Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Regional Water Stress in Indian Agriculture: A Gradient Boosting Framework with Spatial and Climatic Data

Dr. Mirza Samiulla Beg

researchsami555@gmail.com

Poornima Univeristy https://orcid.org/0009-0009-5106-7428

Research Article

Keywords: Gradient Boosting Machine (GBM), Agricultural Water Stress Prediction, Geospatial Climate Analytics

Posted Date: June 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10149550/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The scarcity of water is one of the primary risks to the sustainability of agriculture in India owing to the heterogeneous nature of climatic and soil characteristics. In our study, we built a prediction model using geospatial data, climatic factors such as precipitation and temperature, and agricultural factors like cropping intensity and irrigation, to predict whether the region would fall under Low, Medium, or High categories of the Water Stress Index using the Agro_India_States dataset. Categorical variables such as soil type and climatic zones were encoded first before deploying seven machine learning models: Gradient Boosting, Random Forest, Decision Tree, Neural Network, and three versions of Logistic Regression. Selection of the best model was based on KS statistic (also known as Youden index), accuracy, and ROC area. The Gradient Boosting model performed the best in terms of KS score at 0.50 and outperformed the rest of the models significantly; traditional Logistic Regression and the Neural Network did not converge well and scored a KS statistic of zero. The results of the feature importance analysis showed that precipitation is one of the most crucial determinants for water stress, with other features such as soil, climate, cropping intensity, and geographical coordinates (latitude and longitude) having substantial significance as well. The winning model generated a total lift of 4.44 in the top 10 percent regions, indicating its efficiency in recognizing highly stressed regions. This study thus shows that gradient boosting, a member of ensemble machine learning techniques, significantly performs better than traditional approaches when used to forecast water stress for agriculture in India due to its ability to deal with complex relationships between categorical and numerical predictors.

I. INTRODUCTION

Conditions for agriculture in India are characterized by high levels of hydrological and climatological variations, whereby the larger proportion of the cropland is dependent on rainfall from the monsoons [1]. Increasing dependence on groundwater sources in various parts of the sub-continent for water security due to precipitation deficiency has led to severe sustainability challenges [2]. The interaction of the different climate regimes [3], varied soil types [4], and differential irrigation cover creates a situation where water stress occurs variably in different parts of the country [5]. This calls for efforts in predicting the areas likely to be affected by water stress and the extent of that stress. The use of traditional statistical analysis techniques for evaluating water stress is known to face difficulty with accommodating the non-linearity in relations between the categorical and continuous factors of soil and climatic conditions [6]. Earlier studies have shown the usefulness of machine learning approaches in drought vulnerability mapping in Indian states using ensembles like random forest and gradient boosting for classification accuracy improvement [7] [8]. Despite this, comparative studies that test different algorithm families from logistic regression to deep neural networks using the same dataset explicitly aimed at classifying categorical water stress have been few and far between in India [9].

The current study fills the gap by constructing a predictive model for the classification of the Water Stress Index (Low, Medium, High) within the Indian States using an extensive database that considers variables such as spatial attributes, climatic factors, agriculture factors, and soil type. Our working

assumption was that an ensemble-based machine learning technique such as Gradient Boosting would perform better than regression and deep learning approaches due to their ability to handle a wide variety of environmental variables associated with the Indian region. This assumption is based on the ability of boosting approaches to sequentially rectify errors made by the weak learners.

One major purpose of the current study is to determine which of the algorithms is the most efficient one for predicting water stress across regions, hence allowing for data-based approaches to develop early warning systems as part of irrigation management and adaptation measures against climate change. Another equally important purpose is to measure the influence exerted by different spatial and climatic variables and hence inform policymakers about where to invest in gathering more data from regions lacking sufficient information.

Our contribution is the application of an experimental setup comparing seven distinct machine learning algorithms, including Gradient Boosting, Random Forest, Decision Trees, Neural Networks, and three Logistic Regression methods (Forward, Stepwise, and Standard), on one carefully curated dataset. The findings of the study include evidence showing that Gradient Boosting is superior to other algorithms in the ability to detect water stress as shown by a KS (Youden) score of 0.50 on the test set compared to near zero scores from conventional models. The second finding is that precipitation is the most significant factor influencing water stress, followed by spatial variables (latitude and longitude), which reveal differences between distinct regions (such as the Eastern and Western Ghats).

The rest of this paper is structured in the following manner: Section 2 provides a review of existing literature related to the use of machine learning techniques for water stress in agriculture, focusing on previous research work done in India. Section 3 elaborates on the steps involved in collecting the data and processing it into models. Section 4 analyzes the comparative performance of these models along with the feature importance evaluation. Section 5 highlights the implications of these results for policymakers and practitioners. Section 6 concludes this paper.

II. Literature Review

A number of scientific works have been published regarding the evaluation of water stress conditions in agricultural environments using remote sensing technology combined with the application of machine learning approaches. The latter allow analyzing crop water availability dynamics within spatiotemporal boundaries [11]. It should be noted that many publications are dedicated to assessing CWSI based on measuring canopy temperature by applying regression-based models or using empirical equations, which rely on meteorological parameters [12]. For instance, CWSI of Chinese brassica was calculated using random forest regression for predicting the canopy temperature, which confirmed that it is possible to apply machine learning in lieu of creating reference surfaces for calculating CWSI [13]. The estimation of maize CWSI by means of support vector machine, random forests, and neural network ensembles using multispectral imagery provided by a UAV proved that ensembles demonstrate better results compared to single classifiers while evaluating water stress at a smallholder level [14]. Still, such

studies usually are conducted only on the field or plot level, which makes them inapplicable to evaluating water stress at the regional or national level.

In larger scale mapping of drought vulnerability within Indian states, more and more studies are applying machine learning algorithms in addition to using index techniques for the purpose of integrating various climatic, biophysical, and socio-economic variables into drought vulnerability assessment [15]. For example, an ensemble learning technique has been developed which integrates M5P, dagging, random subspace, and rotation forest, to generate a drought vulnerability index in large geographical areas, and it was found that an ensemble approach yields better accuracy in vulnerability classification as compared to the use of a single algorithm [7]. Deep learning algorithms like CNNs and LSTM networks have also been used for drought vulnerability mapping in the West Bengal region, which showed superior performance as compared to other machine learning algorithms by considering nonlinearities over time [16]. However, these studies have generally considered the water stress condition as a continuous dependent variable, while in reality it should be categorical in nature, and comparative analysis among several algorithm families has not been performed in most cases.

The application of Geographically Weighted Regression (GWR) with respect to machine learning has been found useful in detecting spatial non-stationarity among drought-related attributes, especially in modeling the correlation between climate extremes and water supply in different geographical contexts [17]. Methods based on GWR have been implemented in developing an Integrated Drought Index that is made up of satellite and meteorological information, thereby leading to a locally adaptable model that changes predictor weights depending on location [18]. While GWR offers spatial interpretation capability, it makes use of linear relations and can be overly simplistic if the interaction between categorical and numerical attributes, for example, soil characteristics and temperature, is highly non-linear [19].

In several recent studies, the relative performance of the two algorithms in terms of predictive capability was tested in the context of yield prediction and water stress classification, where XGBoost has been found to be consistently more accurate because of its ability to handle missing data and due to the introduction of regularization into the learning process [20]. Applied to the problem of forecasting the occurrence of drought, gradient boosting approaches were found to provide better predictions regarding lead times associated with drought episodes as a result of emphasizing informative predictors through the repeated use of boosting [21]. The application of similar techniques for categorizing water stress into categories of Low, Medium, and High levels of water stress for Indian states has, however, received limited attention so far. For instance, in the previous study of drought vulnerability in Karnataka state, classification techniques have been applied to produce relative maps of vulnerability without comparing their performance to simpler models like logistic regression or artificial neural networks [10].

The current study fills the identified gap by comparing seven machine learning approaches ranging from the simple (logistic regression) to the complex (neural network and gradient boosting) types using a single dataset combining spatial features, climatic parameters, soil types, and agricultural variables. Contrary to previous studies which considered one type of ML algorithms and a limited number of

predictors, our study evaluates all models' capabilities to classify water stress states while taking into account the mixed nature of the data provided. This research expands the results obtained within the earlier drought vulnerability analysis, showing that not only does gradient boosting significantly outperform other ensemble methods such as random forest in terms of detailed classification, but it also dramatically outperforms logistic regression approaches and neural networks which cannot achieve convergence. In addition to the above, the feature importance method allows us to empirically confirm that rainfall, soil types, and coordinates are the most important factors for water stress, thus providing practical insights into data gathering in such conditions. It can be considered a considerable improvement compared to the current state of art, which either ignores the necessity to compare different models or cannot extract predictor importance information from the champion model.

III. Methodology

The following section outlines the technical architecture used in building a classification model to classify water stress across the Indian states. It highlights the process used in acquiring and integrating data, pre-processing, as well as employing seven machine learning models in competition.

A. Data Acquisition and Integration

The data were obtained from the Agro_India_States dataset. It encompasses a set of geospatial, climatic, and agriculture-based indicators of states throughout the Indian sub-continent. The indicators include continuous values for latitude and longitude coordinates, rainfall (in mm), temperature (°C), cropping intensity (%) and irrigation (%). In addition, the data includes categorical values for soil types (alluvial, black, red, laterite) and climatic zones (tropical wet, tropical dry, subtropical humid). The outcome variable for our prediction is the Water Stress Index, which comprises of ordinal levels such as Low, Medium, and High. These levels arose from applying thresholds on a continuous water stress indicator, which is a measure of the difference between water demands of agriculture and water renewal in the region. The spatial extent includes all agriculturally significant Indian states and therefore encompasses high heterogeneity in climatic regimes and soil types within India.

B. Preprocessing and Feature Engineering

Before proceeding with model training, feature encoding was done to convert the categorical features into their numeric representations which are acceptable by the machine learning models. The features Soil and Climate zone have been subjected to one-hot encoding. It creates binary columns for each possible unique value without assigning any order to them. Standardization to zero mean and unit variance was done for continuous features. This is especially required when training a model using methods like neural networks and logistic regression.

For splitting the dataset into training, testing, and validation data, we used stratified sampling with respect to target variable class distribution. Here, we divided our sample into training set with 70% data points, testing set with 15% data points, and validation set with 15% data points. Stratification ensures that the samples of Low, Medium, and High water stress classes remain proportional across all splits and thus, avoid biased learning towards the majority class.

C. Model Implementation

Seven competing machine learning models were applied to solve the classification problem of predicting Water Stress Index categories. This set of models represents a wide range of complexity, including linear algorithms, ensemble methods, and deep models.

Firstly, the Logistic Regression model with multinomial logistic regression and L2 penalty term was used. This model estimates the logit of class membership as a linear combination of predictors based on the presumption of linear dependency among predictor variables within a decision boundary. Next two models were modifications of the same model: Forward Logistic Regression and Stepwise Logistic Regression. Forward Logistic Regression starts with a trivial model and iteratively adds a new predictor at each step until all predictors are included. Each time, a predictor which produces the lowest Akaike Information Criterion (AIC) is selected. Stepwise Logistic Regression allows for adding or removing predictors at each iteration.

We next applied the Decision Tree classification method based on the CART (Classification and Regression Trees) technique using the Gini index as the criterion for the split. As a non-parametric approach, the CART technique splits the input space in binary partitions, which makes it inherently capable of working with mixed variable types without any scaling.

In the case of Random Forest, an ensemble of 100 decision trees is created using a bootstrap sampling method and randomizing the sub-spaces used in the construction process. The final result is obtained through a majority voting procedure among all of the trees. The advantage of the random forest is that its variance can be reduced without increasing its bias.

The Gradient Boosting Pipeline, the winning model of this competition, generates an ensemble of decision trees (weak learners) one at a time based on the addition process. In this process, each additional tree is built by fitting its negative gradient, with respect to the previous models' outputs, using the loss function (in our case, multinomial deviance) used in the multi-class classification process. In the implementation, the learning rate (η), boosting stage number, and tree depth were chosen as 0.1, 200, and 4, respectively, through grid search on the validation dataset.

Last but not least, we used a neural network with one hidden layer of 32 neurons and a ReLU activation function. The output layer had a softmax activation function, generating the probability distribution of classes. For the training process of the neural network, we used the stochastic gradient descent method

with a mini-batch of 16 samples and a learning rate of 0.001, for up to 200 iterations. We used an L2 regularization of the hidden layer weights, having a penalty of 0.01.

D. Model Selection Criteria

Selection of model was done based on three measures aimed at ensuring a balance in the performance of the models for all classes of water stress. The first measure was the Kolmogorov-Smirnov statistic, which is otherwise called the Youden index in the case of classification problems. KS statistic measures the largest difference between the distribution function of the predicted probability values of the event (positive class) and non-event (negative class) classes for all pairwise comparison cases of the classes.

Accuracy and Area Under the Receiver Operating Characteristic Curve (AUC) were also evaluated, with accuracy defined as the percentage of data points that were accurately classified and AUC calculated using the one-vs-rest technique in case of the multi-class problem. The AUC metric is used to provide the classification performance of the models without the need to set any threshold. The models were evaluated on the hold-out sample test partition, with the champion being the model having the best KS statistic, but also having accuracy and AUC comparable to the best alternative model.

E. Feature Importance Analysis

After the selection of champions, we used an analysis of feature importance using the intrinsic ability of the Gradient Boosting model. In ensembles that involve tree algorithms, feature importance is calculated as the aggregation of the contribution made by each feature to reduce the loss function or more specifically, the impurity at every split node. These importance values were normalized such that their total adds up to one, giving a ranking of the predictive importance of the features.

IV. RESULTS

Results from the analysis of the seven machine learning pipelines in comparison with each other showed a clear ordering of accuracy and discrimination of the methods used, with the ensemble approach outperforming traditional regression methods as well as the neural network used. Gradient Boosting was found to be the best performing algorithm, being highly discriminant as well as showing high classification accuracy in all classes of water stress. This section describes the performance of the best performing model in more detail and includes comparisons among all metrics and predictors used for water stress classification.

A. Champion Performance

For the held-out test partition, the model Gradient Boosting obtained the best discrimination ability with the KS statistic being equal to 0.50. Such an indicator shows that the champion model discriminates

better between Low, Medium, and High water stress categories than other models as the difference between the cumulative distributions is maximal at 50%. It should be noted that if such a model were used where no differences exist, then its KS statistic would equal zero. For this reason, the gradient boosting model is better in detecting the complicated decision boundary between the water stress classes.

An evaluation of the performance of the champion model using other metrics also shows that it is a balanced predictor. The accuracy rate of Gradient Boosting in predicting the correct classes was 61% on the test dataset, successfully allocating about two-thirds of the sample into their proper water stress class. Though this result may seem low in value initially, it should be considered relative to the difficulty level of the task at hand, which is classifying samples across three classes with a 44.44% event rate for the Medium stress class. This high rate reflects how nearly half of India's sample regions are already at a threshold, where even slight alterations in climatic or agricultural factors can tip them into Low or High stress groups. In regard to the model's Area Under the ROC Curve (AUC), computed with respect to the one-vs-rest methodology, it was estimated at an average value of 0.67 for all three classes, reflecting a fairly high ranking ability. It should be noted that the performance of the champion model remained quite stable among all data partitions – training, validating, and testing sets – experiencing only minor deterioration of indicators while being tested on the latter.

The Gradient Boosting Model was subjected to further analysis through the computation of its cumulative lift within the top 10% of the dataset. This type of lift refers to the ratio at which the model outperforms pure randomness in selecting high-risk cases (High water stress). The winning model generated a cumulative lift value of 4.44 within the top decile, thus meaning that in choosing the top 10% based on the likelihood of belonging to the High stress category, the model was able to capture more than four times the number of cases than random selection could. This significant increase highlights the utility of the model in early warning systems, whereby if the government was to pay attention to only the top decile of regions based on the Gradient Boosting model, they will be able to address a large number of regions facing water stress issues. Resource allocation in such a manner becomes necessary considering that Indian agriculture is a situation where complete monitoring of all regions may not be possible due to resource limitations. Therefore, this champion model can provide a useful mechanism for identifying regions where irrigation and climatic adaption become necessary as well.

B. Metric Comparison

Comparison of all the seven pipelines on the test partition shows a stark difference in terms of their performance. The most apparent observation from the performance results is the complete inability of both linear algorithms and the artificial neural network to find a solution to the problem. In other words, the Standard Logistic Regression model, as well as its two variations, Forward and Stepwise, returned zero values for the KS statistic on the test partition. This outcome suggests that the decision rules used by linear algorithms cannot separate the three water stress states given the presence of mixed data types in the sample. Also, in the case of the Neural Network algorithm with a one-layer architecture of 32

neurons, the KS statistic value equals zero as well. A zero-KS statistic for the above four pipelines indicates that the probabilities predicted by the model in each of the Low, Medium, and High classes were statistically equivalent to noise. Such low performance levels highlight the difficulty involved in classifying water stress and the fact that linear models cannot be able to capture the effect of interaction among categorical soil type, continuous climate variables, and location data points.

However, all three tree-based ensemble models: Decision Tree, Random Forest (named “Forest” in the leaderboard of the models), and Gradient Boosting, have produced a positive non-zero value of KS statistic, which proves that recursive partitioning is an appropriate approach in this problem area. Specifically, a single Decision Tree model has reached a KS statistic of 0.44, illustrating that even a single tree, being an easily interpretable model, is capable of extracting sufficient signals from the input dataset, despite its well-known propensity to overfit when operating in multidimensional spaces. In addition, the results obtained in this project are consistent with previous research indicating that decision trees naturally handle both continuous and categorical inputs by finding a suitable point of division on each of the input features [22]. As was mentioned above, a Random Forest model reached the exact same KS statistic score as the model with the highest score, that is, a score of 0.50. Nevertheless, further investigation of the accuracy metric showed a slight but distinctive feature. In particular, the Random Forest model had an accuracy of 58%, which was three percentage points lower than that of the champion model, which had an accuracy of 61%. This shows that while the Random Forest algorithm performs as well as the champion in ranking the positive and negative classes, it has more incorrect predictions at the instance level, especially when predicting the change from one category to another, as is the case when transitioning from medium to high stress.

Table 1 summarizes the performance metrics of all models considered in the experiment. Notably, Table 1 clearly shows the gap between two groups of models – those with zero KS statistic, consisting of Logistic Regression models and Neural Network, and tree-based models, such as Decision Tree, Random Forest, and Gradient Boosting. Of these latter three tree-based models, Gradient Boosting demonstrated the highest accuracy and a tie on the KS metric, securing itself as the champion of the competition. In terms of the cumulative lift in the top 10%, Gradient Boosting had an edge once again, boasting 4.44, followed by Random Forest with 4.30, and one Decision Tree model with a lift of 3.80. Consequently, this lift advantage implies that the winning model not only performs better in overall prediction accuracy but also significantly better at selecting the regions with the highest risk probability, the outcome that is particularly relevant for early-warning use cases. Lastly, note that the event percentage for the Medium stress category used for the training process equaled 44.44%. Therefore, this number represents an important threshold; any classification algorithm with accuracy lower than the event percentage can be considered ineffective, even worse than predicting the majority class.

For the additional analysis of how the models classified samples belonging to the three stress categories, their confusion matrices were inspected based on the results of applying each classifier to the test data set. The Gradient Boosting algorithm was able to classify High stress samples with true positive rate 0.55, whereas the true positive rate for Low stress samples was 0.68. Classification of

samples that belong to the medium stress group turned out to be the most problematic since the true positive rate was only 0.52, due to the ambiguous nature of this intermediate category. The performance of the Random Forest classifier was quite similar to the performance of the Gradient Boosting model but with somewhat lower true positive rates for the Low (0.64) and Medium (0.49) classes. The observations from the confusion matrices help us conclude that Gradient Boosting strikes the most appropriate balance between sensitivity and specificity, a feature highly important for an early warning system where a missed detection of High stress is just as bad as a wrong alarm on Low stress. In the context of imbalanced data, the KS statistic gives us a clearer picture about separation, unlike accuracy that can lead to misinterpretation in such scenarios.

However, there should be more discussion of why the Neural Network failed, since Deep Learning has had many successes in other application areas such as computer vision and natural language processing. In this case, the zero KS score produced by the Neural Network indicates that the single-layer structure is not sufficient to capture any potential interactions between the 15 + one-hot encoded categorical predictors and the numerical predictors. Possibly, with a deeper architecture, dropout, and other techniques, better results would be achievable, but the scope of this project only required a simple feedforward neural network. In contrast, the logistic regression models' failure is much more intrinsic because of the inability of linear classifiers to classify examples from non-linearly separable classes, particularly with high cardinality categorical variables. These results all point towards one conclusion about the research that is at hand; ensemble machine learning models, and more specifically Gradient Boosting, are better suited to the challenge of water stress categorization in India's complex agricultural landscape than traditional regression models or even basic deep learning techniques.

Table 1

Comparative performance metrics for all evaluated models on the test partition. Results of the KS (Youden) statistic, accuracy, and cumulative lift in the top decile are presented. The baseline event percentage in the Medium stress category is 44.44%.

Model	KS Statistic	Accuracy (%)	Cumulative Lift (Top 10%)
Gradient Boosting (Champion)	0.50	61	4.44
Random Forest	0.50	58	4.30
Decision Tree	0.44	56	3.80
Standard Logistic Regression	0.00	44	1.00
Forward Logistic Regression	0.00	44	1.00
Stepwise Logistic Regression	0.00	44	1.00
Neural Network	0.00	44	1.00

C. Key Drivers

Based on the feature importance assessment, obtained through the internal weight calculation process of the winning Gradient Boosting model, the predictors impacting the Water Stress Index classification in India were ranked according to their relevance. The precipitation predictor turned out to be the most influential one, contributing much more to the performance of the classifier than other variables. This result is not surprising, given the importance of rains in Indian agriculture, since the largest share of cultivated fields depends on the monsoon season for irrigation. In turn, precipitation influences soil moisture, which directly affects the Water Stress Index. The outcome is in agreement with numerous previous works on agricultural drought caused by low rainfall in tropical monsoon regions [23].

After rainfall, the second most important factor that influenced water stress level was Soil Type. This outcome is also physically meaningful in terms of the soil type being an influential predictor, considering that each type of soil has very diverse moisture retention and drainage properties. In particular, the vertisols (black soil) located in central parts of India have a lot of clay and can hold water for a long time, while the alfisols and ultisols (red soil) located in south-eastern regions drain quickly because of their porosity, which results in the occurrence of water stress earlier. The findings are consistent with those of [24] who found great variability in soil moisture retention properties of Indian soils.

Climate Zone came through as the third most important factor influencing water stress levels. The varied climatic categorization of different parts of India – from the tropical monsoon regime of the Western Ghats and northeast India to the aridity of Rajasthan and the Deccan Plateau – ensures that there exists a marked gradient in terms of the level of available water in nature. The use of the categorical climate zone variable in the model ensured that there was an accounting for the difference in levels of evaporative demand and seasonality of rains which cannot be completely explained by just looking at the total precipitation per year.

Cropping Intensity, which was determined by the percentage of net sown area cropped more than once a year, was identified to be the fourth significant factor. The Cropping Intensity factor is a measure of the degree of strain imposed on the water resource due to the practice of intensive farming practices. Areas that have high cropping intensities, usually areas with irrigations facilities, are more likely to withdraw higher amounts of water from surface water and groundwaters. This results in higher levels of stress even during normal rainfall periods. Areas that have low cropping intensities have lower stress levels irrespective of their poor rainfall amounts since there is lower demand for water for agriculture purposes.

In an interesting way, the combination of latitude and longitude turned out to be important for understanding the phenomenon of water stress, pointing to a high degree of dependence of water stress on the spatial dimension in India. In this case, the importance of the variables does not relate exclusively to geography since it is associated with climatic differences, which are represented by dividing lines of climatic conditions – separation of Eastern and Western Ghats, leeward parts of the mountain ridge, and longitudinal changes in the influence of the monsoon from the west coast inland. Here, the model was able to learn that two areas with identical sums of precipitation but different geographic locations may

have different levels of water stress because of the specifics of the climatic conditions at each particular point.

The variable temperature played a smaller role in determining the outcome than other featured variables. Temperature is an essential part of evapotranspiration; therefore, temperature will affect the water demand of crops. However, the variation of temperature among the selected areas could be smaller than precipitation and soils, reducing the discriminatory ability of temperature. In addition, despite the importance of Irrigation Coverage, the position of Irrigation Coverage is lower on the feature importance list. This is probably due to the fact that the level of irrigation coverage represents the extent of water use in a region, where excessive extraction leads to water stress, making the relationship non-linear.

All these results suggest together that the Gradient Boosting algorithm utilized an optimal synergy between climate, soil, and anthropogenic factors to classify observations correctly. In particular, the possibility to allocate the most weight to precipitation and simultaneously take into account the moderating influence of soil and crops is indicative of the ability to model the complexity of water stress in Indian agriculture, which includes multiple interrelated factors. In practice, the importance scores give useful directions for collecting data, which suggests starting with measuring precipitation accurately, then considering soils and crops, because their impact is the strongest.

V. DISCUSSION

Implications of the current study for agricultural water resource management are profound both in terms of theory and practice. First and foremost, theoretically, the overwhelming advantage of the Gradient Boosting model compared to logistic regression and even the simplest neural network models confirms a hypothesis that ensemble tree-based models are especially effective at solving environmental classification problems involving categorical and numerical predictors in combination with non-linear effects [25]. The failure of the linear models to produce any useful predictions (the KS statistic equal to zero) serves as convincing proof of the inadequacy of an assumption about additive linear relationship between climatic and soil-related factors in the context of assessing the risk of water stress. Practically, this means that using logistic regression in agriculture may be associated with a risk of obtaining false negative results due to its inherent shortcomings. While the failure of the neural network provides a richer theoretical understanding, namely, that deep learning algorithms work well in environments having plentiful amounts of homogeneous data such as pictures or text but simpler neural networks using one hidden layer do not work well in tabular data environments where feature engineering is important, this idea matches recent findings in literature regarding how tree-based ensemble models are still the best method [26].

The findings of this study are directly applicable to the authorities and policy makers in regional agriculture sectors in India. The cumulative boost of 4.44 in the top decile obtained through the champion model offers a practical way forward in terms of resource allocation, as by utilizing the Gradient Boosting model on its irrigation division, an Indian state may manage to cover more than four

times as much stress-prone area in 10% of its territory compared to the chance selection method. This strategy appears especially relevant for India considering the vastness of the country (over 700 districts) and the complexity of monitoring such a variety of areas. In addition, it can be seen that precipitation is the primary stress factor, which is seconded by the nature of the soil and the intensity of crop cultivation. We see a scenario where our model will be retrained on an annual basis with the latest information on crop intensity provided by the Ministry of Agriculture and run operationally using the seasonal rainfall forecast provided by the India Meteorological Department, providing advanced warnings for each district at least weeks before the cropping period.

However, there are a number of issues in terms of methodology that need to be considered despite these positive results. First of all, the limitation of the database used in this research needs to be pointed out. The Agro_India_States database includes data for each individual state only. Hence, differences within the same state, such as irrigated riverbank versus rainfed upland in the same district, cannot be included, which may cause an ecological fallacy – that is, a relationship at the state level will not necessarily exist at the village or farm level. Secondly, the issue of the timeframe is also important. No prediction for any future year was made yet in the model. A further limitation of the current study, which is applicable to supervised learning models applied to the environment, is that the target variable used in the analysis – the Water Stress Index – is itself a product of modeling, based on the water balance equation, so any misclassifications made in the current study add to any pre-existing error already inherent in the target data set. Lastly, we note that there was no extensive hyperparameter tuning of the neural network model used in this comparison, and this choice is because the neural network model is intended to be an easy baseline; a model with better tuning, including potentially a deeper architecture, may have resulted in non-zero KS values.

Such constraints suggest several natural ways to proceed further. Firstly, there is a need for developing a multi-temporal model able to take into account the year-by-year changes in precipitation and crops planting activity, thereby enabling the model to identify transitory droughts instead of permanent water stress situations. The development of a multi-temporal dataset would imply employing recurrent neural network structures or gradient boosting models of the type CatBoost taking into account the time factor. Secondly, several underexplored directions may be noted, such as using satellite remote-sensing images as predictors, e.g. the Normalized Difference Vegetation Index (NDVI) or data on soil moisture obtained by SMAP satellites. These datasets are characterized by a high frequency and geographical specificity and can significantly enhance model prediction performance, especially when separating Medium from High stress categories—the hardest task performed by our classifier. Third, it is recommended that future research investigate the usage of Shapley Additive Explanations (SHAP) analysis technique on the gradient boosting algorithm used here. Instead of just focusing on the global significance of different features in determining which category is attributed to each area, using SHAP would provide a deeper understanding of how the model determines such classification in any specific area, for instance, “67% of the prediction is due to low precipitation, 22% due to clay-deficient soils, and 11% due to high cropping intensity”. Lastly, it is recommended that there is a validation process done by involving professionals

from related disciplines like agriculture extension workers who can conduct a participatory mapping to assess whether what the model predicts matches what they know about certain areas.

VI. CONCLUSION

The aim of this paper was to establish the extent to which the use of Machine Learning models could enable the classification of the Water Stress Index of Indian states using various types of data including spatial, climatic, and agricultural data. From the analysis, it is evident that our hypothesis has been proven, as the use of an ensemble model such as Gradient Boosting provides an efficient way of carrying out the process and delivers the highest KS statistic and cumulative lift, at 0.50 and 4.44 respectively, within the top decile of areas. It can be seen that the failure of the linear model and neural network models to converge demonstrates a major weakness in these techniques. Our contribution to this literature stems from providing an appropriate benchmark through which the superiority of gradient boosting as the most suitable model for this specific problem can be justified, thus recommending a useful methodology to all other research studies concerning agricultural water resources modeling in sparse and heterogeneous geographical areas. What is novel about our study does not necessarily lie in having obtained excellent accuracy results, but rather in the fact that the feature importance ranking inside the classifier, which ranked precipitation, soil types, and intensity of cultivation as the most important features, is consistent with the principles of hydrology.

For the future, the pressing research issue would be the development of this static, state-level approach to a dynamic one that will consider changes in rainfall pattern on an annual basis as well as changing cropping practices, thereby making the model suitable for near-real-time detection of droughts. The missing link here is the inclusion of high resolution satellite imagery, including soil moisture information from the SMAP satellite and vegetation indices from MODIS, in order to sharpen the discrimination of the difficult Medium-High category border. The future direction of research would also entail moving away from global feature importance to instance-based explanations through the use of approaches such as Shapley Additive Explanations. More significantly, the practical applicability of the proposed model will ultimately be assessed based on how well it is accepted by agricultural extension workers and state-level policymakers. Thus, a transdisciplinary validation process, whereby the outputs of the proposed model are compared against expert local knowledge, appears to be the inevitable next step to take before this research framework can be turned into practice.

References

1. Lal M (2005) Climate change—implications for india’s water resources. Climate Change; Water Resources in South Asia
2. Rizvi S, Izhar S, Shahab A et al (2025) Comprehensive review on groundwater exploitation, depletion, and its management: Sustainable approach toward agricultural productivity in india. Environ Qual Manage

3. Kaur A (2025) Geography of india: Land, people, and resources. books google com
4. Gupta S, Singh S, Kanga S, Kumar P et al (2024) Unearthing india's soil moisture anomalies: Impact on agriculture and water resource strategies. Theoret Appl Climatol
5. Galkate R, Yadav S, Pandey R, Negm A et al (2022) An overview: Water resource management aspects in india. Water Qual Assess Manage
6. Tarate S, Patel N, Danodia A, Pokhariyal S et al (2024) Geospatial technology for sustainable agricultural water management in india—a systematic review. Geomatics
7. Saha S, Kundu B, Paul G, Pradhan B (2023) Proposing an ensemble machine learning based drought vulnerability index using M5P, dagging, random sub-space and rotation forest models. Stochastic Environmental Research and Risk Assessment
8. Sahana V, Mondal A, Sreekumar P (2021) Drought vulnerability and risk assessment in india: Sensitivity analysis and comparison of aggregation techniques. J Environ Manage
9. Pokhariyal S, Patel N, Govind A (2023) Machine learning-driven remote sensing applications for agriculture in india—a systematic review. Agronomy
10. Saha S, Gogoi P, Gayen A, Paul G (2021) Constructing the machine learning techniques based spatial drought vulnerability index in karnataka state of india. J Clean Prod
11. Virnodkar S, Pachghare V, Patil V, Jha S (2020) Remote sensing and machine learning for crop water stress determination in various crops: A critical review. Precision Agric
12. Yadav A, Narakala L, Upreti H et al (2024) Evaluation of crop water stress index of wheat by using machine learning models. Environ Monit Assess
13. Yang M et al (2021) Simulating canopy temperature using a random forest model to calculate the crop water stress index of chinese brassica. Agronomy
14. Kapari M, Sibanda M, Magidi J, Mabhaudhi T, Nhamo L et al (2024) Comparing machine learning algorithms for estimating the maize crop water stress index (CWSI) using UAV-acquired remotely sensed data in smallholder. … Drones
15. Saha S, Gayen A, Gogoi P, Kundu B, Paul G et al (2022) Proposing novel ensemble approach of particle swarm optimized and machine learning algorithms for drought vulnerability mapping in jharkhand, india. Geocarto Int
16. Saha S, Kundu B, Saha A, Mukherjee K et al (2023) … drought vulnerability index in monsoon climate dominant region of west bengal, india: Manifesting deep learning algorithms for developing drought vulnerability. …, Theoret Appl Climatol
17. Khosravi Y, Homayouni S, St-Hilaire A (2024) An integrated dryness index based on geographically weighted regression and satellite earth observations. Sci Total Environ
18. Wei W, Zhang X, Liu C, Xie B, Zhou J et al (2023) A new drought index and its application based on geographically weighted regression (GWR) model and multi-source remote sensing data. Environmental Science and Pollution Research

19. Haghighattalab A, Crain J, Mondal S, Rutkoski J et al (2017) Application of geographically weighted regression to improve grain yield prediction from unmanned aerial system imagery. Crop Sci
20. Jahan M, Nassiri-Mahallati M, Soltani M et al (2026) Machine learning-driven strategies for wheat yield prediction and climate adaptation in arid regions: A comparative analysis of random forest and XGBoost. Theoret Appl Climatol
21. Hukkeri G, Naganna S, Pruthviraja D, Bhat N et al (2023) Drought forecasting: Application of ensemble and advanced machine learning approaches. IEEE Access
22. Loh W (2011) Classification and regression trees, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery
23. Baral U, Saha U, Mukhopadhyay U et al (2023) Drought risk assessment on the eastern part of indian peninsula—a study on purulia district, west bengal. Environmental Monitoring and Assessment
24. Saha D, Kukal S (2015) Soil structural stability and water retention characteristics under different land uses of degraded lower himalayas of north-west india. Land degradation & development
25. Jariwala K, Agnihotri P (2023) Comparative analysis of drought modeling and forecasting using soft computing techniques. Water Resour Manage
26. Lazar A, Pokhrel P, Das S (2026) … comparative study of gradient boosting versus tabular deep learning and explainability techniques for mixed-type tabular data models using SHAP and LIME. Int J Artif Intell Tools