Review of Environment and Earth Sciences
2021 Vol. 8, No. 1, pp. 1-7
ISSN(e): 2313-8440
ISSN(p): 2409-2150
DOI: 10.18488/journal.80.2021.81.1.7

DROUGHT PREDICTION WITH RAW SATELLITE IMAGERY AND ENSEMBLE SUPERVISED MACHINE LEARNING

Owais Raza[1+]
Mohsin Memon[2]
Sania Bhatti[3]
Nazia Pathan[4]

[1,2,3,4]*Department of Software Engineering , Mehran UET, Pakistan.*
Email: owais.leghari@hotmail.com Tel: 923118147605
Email: mohsin.memon@faculty.muet.edu.pk
Email: sania.bhatti@faculty.muet.edu.pk
Email: naziapathan@muetkhp.edu.pk

(+ Corresponding author)

Article History
Received: 26 January 2021
Revised: 17 February 2021
Accepted: 5 March 2021
Published: 1 April 2021

Keywords
Drought
Machine learning
Ensemble machine learning
Satellite imagery
Boosting
Bagging.

ABSTRACT

Drought is one of the biggest challenges that environmentalists face today because of rapidly evolving climate. The negative impacts of drought on the economy, humans and other living organisms endure long after the ending of drought, and with time its intensity also increases. One way to fight the adverse effects of drought is to perform drought prediction and so that appropriate decisions can be made accordingly. Drought prediction can be made considering vegetation and water level in any region therefore, in this research we are using satellite images to predict drought conditions and its various stages, like if it is about to come or it has passed. All these predictions will be helpful for authorities to make informed decisions. We are employing supervised machine learning nevertheless to obtain the best results. We are using boosting and bagging which is ensemble supervised machine learning techniques. The experiments performed proved that bagging is better than boosting classifiers and it is less computationally expensive; and boosting on the other hand is less accurate and computationally expensive.

Contribution/Originality: This research performs drought prediction on tharparkar district using raw satellite imagery and ensemble machine learning techniques.

1. INTRODUCTION

Droughts are geological phenomena that typically begin with a precipitation deficit which can contribute to tremendous socio-economic losses[1]. As droughts have a direct association with water supply, their changing dynamics can have a significant effect on water distress, agricultural development and environmental sustainability due to climate change. Droughts have indeed caused serious economic damage and caused food shortage and water scarcity in Pakistan in the past [2]. Therefore, for early warning, planning and prevention, there is a need to reliably forecast drought conditions in order to mitigate their damaging consequences. Owing to the nature of their sources and spatiotemporal sizes at which they occur, the estimation of droughts has remained problematic for climatologists and environmental scientists [3]. In general, to forecast droughts mathematical, complex and hybrid models are used [4]. Empirical relationships between climate variables and drought indices experimental measurement are used in mathematical forecast models to forecast droughts [4]. Machine learning (ML) algorithms provide capability to learn and advance from previous data automatically without being programmed explicitly. In various hydro - climatic applications such as rainfall prediction, various ML algorithms are used to develop models that can simulate non-linear and linear correlations among predictor variables and predictions [5]. Many ML techniques such as ANN, Random Forest and SVM have been used to model the dynamic nonlinear relationships between drought indices (e.g. standard precipitation index, Enhanced Vegetation Index) and predictors in forecasting droughts [6]. In this research, we are targeting the Tharparkar district of Pakistan which is one of the most vulnerable region to drought in pakistan, drought is directly related to vegetation and water so with the help of raw satellite images, machine learning models are created to make prediction that whether the

picture shows drought condition, pre-drought condition, post drought condition or there is no drought. At any point in time the condition on land must belong to any of these four stages which shows that our model will be providing robust results as all possible classes for drought conditions are covered.

2. RELATED WORK

Many droughts have existed in various areas of the world in the recent years. For instance, the drought in East Africa (2010–2011), the drought in Texas (2012), the drought in the United States of Central Great Plains (2012) and the drought in California (2012−2015) [3] the drought in Australia (1997–2010) [7] the drought in Sahel (2012) and the drought in Pakistan in (1997–2003) [8]. These effects are far more severe in Pakistan, which is among the most vulnerable regions to drought given the enormous reliance on agriculture. In Pakistan, almost 43% of the national workforce is working in agriculture [9]. Prediction of drought is one of the most serious concerns of meteorologists so that various techniques are used to predict drought by different researchers. As in Khan, et al. [10] For the prediction of extreme drought areas considering non-rainfall, severe drought area prediction (SDAP) is suggested. In study [11] a well-known mathematical machine learning tool, Support Vector Machine (SVM), is used to predict seasonal variations of the Uniform Precipitation Index (SPI) in four reservoir basins that supply the water requirements of the capital city of Iran, Tehran for drought prediction. The research in Gomes, et al. [12] investigated the NDVI-LST relationship in a tropical environment in the Tietê River, State of São Paulo, Brazil, through the Vegetation Health Index (VHI) through satellite imagery to determine improvements in vegetation condition in two periods (2000 and 2014). According to research in Khan et al. (2020) it is stated that SVM is one of the common algorithms used for the prediction. However in our research we are not using compact supervised machine learning algorithms but we are using ensemble supervised machine learning i.e. boosting and bagging.

3. AREA OF STUDY

The region chosen for the study is the District of Tharparkar from Sindh Pakistan. Tharparkar is situated between the longitudes and latitudes of 69° 3′ 35″ E and 71° 7′ 47″ E, 24° 9′ 35″ N and 25° 43′ 6″ N. According to the 2017 census, the population of Tharparkar is 165,966,1. Except for a few agricultural areas, Tharparkar is largely desert. In the past, it has been the apex of several droughts, which is the main reason behind the choice of this area of research. In the Tharparkar district, there are 7 talukas, namely Chachro, Dali, Diplo, Islamkot, Kaloi, Mithi and Nagarparkar. Figure 1 shows the map of the study area. Figure 1 is created using GIS for the purpose of this research.

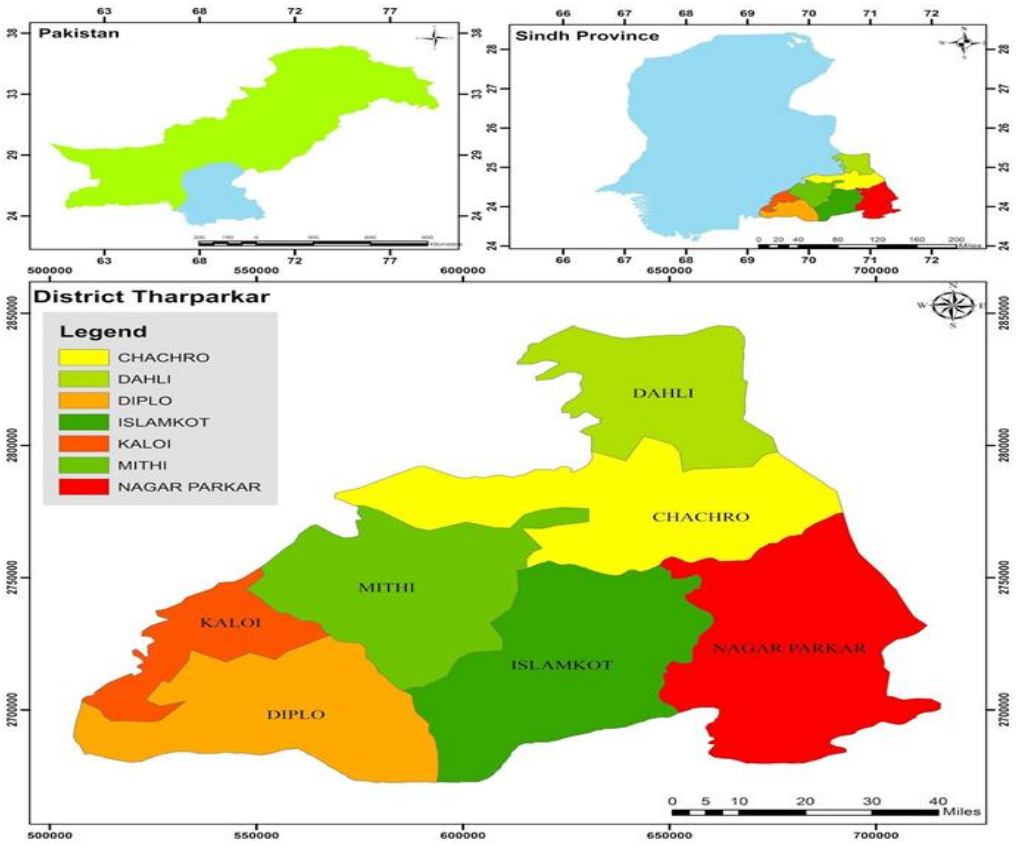

Figure-1. Study area map of Tharparkar.

4. DATASET

Dataset used in this research consists of satellite images which were taken from all 7 talukas of tharparkar district and labeled as pre-drought, drought, post drought and no drought. The Dataset consists of pictures from 2002 to 2020. This dataset is gathered using google earth pro. The Figure 2 shows the frequency of images for each class, there are 323, 432, 53 and 276 images for class no drought, drought, after drought and before drought respectively.

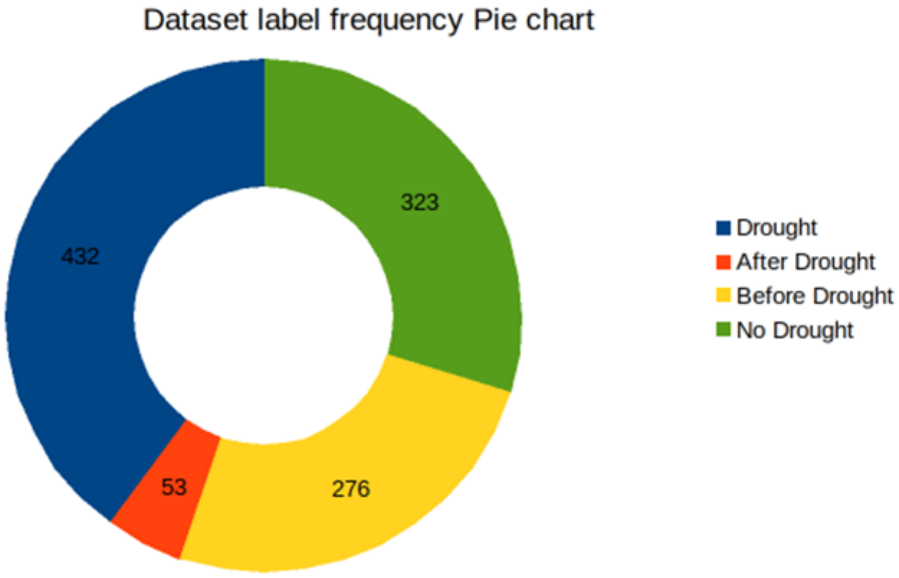

Figure-2. Dataset label frequency Pie chart.

5. METHODOLOGY

Ensemble approaches are designed to enhance the predictive efficiency of a given mathematical learning or model fitting methodology. Rather than using a singular fit method, the basic theory of ensemble methods is to construct a linear combination of certain model fitting methods.

5.1. Bagging Ensemble Technique

With bagging Yaman and Subasi [13] the goal is to fit multiple separate models and find the average of respective predictions in order to develop a model which has lower variance. In bagging, it builds several bootstrap samples such that each new bootstrap sample can act as the other independent dataset extracted from the true distribution. For our research, we have created a bagging classifier from SVM learners, we have iterated the process of bootstrapping samples 10 times. The pseudo code for bagging ensemble learning for our model is given below:

Input: Dataset D, Weak Learner L, Number of Iterations K
Process:
for i=1 to i=K
Di= bootstrap sample of dataset from D
Li=Train Di with L
end for
Output: Learner at the end of K iterations Le

5.2. Boosting Ensemble Learning

The boosting Yaman and Subasi [13] ensemble technique is used to integrate multiple models to understand the dataset by attempting to find models that complement each other by learning from the error of previous models. Boosting consists of sequential fitting of various weak learner in such a way that each model learning from the observation which weren't handled correctly., Eventually we have will have model that not only worked well on easy observation but also the observation which we were difficult to make sense of this, which will result in a strong learner. In this research adaboost boosting technique is used to make the boosting model, adaboost is also called adaptive boosting. In adaptive boosting, the resulting model is good at problems which are difficult to optimize and the adaptive boosting model is the result of constant optimization of weak learners in an iterative way to get the best fit model for both easy and difficult observations. The pseudo code for boosting ensemble learning is shown below:

Input: Dataset D, Weak Learner L, Number of Iterations K
Process:
for i=1 to i=K

Di= bootstrap sample of dataset from D
Li=Train Di with L
ei=Error in Li
Di+1=Adjusting the bootstrapping for error ei
end for
Output: Learner at the end of K iterations Le

In this study two types of models are created one using bagging ensemble learning and other using boosting ensemble learning and the comparison results of both models are presented. Figure 3 represents the steps taken to implement the workflow. Each of the step from Figure 3 is discussed as following:

5.2.1. Dataset Collection

The first step in solving any machine learning problem is to have a relevant dataset, in this case the dataset is created using google earth pro. The dataset contains the images of 7 talukas of tharparkar district from 2002 to 2020. To apply supervised machine learning we need to have a labeled dataset which is discussed in the next step

5.2.2. Dataset Labeling

Each of the images was labeled into class after drought, before drought, drought and no drought. The label of image depends on the region and time of capture of image, for instance, if image is captured during the time of drought; it is labeled as drought, if picture is captured before drought it is labeled as before drought and if picture is taken after the drought it is labeled as after drought, if it is taken during normal condition under no drought then it is labeled as no drought. All the images taken may vary in size and may not be ready to fed to algorithm, therefore in next step our goal is to make image ingestible by machine learning algorithm

5.2.4. Image Manipulation

Now we have labeled the dataset, but it is not ready to be fed to the algorithm. So to make it compatible with the algorithm all the images are altered to occupy a uniform size of 200 x 200 pixels. Then pixel matrices are flattened into arrays to achieve vectors of features. Now the data is ready to be trained by machine learning algorithms.

5.2.5. Splitting the Dataset

We are splitting the dataset into a train and test set to evaluate the performance of both models. We are using 75%-25% split, which means 75% of data is used training and 25% is used during the testing of the model.

5.2.6. Apply Boosting Ensemble

In this step we are applying the boosting ensemble learning algorithm to create a model

5.2.7. Apply Bagging Ensemble

In this step we are applying the bagging ensemble learning algorithm to create a model

5.2.8. Evaluate Models

Both models are individually evaluated based on accuracy, precision, recall, f1 score, confusion matrix and learning curves.

5.2.9. Compare Both Models

After evaluating both models they are compared based on those metrics.

6. RESULTS

During implementation python is used, the libraries used for this purpose are opencv and sklearn. using a virtual machine. The cloud provider for VM used is microsoft Azure, and in VM data science windows image is used. VM supports32 gbs of ram and 128 gb of permanent memory.

In this research the models are being evaluated using following parameters:

1. Accuracy
2. Precision
3. Recall
4. F1 Score

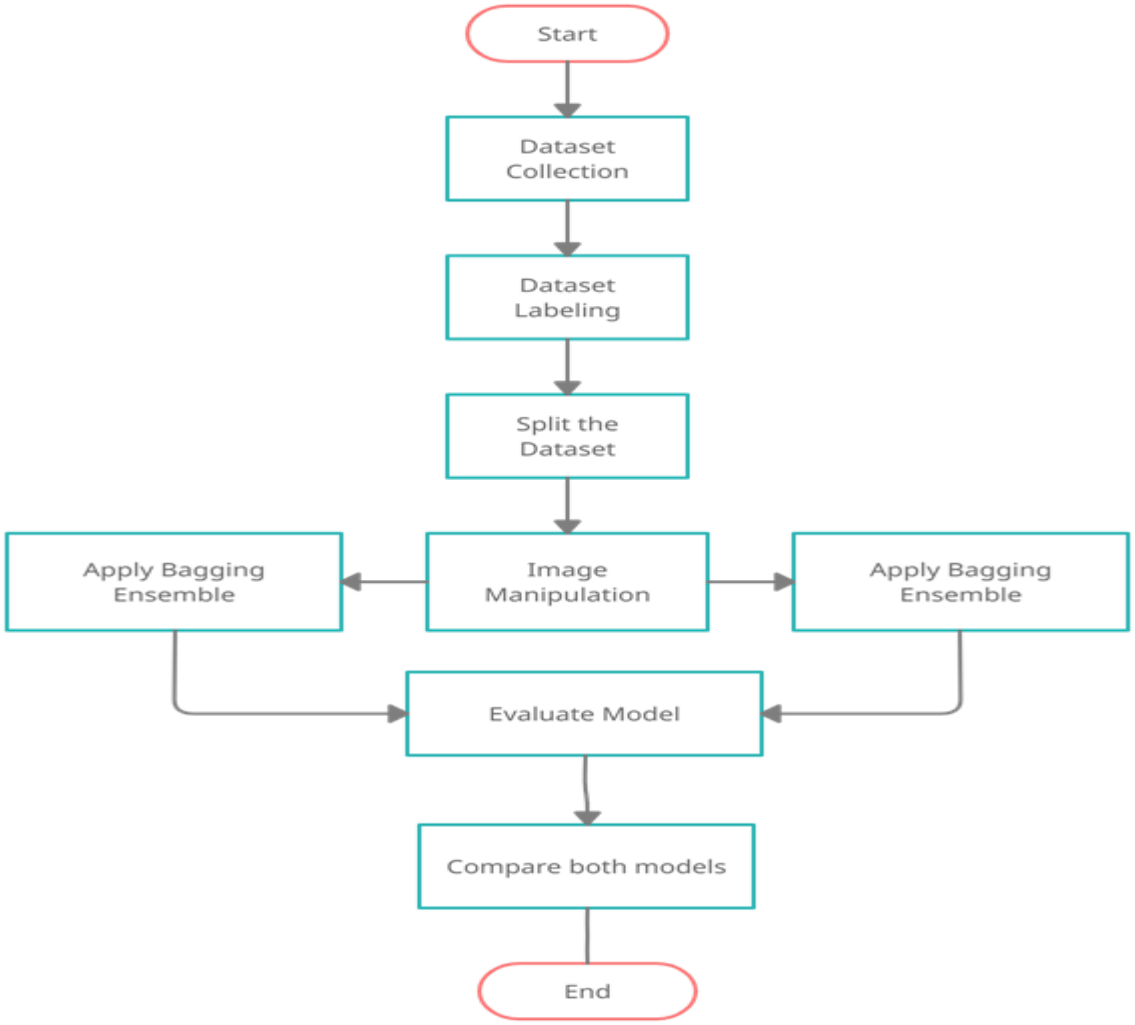

Figure-3. Methodology flowchart.

6.1. Accuracy

Accuracy has been the most widely used classification performance measure. Accuracy is measured among all the labels that algorithm predicted; how many were really correctly predicted. This is the ratio of labels accurately predicted and the amount of original labels. It is represented by Equation 1.

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \qquad (1)Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \qquad (1)$$

6.2. Precision

Precision is also one of the classification evaluation metrics. It represents that out of all the observations predicted as positively labeled how many are actually positive. The precision is represented mathematically by Equation 2.

$$Precision = \frac{Tp}{Tp + Fp} \qquad (2)$$

6.3. F1 Score

The harmonic mean of Recall and Precision is called F-1 score, the range of F1-Score is between 0 and 1. 1 represents the absolute best model and 0 represents the worst model [14]. In some cases precision is important, in others TPR is crucial so when we need to overall understand the accuracy of model F1-Score is one of the best measurements for classification models. It is represented by Equation 3 that shows that F1 score is calculated based on precision and recall. It is twice the product of precision recall divided by sum of precision and recall as shown in Equation 3

$$F1 - Score = 2\frac{Precision * Recall}{Precision + Recall} \qquad (3)$$

6.4. AUC

Area under the ROC curve is the combined performance measurement of the classifier for all the possible classification thresholds. The value of AUC falls between 0 and 1, 1 means all the predictions made by the model are correct and 0 means all the predictions made by the model are wrong [15].

Based on the definition and equations mentioned; accuracy, precision and recall is calculated for the both bagging and boosting technique in . It is observable from the Table 1 that bagging algorithm accuracy, precision and recall are approximately 27% more than boosting algorithms.

Table-1. Evaluation Parameter.

Technique	Accuracy	Precision	F1 Score
Bagging	76.75	75.03	74.38
Boosting	50.10	48.22	48.94

Another way of looking at the model for the purpose of understanding is ROC AUC i.e Area under the curve of ROC which is given in Figure 4. It can be seen from the bar graph that the bagging algorithm is posing a higher value of ROC than boosting algorithm. Thus RUC proves that bagging algorithm is better than boosting algorithm.

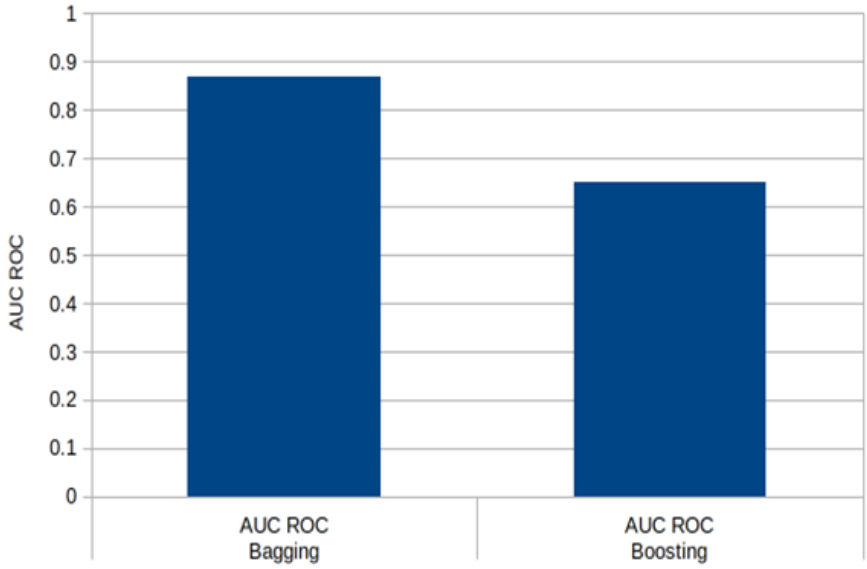

Figure-4. AUC ROC Bar chart of Boosting and Bagging.

7. CONCLUSION

Drought is one of the biggest challenges which environmentalists are facing today and it affects living beings and the economy severely so prediction and early detection of drought is of real importance. From the related work it can be seen that various researchers have used different features and techniques to make drought prediction, but in this research we have used raw satellite images to make drought prediction. We have used two ensemble techniques for classification; bagging and boosting. We have used accuracy, precision and F1 Score as measure of performance for the models. We have also calculated AUC for ROC for both models and plotted ROC. It is apparent from the comparative analysis that bagging performed better than boosting. This model is just the start of research regarding drought using only raw satellite images, in future models for drought prediction will be created by coupling images and other drought, vegetation and water indexes for Tharparkar for better results so that the government and concerned organizations can use it for the prediction of drought and precious lives and resources can be preserved.

Funding: This research is funded in terms of cloud credit by Microsoft's AI for Earth Initiative.
Competing Interests: The authors declare that they have no competing interests.
Acknowledgement: All authors contributed equally to the conception and design of the study.

REFERENCES

[1] H. West, N. Quinn, and M. Horswell, "Remote sensing for drought monitoring & impact assessment: Progress, past challenges and future opportunities," *Remote Sensing of Environment*, vol. 232, p. 111291, 2019. Available at: https://doi.org/10.1016/j.rse.2019.111291.

[2] S. Adnan, K. Ullah, L. Shuanglin, S. Gao, A. H. Khan, and R. Mahmood, "Comparison of various drought indices to monitor drought status in Pakistan," *Climate Dynamics*, vol. 51, pp. 1885-1899, 2018. Available at: https://doi.org/10.1007/s00382-017-3987-0.

[3] Z. Hao, V. P. Singh, and Y. Xia, "Seasonal drought prediction: advances, challenges, and future prospects," *Reviews of Geophysics*, vol. 56, pp. 108-141, 2018. Available at: https://doi.org/10.1002/2016rg000549.

[4] A. Mariotti, S. Schubert, K. Mo, C. Peters-Lidard, A. Wood, R. Pulwarty, J. Huang, and D. Barrie, "Advancing drought understanding, monitoring, and prediction," *Bulletin of the American Meteorological Society*, vol. 94, pp. ES186-ES188, 2013. Available at: https://doi.org/10.1175/bams-d-12-00248.1.

[5] A. Parmar, M. Kinjal, and S. Mithila, "Machine learning techniques for rainfall prediction: A review," presented at the International Conference on Innovations in information Embedded and Communication Systems, 2017.

[6] J. F. Santos, M. M. Portela, and I. Pulido-Calvo, "Spring drought prediction based on winter NAO and global SST in Portugal," *Hydrological Processes*, vol. 28, pp. 1009-1024, 2014. Available at: https://doi.org/10.1002/hyp.9641.

[7] A. I. Van Dijk, H. E. Beck, R. S. Crosbie, R. A. de Jeu, Y. Y. Liu, G. M. Podger, B. Timbal, and N. R. Viney, "The Millennium Drought in Southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society," *Water Resources Research*, vol. 49, pp. 1040-1057, 2013. Available at: https://doi.org/10.1002/wrcr.20123.

[8] I. H. Durrani, A. Shahzada, and M. A. Syed, "Historical and future climatological drought projections over Quetta Valley, Balochistan, Pakistan," presented at the IOP Conference Series: Materials Science and Engineering, 2018.

[9] A. Ullah, D. Khan, and S. Zheng, "Testing long-run relationship between agricultural gross domestic product and fruits production: Evidence from Pakistan," *Ciência Rural*, vol. 48, pp. 1-12, 2018.

[10] N. Khan, D. Sachindra, S. Shahid, K. Ahmed, M. S. Shiru, and N. Nawaz, "Prediction of droughts over Pakistan using machine learning algorithms," *Advances in Water Resources*, vol. 139, p. 103562, 2020. Available at: https://doi.org/10.1016/j.advwatres.2020.103562.

[11] S. Nikbakht, AliReza, Z. Banafsheh, and N. Mohsen, "Seasonal meteorological drought prediction using support vector machine," *Journal of Water and Wastewater; Ab va Fazilab*, vol. 23, pp. 73-85, 2012.

[12] A. C. C. Gomes, B. Nariane, and A. Enner, "Accessing the southeastern Brazil 2014 drought severity on the vegetation health by satellite image," *Natural Hazards*, vol. 89, pp. 1401-1420, 2017. Available at: https://doi.org/10.1007/s11069-017-3029-6.

[13] E. Yaman and A. Subasi, "Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification," *BioMed Research International*, vol. 2019, pp. 1-13, 2019.

[14] Z. DeVries, E. Locke, M. Hoda, D. Moravek, K. Phan, A. Stratton, S. Kingwell, E. Wai, and P. Phan, "Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability," *The Spine Journal*, vol. 2021, pp. 1-8, 2021.

[15] A. C. J. Janssens and F. K. Martens, "Reflection on modern methods: Revisiting the area under the ROC curve," *International Journal of Epidemiology*, vol. 49, pp. 1397-1403, 2020. Available at: https://doi.org/10.1093/ije/dyz274.