Journal of Agriculture and Food Research 18 (2024) 101424

Contents lists available at ScienceDirect

Journal of Agriculture and Food Research

journal homepage: www.sciencedirect.com/journal/journal-of-agriculture-and-food-research

# Advanced prediction of rice yield gaps under climate uncertainty using machine learning techniques in Eastern India

Satiprasad Sahoo [a,b,*], Chiranjit Singha [c], Ajit Govind [a]

[a] International Center for Agricultural Research in the Dry Areas (ICARDA), 2 Port Said, Victoria Sq, Ismail El-Shaaer Building, Maadi, Cairo, 11728, Egypt, 15A
[b] Prajukti Research Private Limited, Baruipur, 743610, West Bengal, India
[c] Department of Agricultural Engineering, Institute of Agriculture, Visva-Bharati (A Central University), Sriniketan, Birbhum, West Bengal, 731236, India

ARTICLE INFO

*Keywords:*
Food security
Rice yield gap
Machine learning
Remote sensing

ABSTRACT

The current study focuses on applying machine learning approaches to forecast future Kharif rice yield gaps in eastern India while accounting for climate change implications. To achieve the United Nations Sustainable Development Goals (SDGs), food security must be prioritized. Rice yield has been predicted using Cubist, GBM, MARS, RF, SVM, and XGB machine learning methods, considering six factors: elevation, soil moisture, precipitation, temperature, soil temperature, and actual evapotranspiration. Climatic change scenarios were generated using the latest climatic Coupled Model Intercomparison Project Phase 6 (CMIP6 MIROC6) Shared Socioeconomic Pathways (SSP) 2–4.5 and SSP5-8.5 datasets between 1990 and 2030. Finally, machine learning algorithms were used to identify rice yield gaps to achieve sustainable agricultural intensification. The rice yield validation was completed with 1889 field-based farmer observation records. The results suggest that Murshidabad and Purba Bardhaman districts had very high rice yields (5.60–3.45 t/ha) when using the Cubist model compared to another model. The findings also reveal a poor rice-yielding zone (1.44–0.39 t/ha) in the western region (Purulia) and a northwestern region (half of the west of Birbhum). The Cubist and RF models' maximum and minimum $R^2$ values were 0.73 and 0.72, respectively. The R2 values were also negligible for the XGB, GBM, SVM, and MARS, machine learning models. Projections for rice production in 2030 indicate that the northern and north-eastern regions (Murshidabad and Purba Bardhaman) as well as the southeastern areas (Jhargram and Paschim Medinipur) have the highest yields, categorized as extremely very high (5.56–3.49 t/ha) and high (3.49–2.49 t/ha). A significant rice yield gap (50-40 %) was found in the center and south-east areas (Bankura, Jhargram, and Paschim Medinipur), the northern region (Murshidabad and Birbhum), and the western region (Purulia). In 2030, the north-western region (Birbhum), as well as the middle and south-eastern regions (Bankura, Jhargram, and Paschim Medinipur districts), had the highest proportion of high (50%–40 %) and very high (60%–50 %) rice yield gap. Our findings can contribute to a new viewpoint on agricultural planning and management for sustainable growth in the face of changing climate circumstances.

## 1. Introduction

Agriculture is the primary source of income for all humans, with agriculture employing 50 % of India's workforce and accounting for 17–18 % of the country's Gross domestic product (GDP). Thus, food security is a crucial concern in a densely populated country. The United Nations has set zero hunger as a sustainable development goal (SDG) for greater future agricultural development. As a result, crop production estimates are critical for food security and achieving zero hunger. The importance of food security and sustainable agriculture goes beyond just supplying food; it involves promoting economic stability, enhancing social well-being, safeguarding the environment, biodiversity conservation, and maintaining ethical principles. Implementing sustainable agricultural practices allows us to build a robust food system that can feed the global population while conserving the planet for future generations [1,2]. The current study's methodological framework focuses on the present and future projection of rice crop yield to calculate yield gap mapping utilizing machine learning approaches on geospatial platforms.

Several studies have attempted to identify yield gaps by predicting

* Corresponding author. International Center for Agricultural Research in the Dry Areas (ICARDA), 2 Port Said, Victoria Sq, Ismail El-Shaaer Building, Maadi, Cairo 11728, Egypt, 15A.
*E-mail address:* satispss@gmail.com (S. Sahoo).

https://doi.org/10.1016/j.jafr.2024.101424
Received 26 April 2024; Received in revised form 1 July 2024; Accepted 16 September 2024
Available online 17 September 2024

historical, current, and future rice crop yields [3–5]. Arumugam et al. [6] used the Decision Support System for Agrotechnology Transfer (DSSAT) to estimate Kharif rice (Oryza sativa L.) production in India from 2001 to 2017, considering weather data, crop mask, sowing dates, and irrigation information. They concentrated on the Pradhan Mantri Fasal Bima Yojana insurance policy, which can be used by Indian farmers for security reasons. Van Klompenburg et al. [7] conducted a thorough literature assessment for agricultural yield projections, considering the algorithm and features. They gathered 567 relevant studies from various internet databases and selected 50 relevant studies for further study. They identified the most utilized features for agricultural yield mapping, such as precipitation, temperature, and soil type. Senthilkumar et al. [8] evaluated rice yield gaps in 2012–13 using quantitative and qualitative data at the farm and field level to improve rice productivity in Ethiopia, Madagascar, Rwanda, Tanzania, and Uganda. Akhter et al. [9] forecast kharif rice yield in Gangetic West Bengal using the Indian Institute of Tropical Meteorology- India Meteorological Department (IITM-IMD) extended range prediction (ERP) method. They examined rice yield data from ERP and the Decision Support System for Agro-technology Transfer (DSSAT). Finally, we conclude that ERP-based rice prediction data outperforms the others Wilson et al. [10] investigated rice crop yield prediction using a variety of machine learning approaches, including K closest neighbor, random forest, linear regression, decision tree, Xgboost, and support vector regression. They considered 15 distinct dataset categories, including soil type, organic carbon, magnesium, boron, potassium, soil temperature, pH, copper, iron, precipitation, and humidity, all of which are used to calculate rice crop output in Kerala. When compared to other approaches, it was discovered that K closest neighbor (KNN) had the highest accuracy at 98.77 %. Debnath et al. [11] forecast future rice yield gap mapping in India from 1981 to 2050 using the RCP 8.5 scenario and the DSSAT. Under future climatic uncertainty, yields were found to fall by 30–60 %. Liu et al. [12] developed a transformer-based model to predict rice yield in the Indian Indo-Gangetic Plains from 2001 to 2016, considering various environmental variables. They employed machine learning and deep learning models to predict rice crop production starting two months before rice maturity, and they created a dependable and simple framework for crop yield prediction. Nayek et al. [4] used stochastic frontier approaches to map rice yield gaps and assess the possibilities for irrigation water reduction in India's Northwestern Indo-Gangetic Plains. However, rice production data from the Global Yield Gap Atlas were used to estimate yield gaps and study nitrogen-use efficiency in varied farmer practices. Yuan et al. [13] mapped rice yield gaps using meteorological and soil data from Southeast Asia, as well as yield data from the Global Yield Gap Atlas. They predicted future rice consumption based on the expected population from the medium fertility option (URL: https://population.un.org). It was discovered that a significant rice yield difference existed in Cambodia, Myanmar, the Philippines, and Thailand, but was much less in Indonesia and Vietnam. Sathya and Gnanasekaran [14] anticipate paddy production using the Multi-layer Representation Learning (MLRL-STM) algorithm with weather and soil data to preserve food security. They employed machine learning and deep learning approaches to estimate agricultural yield and suggested a hybrid model for yield prediction in Thanjavur, Tamil Nadu.

Quille-Mamani et al [15] predict rice yield across the Lambayeque region, Peru by Sentinel-2 imagery of 15 phenological indices from 32 farm plots. The results showed that SVM had better performance ($R^2$ = 0.69, RMSE = 1.23 t/ha, MAE = 1.01) than two other machine learning models (i.e. LR and RF). Liu et al. [16] utilized four machine learning models—support vector regression (SVR), partial least squares regression (PLSR), back propagation neural network (BPNN), and random forest regression (RFR)—to estimate rice yields in China. They used MODIS-derived multi-temporal rice NDVI data spanning from 2001 to 2020. Among these models, the RFR model proved to be the most accurate, with an $R^2$ of 0.65, an RMSE of 388.79 kg/ha, and an rRMSE of 4.48 %.

Most researchers concentrated on statistical rice yield prediction rather than spatially represented spatial distribution across the study areas [4,17]. Many of the researchers only used Analytical Hierarchy Process (AHP), fuzzy, or machine learning (ML) methods to forecast rice yields [18,19], there has been limited focus on achieving the desired accuracy for predicting the rice yield gaps in sustainable food security management. The study addressed the research gap in the study area domain by implementing rice yield gap analysis for the first time using various hydroclimatic parameters, based on future projections from the CMIP6 data sets. In this study, we address the following research questions: (i) What factors are associated with rice yield gap analysis? (ii) Does an ensemble machine learning (ML) approach satisfactorily perform in rice yield gap mapping? and (iii) How to CMIP6 climate modeling ensure the effectiveness of the yield gap methodology in providing insightful aids for future food security assessment in the study area? To answer these questions, the current study concentrated on identifying present and prospective rice yield gaps in Eastern India using CMIP6 climate modeling data and machine learning approaches. This study aimed to develop a scalable methodology for predicting current and future rice yields and yield gaps, along with their underlying causes, in Eastern India's rainfed and irrigated production environments. This approach combined hydroclimatic data, machine learning, and ground information. Compare different ML approach results and assess the sensitivity of explanatory factors contributing to the rice yield gap using the multicollinearity, ordinary least squares regression (OLS), Boruta, and Shapley additive explanations (SHAP) analysis. Specifically, we compare the performance of Random Forest (RF), Extreme gradient boosting (XGB), Gradient boosting machine (GBM), Cubist, multivariate adaptive regression splines (MARS), and Support vector machine (SVM) regressor. Additionally, the Model for Interdisciplinary Research on Climate 6 (MIROC6) SSP2-4.5 and SSP5-8.5-based data were used to predict future rice yield gaps with precipitation and temperature data. This approach ensures food security, economic stability, and the livelihoods of farming communities in the study area.

## 2. Study area

We selected a study of a western lateritic region with low rainfall intensity to challenge rising agricultural productivity. This region primarily confronts climate change, characterized by high temperatures, limited precipitation, and decreased vegetation cover. Geographically, this region spans 21° 56′ N to 24° 52′ N and 88° 44′ E to 85° 45′ E. The research area is in the middle and southern regions of West Bengal (Fig. 1). The total area of this region is 37836.5 $km^2$. It consists of 8 districts: Murshidabad, Birbhum, Purba Bardhaman, Paschim Bardhaman, Bankura, Purulia, Jhargram, and Paschim Medinipur. Murshidabad district is in the heart of West Bengal, with an average elevation of 10 m (30 feet) above mean sea level (MSL). The climate in this district is humid. This district is separated into two parts by the Bhagirathi River, which flows from north to south. The western half, known as 'Radh', has undulating topography and strong reddish clay soil, while the eastern part has extensive marshes and an old riverbed. The eastern section, known as 'Bagri', has fertile alluvial soil that is ideal for farming. Birbhum district is also known as "The Land of Red Soil". The average elevation in this district is 230 feet. The western section of the district is located under the Chota Nagpur Plateau and gradually descends to the lush alluvial plains in the east District Survey Report of Birbhum [20]. The Ajay, Bakreshwar, Brahmani, Bansloi, Dwarka, Hinglo, Kopai, and Mayurakshi rivers flow from east to west in the Birbhum district.

The climate is milder on the eastern side compared to the western side. Purba Bardhaman district is known as the "Rice Bowl of West Bengal". This district's average elevation is 30 m (131 feet). The western section of the district has old alluvial soil, whereas the eastern part has new alluvial soil. This location is in the hot and humid tropical climate zone. Paschim Bardhaman district is situated on the northern bank of the

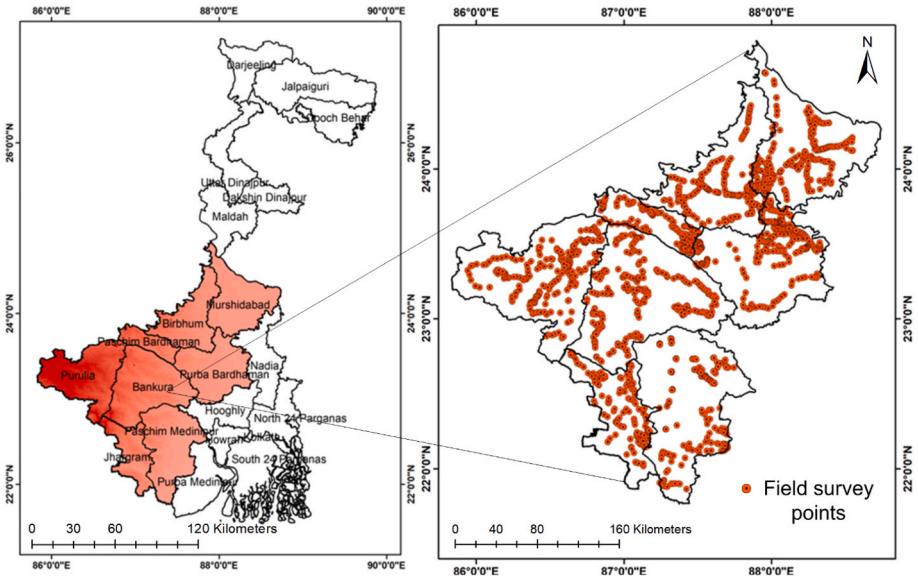


**Fig. 1.** Location of the study area map.

Damodar River. The district has an elevation of 40 m, with a somewhat higher relative relief in the western half compared to the northeast. However, the slope is mild from west to east, and the climate is tropical, both rainy and dry. Laterite Soil is found here that is very permeable, low in organic matter, and acidic. Drought-prone areas Bankura District is in the western section of West Bengal. Damodar The river flows along the district's northern boundary. This area consists of undulating uplands that progressively descend from the Chota Nagpur Plateau. The climate is usually dry and hot in summer, with mild temperatures in winter. This location has red and lateritic soil and an elevation of 78 m (256 feet). Farmers in this district rely heavily on monsoon rains, and irrigation facilities are few. Climate change is seen in the westernmost district of West Bengal.

Purulia is located on the eastern slope of the Chotonagpur Plateau. It has an average elevation of 228 m (748 feet). Purulia's laterite soil includes iron and varies in acidity, alkalinity, and clay type. This district endured hot and humid weather. The district is divided into two physiographic areas: the higher plateau, which includes the Baghmundi and Ajodhya ranges in the west, and the lower plateau, which spans the eastern half. The Baghmundi region has an average altitude of 400–600 m, whereas the Ajodhya range contains peaks above 600 m. Due to undulated geography, over 50 % of precipitation is lost as runoff. Several rivers run through the district, including Damodar, Dwarakeswar, Kangsabati, Kumari, Silabati (Silai), and Subarnarekha. Small dams such as Burda, Futiyary, Gopalpur, Murguma, and Pardi are used to irrigate agricultural fields. Jhargram District is in West Bengal's humid and tropical region, at an average elevation of 81 m above MSL in the red and lateritic zone. This district is located on the Chota Nagpur Plateau, with a gentle eastward slope. This district is situated between the Kangasabati River in the north and the Subarnarekha River in the south. East Medinipur is in West Bengal's southernmost district. It is part of the lower Indo-Gangetic Plain and Eastern Coastal Plains, which are divided into two parts: flat plains to the west, east, and north, and coastal plains to the south. This area is primarily made up of coastal and younger alluvial deposits. The district's elevation is within 10 m of MSL. Tropical wet and dry climates are viewed in this location.

## 3. Material and methods

### 3.1. Data used

#### 3.1.1. Sampling of crop yield

The kharif rice crop yield (t/ha) was collected from 1889 agricultural fields during the survey work (2022–2023) in post-harvesting periods. Before the main survey, a pilot test was conducted with a small group of respondents (including farmers, community members of various ages, fertilizer and seed stores, local senior citizens, and government officials) to identify and address any issues with the questionnaire and potential outcomes. This helps us to design a better survey framework. The field data was gathered during a field survey using a handheld GPS device to geo-tag rice production areas in Eastern India (Fig. 1). The survey collected extensive data on rice yield (in tons per hectare) and input usage from the largest plot cultivated by each farmer in two consecutive Kharif seasons, spanning from 2022 to 2023. During the field survey, the varieties of Kharif rice species growing in the study namely Swarna, Ratna, Lalat, Varam, IR36, IR41, MTU7029, Badsha bhog, Dhanaraj, Chaitali, Maharaj, Minikit, Gtka, Gobindo bhog etc. A random sampling analysis has been steered with a handheld GPS (Garmin Ltd., Olathe, KS, USA) and detected the coordinates at each sampling location. Use random sampling techniques to select farms or plots for data collection. This reduces the likelihood of systematic biases that could skew results. During the survey, detailed information about rice yield information was gathered, including aspects like farming methods, crop disease, fertilizer input, irrigation plans, soil conditions, soil moisture impact of climatic hazards, market trends, and more, integrating the insights and experiences of local farmers for in-depth analysis. The study involved visiting multiple locations in the area to gather samples and document the experiences and observations of local farmers about recent changes in their crop rotation. In this study, outlier detection was performed through box plot analysis. The statistical analysis of rice yield in the study area revealed that the average yield is 1.96 (t/ha), with a standard deviation of 1.48. The minimum recorded yield was 0.017 (t/ha), while the maximum yield reached 6.75 (t/ha). This crop yield database was employed as the dependent variable for the rice yield gap prediction (RYP) analysis. For the ML cross-validation analysis, the yield samples are subset into training (70 %–1322 sample location) and testing (30 %–567 sample location) sets.

After fitting the predictive models to both observed and predicted yields, the predictors were employed to generate actual yield (Ya)

predictions at the plot level, followed by the estimation of potential yield (Yp). The yield gap (Yg) for each farmer was calculated by subtracting the actual yield (Ya) from the potential yield (Yp), which was simulated using vegetation indices (Yg = Yp - Ya) [3]. Subsequently, we calibrated and validated all MLs to predict yield using geo-environmental and hydroclimatic data over the study region. We also identify the key determinants of the yield gap based on the importance values of the variables.

#### 3.1.2. Geo-environmental parameters

Six environmental parameters were applied, based on previous literature [21–25] expert perceptions, and the specific surroundings of the study domain. The six effective parameters are elevation, soil moisture (sm), precipitation (Pr), temperature (temp), soil temperature (stemp), and Actual Evapotranspiration (aet) for spatial prediction of RYP mapping (Supplementary Fig. 1). Table 1 shows the dataset description and sources for the RYP mapping analysis.

In assessing the effects of climate change, the most frequently used factors are temperature and precipitation [22,24,26]. Consistent with previous research, this study focuses on precipitation and temperature as the primary climatic parameters. The information on Pr, temperature, aet, and sm, was obtained from the TerraClimate datasets (1958–2022) through the Google Earth Engine (GEE) cloud. All the climatic variables within the study boundaries were averaged then the IDW (Inverse Distance Weightage) interpolated method was applied with the ArcGIS software v10.7. Topographical variation of the study area extracted from the Shuttle Radar Topography Mission (SRTM) DEM data (USGS/SRTMGL1_003, 30m spatial resolution). The stemp data were derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) Climate Reanalysis (Temperature of the soil in layer 2 databases at 15–30 cm depth (1958–2022). All the assigned six geo-environmental parameters are resampled (30m × 30m) with the bilinear interpolation method in R software v.3.4.2 [27,28].

Comprehending crop yield and water demands will adapt to future climatic conditions on a local level is crucial for formulating more accurate and effective adaptation approaches for climate resilience agricultural domain [29,30]. The current study employed the temperature and precipitation data with MIROC6 global circulation model datasets from project phase 6 of the Coupled Model Intercomparison Project (CMIP6) (Supplementary Fig. 2). Shared Socioeconomic Pathways (SSPs) of future periods spanning 1990 and 2030 under SSP2-4.5 and SSP5-8.5 scenarios were employed to crop yield prediction as well as RYP mapping in the future.

### 3.2. Methodology

The research methodology is summarized as follows: Initially, the study utilized a crop yield database alongside six crop yield predictors to prepare datasets for training, and testing phases of machine learning (ML) models, preceded by a statistical Ordinary Least Square Regression (OLS), multicollinearity and Boruta analysis. Subsequently, RYP maps were produced by analyzing the spatial relationships between crop yield predictors and crop yield coupled with evaluating the ML model's accuracy through selected metrics (i.e. $R^2$, RMSE, MAE, and MSE).

Lastly, the most effective model was identified, and an in-depth analysis of the crop yield predictors' significance was conducted using an explainable AI (XAI) technique, specifically Shapley Additive exPlanations (SHAP). The overall methodological framework is shown in Fig. 2.

#### 3.2.1. OLS and VIF analysis

The generalized linear modeling approach utilized in this research is known as ordinary least squares (OLS) regression. It is designed to model a response or dependent variable as yield, offering a comprehensive prediction model. This method permits the inclusion of either a single or multiple explanatory variables in the RYP analysis. The mathematical calculation of the OLS was executed with Eq.1 [27,28]. The variation inflation factors (VIF) justified the collinearity check of the yield predictors. Moreover, the study considered the cut-off of VIF score strictly <5.

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_n x_n + \varepsilon \ldots 1$$

where y denotes the dependent parameters; βo represents the intercept; β1, β2, β3 … βn are the coefficient score of the independent parameters x (x1, x2, x3 … xn); and ε denote the error term.

#### 3.2.2. Boruta feature selection analysis

This research employed random forest-based Boruta feature selection techniques to assess the efficacy of the proposed RYP model. The Boruta algorithm performed 10 iterations in 4.938149 s through R software v.3.4.2. through random forest model. The ranking of each predictor is confirmed with the mean importance value [35].

#### 3.2.3. Machine learning regression application

Numerous research works have highlighted the significance of machine learning as a key instrument in supporting decision-making for crop yield predictions [36–38]. Machine learning serves as an invaluable aid to farmers by providing comprehensive advice and insights into crop management, thereby aiding in minimizing agricultural losses. The machine learning models examined in this study include the Random Forest (RF), Extreme gradient boosting (XGB), Gradient boosting machine (GBM), Cubist, multivariate adaptive regression splines (MARS), and Support vector machine (SVM) regressor respectively. The selection of these methodologies was influenced by the quantitative characteristics of the predictive data and the extent of the dataset involved. All the employed ML model tuning parameters are summarized in Supplementary Fig. 3 and Supplementary Table 1.

#### 3.2.4. Random forest (RF)

The RF algorithm stands out as a key supervised machine learning technique, adept at handling both classification and regression challenges. The Random Forest model is a powerful and versatile tool for

**Table 1**
Description of data sources for rice crop yield prediction.

| Parameters Group | Description | Source | Reference |
|---|---|---|---|
| Elevation (degree) | SRTM DEM (30 m) | (URL: usgs.gov.in) | [31] |
| Soil temperature (°C) | ECMWF Climate Reanalysis (Temperature of the soil in layer 2), 11132m, (15–30 cm depth), (annual average 1958–2022) | (URL: https://cds.climate.copernicus.eu) | [32] |
| Soil moisture (mm) | IDAHO_EPSCOR/TERRACLIMATE, 4638.3 m, (annual average 1958–2022), (15–30 cm depth) | (URL: https://www.climatologylab.org/terraclimate) | [33] |
| Precipitation (mm) | IDAHO_EPSCOR/TERRACLIMATE, 4638.3 m, (annual average 1958–2022) | | |
| Temperature (°C) | IDAHO_EPSCOR/TERRACLIMATE, 4638.3 m (annual average 1958–2022) | | |
| Actual Evapotranspiration (mm) | IDAHO_EPSCOR/TERRACLIMATE, 4638.3 m (annual average 1958–2022) | | |
| CMIP6 Precipitation (mm) | NASA/GDDP-CMIP6, 27830 m, (annual average 1990–2030), (ssp245, ssp585) | (URL: https://registry.opendata.aws/nex-gddp-cmip6/) | [34] |
| CMIP6 Temperature (°C) | | | |

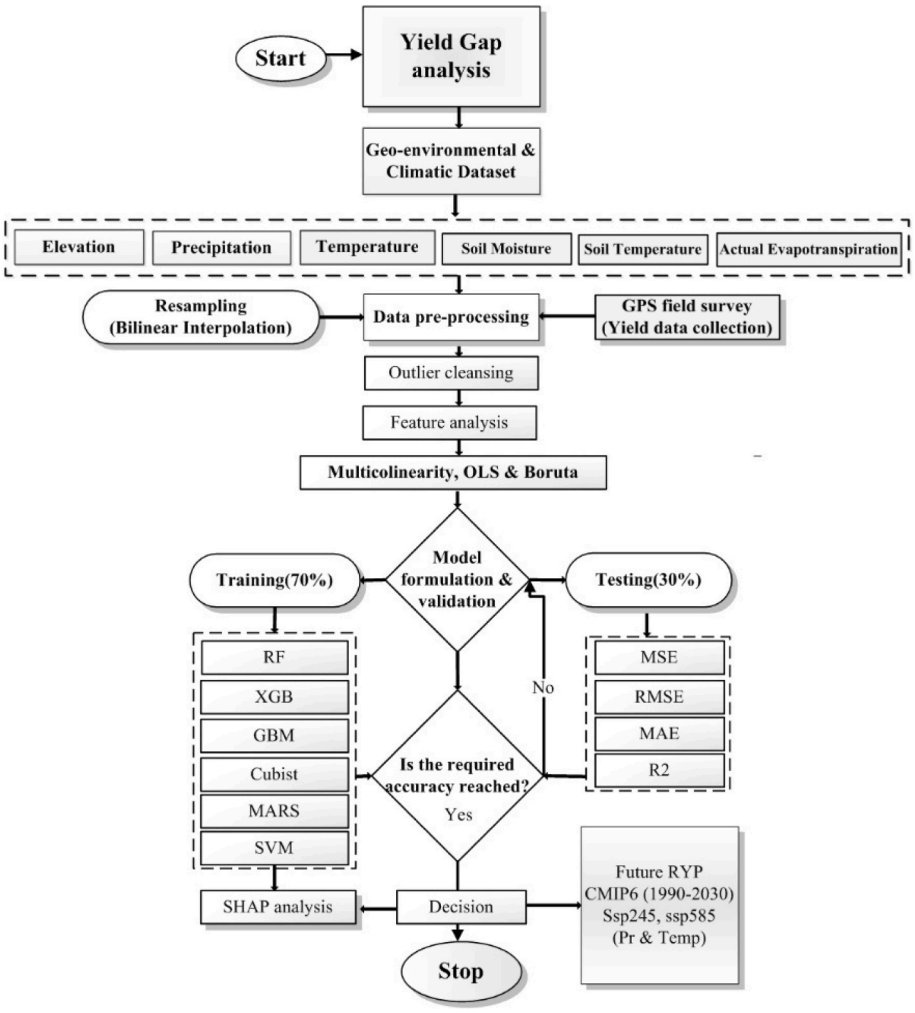


**Fig. 2.** Overall methodological framework for rice yield prediction.

various predictive modeling tasks with there accuracy, robustness to overfitting, handling of missing values, feature importance, variability, scalability, etc. Its ability to produce high-accuracy predictions and robustness against overfitting make it a popular choice among data scientists. It operates by employing a multitude of decision trees (DT), leveraging the bootstrap technique, and implementing aggregation [39]. It starts at the base and navigates through various bifurcations based on variable outcomes, ultimately culminating at a leaf node that reveals the final decision. In these trees, a starting feature, say Feature A, initiates a split based on a predefined criterion. Depending on the response being affirmative or negative, the tree follows a corresponding path, repeating this process until it arrives at the leaf node where the decision is finalized. Bootstrapping involves randomly selecting data subsets for multiple iterations and variables. The aggregation, or ensemble approach, combines multiple models trained on the same dataset, averaging their outputs to enhance the overall predictive or classification accuracy. In this study, the RF model was configured with the following parameters: mtry set to 1, number of trees set to 500, proximity enabled, and fit Best disabled. However, its complexity, computational requirements, and potential biases must be considered when selecting it for specific applications. Understanding its strengths and limitations helps in effectively utilizing the RF model to achieve reliable and interpretable results.

### 3.2.5. *Extreme gradient boosting (XGB)*

The *XGB* algorithm is an advanced tree optimization ML tool that has gained widespread usage in various data analysis domains. Uniquely intended as an applied gradient-boosting machine, especially for regression and classification trees. It employs the boosting concept. This concept integrates the predictions of weak learners through additive training methods to form a robust learner, aiding in preventing overfitting and enhancing mathematical proficiency. The XGB architecture, demonstrating the simplification of objective functions by combining prediction and regularization terms, minimizes loss function while maintaining optimal processing speed [40]. XGB is an advanced implementation of the gradient boosting framework designed to optimize both performance and efficiency. Here's the rationale behind using the XGB model: Gradient Boosting Framework, Regularization functionality, capabilities of missing values handling, scalability, and custom Objective functions. Its ability to handle large datasets and deliver accurate predictions makes it a popular choice for crop yield prediction analysis. However, its complexity, computational requirements, and the need for careful hyperparameter tuning must be considered when choosing it for specific tasks. In this study, the XGB model was tuned using the following optimal parameters: eta:0.3, maximum depth: 3, gamma: 0.001, coal sample by tree: 0.8, minimum child weight:1, subsample:1 and objective function:reg: squared error respectively.

### 3.2.6. *Gradient boosting machine (GBM)*

The GBM emerges as a highly influential machine learning algorithm, garnering significant attention across diverse environmental applications such as agriculture, climatology, and soil studies [41,42].

GBM is a machine learning technique that builds predictive models in a stage-wise fashion, combining the predictions of multiple weak learners (typically decision trees) to create a strong learner. Here's the rationale behind using the GBM model: sequential learning ability, gradient descent optimization funncanality, regularization, and ensemble learning capability.

Renowned for its robust decision-making capabilities using tree-like structures, GBM is particularly recognized for its effectiveness in crop yield estimation, acknowledged by experts in the field. An inherent advantage of GBM lies in its versatility, capable of addressing both regression and classification tasks, thus facilitating effective decision-making in various contexts. The ensemble learning algorithm adopts an iterative approach, sequentially building upon weak trees, thereby progressively enhancing the performance of previous models. This iterative process results in powerful predictive models that excel in delivering accurate estimates. However, its computational demands, sensitivity to noisy data, and the need for careful parameter tuning should be considered when applying GBM to real-world problems. The current study utilized the following parameters for GBM modeling i.e., n.trees: 150, interaction. depth: 3, shrinkage: 0.1, and n.min-obsinnode:10. A gradient with loss function: gaussian boosted model, and iterations:150.

### 3.2.7. Cubist

The Cubist machine-learning nonparametric analysis method is an approach rooted in the construction of regression trees. Initially, a tree is constructed by defining rules that partition the data into reasonably homogeneous groups concerning the variable of interest, such as productivity, about the predictor variables. According to Ref. [43], the prediction-oriented regression model referred to as the cubist model stands out for its distinctive approach. A key advantage of the cubist method lies in its incorporation of multiple training committees, a feature designed to balance case weights effectively. The Cubist model is a rule-based machine learning technique that combines regression trees with linear models to make predictions. Here's the rationale behind using the Cubist model: rule-based approach, ensemble learning, and variable selection. This model has various limitations i.e., heavily depends on parameter settings, is limited for regression tasks, and is computationally intensive. The Cubist model offers a blend of interpretability and predictive accuracy, making it suitable for scenarios where understanding the decision-making process is as important as accurate predictions. While it excels in transparency and non-linear relationship modeling, researchers and practitioners should consider its complexity, parameter sensitivity, and suitability for specific tasks when choosing it for machine learning applications. This study utilized different Cubist tuning parameters for analyzing the rice yield gap, including committees: 20, neighbors: 9, and several rules per committee: 18, 13, 12, 12, 14, 8, 17, 12, 10, 10, 12, 8, 13, 12, 13, 11, 12, 10, 14, 10 for rice yield gap analysis.

### 3.2.8. Multivariate adaptive regression splines (MARS)

MARS, a powerful technique, excels in creating a strong predictive model for a response variable [44]. It is particularly useful in agriculture, where it estimates yields or other characteristics based on agronomic traits. Both simple and multiple linear regression methods are employed for determining plant traits. However, it's important to note that deviating from distributional assumptions can negatively impact the reliability of these methods. MARS is a non-parametric regression technique that builds models by partitioning the input space into regions characterized by linear functions. It offers a balance between flexibility in modeling non-linear relationships and interpretability through segment-based linear models. The rationale behind using the MARS model includes piecewise linear modeling, adaptive basis functions, and automatic interaction detection. While its strengths include adaptability and transparency, researchers and practitioners should consider its limitations, such as sensitivity to noise, overfitting risk, limited to regression, and computational demands, when applying MARS to real-world regression problems. Understanding these trade-offs is essential for effectively leveraging MARS to derive meaningful insights from data while ensuring robust model performance. This study employed the various best tuning paramtersfor MARS modelling including, nprune: 14 and degree:1.

### 3.2.9. Support vector machine (SVM)

SVM is a supervised machine learning method employed to analyze the linear association between two continuous variables. The rationale behind using the SVM model includes maximum margin classifier, kernel trick for non-linear separability, regularization parameter, and effective in High-dimensional spaces. The core concept of SVM involves identifying an optimal fit line while maintaining the fitting error within a specific threshold. These optimal fits aim to estimate the best value within a defined hyperplane margin. In SVM, the most effective hyperplane is the one that encompasses the maximum number of points [45]. Its strengths lie in its robustness to overfitting, versatility in kernel functions, and ability to find global optimum solutions. However, practitioners should consider its computational intensity, sensitivity to kernel choice, and limitations in interpretability when applying SVMs to real-world problems.In this study the different hyperparameter were used i.e., epsilon: 0.1 cost C:1, kernel: Gaussian Radial Basis function,: sigma:0.469, number of Support Vectors: 709, and objective function value: −268.358.

### 3.2.10. Model validation

In this study, the 10-fold cross-validation was considered for their model validation efficiency. The various statistical metrics namely mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and $R^2$ employed for evaluating ML regression models.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{(y_i \sum \hat{y}_i)^2}{n}} \tag{3}$$

$$\mathrm{MAE} = \sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{n} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{5}$$

where in Equations (2)–(5), y represents the predicted yield, and $\hat{y}$ is the observed yield.

### 3.2.11. SHapley additive exPlanations (SHAP) analysis

Assessing feature importance for all machine learning models involved utilizing the Tree Explainer from the shape package in Python v3.7.0., aiming to identify the most influential predictors for RYP analysis. The Tree Explainer method utilizes Shapley values to portray both the global importance of features and their ranking, as well as the local impact of each feature on the model output. Utilizing SHAP, an interpretable machine learning (IML) approach grounded in game theory [46], the study employed this methodology to ascertain the contribution of each variable to the modeled yield across the study area. The variable causing the most significant decrease in yield at each point was identified and mapped throughout the study area, with the spatial extent represented by each variable quantified. Additionally, SHAP values for each predictor variable were extracted and mapped for a case study field. This mapping demonstrated the magnitude of the impact of each variable on yield within a spatial context, presented in easily

interpretable units (t/ha). This analysis was conducted on the model predictions using a representative sample from the testing dataset.

## 4. Results

### 4.1. OLS and multicollinearity test

The precipitation, temperature, and elevation demonstrate a high level of 99 % ($p < 0.001$), agreement, reaching regarding rice yield prediction in the study area. On the other hand, aet, sm and stem exhibit significance levels of 95 % ($p < 0.05$). Statistical validation of these rice yield predictors was conducted using ML analysis. According to the multicollinearity analysis (Table 1), it is found that the elevation factor has the maximum score of variance inflation factor (VIF) (3.43) while the minimum value of VIF is temperature (1.10) (Table 2). This result indicates that, there is no significant collinearity issues among the predictors for rice yield gap analysis. Allowing all 6 RYPs to be considered for utilization in the Rice Yield Prediction (RYP) modeling process.

### 4.2. Boruta feature selection

According to the Boruta feature selection results (Fig. 4), stemp factors achieved maximum mean importance (55.30) (Table 3). Finally based on that result, all the crop yield parameters are confirmed for the parameters sensitivity analysis.

Various ML techniques (cubist, GBM, MARS, RF, SVM, and XGB) were used to predict Kharif (monsoon) rice yield in Murshidabad, Birbhum, Paschim Bardhaman, Purba Bardhaman, Bankura, Purulia, Paschim Medinipur, and Jhargram districts of West Bengal, considering five parameters such as elevation, soil temperature, soil moisture, actual evapotranspiration, precipitation, and temperature. Future rice yield projection was made using Coupled Model Intercomparison Project Phase 6 (CMIP6) climate model-based precipitation and temperature data under SSP2–4.5 and SSP5-8.5 by ML. Finally, current, and projected Kharif rice yield gap mapping has been constructed using simulated and field-based rice yield data. Furthermore, in 1889, farmers were surveyed using data used for ML model training (70 %) and testing (30 %) purposes.

### 4.3. Analysis of rice yield predictions

The six ML models including cubist, GBM, MARS, RF, SVM, and XGB were used to calculate the current Kharif rice yield for 2023. The field-based Kharif rice yield data were used for calibration and validation purposes. The findings indicate that Murshidabad and Purba Bardhaman districts have exceptionally good rice yields due to adequate water availability. Using the Cubist model, we noticed that the highest level of rice yield (5.60–3.45 t/ha) is seen in the northern and north-eastern regions (Murshidabad, East Burdwan districts, and some of the north-eastern and southern parts of Birbhum district, as well as some of West Burdwan district), and partially in the western and south-eastern regions (western part of West Burdwan and the border area of Jhargram and East Medinipur Districts). Rice yields are high in the majority of the south and south-eastern regions (Jhargram, Paschim Medinipur districts, and a small piece of the border area between Jhargram and Bankura districts). It can be seen in the northern region (some of the western and northern parts of Murshidabad, as well as the eastern and southern parts of Birbhum), the eastern region (the eastern and southern parts of East Burdwan, the northern part of Bankura), and a very small portion of the western region. This study area has a scattered distribution of moderate rice yields (2.37–1.44 t/ha), similar to the eastern, southern, and western parts of Birbhum, some of West Burdwan, partially eastern and north-western parts of Bankura, scattered distribution of Jhargram and the northern part of East Medinipur, and scattered distribution of Purulia districts. The western region of the study area (overall Purulia and partial western part of Birbhum) is in a low rice-yielding zone (1.44–0.39 t/ha), while the middle portion of the study area (overall Bankura districts) and a portion of the north-western region (western part of Birbhum district) have very low rice yields (0.39–0.10 t/ha).

Using GBM model, we noticed that very high rice yields (5.60–3.45 t/ha) are observed in the maximum portion of the northern and north-eastern regions (Murshidabad, East Burdwan districts, and partially viewed north-eastern and southern parts of Birbhum), and very few portions of the western region (western part of West Burdwan), and southern region (border part of Jhargram and East Medinipur districts) (Fig. 3). The south-eastern region (overall East Medinipur district and partially Jhargram district) has the highest proportion of high rice yields (3.45–2.37 t/ha), while the north-eastern region (eastern and western parts of Murshidabad, scattered distribution of East Burdwan, partially eastern and south-eastern parts of Birbhum), and western region (western part of West Burdwan districts) are scattered distributed under high rice yield zones. The moderate rice production (2.37–1.44 t/ha) is the result of discrepancies in all districts save the northeastern region. In this study area, western and north-western region (overall Purulia district and partially Birbhum district, scattered West Burdwan district) and south-middle part (partially scattered distribution of Bankura district, some of northern part of Jhargram and East Medinipur districts) of the study area show that low rice yield (1.44–0.39 t/ha) and in the middle portion (overall Bankura) of the study area and partially north-western region (west Birbhum districts, scattered distribution.

Using the MARS model, we discovered that the highest level of rice yield (5.60–3.45 t/ha) was observed in the north-eastern (partial north-eastern Murshidabad and very low portion of eastern East Burdwan) and south-eastern regions (border part of Jhargram and East Medinipur districts and partially southern part of East Medinipur districts). Another observation was that the north-eastern and eastern regions (the majority of Murshidabad and East Burdwan), as well as the western half of West Burdwan and the southern section of East Medinipur and Jhargram districts, have high rice yields (3.45–2.37 t/ha). The moderate rice yield (2.37–1.44 t/ha) exists in the northern region (scattered distributed of Murshidabad, the eastern and southern part of Birbhum), the eastern region (south-eastern and south-western part of East Burdwan), the southern eastern region (northern and eastern part of East Medinipur, the northern portion of Jhargram), and some of the western region. Low rice yield (1.44–0.39 t/ha) is observed in the middle and western and north-western regions of the study area (maximum portion of Purulia and Birbhum districts, average distribution of Bankura district, very low portion of West Burdwan), as well as in some of the south-eastern regions (northern and middle portion of East Medinipur). Finally, the central area (generally Bankura) and a small fraction of the western and north-western regions (partially western and southern Purulia and a very small amount of the south-western portion of Birbhum districts) have very poor rice yields (0.39–0.10 t/ha).

In this model (RF), we observed that the highest level (5.60–3.45 t/ha) of rice yield area viewed in the northern and north-eastern region (maximum portion of Murshidabad, overall East Burdwan except some of the southern part and north-western part, some of the southern part of Birbhum), and a partial portion of the western region (western part of

**Table 2**
Results for OLS and VIF analysis in crop yield prediction

| Variable | VIF | β | std err | t | P>\|t\| |
|---|---|---|---|---|---|
| temp | 1.10 | 0.397 | 0.041 | 9.744 | 0.00*** |
| aet | 1.98 | −0.086 | 0.030 | −2.894 | 0.004** |
| ele | 3.43 | 0.004 | 0.001 | 6.668 | 0.00*** |
| pr | 1.65 | −0.093 | 0.006 | −14.569 | 0.00*** |
| sm | 2.49 | 0.001 | 0.001 | 3.054 | 0.002** |
| stemp | 1.58 | −0.001 | 0.001 | −3.077 | 0.002** |
| R-squared = 0.79, F-statistic = 1098, DurbinWatson = 1.24, Prob > chi$^2$ = 0.002 | | | | | |

β ~Coefficient; t ~ *t*-test std err ~ Standard error, Robust standard errors~ *p < 0.05, **p < 0.01, and, ***p < 0.001,F ~ Statistical, $R^2$~ Linear regression.

**Table 3**
Boruta feature importance summary for crop yield parameters selection

| Variable | meanImp | medianImp | minImp | maxImp | normHits | decision |
|---|---|---|---|---|---|---|
| temp | 34.18584 | 33.91949 | 32.72801 | 36.8138 | 1 | Confirmed |
| aet | 38.58622 | 38.62706 | 35.85058 | 41.40367 | 1 | Confirmed |
| ele | 51.5107 | 51.67904 | 48.56954 | 53.77524 | 1 | Confirmed |
| pr | 53.16091 | 53.16332 | 51.37621 | 55.0515 | 1 | Confirmed |
| sm | 37.44425 | 37.68965 | 34.83757 | 40.30518 | 1 | Confirmed |
| stemp | 55.30331 | 55.43543 | 50.87711 | 57.75792 | 1 | Confirmed |

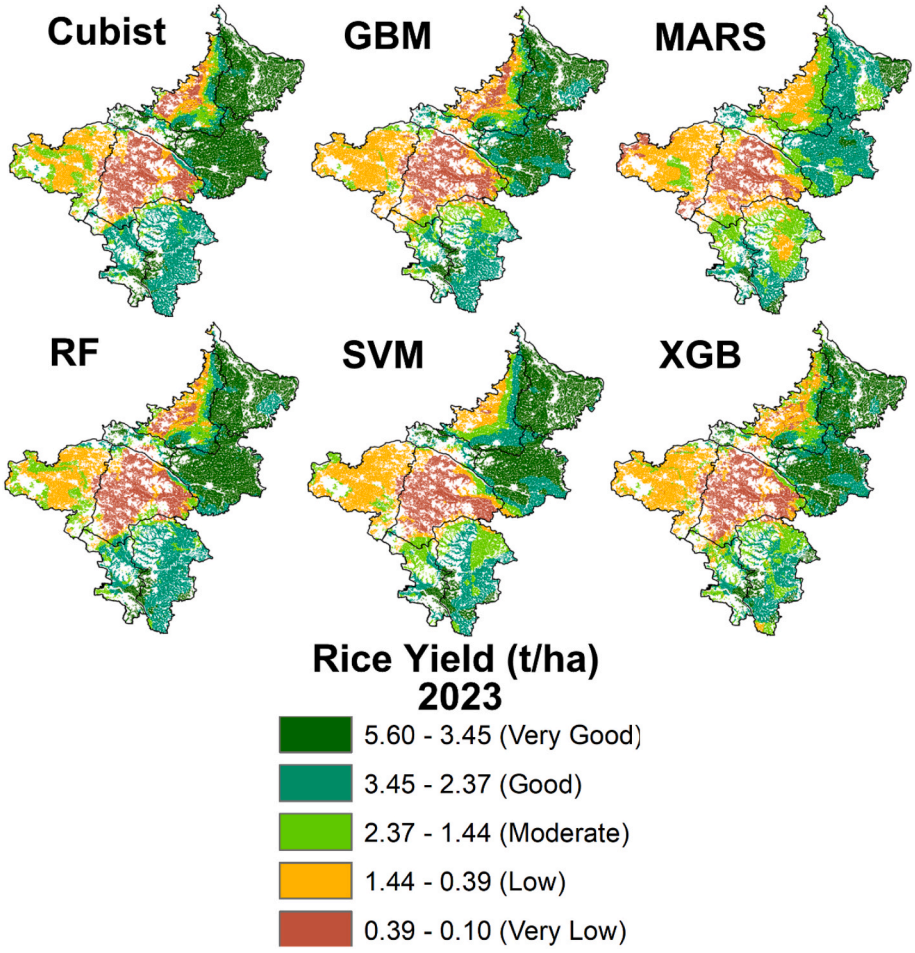


**Fig. 3.** Rice yield mapping by six machine learning techniques of 2023.

West Burdwan) and also southern region (border of Jhargram and East Medinipur, some of the eastern part of East Medinipur). We see that high rice yield (3.45–2.37 t/ha) is viewed in the maximum portion of the south-eastern region (overall East Medinipur and Jhargram), and partially in the eastern region (southern part of East Burdwan), western region (very few of the northern part of Purulia, scattered distribution of West Burdwan), north-eastern region (east Murshidabad, some of the western part of Murshidabad), and north middle region (eastern and southern part of Birbhum). We also notice that intermediate rice yields (2.37–1.44 t/ha) are dispersed, except for the northeastern region. The study area includes a low rice-yielding zone (1.44–0.39 t/ha) in the western region (overall Purulia) and a north-western region (the western portion of Bibhum), with some spread throughout Bankura and eastern Birbhum districts. Finally, in the middle of the research area (Bankura districts) and a portion of the north-western and western regions (south-western part of Birbhum and some of West Burdwan), rice yields are quite low (0.39–0.10 t/ha).

In the SVM model, we observed that the highest level (5.60–3.45 t/ha) of rice yields are viewed in the northern and north-eastern region (overall Murshidabad and maximum portion of East Burdwan, some of north-eastern and south-eastern part of Birbhum district) and partially situated in the western region (western part of West Burdwan) and south-eastern region (border of Jhargram and East Medinipur and some of south-eastern part of East Medinipur). We see that high rice yields (3.45–2.37 t/ha) are viewed in the maximum portion of the southern and south-eastern region (a maximum portion of East Medinipur and the southern part of Jhargram), and partially viewed in the eastern region (southern, south-eastern part, and northern part of East Burdwan), western region (scatter distribution of West Burdwan), north-eastern region (very few of east Murshidabad), and north-middle region (north-eastern and south-eastern part of Birbhum. And we also see that moderate types of rice yield (2.37–1.44 t/ha) are viewed north-western region (middle portion of the north to south Birbhum districts), southern and south-eastern region (northern and some of scattered distributed East Medinipur and Jhargram districts), and scattered distributed everywhere (middle of West Burdwan, western and south-eastern part of

Purulia, northern part of Bankura, and southern part of East Burdwan), except the north-eastern region (maximum portions research area has poor rice producing zones (1.44–0.39 t/ha) in the western region (overall Purulia), central region (scattered distributed of Bankura, northern part of East Medinipur), north-western region (western part of Bibhum), and eastern region (southern part of East Burdwan). Finally, in the central portion of the research area (Bankura districts) and a portion of the north-western region (some western parts of Birbhum), rice yields are extremely poor.

In the XGB model, we observed that the highest level (5.60–3.45 t/ha) of rice yields are viewed in the northern and north-eastern regions (Murshidabad, East Burdwan, and a few areas of north-eastern and southern parts of Birbhum districts) and partially distributed in the south-eastern region (border region of Jhargram and Purba Medinipur and some of the eastern part of Purba Medinipur), western region (western part of Paschim Bardhaman) (Fig. 3). And we observed that high rice yield (3.45–2.37 t/ha) is viewed in the maximum portion of the southern and south-eastern regions (maximum portion of Purba Medinipur and some of Jhargram), and partially in the western region (northern part of Purulia and scatter distribution of west of Paschim Bardhaman), and the eastern region (scattered distributed of Purba Bardhaman, Murshidabad, the eastern and southern part of Birbhum). The moderate forms of rice yield (2.37–1.44 t/ha) are scattered throughout the research area, with the highest proportion observed in the south-eastern region (Jhargram and Purba Medinipur) and the north-western region (eastern and western Birbhum). Low rice-yielding zones (1.44–0.39 t/ha) can be found in the western and north-western regions (overall Purulia, the western part of Birbhum, and scattered distributed Paschim Bardhaman), as well as in the middle region (scattered distributed Bankura) and the south-eastern region. Finally, very low rice yields (0.39–0.10 t/ha) are observed in the middle portion of the study area (overall Bankura districts) and a small portion of the scattered western and north-western regions (some eastern part of Purulia, the middle portion of Paschim Bardhaman, and scattered distribution of Birbhum districts).

Except for the MARS model, the cubist, RF, GBM, XGB, and SVM models all exhibit a similar tendency for high rice yields. High rice yields were also seen in several areas of Murshidabad, Purba Bardhaman, Paschim Medinipur, and Jhargram districts. Rice yields were reported to be moderate in the eastern half of Birbhum and relatively scattered areas of Purulia, Paschim Medinipur, and Jhargram districts. Low rice yields were also discovered in western Birbhum, entire Purulia, and very minor portions of Paschim Medinipur and Jhargram districts. Water stress caused extremely low yields in tiny sections of Birbhum, Purba Bardhaman, and entire areas of Bankura, as well as very small parts of Purulia districts. The highest and minimum $R^2$ values for the cubist and RF models were found to be 0.73 and 0.72, respectively (Fig. 4a & d). The $R^2$ value was likewise found to be negligible for the XGB (0.710), GBM

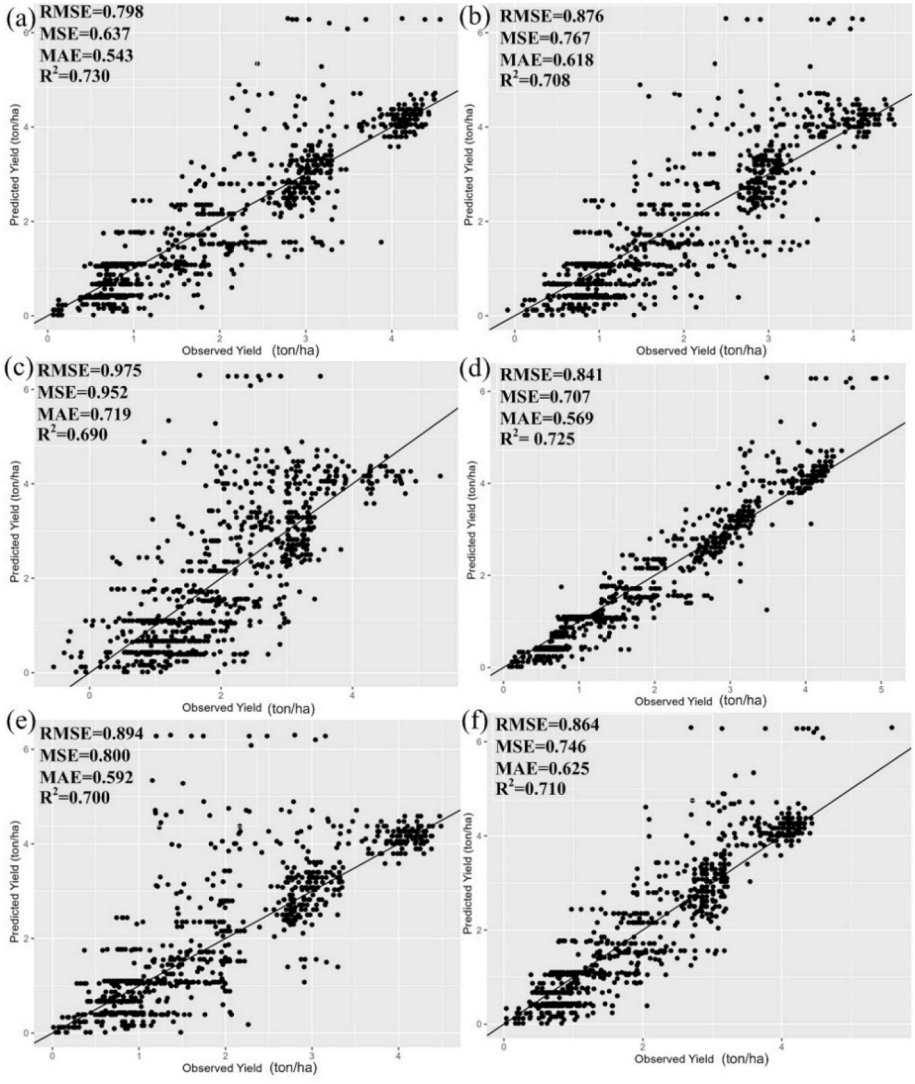


**Fig. 4.** Scatter plots representation (a) cubist, (b), GBM, (c) MARS (d) RF, (e) SVM and (f) XGB for validation purposes.

(0.708), MARS (0.690), and SVM (0.700), machine learning models (Fig. 4b, c, 4e, & 4f). The RMSE values for cubist and MARS models ranged from 0.79 to 0.97. In the model predictive accuracy as the maximum MSE value showed in the MARS model around 0.952, followed by SVM (0.800), GBM (0.767), XGB (0.746), RF (0.707), and Cubist (0.637). According to the MAE analysis, the highest value was shown by the MARS model at approximately 0.719, followed by XGB (0.625), GBM (0.618), SVM (0.592), RF (0.569), and Cubist (0.543). Finally, the Cubist model achieved the highest $R^2$ (0.730) and the lowest RMSE (0.798) values, indicating it is the best model for predicting rice yield compared to the other models.

### 4.4. Analysis of future rice yield predictions

Six machine learning models were used to forecast future rice yields based on five factors. The CMIP6-based precipitation and temperature data were used to generate projected rice yields for SSP2-4.5 and SSP5-8.5. Shared Socioeconomic Pathways (SSPs) explain future society's socioeconomic evaluation, adaptation, vulnerability, and natural ecosystems [47]. This study considers SSP2-4.5 and SSP5-8.5, which involve balance and extreme weather conditions. Cubist, GBM, RF, SVM, and XGB, excluding MARS, revealed very high Kharif rice yields in Murshidabad, Purba Bardhaman, and a tiny portion of Paschim Medinipur and Jhargram districts under SSP2-4.5 of 2030 (Fig. 5). It was also discovered that very high Kharif rice yields occurred in tiny parts of Murshidabad, the majority of Purba Bardhaman, and very small parts of Paschim Medinipur and Jhargram under SSP5-8.5. poor and very poor Kharif rice yields were discovered in various portions of Birbhum, as well as in the majority of Bankura and Purulia districts, employing all machine learning methods of 2030 under SSP2-4.5 and SSP5-8.5, respectively. Moderate Kharif rice output was observed in tiny areas of Murshidabad, Birbhum, Paschim Bardhaman, Paschim Medinipur, and Jhargram districts. However, ML-based results are widely accepted for future rice yield estimates using several statistical markers (e.g., $R^2$, RMSE, MAE).

Using several machine learning approaches, we examined this research area and forecasted rice yield (t/ha) under SSP2-4.5 in 2030. We revealed that in the Cubist, GBM, MARS, RF, SVM, and XGB approaches, the highest portion of rice yield is extremely high (5.56–3.49 t/ha) and high (3.49–2.49 t/ha) in the northern and north-eastern regions (Murshidabad and Purba Bardhaman) and southeastern region (Jhargram and Purba Medinipur). In this region, we show that the output of rice yield will increase under SSP2-4.5 in 2030 due to climate change, adequate water supply, the use of new high-yielding rice varieties, better pest management, proper irrigation systems, the use of good fertilizer, good seeds, advanced tools, and machinery, developing farmer literacy and yielding skills, and so on. There are moderate rice yields (2.49–1.67 t/ha) scattered all over the region, especially Cubist, GBM, SVM, and XGB techniques, we see the maximum portion in the south-eastern region (Paschim Medinipur), eastern region (east and western part of Purba Bardhaman), north-western region (a scattered portion of Birbhum), and MARS and RF techniques, we see the middle-eastern region (southern part of Purba Bardhaman, south-eastern part of Bankura), north-western region (southern part of Bankura). We determined that the maximum rice yield zone in the study area will become low (1.67–1.04 t/ha) and very low (1.04–0.05 t/ha) due to climatic

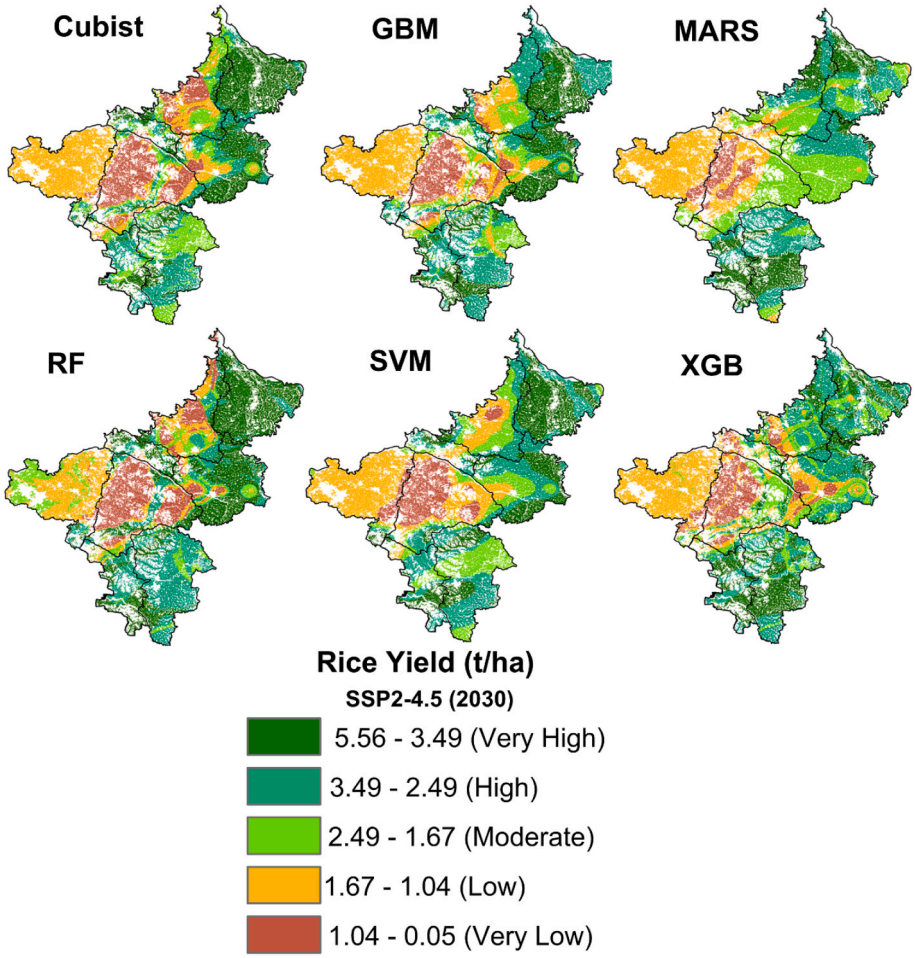


**Figure: 5.** Future rice yield mapping by six machine learning techniques of 2030 under SSP2-4.5 of the MIROC6 model.

change, water scarcity, and solar radiation.

In this study area, using all these techniques (Cubist, GBM, MARS, RF, SVM, and XGB), we observed that high (3.51–2.78 t/ha) and very high (5.09–3.51 t/ha) levels of rice yield show maximum portion in the eastern region (maximum portion of Purba Bardhaman district and a partial portion of eastern part of Paschim Bardhaman district), north-eastern (maximum portion of Murshidabad district and some of south-eastern part of Birbhum district), and south-eastern region (a maximum portion of Jhargram. We also examined the western region (parts of Purulia and Bankura districts) using GBM and RF approaches. In this region, we show that rice yield will increase under SSP5-8.5 in 2030 due to climate change, adequate water supply, use of new high-yielding rice varieties, better pest management, proper irrigation system, use of good fertilizer, good seeds, advanced tools and machinery, development of farmer literacy and yielding skills, and so on. Rice yields at modest levels (2.78–2.01 t/ha) are widely distributed (Fig. 6). Using Cubist, GBM, RF, SVM, and XGB techniques, we show that the eastern and south-eastern regions (south-eastern Birbhum, a portion of Purba Bardhaman, and scattered Medinipur districts) have the highest rice yield, followed by the western region (some Purulia districts), and the northern and northeastern regions. However, we observed that there are no moderate rice yield zones in the western region when applying MARS methodologies. Finally, low (2.01–1.20 t/ha) and very low (1.20–0.06 t/ha) regions are viewed in the western region (maximum portion of Purulia, some of Paschim Bardhaman), north-western (Birbhum district), and middle region (Bankura district), and partially in the north-eastern, eastern, and south-eastern regions (scatter portion of Murshidabad, Purba Bardhaman, Jhargram, and Paschim Medinipur districts) of the study area due to climatic change, water scarcity, and solar radiation.

### 4.5. Analysis of rice yield gap mapping

The rice yield gap mapping is very important for sustainable agriculture production to meet food security problems considering climate change impacts [11]. It is very crucial for agriculture intensification on regional to global scales. However, the first step is to estimate the rice yield considering different influencing factors by six machine learning methods. Then the rice yield gap was mapped from simulated and field-based rice yield data using ML from 2023 to 2030. It was observed that a very high rice yield gap (50–60 %) showed in some parts of the Birbhum, western parts of Bankura, Purulia, Paschim Medinipur, and Jhargram districts by all machine learning models because of insufficient water supply and soil quality. It was also observed that high rice yield showed in small parts of Murshidabad, small parts of the Birbhum, western parts of the Purulia and Bankura, and small areas of Paschim Medinipur and Jhargram.

This study topic uses a variety of machine-learning approaches to show the rice yield disparity (%) in 2023. Cubist, GBM, SVM, and XGB techniques have been used to identify a very high (60-50 %) rice yield gap in the middle-northern region (south-eastern part of Birbhum), north-western region (northern and south-western part of Birbhum), south-eastern region (a maximum portion of Jhargram and Purba Medinipur districts), and some of the eastern (Purba Bardhaman) and western regions (Purulia), but except for the extreme northern region, the MARS technique has a very high rice yield gap that is evenly distributed (Fig. 7). On the other hand, RF approaches have a very small

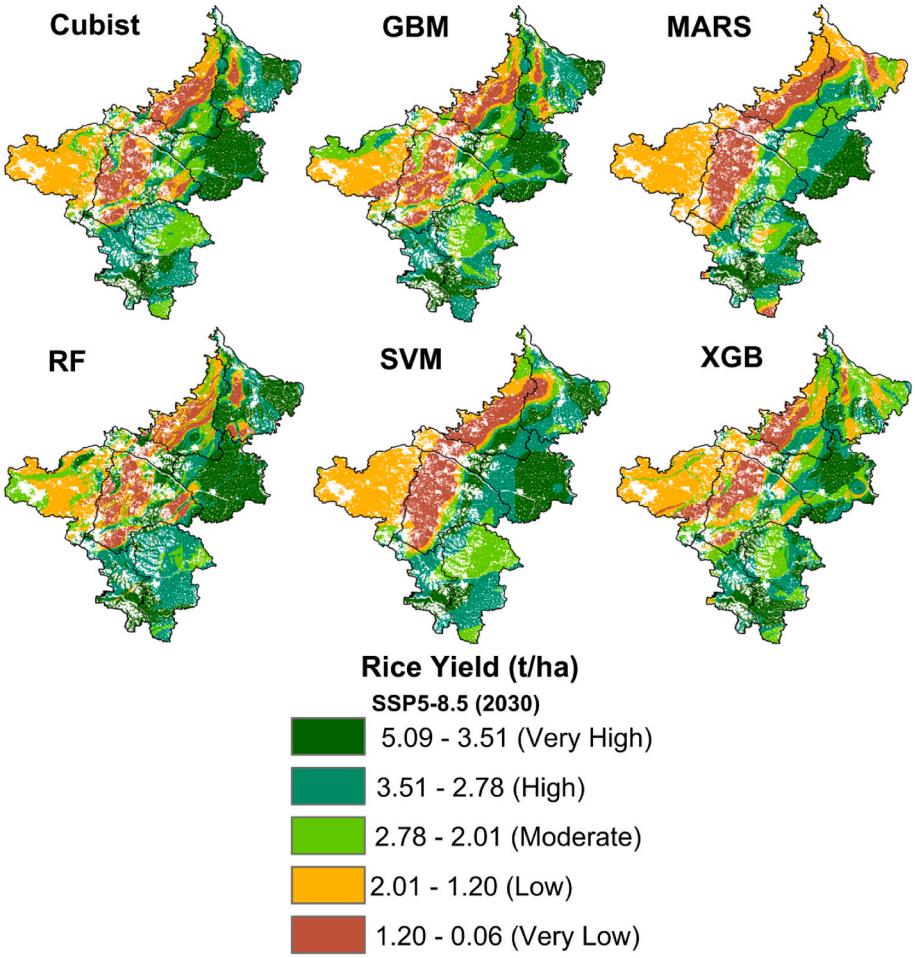


**Figure:6.** Future rice yield mapping by six machine learning techniques of 2030 under SSP5-8.5 of MIROC6 model.

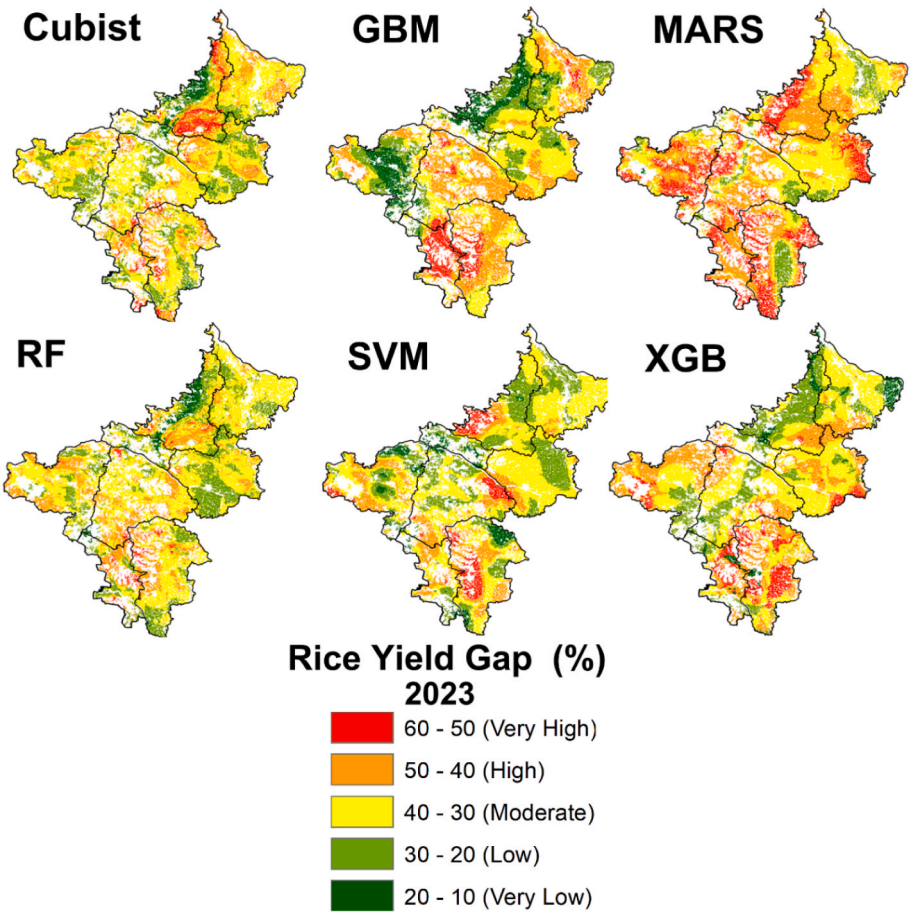


**Fig. 7.** Rice yield gap mapping by six machine learning techniques of 2023.

rice yield gap that is unevenly distributed. It was shown that this region experienced the biggest rice production gap due to tremendous issues with less water availability and several sorts of environmental concerns such as climate change, and land degradation. There is a high rice yield gap (50-40 %) scattered throughout the area, but the maximum we identified (using GBM and MARS techniques) in the middle and south-east regions (Bankura, Jhargram, and Paschim Medinipur districts), northern region (Murshidabad and Birbhum districts), and western region (Purulia district). It was discovered that a moderate rice yield gap (40-30 %) covered almost all regions, while a low rice yield gap (30-20 %) was distributed across most districts, with a very low rice yield gap (20-10 %) affecting most of the north-western region (western part of Birbhum and northern part of Paschim Bardhaman districts) and western region (Purulia district).

The percentage changes in the rice yield gap mapping study were expressed using the Cubist method (Fig. 8). According to this method, 20.18 % changes in Murshidabad and 32.59 % changes in Purba Bardhaman districts indicate a high rice yield gap; 35.66 % changes in Birbhum, 32.10 % changes in Paschim Bardhaman, 33.96 % changes in Bankura, and 43.31 % changes in Jhargram districts indicate a moderate

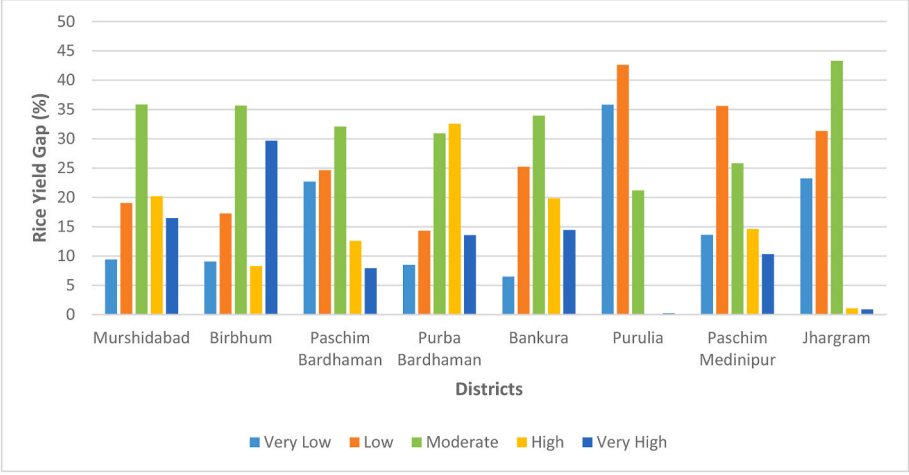


**Fig. 8.** Percentage changes of rice yield gap mapping by a Cubist ML method.

rice yield gap; and finally, 42.61 % changes in Purulia and 35.6 % changes in Paschim Medinipur districts indicate a low rice yield gap. As a result, in all districts, 16.48 % changes in very high rice yield gaps were observed in Murshidabad district due to changing climate and variability, while 23.25 % changes in very low rice yield gaps were observed in Jhargram district due to agricultural input availability, agricultural efficiency, climate change, rainfed rice crop growth, and converting agricultural fallow lands into cultivable lands (Fig. 9).

### 4.6. Analysis of future rice yield gap mapping

Using all of the ML techniques (Cubist, GBM, MARS, RF, SVM, and XGB), the maximum portion of high (50%–40 %) and very high (60%–50 %) rice yield gaps increase in the north-western region (Birbhum district), middle and south-eastern regions (Bankura, Jhargram, and Paschim Medinipur districts) in 2030 under the SSP2-4.5 scenario due to water availability challenges, various types of environmental problems, and so on (Fig. 10). A moderate rice yield disparity (40%–30 %) was seen in all districts, particularly in the Cubist and RF approaches we identified, with most of the moderate rice yield gap evenly distributed throughout the region. It observed that in Cubist, GBM, MARS, RF, SVM, and XGB techniques, the maximum portion of low (30%–20 %) and very low (20%–10 %) percentage rice yield gap in the eastern, north-eastern region (Purba Bardhaman, Murshidabad districts), western region (Paschim Bardhaman district), and scattered it on the north-western region (Birbhum district), western region (Purulia and Bankura districts), and south-eastern region (scatter portion of Jhargram and Paschim Medinipur districts).

This map examined the rice yield gap (%) in 2030 using various ML algorithms (Cubist, GBM, MARS, RF, SVM, and XGB) based on the SSP5-8.5 scenario (Fig. 11). The rice yield gap in the middle-eastern region (overall Bankura), north-middle region (south-eastern part of Birbhum), western region (northern part of Purulia), and south-eastern region (spreadly distributed of Jhargram and Paschim Medinipur districts) was very high (60-50 %) and high (50-40 %) compared to the north-western region (southern and western part of Birbhum) and eastern region (north-eastern part of Purba Bardhaman), but in the MARS, SVM, and XGB techniques. As a result, it was discovered that the Bankura district has a significant rice yield gap due to inadequate crop management, dry and drought-prone locations, soil concerns, poor water management, high-temperature challenges, and so on. It was shown that a moderate rice yield gap (40-30 %) was scattered throughout the region. According to SVM methodologies, the western area (Purulia and the western half of Bankura districts), the south-east region (Jhargram and Paschim Medinipur districts), and the eastern region (Purba Bardhaman district) had the greatest mild rice yield gap. In the northern and north-eastern regions (Murshidabad district), north-western, western, and middle-western regions (partially viewed Birbhum, Paschim Bardhaman, and the eastern part of Purba Bardhaman districts), where the rice yield gap is low (30-20 %) and very low (20-10 %), the XGB technique was also found in the eastern region (Purba Bardhaman district) (Fig. 11). However, Murshidabad and Paschim Bardhaman districts had low and very low rice yield gaps, which resulted in improved crop management, greater use of balanced fertilizer, planting healthy seeds, pest and

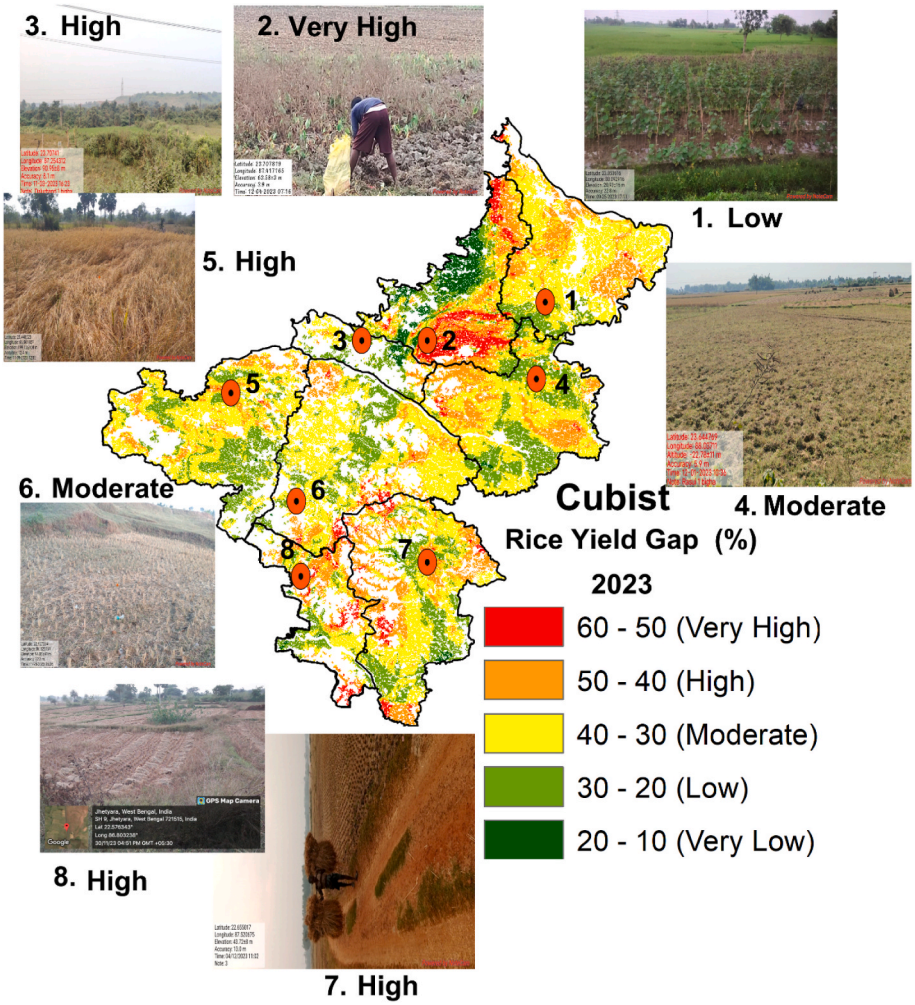


**Fig. 9.** Validation of the best machine learning model of Cubist with field observation photographs.

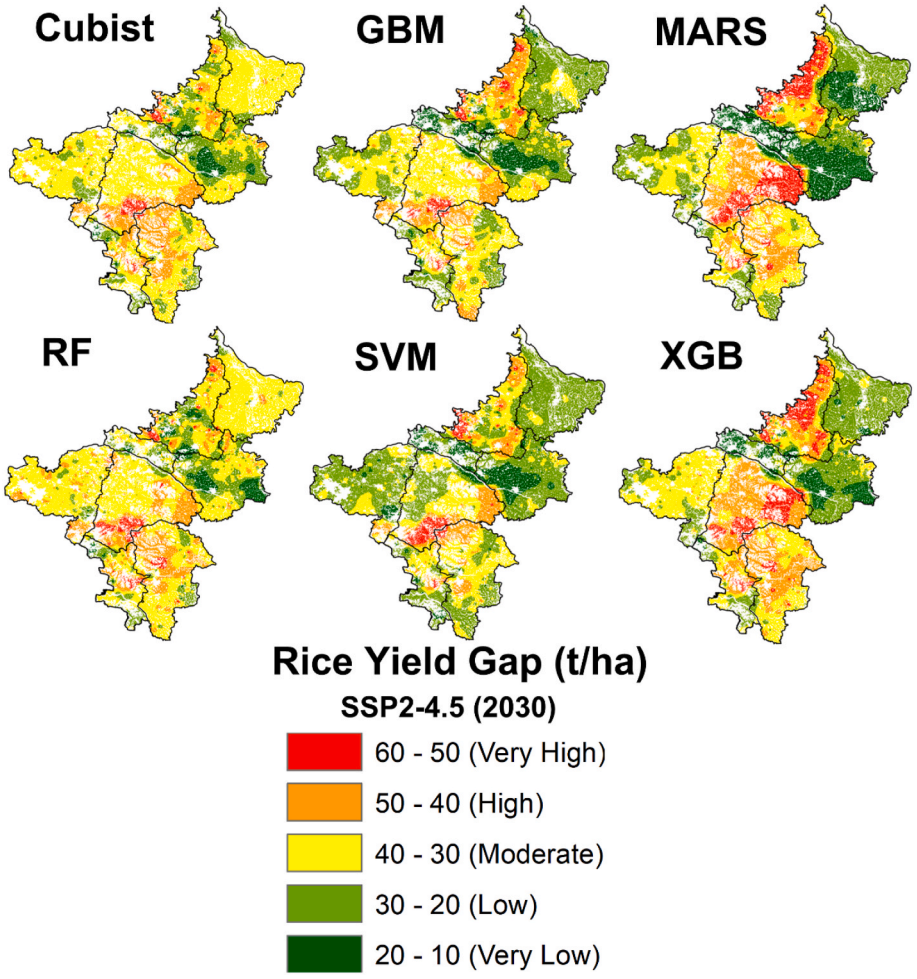


**Fig. 10.** Future rice yield gap mapping by six machine learning techniques of 2030 under SSP2-4.5 of the MIROC6 model.

disease control, rational irrigation, and water supply, among other things.

## 5. SHAP analysis

This analysis aimed to determine the importance of different features in the model's decision-making process across three cities, as depicted in Fig. 12a–f. The mean absolute SHAP value was calculated to assess how much each feature impacts the model's prediction. This analysis aimed to highlight the relative importance of features in the model's decision-making process. The parameters are arranged in descending order, indicating that features at the top exert a more substantial influence on the final prediction compared to those lower down. Fig. 12a–f displays SHAP values for input factors, illustrating their impact on trends (i.e., SHAP summary (left side) and bar plot (right side). The bar plot function is passed a matrix of SHAP values, it generates a global feature importance plot. This plot represents the overall importance of each feature, calculated as the mean absolute value across all provided samples. A SHAP bar plot visualizes the feature importances calculated using SHAP values. According the SHAP bar plot features are most influential in making predictions and how they contribute to the overall model behavior. It provides a clear and interpretable way to assess feature importance in machine learning models (Fig. 12 right side). The SHAP values are shown on the x-axis, representing the influence of each conditioning factor on the prediction. On the y-axis, each dot represents a sample, colored from blue for lower values to pink for higher values of a factor. The horizontal position of the dot indicates whether the conditioning factor positively or negatively influences the prediction. In the cubist analysis, the key parameters identified are Pr, sm, elevation, and stemp (Fig. 12a left side). Among these, Pr emerges as a primary contributor to the RYP analysis. According to the SHAP bar plot (Fig. 12a right side), analysis, elevation, stemp, and Pr are identified as the most important factors influencing the RYP analysis. GBM based SHAP analysis showed most influential parameters for RYP analysis are ranked as elevation, aet, Pr, stemp, sm, and temp (Fig. 12b, left side). Conversely, the SHAP bar plot indicates that Pr and aet are prioritized as the most significant factors for RYP analysis, with temp having the least impact (Fig. 12b, right side). Fig. 12c (left side) illustrates that the descending order of the most important parameters in the RYP analysis, according to the SHAP values of the MARS model, are aet, Pr, stemp, elevation, sm, and temp. Conversely, the SHAP bar plot ranked the parameters as elevation, stemp, aet, Pr, sm, and temp (Fig. 12a, right side). The SHAP summary and SHAP bar plots for the RF, SVM, and XGB models (Fig. 12d–f) indicated that Pr and elevation are the top contributors. Conversely, stemp and sm have the least impact on the RYP analysis. Finally, the SHAP summary plot revealed that Pr, elevation, and aet were the most significant predictors of RYP analysis based on the all-ML results. Factors like temp, and stemp exerted a moderate impact, whereas sm had a minimal effect on RYP analysis (Fig. 12, left). Unlike the SHAP bar plot, the predicted RYP by the All ML model showed a different order in the important features range value as pr (0.21–0.49) > ele (0.23–0.46) > aet (0.14–0.37) > sm (0–0.35) > stemp (0.01–0.32) > temp (0.09–0.16) (Fig. 12, right).

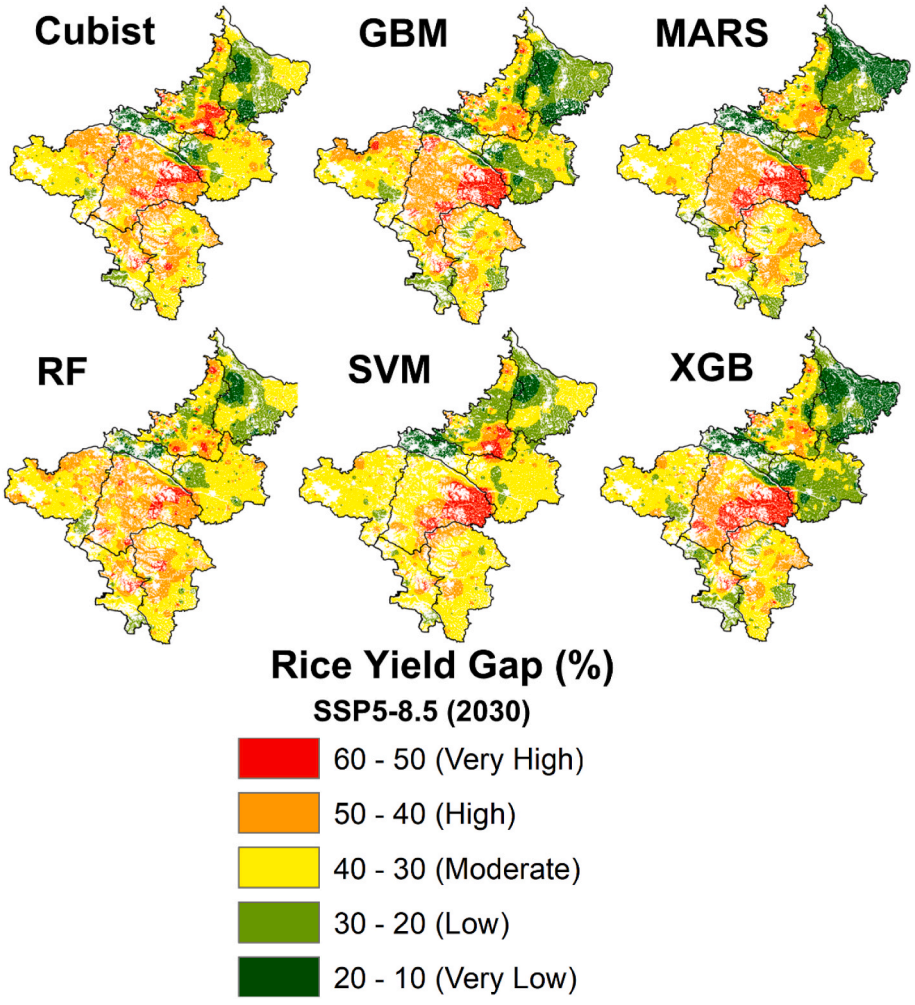


**Fig. 11.** Future rice yield gap mapping by six machine learning techniques of 2030 under SSP5-8.5 of the MIROC6 model.

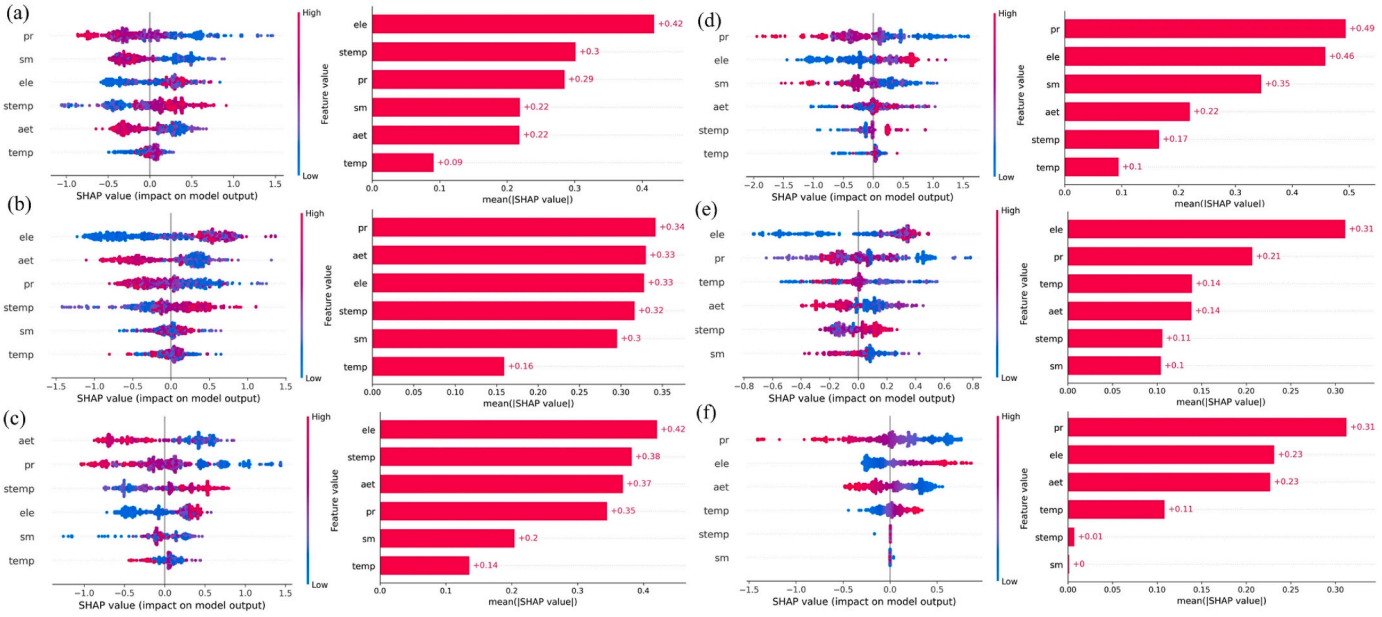


**Fig. 12.** SHAP summary (left side) and bar plot (right side) for Rice yield mapping analysis (a) cubist, (b), GBM, (c) MARS (d) RF, (e) SVM, and (f) XGB for validation purposes.

## 6. Discussion

Investigating environmental and hydrometeorological parameters for predicting rice yields highlights a sophisticated understanding of how remote sensing integrated ML can improve agricultural productivity. Our analysis across the study region, utilizing six ML models, revealed that geo-environmental parameters related to elevation, soil moisture (sm), precipitation (Pr), temperature (temp), soil temperature (stemp), and Actual Evapotranspiration (aet) all performed comparably in predicting RYP under the tropical climate conditions of Eastern India (Table 1).In this study Cubist model, followed by RF and XGB, proved to be the top performers due to their superior $R^2$ values and lower RMSE values (Cubist < RF < XGB < GBM < SVM < MARS), highlighting their potential as reliable MLs. The study found a significant 99 % agreement ($p < 0.001$) among precipitation, temperature, and elevation variables in predicting rice yield in the area. A related study investigated climate-smart farming methods. According to Rockstrom et al. [48], yields can be greatly increased in the face of fluctuating precipitation conditions when technical innovation is combined with smart water management. The Boruta feature selection results (Table 3) showed that stemp factors had the highest mean importance (55.30), followed by Pr (53.16), elevation (51.51), aet (38.58), temp (34.18), and sm (37.44). These predictors were confirmed for sensitivity analysis of RYP mapping parameters. The SHAP summary plot indicated that pr, elevation, and aet were the most crucial predictors in the RYP analysis across all machine learning results. Temp and stemp had a moderate influence, while sm had a minimal impact on the RYP analysis (Fig. 13). It is commonly known that plant transpiration, soil hydroclimatic conditions, and the physiological processes by which soil characteristics influence crop productivity are all related [49].

Previous research shows how machine learning (ML) is used for the prediction of crop yield [7]. A systematic literature review to extract and synthesize the algorithms and features. They investigated selected studies carefully, analyzed the methods and features used, and provided suggestions for further research. According to this research analysis, temperature, precipitation, and soil type are mostly used features and the most applied algorithm is Artificial Neural Networks. They also performed an additional search in electronic databases to identify deep learning-based studies and extracted the applied deep learning algorithms. According to this additional analysis, the most widely used deep learning algorithm is Convolutional Neural Networks (CNN), and others also used are Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN). On the other way, our analysis specified only rice yield gap prediction of specified districts of West Bengal in present and future scenarios using different types of ML models like Cubist, GBM, MARS, RF, SVM, and XGB and interpreted which model shows high rice yield gap and which model shows low. According to our analysis, we interpret the output results using different types of features criteria like climate change, soil type and texture, water availability, varieties of new high-yielding rice, pest management, irrigation system, fertilizer, and seeds, advanced tools and machinery, skills, and literacy of agricultural labor etc. Recognizing potential yields and yield gaps is vital for maintaining high crop productivity, improving livelihoods, and reducing environmental impact.

In Eastern India, with its rainfed and irrigated agricultural systems, the observed yield variability highlights the urgent need for adaptive strategies to mitigate the impacts of climate change and variability. Innovations such as rainwater harvesting, conservation tillage, drought-tolerance crop varieties, best crop rotation analysis, and precision irrigation systems are crucial components of a comprehensive strategy to reduce yield gaps and improve food security [50,51]. To effectively resolve yield gaps, Ray et al. [52] argued against concentrating only on climatic conditions and instead supported a broader approach that also takes soil suitability and management techniques into account.

Nayak et al. [4] revealed a rice yield gap and analyzed nitrogen inputs without compromising rice productivity by utilizing a large field database of individual farmers for rice production in India's northwestern Indo-Gangetic plain. Another study conducted in India used the DSSAT (Decision Support System for Agrotechnology Transfer) model to examine the rice production gap under projected climate change scenarios [11]. However, our data revealed a rice yield gap in the present and future scenarios using various types of machine learning models in eight districts of West Bengal like Murshidabad, Birbhum, Purba Bardhaman, Paschim Bardhaman, Bankura, Purulia, Jhargram, and Paschim Medinipur).

The study demonstrates that crop yield prediction at an early stage can be achieved using machine learning approaches such as the Statistical Model of Multi Linear Regression, Back Propagation Neural Network (BPNN), Support Vector Machine (SVM), and General Regression Neural Networks (GRNN) over 18 years in 28 districts of Tamil Nadu [17]. This study examined rice yield prediction in the Tamil Nadu delta region and focused on integrating crop, meteorological, and soil data from agricultural datasets to evaluate yield prediction behavior using the MLR-LSTM (Multiple Linear Regression and Long Short-Term Memory) model, which is a hybrid architecture. The results are compared to classic machine learning approaches such as support vector machines (SVM), long short-term memory (LSTM), and random forests (RF) [14]. This study offered a model interpretation and prediction of

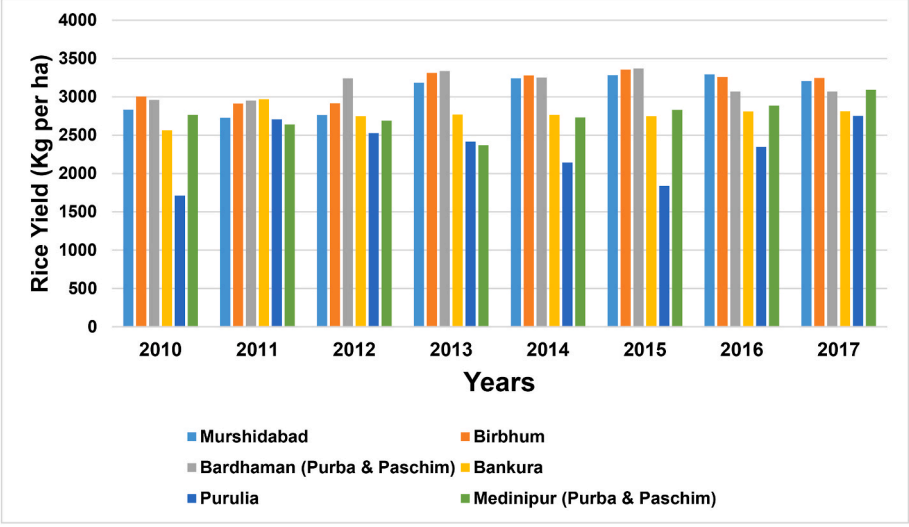


**Fig. 13.** District-wise rice yield data from 2010 to 2017.

rice production over the Indo-Gangetic Plains in India by combining time-series satellite data, environmental variables, and rice yield records from 2001 to 2016 using a Transformer technique [12]. This research forecasted rice yield in Kerala, India using Machine Learning and K Nearest Neighbour Regression with the best accuracy of 98.77 %.

However, the present research focused on future prediction of the rice yield gap mapping considering six different influencing parameters in the eight districts of West Bengal. It was observed that the highest $R^2$ is 0.73 by the Cubist model. Fig. 13 depicts district-specific rice yield data from 2010 to 2017 (http://data.icrisat.org). Murshidabad, Birbhum, and Purba Bardhaman districts produced more Kharif rice than other districts in 2017 due to reliable irrigation facilities, high-yielding cultivars, enough fertilizer consumption, agricultural instruments, and machinery. These findings were closely associated with our estimated current and future rice yield statistics. As a result, our machine learning predictions are very acceptable and correspond well with field observation datasets. The research's limitations include a lack of field observations in locations in which transportation is very unavailable. Furthermore, the future direction of the research will incorporate deep learning and artificial intelligence for rice yield and gap monitoring. Using approaches such as Cubist, GBM, MARS, RF, SVM, and XGB, the highest rice yields, categorized as extremely high (5.56–3.49 t/ha) and high (3.49–2.49 t/ha), are projected for the northern and north-eastern regions (Murshidabad and Purba Bardhaman) and the southeastern region (Jhargram and Purba Medinipur). In these regions, rice yield is expected to increase under the SSP2-4.5 scenario by 2030 due to climate change, adequate water supply, the adoption of new high-yield rice varieties, improved pest management, proper irrigation systems, effective fertilizer use, quality seeds, advanced tools and machinery, and enhanced farmer education and skills. Similarly, the Bankura, Purulia, and West Birbhum districts experience a substantial rice yield gap due to factors such as inadequate crop management, dry and drought-prone conditions, soil issues, poor water management, low soil organic carbon, and high-temperature challenges. The study found that a moderate rice yield gap (40-30 %) is widespread throughout these regions.

## 7. Adaptation strategies and policy recommendations

Adaptation techniques and policy proposals are critical components of a sustainable agricultural development framework that takes climate change consequences into account. Rice is extremely susceptible to decreased water availability because it requires a lot of water for production. Thus, water-saving irrigation systems are critical for soil and water conservation efforts. Various adaptation tactics, such as crop spacing, transplanting timing, N-fertilizer use, and cultivars, can be employed to maximize optimal rice yield. Various policies [53] can be


**Fig. 14.** Representation of sustainable development goals correlated with photographs of field surveys in several districts.

implemented, such as the construction of farm ponds to store rainwater, the adoption of water-saving cultivation methods, protective irrigation during critical stages, the conversion of rice fields to other purposes, and crop insurance for rainfed and irrigated land. However, these types of adaptation tactics and policy consequences are difficult to implement because farmers' knowledge, expertise, and resource availability vary by place. Thus, training programs and demonstrations of agricultural technologies for local government officials are critical in adapting to climate catastrophe. In addition, our research focuses on Kharif rice output and gap mapping for future agricultural water management planning to meet the UN's sustainable development goals 1, 2, 8 & 13 (https://sdgs.un.org/goals). Several district-specific field survey photographs were combined with SDGs to establish a relationship between the current research's actual livelihood situations (Fig. 14). However, it is critical to take into account a broad range of technical improvements in rainfed dry circumstances with variable precipitation, such as genetic crop upgrades, soil moisture preservation strategies, soil health indicators and the use of precision agriculture approaches [50,58]. These technologies may improve productivity and resilience [54].

The study highlights the importance of adaptive management and suggests that strategies and practices like precision agriculture, soil resilience, drought resilience, seed resilience, live stock system adaptation, crop variety adaptation, conservation tillage, and enhanced water use efficiency are crucial for reducing yield gaps [54–56]. Furthermore, the adoption of the Climate-Resilient Agriculture (CRA) approach is an alternative that is critical to food security and reduces the yield gap in the face of changing climate circumstances. For rural societies to flourish sustainably, it may be beneficial [27,28]. Due to variations in precipitation, temperature, farming practices, and soil characteristics across the study area, specific planting and harvesting schedules are necessary for crops, especially rice, which also contend with challenges during the dry summer and winter seasons. Given diminishing water resources and heavy reliance on rainfed and irrigated rice production, the agricultural policies of the study region emphasize the urgent requirement for water-efficient practices, including restrictions on irrigation for rice cultivation. This study emphasizes maintaining soil health through the application of organic fertilizers, strategic irrigation planning, effective management of rice fallow periods with suitable alternative crops, and enhancing managed aquifer systems to improve soil moisture conditions during the dry season [57]. These predictive methodologies are essential for policymakers to proactively plan and implement risk minimization and adaptation strategies at both regional and national levels.

## 8. Conclusions

The current study focuses on Kharif rice yield and gap mapping utilizing ML algorithms with field observation datasets. Various machine learning methods, including Cubist, GBM, MARS, RF, SVM, and XGB, were used to predict present and future rice yield (t/ha) and yield gap (%) monitoring considering SSP2-4.5 and SSP5-8.5 climate data in 8 districts of West Bengal, India, in 2023 and 2030. Among the six MLs, the Cubist model, followed by RF and XGB, proved to be the top performer, showcasing superior $R^2$ values and lower RMSE values (in descending order of $R^2$: (Cubist > RF > XGB > GBM > SVM > MARS), with minimal distinctions in RMSE values across these MLs. The Boruta and SHAP summary plots indicated that pr, elevation, and aet were the most crucial predictors in the RYP analysis across the study region. The machine learning models were compared and classified into five zones (very high, high, moderate, low, and very low). The study findings highlight that Murshidabad and Purba Bardhaman districts exhibit high rice yields due to ample water availability. Using the Cubist model, it was determined that the highest rice yields (5.60–3.45 t/ha) are predominantly found in the northern and northeastern regions, including Murshidabad, Purba Bardhaman, parts of Birbhum, and some areas of Paschim Bardhaman. Moderate rice yields (2.37–1.44 t/ha) are scattered across various parts of Birbhum, Paschim Bardhaman, Bankura, Jhargram, northern Paschim Medinipur, and Purulia districts. Conversely, the western region of the study area, encompassing Purulia and western Birbhum, shows low rice yields (1.44–0.39 t/ha), while the middle portion and northwestern region (Bankura districts and western Birbhum) have very low yields (0.39–0.10 t/ha). The analysis discovered that the north-eastern regions of Murshidabad and Purba Bardhaman had high rice yields in 2023, while the south-eastern regions of Jhargram and Paschim Medinipur had a rise by 2030. In 2023, the Cubist approach shows a significant rice yield disparity in the eastern, western, and southern portions of the research area. By 2030, this gap is expected to reduce to some extent in Birbhum and Bankura districts. Rice crop growth in many districts remained unpredictable due to their reliance on precipitation. Climate change, soil issues, water stress, hard rock, overuse of agrochemicals, and farmers' inefficiency with sophisticated technologies are all factors limiting rice production. If predictors other than elevation are of low resolution, the reliability of yield forecasts could diminish. The study area exhibits diverse climatic conditions from east to west, varying agricultural practices, soil qualities, and different crop varieties, all of which influence the reliability of rice yield gap mapping. While remote sensing-derived estimates of yield and yield gaps may not always achieve the precision of field-based assessments, the extensive spatial and temporal coverage offered by remote sensing technologies often offsets these shortcomings, rendering it a valuable tool for diverse applications.

To preserve rice yields in the future, it is recommended to identify and implement climate change adaptation alternatives such as changing sowing dates and seedling ages, specific crop varieties, agricultural system (i.e. rainfed and irrigated system) fertilizer application timing, and irrigation management practices. The study further utilized crop scouting alongside a multi-year database gathered across various seasons and specific crop varieties. In the future modeling approach, advanced deep learning techniques and high-resolution Unmanned Aerial Vehicle (UAV), PlanetScope data will be integrated with soil parameters, farming practices, and socioeconomic conditions of farmers, along with remote sensing indices, to conduct a comprehensive analysis of rice yield gaps at the regional level. However, these methodological frameworks can be utilized with or without adjustments in any subject field to achieve the UN's sustainable development goals.

## CRediT authorship contribution statement

**Satiprasad Sahoo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chiranjit Singha:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Ajit Govind:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

We acknowledge the project "Integration of Digital Augmentation for Sustainable Agroecosystem in Western Lateritic Zone under National

Hydrology Project, West Bengal" under which this work is mapped. The author also conveys special thanks to the International Centre for Agricultural Research in the Dry Areas (ICARDA) for supporting the necessary logistics for this research work. We would like to express our heartfelt appreciation to all enumerators, farmers, professionals, and anyone who contributed to this research.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jafr.2024.101424.

## References

[1] S. Umesha, H.M. Manukumar, B. Chandrasekhar, Sustainable Agriculture and FoodSecurity, Elsevier eBooks, 2018, pp. 67–92, https://doi.org/10.1016/b978-0-12-812160-3.00003-9.
[2] N.K. Arora, Agricultural sustainability and food security, Environ. Sustain. 1 (3) (2018) 217–219, https://doi.org/10.1007/s42398-018-00032-2.
[3] K.P. Devkota, A. Bouasria, M. Devkota, V. Nangia, Predicting wheat yield gap and its determinants combining remote sensing, machine learning, and survey approaches in rainfed Mediterranean regions of Morocco, Eur. J. Agron. 158 (2024) 127195, https://doi.org/10.1016/j.eja.2024.127195.
[4] H.S. Nayak, J.V. Silva, C.M. Parihar, S.K. Kakraliya, T.J. Krupnik, D. Bijarniya, T. B. Sapkota, Rice yield gaps and nitrogen-use efficiency in the Northwestern Indo-Gangetic Plains of India: evidence based insights from heterogeneous farmers' practices, Field Crops Res. 275 (2022) 108328.
[5] Y. Sulaeman, V. Aryati, A. Suprihatin, P.T. Santari, Y. Haryati, S. Susilawati, D. R. Siagian, V. Karolinoerita, H. Cahyaningrum, J. Pramono, H.S. Wulanningtyas, L. Fauziah, B. Raharjo, S. Syafruddin, D. Cahyana, W. Waluyo, B. Susanto, R. Purba, D.O. Dewi, M. Yasin, Yield gap variation in rice cultivation in Indonesia, Open Agriculture 9 (1) (2024), https://doi.org/10.1515/opag-2022-0241.
[6] P. Arumugam, A. Chemura, B. Schauberger, C. Gornott, Near real-time biophysical rice (oryza sativa L.) yield estimation to support crop insurance implementation in India, Agronomy 10 (11) (2020) 1674.
[7] T. Van Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: a systematic literature review, Comput. Electron. Agric. 177 (2020) 105709.
[8] K. Senthilkumar, J. Rodenburg, I. Dieng, E. Vandamme, F.S. Sillo, J.M. Johnson, K. Saito, Quantifying rice yield gaps and their causes in Eastern and Southern Africa, J. Agron. Crop Sci. 206 (4) (2020) 478–490.
[9] J. Akhter, R. Mandal, R. Chattopadhyay, S. Joseph, A. Dey, M.M. Nageswararao, A. K. Sahai, Kharif rice yield prediction over Gangetic West Bengal using IITM-IMD extended range forecast products, Theor. Appl. Climatol. 145 (3–4) (2021) 1089–1100.
[10] A. Wilson, R. Sukumar, N. Hemalatha, Machine learning model for rice yield prediction using KNN regression, agriRxiv (2021) (2021) 20210310469.
[11] S. Debnath, A. Mishra, D.R. Mailapalli, N.S. Raghuwanshi, V. Sridhar, Assessment of rice yield gap under a changing climate in India, J. Water Climate Chan. 12 (4) (2021) 1245–1267.
[12] Y. Liu, S. Wang, J. Chen, B. Chen, X. Wang, D. Hao, L. Sun, Rice yield prediction and model interpretation based on satellite and climatic indicators using a transformer method, Rem. Sens. 14 (19) (2022) 5045.
[13] S. Yuan, A.M. Stuart, A.G. Laborte, J.I. Rattalino Edreira, A. Dobermann, L.V. N. Kien, P. Grassini, Southeast Asia must narrow down the yield gap to continue to be a major rice bowl, Nature Food 3 (3) (2022) 217–226.
[14] P. Sathya, P. Gnanasekaran, Paddy yield prediction in tamilnadu delta region using MLR-LSTM model, Appl. Artif. Intell. 37 (1) (2023).
[15] J.A. Quille-Mamani, L.A. Ruiz, L. Ramos-Fernández, Rice crop yield prediction from sentinel-2 imagery using phenological metric, Environ. Sci. Proc. 28 (2023) 16, https://doi.org/10.3390/environsciproc2023028016, 2023.
[16] Z. Liu, H. Ju, Q. Ma, C. Sun, Y. Lv, K. Liu, T. Wu, M. Cheng, Rice yield estimation using multi-temporal remote sensing data and machine learning: a case study of Jiangsu, China, Agriculture 14 (2024) 638, https://doi.org/10.3390/agriculture14040638.
[17] S.V. Joshua, A.S.M. Priyadharson, R. Kannadasan, A.A. Khan, W. Lawanont, F. A. Khan, M.J. Ali, Crop yield prediction using machine learning approaches on a wide spectrum, Comput. Mater. Continua (CMC) 72 (3) (2022) 5663–5679.
[18] M.D. Islam, L. Di, F.M. Qamer, S. Shrestha, L. Guo, L. Lin, T.J. Mayer, A.R. Phalke, Rapid rice yield estimation using integrated remote sensing and meteorological data and machine learning, Rem. Sens. 15 (2023) 2374, https://doi.org/10.3390/rs15092374.
[19] N. Gandhi, L.J. Armstrong, O. Petkar, A.K. Tripathy, Rice crop yield prediction in India using support vector machines,", in: *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Khon Kaen, Thailand, 2016, pp. 1–5, https://doi.org/10.1109/JCSSE.2016.7748856.
[20] Distrct Survey Report of Birbhum (DSI 2019). RSP Green Development & Laboratories PVT. LTD, West Bengal, India,pp.1-132.
[21] F.M. Talaat, Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes, Neural Comput. Appl. 35 (2023) 17281–17292, https://doi.org/10.1007/s00521-023-08619-5.
[22] P.R. Jena, B. Majhi, R. Kalli, et al., Prediction of crop yield using climate variables in the south-western province of India: a functional artificial neural network modeling (FLANN) approach, Environ. Dev. Sustain. 25 (2023) 11033–11056, https://doi.org/10.1007/s10668-022-02517-x.
[23] S. Khaki, L. Wang, Crop yield prediction using deep neural networks, Front. Plant Sci. 10 (2019) 1664, https://doi.org/10.3389/fpls.2019.00621, 462X.
[24] T. Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: a systematic literature review, Comput. Electron. Agric. 177 (2020) 105709, https://doi.org/10.1016/j.compag.2020.105709.
[25] P. Filippi, E.J. Jones, N.S. Wimalathunge, P.D.S.N. Somarathna, L.E. Pozza, S. U. Ugbaje, T.F.A. Bishop, An approach to forecast grain crop yield using multilayered, multi-farm data sets and machine learning, Precis. Agric. 1–15 (2019), https://doi.org/10.1007/s11119-018-09628-4.
[26] T.A. Carleton, Crop-damaging temperatures increase suicide rates in India, Proc. Natl. Acad. Sci. USA 114 (33) (2017) 8746–8751.
[27] C. Singha, S. Gulzar, K.C. Swain, D. Pradhan, Apple yield prediction mapping using machine learning techniques through the Google Earth Engine cloud in Kashmir Valley, India, J. Appl. Remote Sens. 17 (1) (2023) 014505, https://doi.org/10.1117/1.JRS.17.014505, 18 January 2023.
[28] C. Singha, S. Sahoo, A. Govind, B. Pradhan, S. Alrawashdeh, T.H. Aljohani, H. Almohamad, A.R.M.T. Islam, H.G. Abdo, Impacts of hydroclimate change on climate-resilient agriculture at the river basin management, J. Water Climate Chan. 15 (1) (2023) 209–232, https://doi.org/10.2166/wcc.2023.656.
[29] K. Alsafadi, S. Bi, H.G. Abdo, et al., Modeling the impacts of projected climate change on wheat crop suitability in semi-arid regions using the AHP-based weighted climatic suitability index and CMIP6, Geosci. Lett. 10 (2023) 20, https://doi.org/10.1186/s40.
[30] N. Arunrat, S. Sereenonchai, W. Chaowiwat, C. Wang, Climate change impact on major crop yield and water footprint under CMIP6 climate projections in repeated drought and flood areas in Thailand, Sci. Total Environ. 807 (2) (2021) 150741, https://doi.org/10.1016/j.scitotenv.2021.150741.
[31] T.G. Farr, P.A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, L. Roth, D. Seal, S. Shaffer, J. Shimada, J. Umland, M. Werner, M. Oskin, D. Burbank, D.E. Alsdorf, The shuttle radar topography mission, Rev. Geophys. 45 (2) (2007), https://doi.org/10.1029/2005RG000183. RG2004, at.
[32] S.J. Muñoz, ERA5-Land monthly averaged data from 1981 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (2019), https://doi.org/10.24381/cds.68d2bb30, 01 January 2024.
[33] J.T. Abatzoglou, S.Z. Dobrowski, S.A. Parks, K.C. Hegewisch, Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015, Sci. Data 5 (2018) 170191, https://doi.org/10.1038/sdata.2017.191.
[34] B. Thrasher, E.P. Maurer, C. McKellar, P.B. Duffy, Technical Note: bias correcting climate model simulated daily temperature extremes with quantile mapping, Hydrol. Earth Syst. Sci. 16 (9) (2012) 3309–3314, https://doi.org/10.5194/hess-16-3309-2012.
[35] S. Fei, L. Li, Z. Han, et al., Combining novel feature selection strategy and hyperspectral vegetation indices to predict crop yield, Plant Methods 18 (2022) 119, https://doi.org/10.1186/s13007-022-00949-0.
[36] C. Singha, K.C. Swain, H. Jayasuriya, Growth and yield monitoring of potato crop using Sentinel-1 data through cloud computing, Arabian J. Geosci. 15 (2022) 1567, https://doi.org/10.1007/s12517-022-10844-6.
[37] C. Singha, K.C. Swain, Rice and potato yield prediction using artificial intelligence techniques, in: P.K. Pattnaik, R. Kumar, S. Pal (Eds.), Internet of Things and Analytics for Agriculture, Volume 3, Studies in Big Data, vol. 99, Springer, Singapore, 2022, https://doi.org/10.1007/978-981-16-6210-2_9.
[38] C. Singha, K.C. Swain, Rice crop growth monitoring with sentinel 1 SAR data using machine learning models in google earth engine cloud, Remote Sens. Appl.: Society and Environment 32 (2023) 101029, https://doi.org/10.1016/j.rsase.2023.101029.
[39] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[40] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, Xgboost: Extreme Gradient Boosting; R Package Version 0.4-2, vol. 1, The R Foundation, Indianapolis, IL, USA, 2015, 2015.
[41] M. Kganyago, P. Mhangara, C. Adjorlolo, Estimating crop biophysical parameters using machine learning algorithms and sentinel-2 imagery, Rem. Sens. 13 (2021) 4314.
[42] G. Sahbeni, B. Székely, P.K. Musyimi, G. Timár, R. Sahajpal, Crop yield estimation using sentinel-3 SLSTR, soil data, and topographic features combined with machine learning modeling: a case study of Nepal, AgriEngineering 5 (2023) 1766–1788, https://doi.org/10.3390/agriengineering5040109.
[43] J.R. Quinlan, Combining instance-based and model-based learning, in: Proceedings of the 10th International Conference on Machine Learning, Amherst, MA, USA, 27–29 June 1993, 1993, pp. 236–243.
[44] E. Eyduran, M. Akin, S. Eyduran, Application of Multivariate Adaptive Regression Splines through R Software, Nobel Academic Publishing, Ankara, Turkey, 2019.
[45] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.
[46] E.J. Jones, T.F.A. Bishop, B.P. Malone, P.J. Hulme, B.M. Whelan, P. Filippi, Identifying causes of crop yield variability with interpretive machine learning, Comput. Electron. Agric. 192 (2022) 106632, https://doi.org/10.1016/j.compag.2021.106632.
[47] W.Z. Othman, N.N.A. Tukimat, Assessment on the climate change impact using CMIP6, in: IOP Conference Series: Earth and Environmental Science, vol. 1140, IOP Publishing, 2023, February 012005, 1.

[48] J. Rockstrom, J. Williams, G. Daily, A. Noble, N. Matthews, L. Gordon, H. Wetterstrand, F. DeClerck, M. Shah, P. Steduto, C. de Fraiture, N. Hatibu, O. Unver, J. Bird, L. Sibanda, J. Smith, Sustainable intensification of agriculture for human prosperity and global sustainability, Ambio 46 (2017) 4–17, https://doi.org/10.1007/s13280-016-0793-6.

[49] A.J. Teuling, R. Stockli, S.I. Seneviratne, Bivariate colour maps for visualizing climate data, Int. J. Climatol. 31 (2011) 1408–1412.

[50] S. Sahoo, A. Govind, Understanding changes in the hydrometeorological conditions towards climate-resilient agricultural interventions in Ethiopia, Agronomy 13 (2023) 387, https://doi.org/10.3390/agronomy13020387.

[51] C. Singha, K.C. Swain, S.K. Swain, Best crop rotation selection with GIS-AHP technique using soil nutrient variability, Agriculture 10 (2020) 213, https://doi.org/10.3390/agriculture10060213.

[52] D.K. Ray, N.D. Mueller, P.C. West, J.A. Foley, Yield trends are insufficient to double global crop production by 2050, PLoS One 8 (2013) e66428.

[53] V.O. Sadras, K. Cassman, P. Grassini, W.G.M. Bastiaanssen, A.G. Laborte, A. E. Milne, P. Steduto, Yield Gap Analysis of Field Crops: Methods and Case Studies, 2015.

[54] J.A. Foley, N. Ramankutty, K.A. Brauman, E.S. Cassidy, J.S. Gerber, M. Johnston, N.D. Mueller, C. O'Connell, D.K. Ray, P.C. West, Solutions for a cultivated planet, Nature 478 (2011) 337–342.

[55] L. Lipper, P. Thornton, B.M. Campbell, T. Baedeker, A. Braimoh, M. Bwalya, P. Caron, A. Cattaneo, D. Garrity, K. Henry, R. Hottle, L. Jackson, A. Jarvis, F. Kossam, W. Mann, N. McCarthy, A. Meybeck, H. Neufeldt, T. Remington, E. F. Torquebiau, Climate-smart agriculture for food security, Nat. Clim. Change 4 (12) (2014) 1068–1072, https://doi.org/10.1038/nclimate2437.

[56] J. Pretty, T.G. Benton, Z.P. Bharucha, L.V. Dicks, C.B. Flora, H.C.J. Godfray, D. Goulson, S. Hartley, N. Lampkin, C. Morris, G. Pierzynski, P.V.V. Prasad, J. Reganold, J. Rockstrom, P. Smith, P. Thorne, S. Wratten, Global assessment of agricultural system redesign for sustainable intensification, Nat. Sustain. 1 (2018) 441–446, https://doi.org/10.1038/s41893-018-0114-0.

[57] A. Bendidi, K. Daoui, A. Kajji, L. Bouichou, M. Bella, M. Ibriz, R. Dahan, Response of bread wheat to sowing dates and the genotypes in Morocco, J. Exp. Agric. Int 14 (2016) 1–8, https://doi.org/10.9734/jeai/2016/30216.

[58] M. Devkota, K.P. Devkota, S. Kumar, Conservation agriculture improves agronomic, economic, and soil fertility indicators for a clay soil in a rainfed Mediterranean climate in Morocco, Agric. Syst. 201 (2022) 103470, https://doi.org/10.1016/j.agsy.2022.103470.