CORPUS PUBLISHERS

Environmental Sciences and Ecology: Current Research (ESECR)

ISSN: 2833-0811

Volume 6 Issue 1, 2025

Article Information

Received date : February 24, 2025

Published date: February 28, 2025

*Corresponding author

Muhammad Imran, College of Hydrology and Water Resources, Hohai University, Nanjing 210098, P.R. China

Keywords: Climate change, Precipitation prediction, KNN, SVM, RF, Upper Indus Basin

Abbreviations: IPCC: Intergovernmental Panel on Climate Change; SVM: Support Vector Machine; KNN: K-Nearest Neighbor; RF: Random Forest; UIB: Upper Indus Basin; PMD: Pakistan Meteorological Department

DOI: 10.54026/ESECR/10108

Research Article

Assessing Machine Learning Models for Precipitation Prediction in the Upper Indus Basin: A Comparative Analysis

Muhammad Imran[1]*, Nur E Jannat Mishu[2], Hamza Khaliq[3] and Faiza Shahzad[3]

[1]*College of Hydrology and Water Resources, Hohai University, China*

[2]*College of Information Science and Engineering, Hohai University, China*

[3]*College of Environment, Hohai University, China*

Abstract

Precipitation plays a critical role in the effective management of water resources and the maintenance of reservoir water levels. However, climate change has significantly altered precipitation patterns, leading to extreme hydrological events such as droughts and floods, which have profound socioeconomic and environmental impacts. This study focuses on predicting precipitation events in the Upper Indus Basin (UIB) using machine learning models. In this study three widely used machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) were employed to forecast precipitation events in the UIB. The dataset was divided into training (80%) and testing (20%) subsets for model evaluation. Among the algorithms tested, KNN demonstrated the best predictive performance, yielding a mean absolute error (MAE) of 2.662, a root means square error (RMSE) of 16.3, and an R^2 score of 0.879, with an overall accuracy of 83.16%. The results indicate that the KNN algorithm is the most effective machine-learning model for precipitation prediction in the UIB. The findings of this study contribute to improving early warning systems and facilitating efficient water resource management in the face of climate variability and extreme weather events.

Introduction

Many Asian countries, including Pakistan, heavily rely on agriculture, with a large share of Pakistan's GDP rooted in this sector [1]. Pakistan's agricultural productivity is heavily dependent on irrigation, primarily sourced from the Upper Indus Basin [2]. However, climate change poses a serious threat to the hydrological cycle, impacting both water availability and precipitation patterns. According to the Intergovernmental Panel on Climate Change (IPCC), global temperatures have risen by approximately 0.72°C between 1951 and 2012, with projections indicating a further increase of 1°C to 3°C by 2050 and 2°C to 5°C by 2100, depending on greenhouse gas emission scenarios (IPCC, 2013). Precipitation, a fundamental component of the Earth's climate system and a crucial resource for agriculture and water management is becoming increasingly unpredictable due to climate change [3]. Variability in precipitation patterns can lead to extreme hydrological events such as droughts and floods, which directly affect food security, livelihoods, and water resource management. In this context, the development of accurate precipitation prediction models is essential for mitigating the adverse impacts of climate change and ensuring sustainable water resource planning. Recent advancements in machine learning have provided promising solutions for improving precipitation forecasting accuracy. Machine learning models can effectively analyze complex climate patterns and enhance predictive capabilities by integrating large-scale precipitation datasets [4]. In this study, three widely used machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) are employed to predict precipitation events in the UIB [5]. By leveraging these algorithms, this study aims to improve precipitation forecasting, which is critical for informed decision-making in agriculture and water resource management.

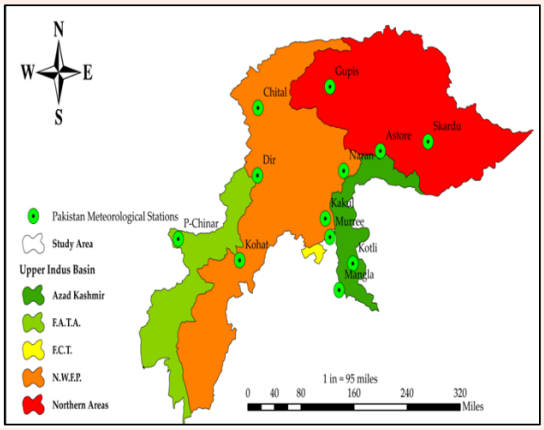
Materials and Methods

The methodology for this study involves the collection, preprocessing, and analysis of meteorological data using machine learning techniques to predict precipitation events in the Upper Indus Basin (UIB). The following steps outline the process:

Study area

This study was conducted in the Upper Indus Basin (UIB), Pakistan, which extends between 32°25' to 37°40' N latitude and 70°55' to 81°35' E longitude. The UIB, with a total catchment area of 164,867 km^2, plays a crucial role in regulating water resources for agricultural and hydropower needs in the region. The basin's topography was analyzed using the Shuttle Radar Topography Mission Digital Elevation Model (SRTM DEM), providing detailed insights into elevation variations and hydrological characteristics (Ahmad et al., 2024). The spatial extent of the study area is illustrated in Figure 1.

Figure 1: Study Area-Upper Indus Basin-Pakistan.



Data Collection

This study utilized meteorological data comprising various climatic parameters as input variables, including minimum and maximum temperatures and precipitation, while rainfall was considered as output variables. The dataset was obtained from multiple meteorological stations under the Pakistan Meteorological Department (PMD), as detailed in Table 1. The collected dataset spans 30 years (1991–2020) and consists of 1,800 records across three variables, providing a comprehensive temporal representation of climate trends in the Upper Indus Basin (UIB).

Table 1: Different Climatological Stations.

Serial #	Station Name	Extreme Temperature (°C)	Least Temperature (°C)	Avg. Precipitation (mm)	Rainfall (Output)
1	Kotli	49.8	38	210.57	
2	Murree	34.5	24	255	Rain
3	Mangla	49.9	37.8	195.6	or
4	Narran	45.8	33.7	177.8	No Rain
5	Dir	46	29	166	

Data Preprocessing

The dependent variable for this study is rainfall, which was classified into two categories: True (1) for rain days and False (0) for no-rain days. Various machine learning algorithms, including K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM), were employed to construct predictive models. The dataset was divided into two subsets: 80% for training and 20% for testing. The training set is used to teach the model by optimizing weights and biases through labeled examples. The model is trained to minimize loss by analyzing the data and finding the optimal parameters. The test set is independent of the training data but follows the same probability distribution; it is used to evaluate the model's performance and estimate the generalization error [6].

Hyperparameter Optimization: Grid Search CV

The algorithms' parameters were optimized using Grid Search CV, a method in the sklearn library. It automates the process of fitting models to the training data by searching through predefined hyperparameters [7]. Grid Search CV evaluates each combination of hyperparameters using cross-validation and selects the best-performing set to optimize model accuracy. From the observation of the dataset, the algorithms easily find out the best correction factor by using Grid Search CV for precipitation predictions in the future as shown in the given below Table 2.

Table 2: Best Hyperparameters of ML Models.

Model	C-Value	Gamma	Kernel
SVM	1000	1	RBF
KNN	K-Neighbor Values		
	15		
RF	max_features		n_estimators
	Auto		100

Methods

In this study, three machine learning algorithms such as K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) were employed to predict precipitation events in the Upper Indus Basin (UIB). Below is a detailed explanation of each algorithm:

K-Nearest Neighbor (KNN): K-Nearest Neighbor (KNN) is a supervised learning algorithm used for classification. It operates by identifying the K nearest neighbors to a given data point in the feature space and assigning the most frequent class label (rain or no-rain) among these neighbors [8]. The distance between data points is calculated using metrics like Euclidean distance. KNN does not explicitly learn a model during training but stores the entire training dataset and performs predictions based on proximity to labeled examples.

Random Forest (RF): Random Forest (RF) is an ensemble learning method that constructs a collection of decision trees and combines their predictions. Each tree is trained on a random subset of the data, and predictions are made by aggregating the individual outputs of all trees, using majority voting for classification [9]. The key parameters in Random Forest include the number of trees and the maximum depth of trees. By averaging the results from multiple trees, RF reduces the risk of overfitting and increases predictive accuracy.

Support Vector Machine (SVM): Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification tasks [10]. SVM works by finding the optimal hyperplane that separates data points belonging to different classes (rain or no rain). It aims to maximize the margin between the classes and uses support vectors (data points closest to the hyperplane) to define the optimal boundary. SVM can handle both linear and non-linear classification problems by applying different kernel functions (e.g., linear, polynomial, or radial basis function (RBF)).

Model Evaluation

Each of the above algorithms was trained using the 80% training data, and their performance was evaluated on the 20% test data. The models were compared based on accuracy, precision, recall, and F1-score to identify the most effective model for predicting precipitation events in the UIB. The flow of the models is shown in Figure 2.

Results and Discussion

Figure 3 presents a comparison of the precipitation predictions made by three machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF). The performance of each model was evaluated based on its ability to predict both no precipitation (no-rain) and precipitation (rain) day.

1. The SVM model accurately predicted 586 no-rain days and 277 rain days, but it made one false rain prediction, where the model incorrectly forecasted rain on a day when no precipitation occurred. This indicates that SVM demonstrated good predictive capability but with a slight margin of error in terms of false positives.
2. The KNN model was able to correctly predict 578 no-rain days and 263 rain days, but it had a higher number of errors in comparison to SVM. Specifically, the KNN model made 15 false rain predictions (false positives), where it predicted rain on days that were dry and missed 8 rain events (false negatives), where it failed to predict precipitation on days that experienced rainfall. Despite these errors, KNN's performance was relatively strong overall.
3. The RF model showed excellent prediction of 586 no-rain days and 278 rain days, with a very small margin of error. However, it still exhibited a slightly lower performance compared to KNN, especially in correctly identifying the rain events, as indicated by the total number of accurate predictions.

In terms of overall accuracy, the KNN model outperformed the other two models, achieving an accuracy rate of 83.18%, as shown in Figure 4. The SVM model scored an accuracy of 46.74%, while the RF model achieved 42.13% accuracy. These results indicate that KNN was the most effective machine learning algorithm for predicting precipitation events in this study, offering the highest precision in identifying both no-rain and rain days and providing a more reliable framework for precipitation prediction.

Figure 2: Flow chart of ML Models.

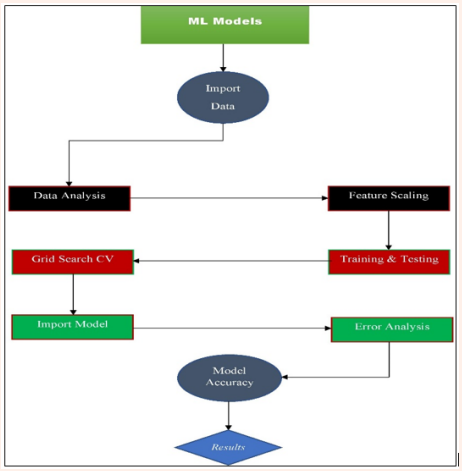


Figure 3: Comparison of the precipitation predictions.

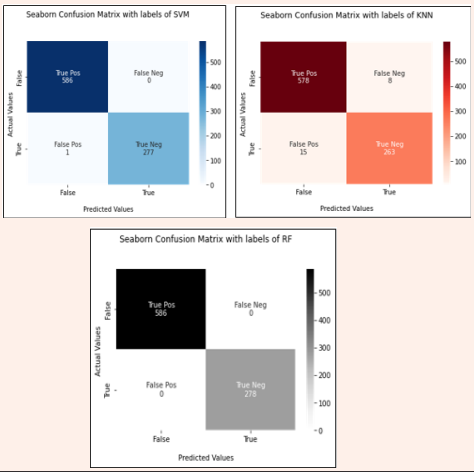
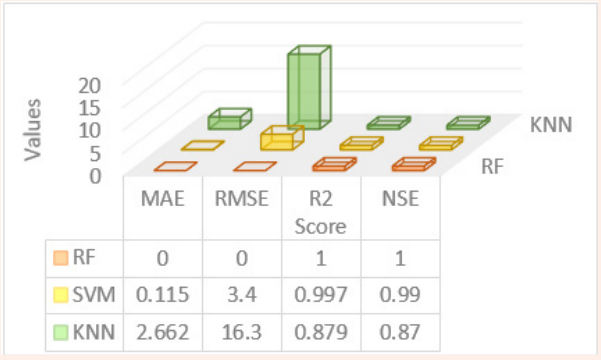
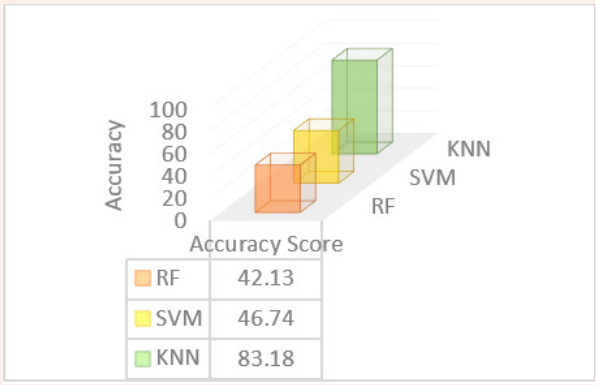


Figure 4: Performance Evaluation Metrics and Accuracy of ML models.



	MAE	RMSE	R2 Score	NSE
RF	0	0	1	1
SVM	0.115	3.4	0.997	0.99
KNN	2.662	16.3	0.879	0.87



Accuracy Score	
RF	42.13
SVM	46.74
KNN	83.18

Conclusion

Pakistan, as one of the countries most vulnerable to climate change, faces significant challenges related to climate-induced disasters and water stress. Its water resources, crucial for agriculture, industry, and daily life, are particularly susceptible to the impacts of climate variation. In this study, we examined the potential effects of climate change, particularly changes in precipitation patterns, on water reserves in the Upper Indus Basin (UIB) using machine learning techniques. The ability to accurately predict precipitation events in this region is essential for managing water resources, mitigating flood and drought risks, and ensuring sustainable agricultural practices.

Among the machine learning algorithms tested in this study-Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF), the KNN algorithm was found to be the most effective in predicting precipitation events, with an accuracy of 83.18%. The KNN model demonstrated the best performance in correctly predicting both no-rain and rain days, outperforming both SVM and RF models, which achieved significantly lower accuracies of 46.74% and 42.13%, respectively. The KNN model's superior performance highlights its suitability for precipitation forecasting, making it an excellent choice for enhancing early warning systems and facilitating water resource management in regions susceptible to climate variability.

The findings of this study have important implications for water resource management in Pakistan, where effective prediction of precipitation events is critical to ensuring adequate water availability and mitigating the adverse impacts of climate extremes such as floods and droughts. The KNN algorithm, with its high accuracy,

can assist policymakers and water management authorities in making more informed decisions regarding the management of water reserves, irrigation systems, and flood control measures.

While KNN performed well in this study, it is important to acknowledge the limitations of the current approach. For instance, the accuracy of predictions may vary depending on the quality and resolution of the meteorological data used. Additionally, incorporating more variables, such as wind speed, humidity, and historical climatic trends, could further improve model performance. Future research could also explore the integration of ensemble methods or deep learning models to enhance prediction accuracy and robustness. In conclusion, this study demonstrates the potential of machine learning algorithms, particularly KNN, for predicting precipitation events in the Upper Indus Basin, and offers valuable insights for better water resource management in Pakistan. As climate change continues to affect precipitation patterns, these predictive models can play a crucial role in adapting to changing climate conditions and ensuring sustainable water management in the future.

References

1. Baocheng H, Jamil A, Bellaoulah M, Mukhtar A, and Clauvis NK (2024) "Impact of climate change on water scarcity in Pakistan. Implications for water management and policy". J Water Clim Chang 15(8): 3602-3623.
2. Khan MZ, Abbas H and Khalid A (2022) "Climate vulnerability of irrigation systems in the Upper Indus Basin: insights from three Karakoram villages in northern Pakistan". Clim Dev 14(6): Ppp. 499-511.
3. Zhu H, Chen K, Chai H, Ye Y and Liu W (2024) "Characterizing extreme drought and wetness in Guangdong, China using global navigation satellite system and precipitation data". Satell Navig 5(1): P. 1.
4. Salcedo-Sanz S (2024) "Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: a review". Theor Appl Climatol 155(1): Pp. 1-44.
5. Shabani S (2020) "Modeling Pan Evaporation Using Gaussian Process Regression K-Nearest Neighbors Random Forest and Support Vector Machines; Comparative Analysis".
6. Dobbin KK and Simon RM (2011) "Optimally splitting cases for training and testing high dimensional classifiers". BMC Med Genomics 4(1): Pp. 31.
7. Gairola S, Bhatt R, Gopal L, Garg N, Singh S, et al. (2023) "Enhancing Fertilizer Prediction: A Comprehensive Analysis with Grid Search CV and Multiple Machine Learning Algorithms". International Conference on Sustainable Communication Networks and Application (ICSCNA). Pppp. 1346-1352.
8. Zhang S, Zong XLiM, Zhu X and Wang R (2018) "Efficient kNN Classification with Different Numbers of Nearest Neighbors". IEEE Trans Neural Networks Learn Syst 29(5): Pppp. 1774-1785.
9. Wei W (2022) "Seasonal prediction of summer extreme precipitation over the Yangtze River based on random forest". Weather Clim Extrem 37 P. 100477.
10. Du J, Liu Y, Yu Y, and Yan W (2017) "A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms".