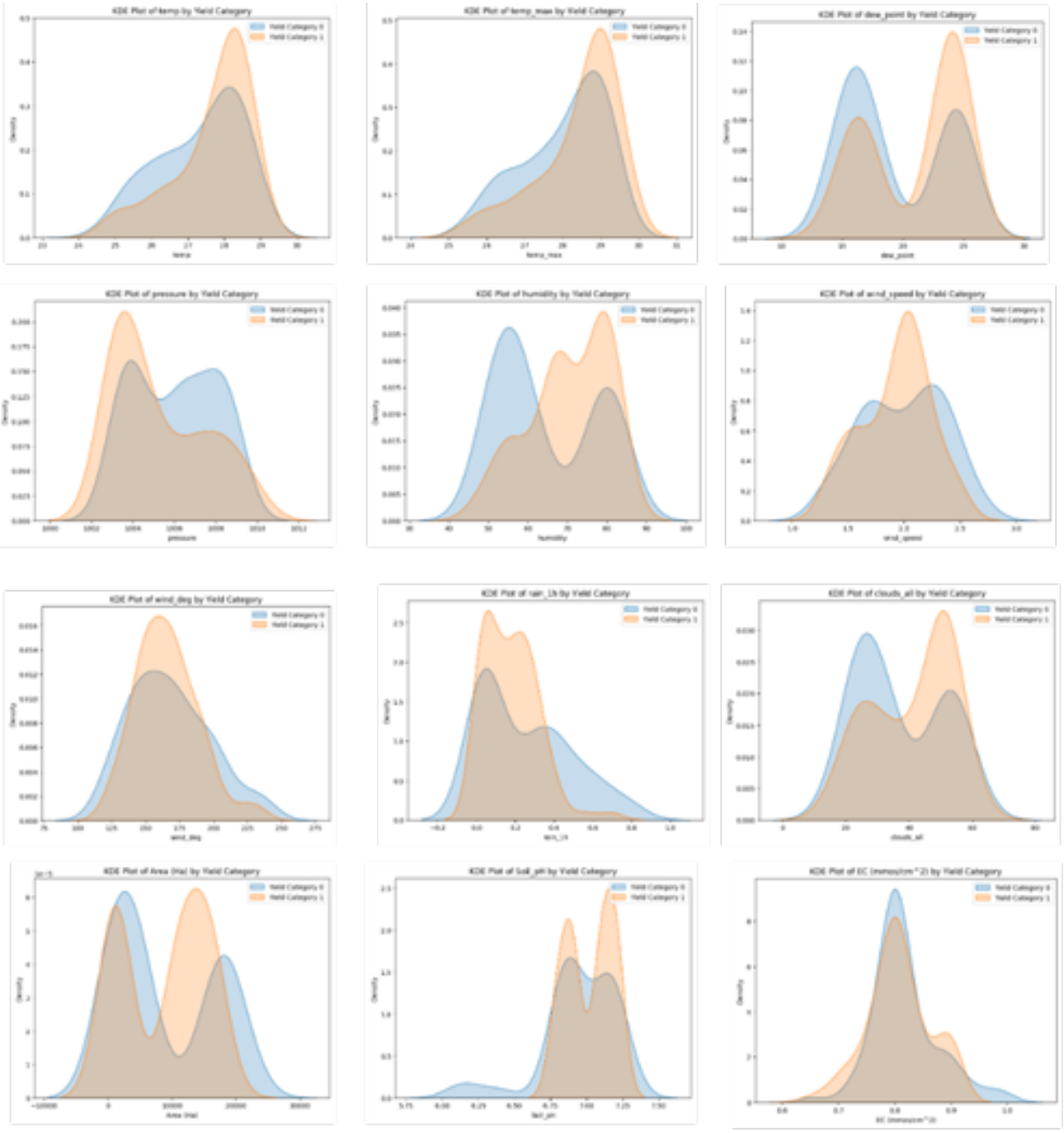
Weather Data: obtained from OpenWeatherMap.org (https://openweathermap.org/), this type contains a group of meteorological parameters that are important to predict yields. The parameters consist of average temperature, dew point, minimum and maximum temperature, wind direction, atmospheric pressure, wind speed, rainfall within the past hour, humidity, and cloud cover.

Soil Data: Collected from the Soil Test Laboratory in Rudrapur, the soil data comprises essential chemical and physical properties. These include soil pH, electrical conductivity (EC in mmos/cm²), soil organic carbon percentage (Soil_O/C), available phosphorus (kg/ha), available potash (kg/ha), soil zinc concentration (ppm), and mineralizable nitrogen content (kg/ha).

Crop Data: Crop information on the yield and the area of cultivation was retrieved in the Statistical Yearbook of the District found in Vikas Bhawan. This section contains information about yields (quintals per hectare) and the hectares that are under cultivation.

Results and discussion

The analysis of Kernel Density Estimation (KDE), Figure 4. showed that there were visible differences in the distributions across high- and low-yield groups in relation to a number of soil and climatic variables. There were consistent high levels of soil organic carbon, available phosphorus and mineralizable nitrogen found in the high-yield observations, which demonstrated the high level of association between the status of soil fertility and the productivity of rice.



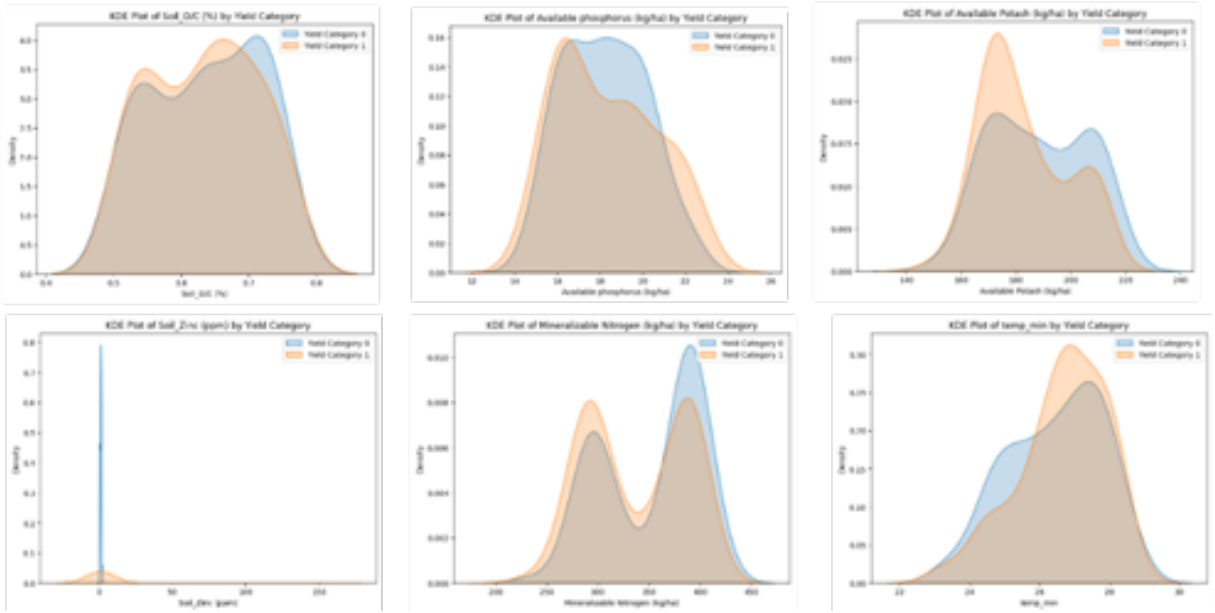


Figure 4 – KDE plots

Conversely, the overlaps of soil pH and electrical conductivity among yield classes were highly significant showing that the parameters did not significantly exceed the agronomically applicable limits and did not have a discriminative effect on the yields in the study area. Climatic variables including features of temperature exhibited significant distributional changes in yield categories, which indicates the sensitivity of rice yield to thermal factors. The results are similar to other machine learning-based rice production works which show that variability of temperature and nutrient availability are the most common influencers of yield prediction accuracy through analysis of feature importance and model performance (Chandrakumar et al., 2023; De Clercq et al., 2024). These methods are used in combination, and the regression-based feature importance used to measure predictive effect, and KDE used to give intuitive information about how agronomic and climatic factors vary between yield categories.

The frequency distributions of the 19 input variables are shown in Figure 5. Crop yield was mostly concentrated in the lower range (10–40 quintal/hectare), with fewer fields recording higher yields. Temperature variables followed fairly regular distributions, reflecting the relatively stable climatic conditions of the region. Dew point showed two distinct peaks, likely capturing the contrast between drier and more humid phases of the crop season. Rainfall was sparse for most observations, with occasional high-intensity events pulling the distribution rightward — a pattern quite common in the Terai belt. Soil pH was largely near-neutral (6.75–7.0), which is typical for the agricultural soils of Udham Singh Nagar, while EC and Soil Zinc were more variable, pointing to uneven soil fertility across sampled fields. The spread in Available Potash and Mineralizable Nitrogen likely reflects differences in nutrient management practices among farmers.

Sharma et al. (2024) predicted rice yield using only five meteorological variables — maximum and minimum temperature, relative humidity, rainfall, and sunlight hours – organized as fortnightly averages based on Standard Meteorological Weeks for a single district in Haryana, with no consideration of soil properties or field-level parameters. The present study, in contrast, analyzed distributions of 19 variables spanning meteorological conditions, soil health indicators, and agronomic parameters, revealing key data characteristics such as bimodal dew point, right-skewed rainfall, and multimodal soil nutrient distributions – offering a more comprehensive empirical basis for rice yield prediction in Udham Singh Nagar district.

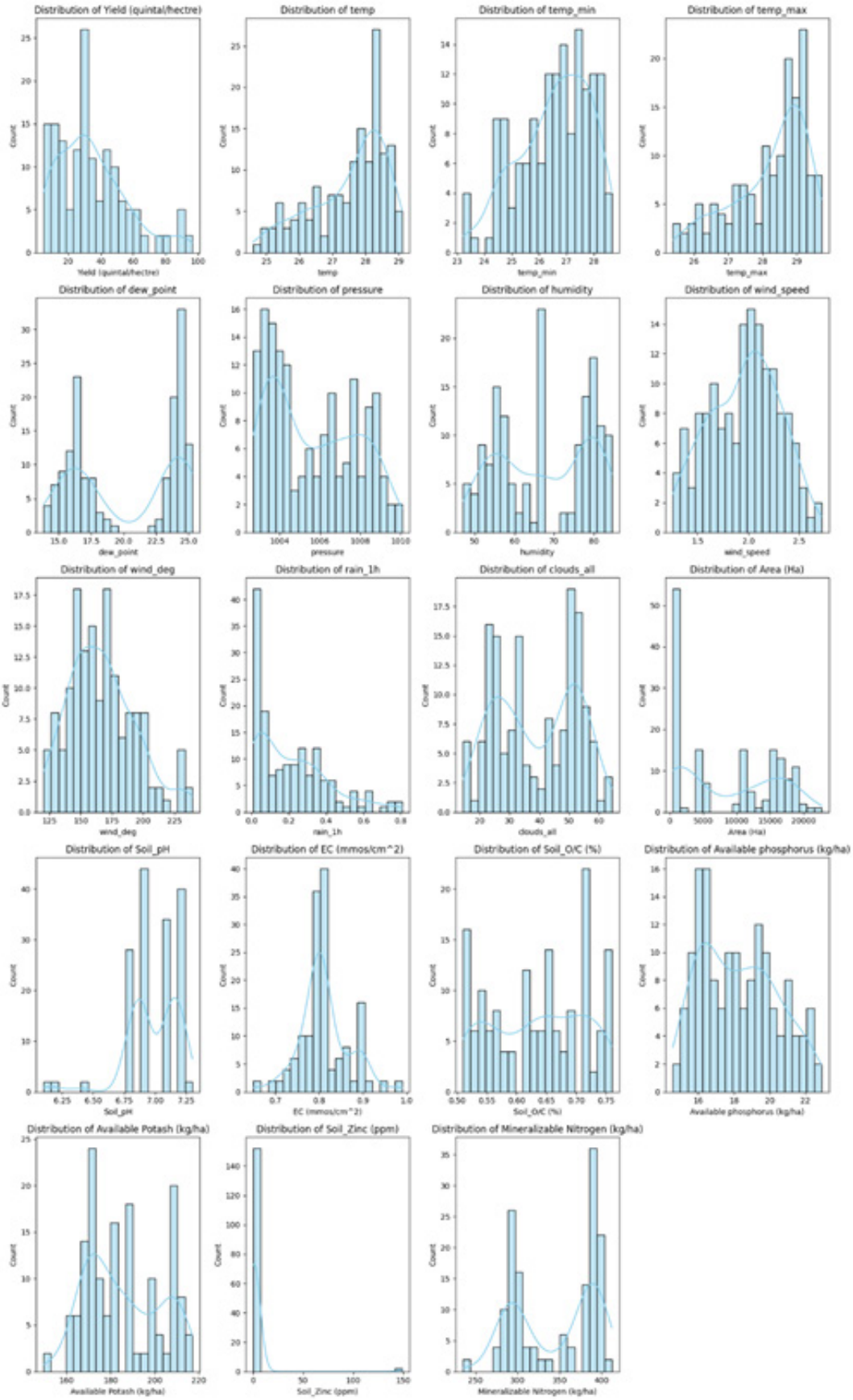


Figure 5 – Distribution of Input Variables

Figure 6. presents the pairwise relationships among the meteorological variables. As expected, temperature, minimum temperature, and maximum temperature moved closely together, which is worth keeping in mind when selecting model inputs to avoid redundancy. Dew point and pressure showed a mild inverse trend, while humidity had little consistent relationship with most other variables. Wind speed and direction appeared largely independent of the temperature-related features. Rainfall was near-zero for the bulk of observations, which aligns with its skewed distribution seen in Figure 5.

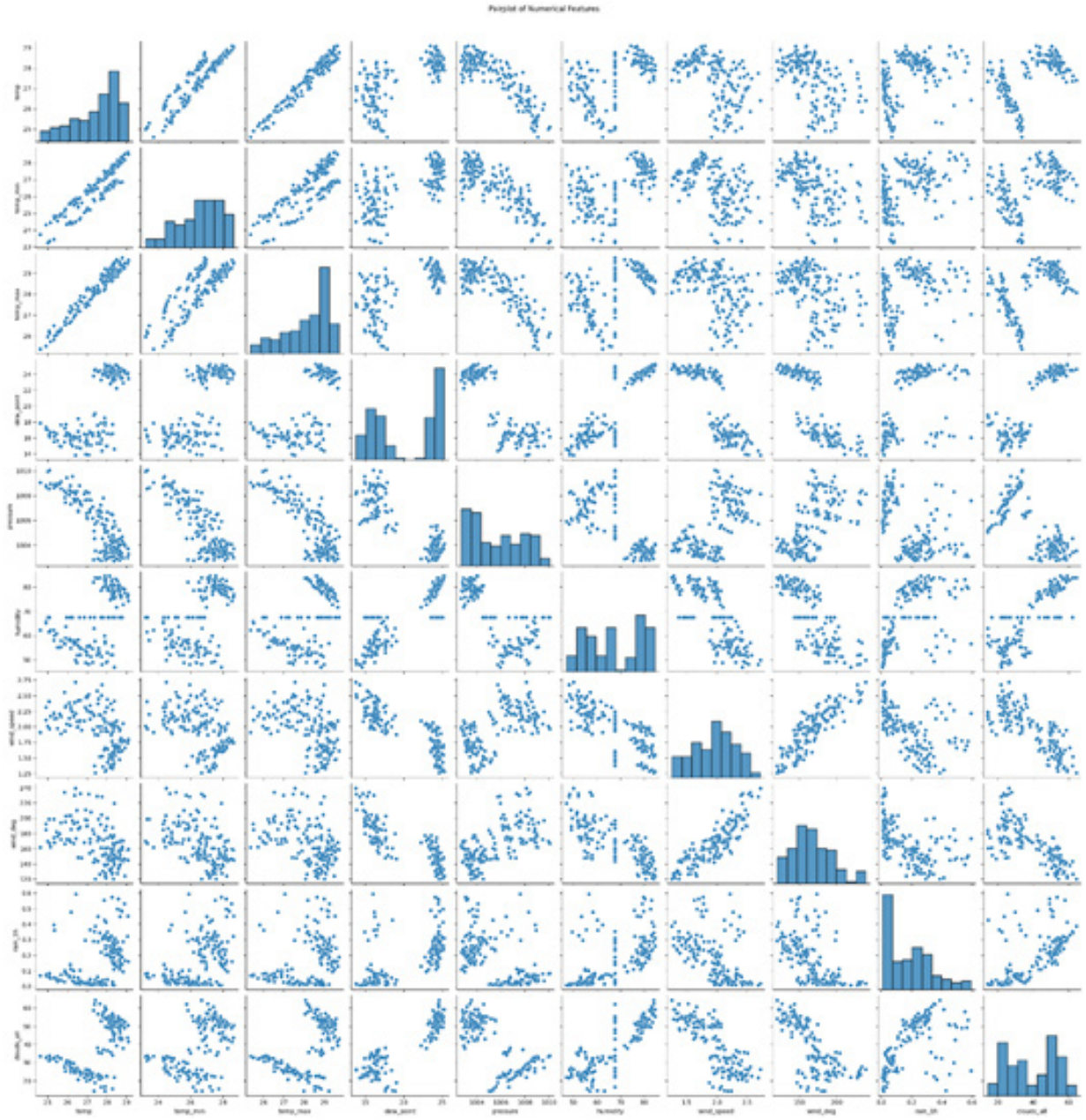


Figure 6 – Pairplot of Numerical Features

Naseer et al. (2025) looked at how soil and weather variables relate to each other, but their pairwise analysis covered only seven features – N, P, K, pH, temperature, humidity, and rainfall – where all features showed positive correlations, with potassium and phosphorus being the most strongly linked at 0.74. Notably, variables like dew point, atmospheric pressure, wind speed, wind direction, and cloud cover were not part of their analysis. In the present study, pairwise relationships

were examined across the meteorological features, and the patterns were noticeably more varied – temperature, temp_min, and temp_max moved closely together as expected, while dew point and pressure showed a mild opposing trend, and wind-related variables had little to no consistent relationship with temperature. This mixed correlation structure is something the seven-variable framework of Naseer et al. (2025) did not capture, pointing to the need for more careful feature selection when building yield prediction models for rice in the Terai agro-climatic zone.

Figure 7. shows that the correlation among climatic variables is high and linear, especially between minimum temperature, maximum temperature, and dew point, reflecting clustered thermal behaviour among the weather-related predictors. Attributes of soil fertility, such as available phosphorus and mineralizable nitrogen, are strongly correlated with soil organic carbon, an indicator of combined nutrient processes within the soil system. In comparison, rice yield shows a tendency to exhibit weak pairwise linear relationship with individual environmental and soil variables, which indicates that simple linear relationships can only explain a small amount of yield variability. The absence of significant linear yield-feature relationships motivates the use of non-linear machine learning models to characterise higher-order interactions and intricate relationships. In this regard, the current results are conceptually aligned with the synthesis presented by Kamilaris and Prenafeta-Boldu (2018), who state that crop yield mechanisms are not often determined by simple linear relationships and are better represented in the context of non-linear learning frameworks, based on a wide survey of agricultural studies.

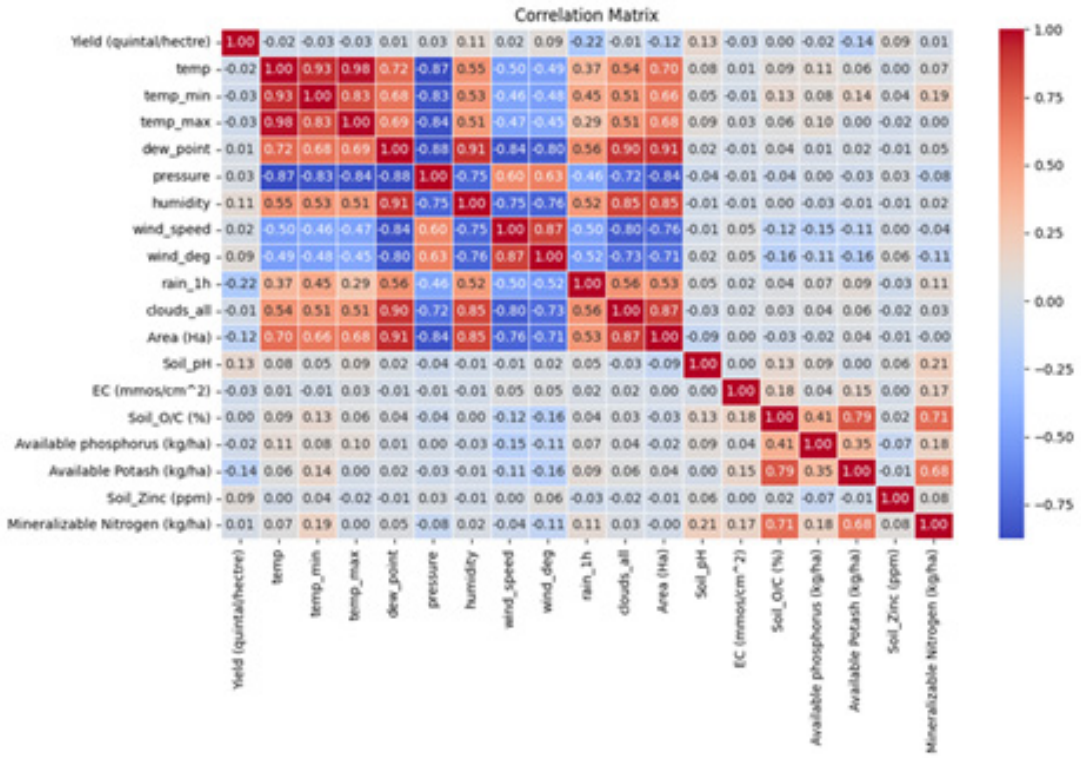


Figure 7 – Correlation Matrix

Results of the current study are generally in agreement with previous classification-based research of rice yield-prediction by Akula et al. (2021) in the Journal of Agrometeorology, but further expands the line of research by using more recent ensemble learning models and measures of performance. Both articles show the appropriateness of classification systems to place the outcomes of rice yield in categories, which can provide viable interpretability in agricultural decision support. In the case of the study carried out in Ranga Reddy district, Telangana, Akula et al. (2021) stated that the Multilayer

Perceptron (MLP) classifier demonstrated the best classification accuracy (about 74.19 percent), exceeding that of the traditional tree-based (J48/C4.5) and rule-based classifiers, which proved the importance of modeling non-linear relationships between weather variables and yield category.

By comparison, the current research gets much better classification (Table 2), with CatBoost reaching 80.9% and AUC of 0.90, then Gradient Boosting and XGBoost. These gradient boosting ensembles have a better performance as compared to the Logistic Regression, Random Forest, SVM, and KNN which shows that they have a greater ability to model high-dimensional, complex interactions between weather in relation to soil and crop-related variables and thus shows the advantages of using the modern ensemble methods in rice yield classification.

Table 2 – Performance metrics of base classifiers

Base Model	Accuracy	Precision	Recall	F1 Score
CatBoost	80.85%	82.00%	81.00%	81.00%
Gradient Boosting	78.72%	81.00%	79.00%	79.00%
KNN	53.00%	54.00%	53.00%	53.00%
Logistic Regression	76.60%	77.00%	77.00%	77.00%
Random Forest	68.09%	69.00%	68.00%	68.00%
Support Vector Machine	70.21%	72.00%	70.00%	70.00%
XGBoost	76.60%	78.00%	77.00%	76.00%

The results of the current paper are concomitant to those of Solanki (2024) who compared several machine learning classifiers to the classification of rice crop stages and proved that XGBoost and SVM were highly effective in case applied to feature-enriched data sets. On the other hand, the current paper is centered on district-level weather-soil-crop data in order to categorize rice yield and finds CatBoost the most stable standalone data classifier based on various evaluation metrics, with SVM and Random Forest having a relatively low level of discriminative ability. This variance highlights how both stage of prediction and data structure affect model behaviour with crop stage classification and yield categorization having completely different learning problems. However, in both papers, gradient boosting-based techniques prove to be trustworthy and flexible rice-related classifiers, which underlies their applicability to various precision agronomy activities such as phenological stage classification, as well as yield-related decision support.

In the given work, ROC-AUC analysis also proves the high discriminative power of CatBoost (AUC = 0.90), then XGBoost (0.89), Gradient Boosting (0.86), but Random Forest (0.78), SVM (0.72), and KNN (0.57) have lower rates of classification, Figure 8.

The results of the current research are in harmony with the existing literature that proves the superiority of hybrid ensemble classifiers in comparison with single base models in the process of agricultural classification. Ge et al. (2021) offer a representative example and researched the issue of rice phenological stage detection with the help of UAV-RGB images, which stated that ensemble approaches, including soft/weighted voting and stacking, were more critical in classification accuracy than individual classifiers. In particular, stacking was the most accurate (around 96.2%), then soft and hard voting (94.7 and 93.1) and demonstrates that complementary predictions of a model can be effectively combined. Probabilistic aggregation of the classifier outputs was considered as the superior outflow of the soft voting whereas stacking was favorable to learn the optimal combination of different base learners leading to the enhanced robustness in the capability of classifying intermediate phenological phases.

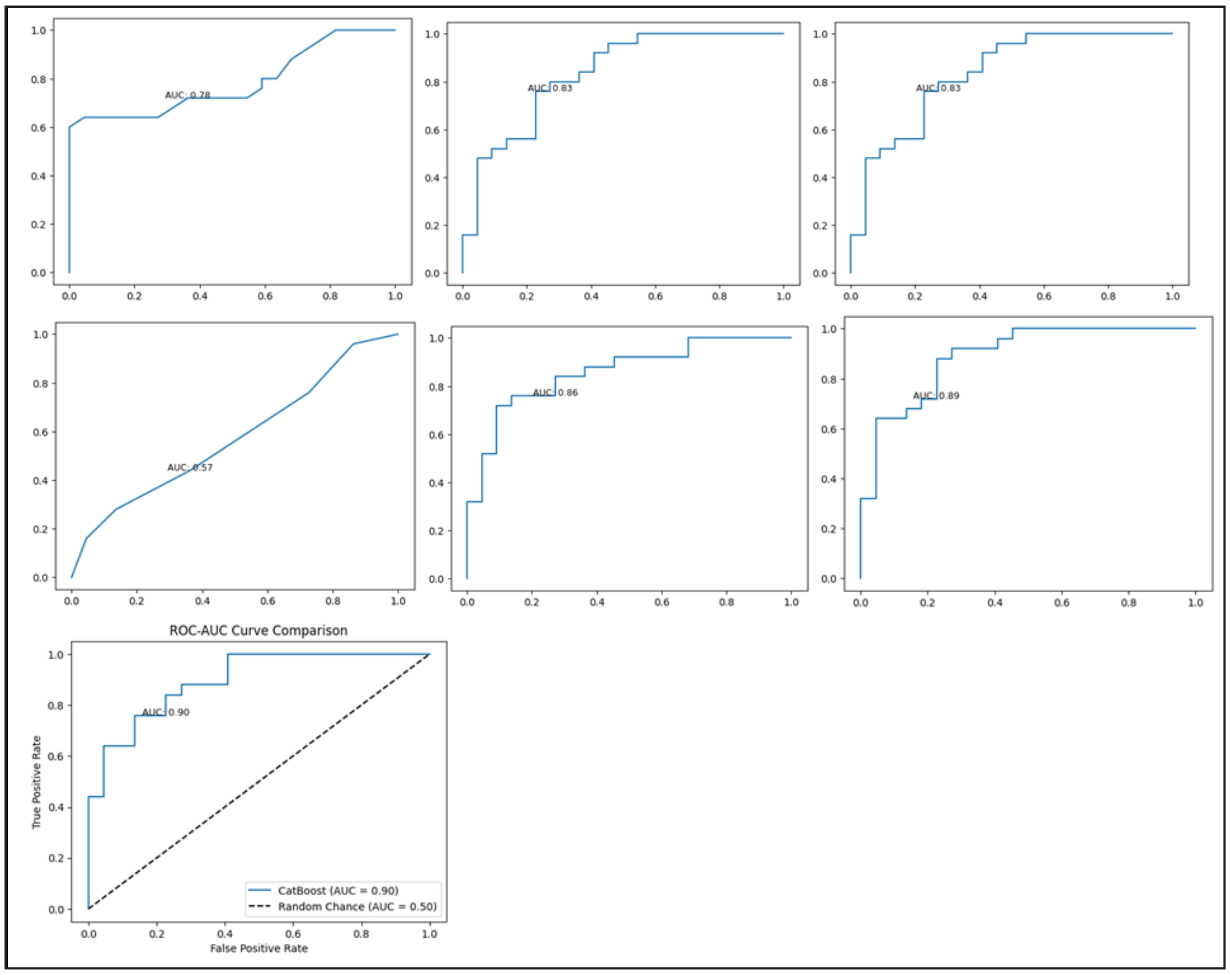


Figure 8 – ROC-AUC score comparison of Baseline Models

Similar advantage of the hybrid ensemble learning, however under different prediction task with different data format, can be seen in the current study on paddy yield classification. Table 3. illustrates that the weighted hard voting classifier, in which more importance is laid on CatBoost, had the highest overall performance (97.37%), which was better than those of soft voting and stacking. The outcome points to the importance of attention to selecting models and optimizing weights, which can further improve the performance of ensembles in the classification of yields. Taken together, these results confirm the bigger picture that ensemble methods, especially weighted voting and stacking provide better generalization and robustness in case of application to heterogeneous agricultural data, thus justifying their applicability to a variety of precision agriculture tasks.

Table 3 – Performance comparison of various hybrid ensemble models on classification metrics

Hybrid Models	Accuracy (%)	Precision (%)	Recall (%)	F1Score (%)
Stacking Classifier (Random Forest)	95.6	95.6	95.6	95.5
Weighted Hard Voting Classifier	97.3	97.3	97.3	97.3
Stacking Classifier (CatBoost)	95.6	94.6	98.6	96.6
Weighted Soft Voting Classifier	95.6	95.6	95.6	95.5
Soft Voting Classifier	95.6	94.6	98.6	96.6

The current research Islam et al. (2023) used regression models, such as a stack-ensemble hybrid, to predict rice yield (kg/ha) and had its best fit of RMSE ≈ 328 kg/ha and MAE ≈ 317 kg/ha, which is an improvement of 20-30% over base models. This shows that hybrid ensembles are good in capturing non-linear associations between environmental factors like NDVI, rainfall and soil moisture, even though the variability in space influenced the precision in certain districts. Whereas, the current work Figure 9. used classification models to predict high and low yield classes, the weighted hard voting classifier obtained Accuracy, Precision, Recall, and F1-score of 97.3%, which is significantly better than base classifiers (53-81%). It means that complex feature interactions can be effectively modeled by hybrid classifiers and they will be able to achieve high performance on all datasets. Collectively, the two studies show that ensemble methods are very effective in increasing the predictive quality, and that regression ensembles provide a better estimation of continuous yields and the classification ensembles can provide a highly accurate categorical forecast, particularly when multi-source environmental data is used.

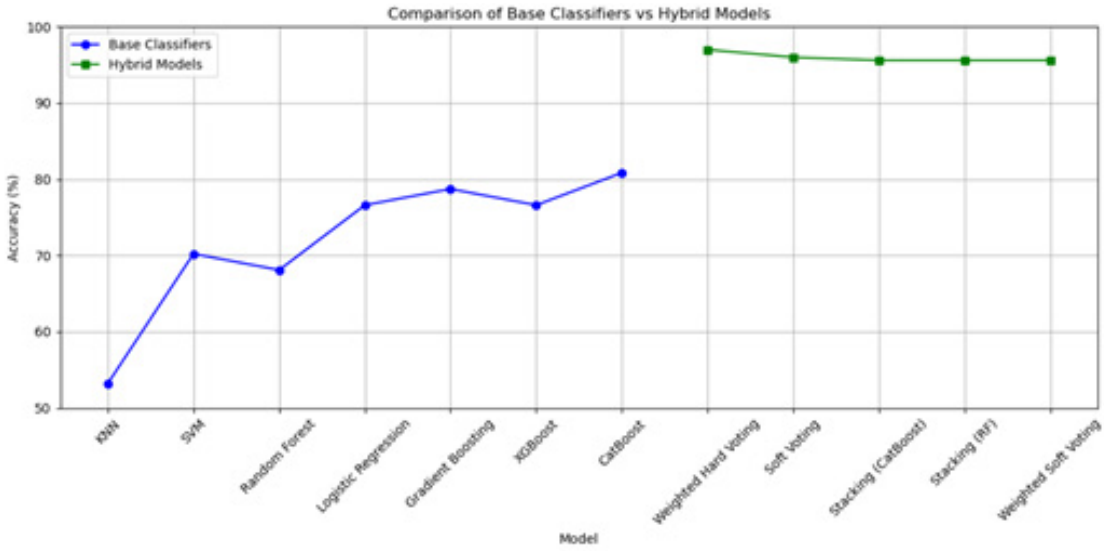


Figure 9 – Comparison of model accuracy between base classifiers and hybrid ensemble models

The current work builds on the previous ensemble-related rice classification studies, especially the one conducted by Ge et al. (2021) by an explicit and quantitative ROC-based analysis of hybrid models. They showed that stacking, soft voting, and hard voting ensembles are efficient in rice phenological stage classification and that stacking, soft voting, and hard voting ensembles have high classification accuracies (around 96.2 percent, 94.7 percent and 93.1 percent, respectively) and strong classifier discrimination represented by ROC curves. Their analysis, however, was based on visual analysis of ROC curves, without providing numerical values of ROC-AUC.

Conversely, Figure 10. the current paper measures the categorization of district-level rice yields with conventional hybrid classifiers and provides clear ROC-AUC values and ROC curves, which allow a more stringent and straightforward measure of the discriminative efficiency. The stacking classifier using CatBoost meta-learner scored 0.9908, stacking model using Random Forest meta-learner scored 0.9951, and weighted soft voting classifier scored 0.9941, a high level of classification was seen to be possible using all the three ensemble configurations. Although the two papers confirm the usefulness of ensemble learning in rice-related classification tasks, explicit ROC-AUC measures used in the current study enhance the performance of comparability and offer a more detailed method of evaluation to the yield-based decision support application.

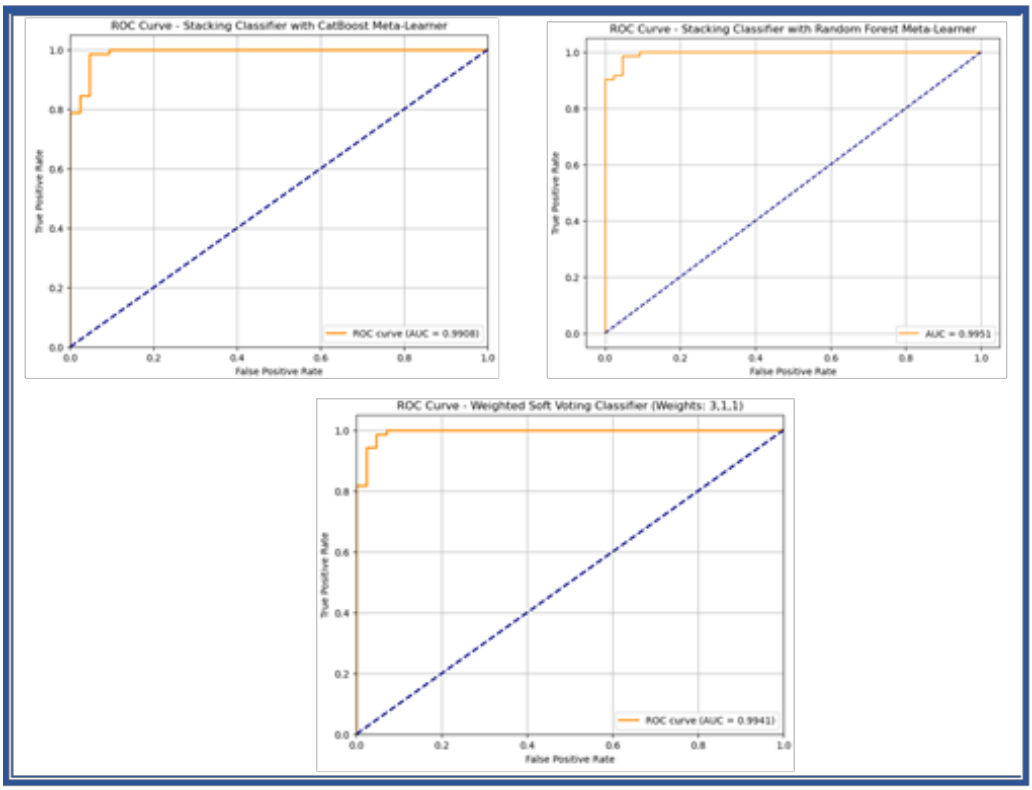


Figure 10 – ROC AUC curves of hybrid ensemble models

The study therefore reveals that there was a pronounced steady improvement in prediction performance when individual base classifier models are replaced by hybrid ensemble models to classify rice yield. The individual classifiers, or KNN, SVM, Random Forest, Logistic Regression, Gradient Boosting, XGBoost, and CatBoost, demonstrate moderate and high accuracy, with the lowest value of around 53 percent to the highest of about 81 percent, respectively. In line with the previous results (Akula et al., 2021; Solanki, 2024), tree-based and boosting models designed to be used with ensembles show better performance compared to distance and linear-style ones, which indicates their capability to learn non-linear relations between soil, weather-related, and crop-specific variables. Still, individual models are limited in their predictive power in terms of model-specific bias and the failure to capitalize on complementary decision patterns.

By comparison, the hybrid ensemble models, such as the weighted hard voting, soft voting, stacking, and weighted soft voting, record significantly greater accuracies with the range of about 95-97 that is clearly higher than that of all single classifiers. The mechanisms behind this performance increase are successfully compensating errors in a variety of learners, better generalization by means of diversity in an ensemble, and probabilistic aggregation that is used in the soft and weighted voting schemes. Out of the tested hybrids, CatBoost-based stacking and weighted voting strategies show a persistently high performance, which proves the success of meta-level learning and the optimal model weighting.

The current research contributes to the existing literature on the subject because it goes beyond the established rice-related prediction literature where most studies concentrate on the regression-based yield estimation or single-model classification and usually consider the accuracy or the qualitative ROC analysis as the final objective. The presented quantitative results improvements support the strength and the efficient applicability of the concept of ensemble learning to the classification of rice yield and decision support in a district level.

Conclusion

This paper has shown the promise of hybrid ensemble models that are machine learning driven to enhance prediction of paddy yield in the Udham Singh Nagar district of Uttarakhand. The results have shown that ensemble methods, in particular, stacking and voting-based classifiers including the boosted learners, are more accurate and robust in prediction than the traditional standalone model.

The proposed approach will enhance the development of precision agriculture, as it will use data and consequently make decisions about the activities of agriculture based on the use of soil, climatic, and historical crop data. The findings support the use of combination of various learning algorithms to accommodate the complexity and non linearity of agricultural data. In a practical sense, the suggested framework can affirm sustainable farming by facilitating on-time and dependable yield evaluation hence contributing to food security and resource management in rice-cultivating areas.

Future work may focus on extending the framework to multi-seasonal and multi-regional datasets, integrating remote sensing variables, and exploring deep learning–based ensemble architectures to further enhance scalability and real-world applicability.

Acknowledgement

The authors would like to express their sincere gratitude to the following individuals: Prof. Ajay Kumar Srivastava, Department of Agronomy, and Dr. S.K Gangwar, Professor, Department of Soil Science, GOVIND BALLABH PANT UNIVERSITY OF AGRICULTURE & TECHNOLOGY, Pantnagar, Uttarakhand, for their valuable support and guidance during this study.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research.

Funding Agency

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1 State Agriculture Statistics Data. Agriculture Department Uttarakhand. Available at: https://agriculture.uk.gov.in/document-category/state-agriculture-statistics-data/ (accessed 2025).

2 Sharma, M., Sharma, N., and Sachdeva, S. Ground Water Quality Assessment in Udham Singh Nagar, Uttarakhand, India. International Journal of Lakes and Rivers, 16 (2), 173–183 (2023). https://doi.org/10.37622/ijlr/16.2.2023.173-183

3 Tan, C., et al. Stacked and optimized machine learning for rice yield prediction in Asia. Agricultural Systems, 194, 103259 (2021). https://doi.org/10.1016/j.agsy.2021.103259

4 You, Y., Cao, J., and Zhou, W. A Survey of Change Detection Methods Based on Remote Sensing Images for Multi-Source and Multi-Objective Scenarios. Remote Sensing, 12 (15), 2460 (2020). https://doi.org/10.3390/rs12152460

5 Chandrakumar, T., Avanthica Sri, M.M., Mirdula, K., and K., M. Paddy Yield Forecasting using Regression Techniques. Proceedings of the IEEE Delhi Section Conference (DELCON), 1–6 (2023). https://doi.org/10.1109/DELCON57910.2023.10127256

6 Renju, R.S., Deepthi, P.S., and Chitra, M.T. A Review of Crop Yield Prediction Strategies based on Machine Learning and Deep Learning. Proceedings of the International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS) (2022). https://doi.org/10.1109/IC3SIS54991.2022.9885325

7 Joshua, V., Priyadharson, S.M., and Kannadasan, R. Exploration of Machine Learning Approaches for Paddy Yield Prediction in Eastern Part of Tamilnadu. Agronomy, 11 (10), 2068 (2021). https://doi.org/10.3390/agronomy11102068

8 De Clercq, D., and Mahdi, A. Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis data. arXiv:2403.07967 (2024). Available at: https://arxiv.org/abs/2403.07967

9 Yewle, A.D., Mirzayeva, L., and Karakuş, O. Multi-modal Data Fusion and Deep Ensemble Learning for Accurate Crop Yield Prediction. arXiv:2502.06062 (2025). Available at: https://arxiv.org/abs/2502.06062

10 Kamilaris, A., and Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90 (2018). https://doi.org/10.1016/j.compag.2018.02.016

11 Manjunath, M.C., and Palayyan, B.P. An Efficient Crop Yield Prediction Framework Using Hybrid Machine Learning Model. Revue d'Intelligence Artificielle, 37 (4), 1157–1167 (2023). https://doi.org/10.18280/ria.370428

12 Chandraprabha, M., and Rajesh Kumar Dhanraj. Ensemble Deep Learning Algorithm for Forecasting of Rice Crop Yield based on Soil Nutrition Levels. ICST Transactions on Scalable Information Systems, 10 (3), e7–e7 (2023). https://doi.org/10.4108/eetsis.v10i3.2610

13 TNN. PAU-BITS Pilani tie up to marry agri with tech. The Times of India (May 27, 2025). Available at: https://timesofindia.indiatimes.com/city/ludhiana/pau-bits-pilani-tie-up-to-marry-agri-with-tech/articleshow/121446118.cms

14 Guruprasad, R.B., Saurav, K., and Randhawa, S. Machine Learning Methodologies for Paddy Yield Estimation in India: A Case Study. Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 5447–5450 (2019). https://doi.org/10.1109/IGARSS.2019.8900339

15 Photon Foundation. The Journal of Ethnobiology and Traditional Medicine. Available at: https://sites.google.com/site/photonfoundationorganization/home/the-journal-of-ethnobiology-and-traditional-medicine (accessed March 13, 2024).

16 Karthik Yasaswy, M.Y.S., Manimegalai, T., and Somasundaram, J. Crop Yield Prediction in Agriculture Using Gradient Boosting Algorithm Compared with Random Forest. Proceedings of the International Conference on Cyber Resilience (ICCR) (2022). https://doi.org/10.1109/ICCR56254.2022.9995829

17 Badshah, A., Alkazemi, B.Y., Din, F., Zamli, K.Z., and Haris, M. Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability. IEEE Access, 12, 162799–162813 (2024). https://doi.org/10.1109/ACCESS.2024.3486653

18 Chen, T., and Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794 (2016). https://doi.org/10.1145/2939672.2939785

19 Yandex. CatBoost: State-of-the-Art Open-Source Gradient Boosting Library with Categorical Features Support. Available at: https://catboost.ai/ (2019).

20 Rajesh Yamparla, Harisa Sultana Shaik, Naga, M.P., and Srilakshmi Nallamothu. Crop Yield Prediction using Random Forest Algorithm. Proceedings of the 7th International Conference on Communication and Electronics Systems (ICCES) (2022). https://doi.org/10.1109/ICCES54183.2022.9835756

21 Majnik, M., and Bosnić, Z. ROC Analysis of Classifiers in Machine Learning: A Survey. Intelligent Data Analysis, 17 (3), 531–558 (2013). https://doi.org/10.3233/IDA-130592

22 Ahmed, I. What is Hard and Soft Voting in Machine Learning? Medium (May 31, 2023). Available at: https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32

23 Lin, T.Y., Han, P.Y., Yin, O.S., How, K.W., and San, H.F. Stacking Ensemble Approach for Churn Prediction: Integrating CNN and Machine Learning Models with CatBoost Meta-Learner. Journal of Engineering Technology and Applied Physics, 5 (2), 99–107 (2023). https://doi.org/10.33093/JETAP.2023.5.2.12

[1]*Кулял М.,
ORCID ID: 0000-0002-2367-0897,
*e-mail: malika_21@rocketmail.com
[2]д-р. Уманг,
ORCID ID: 0000-0002-9458-5817,
e-mail: anilumang@yahoo.co.in
[1]д-р. Саксена П.,
ORCID ID: 0009-0007-0544-593X,
e-mail: parul_saxena@yahoo.com
[3]д-р. Пант Дж.,
ORCID ID: 0000-0003-2279-5556,
e-mail: geujay2020@gmail.com

[1]Компьютерлік ғылымдар кафедрасы, Soban Singh Jeena университеті,
Уттаракханд қ., Үндістан
[2]Компьютерлік қолданбалар кафедрасы, Kumaun университеті,
Уттаракханд қ., Үндістан
[3]Компьютерлік ғылымдар және инженерия кафедрасы, Graphic Era Hill университеті,
Уттаракханд қ., Үндістан

МАШИНАЛЫҚ ОҚЫТУҒА НЕГІЗДЕЛГЕН КҮРІШ ӨНІМДІЛІГІН БОЛЖАУ: БАЗАЛЫҚ ЖӘНЕ АНСАМБЛЬДІК МОДЕЛЬДЕРДІ САЛЫСТЫРМАЛЫ БАҒАЛАУ (УТТАРАКХАНД ШТАТЫ, УДХАМ-СИНГХ-НАГАР АУДАНЫ)

Аңдатпа

Күріш Үндістандағы азық-түлік қауіпсіздігінің негізі болып табылады, миллиондаған адамның тіршілігін және ұлттық экономиканы қолдайды. Алайда тұрақсыз климаттық жағдайлар күріш өнімділігін болжап болмайтын деңгейге жеткізеді. Бұл зерттеу Уттаракханд штатындағы Удхам Сингх Нагар ауданында күріш өнімділігін болжау үшін ауа райы, топырақ және егін деректерін біріктіретін машиналық оқыту жүйесін әзірлейді. Негізгі классификаторлар арасында CatBoost ең жақсы нәтижелерді көрсетті: дәлдігі 80,85% және ROC-AUC көрсеткіші 0,90. Өнімділікті одан әрі жақсарту үшін Optuna көмегімен бапталған CatBoost, XGBoost және LightGBM модельдері гибридті ансамбльдерге біріктірілді. CatBoost моделіне көбірек салмақ беретін WGB классификаторы ([3,1,1]) ең жоғары дәлдікке – 97,37%-ға қол жеткізді, одан кейін Stacking ансамбльдері (95,6%) және жұмсақ дауыс беру ансамбльдері (96%-ға дейін) орналасты. Бұл нәтижелер жоғары ROC-AUC көрсеткіштерімен расталды. Жалпы алғанда, зерттеу мұқият оңтайландырылған ансамбль модельдерінің өнімділікті болжау дәлдігін айтарлықтай арттыра алатынын көрсетеді, бұл Үндістанның климатқа сезімтал аймақтарында күріш өсіруді дәлірек әрі тұрақты етуге арналған практикалық құрал ұсынады.

Түйін сөздер: CatBoost, градиентті күшейту, гибридті ансамбль, машиналық оқыту, кездейсоқ орман, XG күшейту, өнімділік.

[1]* Кулял М..,
ORCID ID: 0000-0002-2367-0897,
*e-mail: malika_21@rocketmail.com
[2]д-р. Уманг,
ORCID ID: 0000-0002-9458-5817,
e-mail: anilumang@yahoo.co.in
[1]д-р. Саксена П.,
ORCID ID: 0009-0007-0544-593X,
e-mail: parul_saxena@yahoo.com
[3]д-р. Пант Дж.,
ORCID ID: 0000-0003-2279-5556,
e-mail: geujay2020@gmail.com

[1]Кафедра компьютерных наук, Университет Soban Singh Jeena, Алмора 263601, Уттаракханд, Индия
[2]Кафедра компьютерных приложений, Университет Kumaun, Найнитал, Уттаракханд, Индия
[3]Кафедра компьютерных наук и инженерии, Университет Graphic Era Hill, кампус Бхимтал, Уттаракханд, Индия

ПРОГНОЗИРОВАНИЕ УРОЖАЙНОСТИ РИСА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ БАЗОВЫХ И АНСАМБЛЕВЫХ МОДЕЛЕЙ НА ПРИМЕРЕ ОКРУГА УДХАМ-СИНГХ-НАГАР, ШТАТ УТТАРАКХАНД (ИНДИЯ)

Аннотация

Рис является краеугольным камнем продовольственной безопасности в Индии, поддерживая миллионы средств к существованию и национальную экономику. Однако нестабильные климатические условия делают урожайность риса все более непредсказуемой. В данном исследовании разрабатывается система машинного обучения для прогнозирования урожайности риса в районе Удхам Сингх Нагар, Уттаракханд, путем интеграции данных о погоде, почве и севах. Среди базовых классификаторов CatBoost показал лучшие результаты с точностью 80,85% и ROC-AUC 0,90. Для дальнейшего повышения производительности Optuna-tuning модели CatBoost, XGBoost и LightGBM были объединены в гибридные комплекты. Классификатор взвешенного жесткого голосования, придающий больший вес CatBoost ([3,1,1]), достиг наивысшей точности – 97,37%, за ним следуют ансамбли Stacking (95,6%) и ансамбли мягкого голосования (до 96%). Эти результаты были подтверждены высокими баллами ROC-AUC. В целом исследование показывает, что тщательно оптимизированные ансамблевые модели могут значительно повысить точность прогнозирования урожайности, предоставляя практический инструмент для более точного и устойчивого рисового выращивания в климатически чувствительных регионах Индии.

Ключевые слова: CatBoost, усиление градиента, гибридный ансамбль, машинное обучение, случайный лес, усиление XG, доходность.