Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 02
2025

Research Article
https://doi.org/10.56979/902/2025

Performance Evaluation of Machine Learning and Deep Learning Models for Rainfall Classification Using Climatic Datasets of Pakistan

Hira Farman[1,3], Arif Hussain[1], Jahangir Baig[2] , Sharaf Hussain[3], Alisha Farman[3], and Qurat-ul-ain Mastoi[4]

[1]Department of Computer Science, Karachi Institute of Economics and Technology, Pakistan.
[2]Department of statistic, University of Karachi, Pakistan.
[3]Department of Computer Science, Iqra University, Karachi, Pakistan.
[4]School of computer Science and Creative technologies, University of the west of England Bristol, United Kingdom.
[*]Corresponding Author: Hira Farman. Email: hira.farman@iqra.edu.pk

Received: June 23, 2025 **Accepted:** August 28, 2025

Abstract: This research seeks to improve rainfall forecasting, which is essential to agricultural operations, water management, and disaster preparedness, especially in floods and droughts. Proper forecasting of rainfall is critical in the sustainable development by preventing the effects of extreme weather conditions like flooding that may cause loss of life, health problems, and economic disturbances. Nevertheless, because of the unpredictable character of rainfall, the traditional forecasting models have proven to be a great problem since in most cases they lack the ability to understand the complicated set of interactions that determine the formation of meteorological patterns. To solve this, the research uses different machine learning (ML) algorithms, such as Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), AdaBoost, Gradient Boosting, K-Nearest Neighbors (KNN), and Naive Bayes (NB) to offer better predictions in Karachi, Pakistan. The dataset used is the one that was provided by Visual Crossing and had 33 weather-related variables, including temperature, humidity, the speed of wind, and air pressure, and 4,778 observations between 2011 and 2023. A thorough process of data preprocessing such as data cleaning, transformation, and selecting features was applied prior to the division of dataset into a training set and a test set. The model was evaluated using the 5-fold cross-validation and the performance was assessed as precision, recall, accuracy and ROC curves. Random Forest has proved to be the most accurate with 99 percent of them and Naive Bayes has reported the overfitting nature of all models. AdaBoost and Gradient Boosting had a similar performance whereby both dealt with the problem of overfitting. Moreover, a deep learning network (BiLSTM) was used to identify temporal correlations in the sequence of rainfall which also demonstrated a test accuracy of 99.8 that confirms the reliability of deep learning models in addition to conventional ML models. The results show that machine learning, as well as deep learning algorithms, can learn and comprehend complex climate patterns and can considerably improve the accuracy of the weather predictions. These models can be utilized to make more informed decisions using the data on climate resilience, disaster preparedness, and sustainable environmental management.

Keywords: Forecasting; Machine Learning; Classification

1. Introduction

Rain, snow and or sleet are capable of having an impact on activity, as these are forms of precipitation that affect many activities that happen outdoors. It is very important that rainfall be predicted to the closest best to enable those in various fields make the right decisions of whether or not to engage in certain activities. Flood disasters resulting from heavy rainfall events have significant societal and economic impacts [2]. There have been several devastating flash flood disasters in Asia and the Pacific, both in urban and rural areas. A number of the countries affected by the recent floods that have attracted public notice are Pakistan, China, India, Indonesia, Mongolia, and Nepal. The biggest flooding in Pakistan's history occurred in 2022, when the country was completely submerged under water. The (United Nations International Children's Emergency Fund) UNICEF reported the floods occurred in 2022 that damaged more than 84 districts. 5 million children. Among an estimated 20 million people needing emergency aid now as a consequence of the heavy flood occurrences, there are approximately 10 million children. Due to the tremendous levels of flooding, millions of people battle for their lives as well as illnesses like cholera, dengue, typhoid, malaria, and acute diarrhea carried on by stagnant and unclean water. Besides, human lives are lost, with significant implications on infrastructures in terms of property, loss of agricultural infrastructure [7, 32, 44], and political stability. Therefore, in order to optimize resource management, increase readiness, and make wise judgments amid severe weather conditions, accurate and dependable rainfall forecasting is necessary [3-4]. Effective disaster monitoring and management is a worldwide concern. One of the most challenging scientific and technological challenges of the past century has been global rainfall prediction [5, 6, and 8]. Among the ML models RF, MLR, and XG Boost were also designed and tested by [9] using the same environmental data to calculate rainfall amounts that would probably fall during a particular day. A deep neural network in this work [13] which explores the opportunities for flooding, it was taken into consideration multiple different factors such as rain volume and temperature. The wavelet theory with MTLNN, in this model, are proposed as an efficient prediction method for rainfalls [14]. AI-based ensemble technique [26] proved further the accuracy of the forecast than the conventional method. Stated experimental results [37] pointed out that the hybrid models were more suited to create generalization with less errors and less computation costs in prediction of rainfall; also they converge to targeted values faster compared to single model. By implementing Random Forest (RF), Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Multivariable Adaptive Regression Splines (MARS), the average weekly and daily rainfall at Rachini station [38] was forecasted

Data mining techniques [40] have proven more effective than traditional statistical models for rainfall forecasting, as shown in studies using climatic data from southern Saudi Arabia. A study on the West African coast [45] (1981-2015) emphasized the significant role of intense rainfall frequency and inadequate distribution in causing severe floods, which affect the ecosystem, agriculture and human settlements. Precise forecasting of rainfall is therefore an essential aspect of early warnings of floods as well as disaster preparedness. Artificial Intelligence (AI) has become an effective instrument that can enhance the accuracy and effectiveness of rainfall prediction. The paper combines several hybrid AI methods, such as the use of neural networks with optimization algorithms such as ACO, PSO and MRA to come up with superior models on both short-term and long-term rainfall forecasting [55]. More sophisticated systems, like streamlined flood prediction systems [63, 64], and predictors, such as the logistic regression and SVM [77], improve predictive accuracy. On the whole, the hybrid ML-DL approach will enhance the severity of the initial warning systems, which will aid the community and decision-makers in reducing the effects of floods in (Fig. 1).

Floods are also one of the most devastating natural disasters that lead to loss of lives, destruction of infrastructure and short term and long term economic and environmental effects. The rise in water is very high and can cause deaths and massive displacement as experienced in China, Mekong Delta Vietnam and other flood-prone areas [7883]. Diseases like cholera and dengue, water contamination, and poisonous pollution are other diseases that are spread as a result of floodwaters that pose serious threats to the health of the population [79, 81]. Floods interfere with industries, agriculture, and transport networks and cause permanent financial and ecological losses. Pakistan experiences extreme flooding during the monsoon season, which occurs between June and September, and is a consequence of wet winds blowing through the Arabian Sea. Accurate rainfall prediction models are thus important in preparing in time against floods, proper resource planning and reducing the socio-economic and environmental impacts.

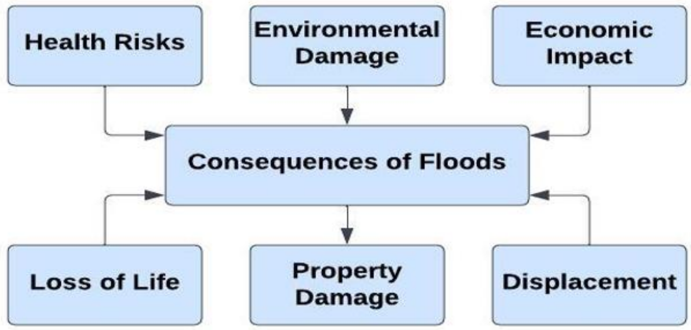


Figure 1. Consequences of Flood

Pakistan is one such place which receives heavy rainfall and floods take the lives of many people in this country every monsoon season. Research and efforts in rainfall prediction of Pakistan have continued to be done. There has been an active effort by research scientists and organizations dealing with meteorology in enhancing the precision of rainfall forecast, particularly during critical monsoon seasons. It is important to note that it is not an easy task to predict the amount of rainfall accurately particularly in a country that is diverse and geographically diverse such as India. Unpredictability of monsoon patterns, local conditions, and weather change place an extra burden of trouble on the process of prediction. This paper compares numerous conventional machine learning (ML) methods in order to forecast rainfall. The second step is to identify the best models that can be used to predict Pakistan rainfall. Finally, an extended experiment of the suggested model is undertaken, and it is well-researched in the present research.

Nevertheless, the majority of rainfall prediction studies that are currently available in Pakistan either use classical statistical models or single-city datasets, and they do not consistently assess several machine learning classifiers under the same circumstances. In order to identify the most dependable and practically feasible method for localized rainfall prediction, there is still a research vacuum in the evaluation and comparison of ensemble-based models across several meteorological variables.

The final experiments solely concentrate on machine learning (ML) models, even though the original goal of this work was to compare both ML and DL architectures for rainfall categorization. In order to guarantee computational efficiency and interpretability with a small dataset size, DL models were purposefully left out of the results. However, because they guide future model extensions, the literature evaluation incorporates DL techniques for context and completeness.

Research questions of the study are listed below

RQ1: This study compares and contrasts different Machine Learning Algorithms to identify the best suited for recommendation system applications.

RQ2: Impact of the variable selection method of the different approaches influencing the precision of Rainfall forecasting in Pakistan?

RQ3: The research being conducted will concentrate on the effects and viability of utilizing machine learning and data mining specifically for Pakistan in order to predict weather for agricultural and disaster management purposes.

RQ4: To what extent can data-driven machine learning models enhance early-warning systems for rainfall and flood preparedness compared to traditional statistical or rule-based forecasting methods in Pakistan?

RQ5. How does feature selection and preprocessing (including normalization, encoding, and imbalance handling) impact the accuracy and generalization performance of rainfall classification models?

The research aids in resource allocation and proactive disaster management by improving early warning systems and rainfall prediction accuracy. The remaining portions of this work are separated into the following categories: Section 2 examines the theoretical foundations of early rain prediction using multiple machine learning approaches and provides a comprehensive literature assessment of the tools and resources employed. The techniques and algorithms are discussed in Section 3. From Section 4-6, study discusses the methodology, results and learnings made in order to predict early rains using these predictive machine learning algorithms and working with a dataset.

2. Background and Significance

Over the recent years, it has come to the realization that a lot of experimentation has resulted in the creation of a highly effective Rainfall prediction system. Various methods and technologies are usually used to develop efficient rainfall prediction systems, such as meteorological data analysis, Deep learning, machine learning, and numerical weather prediction models. The outlook of the thrilling research in the near future implies that researchers are hopeful of the further developments in the field. The future research areas may include the optimization of existing models, the inclusion of additional data sources, and the increase in the temporal and spatial accuracy of predictions, and the development of understanding of complex atmospheric processes that affect precipitation. Working together, meteorologists, climatologists, and data scientists will probably be essential to expanding our understanding and creating more accurate prediction systems as the subject of rainfall prediction develops. The literature reviews of previous works are included here.

Data-driven rainfall studies span diverse regions and methods: Australia's monthly forecasting compares ANN, KNN, MLR, and SVR on 24 stations (1970–2014) with five climate factors [1]. A dual framework combines sensor-driven ANN/DT rain classification with RNN/ES-LSTM seasonal–hourly prediction to aid flood response [2]. For South Africa's April-2022 COL floods, MaxEnt highlights elevation and land-use/cover as key flood drivers and maps province-scale susceptibility [3]. Daily prediction in Vietnam evaluates SVM, ANN, and PSO-ANFIS [4], while Kerala's seasonal forecasts (2011–2016) test KNN/ANN/ELM, with ELM yielding lowest MAPE in monsoon periods [5]. LSTM/RNN with six variables model annual/monthly totals [6]; a LoRaWAN node with logistic regression is reliable only up to 2-day lead times [7]. A stacking ensemble (KNN, XGB, SVR, ANN) improves monthly prediction on average but varies by site in the Taihu Basin [8]. For daily totals, RF, MLR, and XGBoost achieve MAE≈4.49/4.97/3.58 and RMSE≈8.82/8.61/7.85, respectively [9].

The study [10] may also highlight persistent bugs in the rainfall prediction using a machine learning model, that is, a scarcity of data, requirement for real-time alerts, and complexities of atmospheric phenomena. It could additionally deliberate over the shortcomings of the current models and possible ways to make them better. In this experiment, k-means clustering was implemented [11] for the purpose of data classification of the model forecast data. Short-term memory modeling recalled, then passed through several modalities of rainfall (LSTM). Initially, the samples were classified into four parts by using the K-means clustering algorithm. After that, the LSTM was used to create models for the various types of data. Eight different types of meteorological characteristics, including the model-forecast rainfall, were used as inputs, and the output was the difference between the actual and model-forecast rainfall. The developed model was then used to modify the rainfall that the model anticipated. The extent of how well machine learning models are able to capture the intricacies of flood dynamics cannot be determined without an appreciation of the input features [12]. The assessment would involve the different types of data involved in the flood prediction literature such as hydrological, meteorological, topographical and other pertinent aspects.

Deep Neural Network (DNN) was used to predict the probability of flooding based on temperature and rainfall and its performance was compared with SVM, KNN, and Naive Bayes model [13]. Nevertheless, the model was based on the few meteorological variables, and was not hydrologically validated. A Wavelet-coupled Multi-order Time Lagged Neural Network (WMT-LNN) was also introduced to the field of rainfall prediction that advances performance by decomposing time-series with the wavelet algorithm [14], yet it is very data-specific and computationally intensive to run in real-time. In the case of Pakistan, 50-year rainfall variability was the focus of Bayesian kriging regression to plan and manage [15], although accuracy of spatial interpolation is strongly dependent on the density of gauges and uniformity of quality data. The Sliced Functional Time Series (SFTS) model uses local time-dependent rain patterns in the short term [16], but does not work well in abrupt monsoon transitions or at sparsely covered areas. A smart-city rainfall prediction system that was based on SVM, DT, Naïve Bayes, and KNN was also introduced with the view of real-time implementation [17]; nevertheless, the reliability of its forecasts remains very low in the situation when the sensor data are unavailable or distorted, which impacts the performance of the system.

The study [18] investigates how machine learning algorithms could be integrated with remote sensing data to improve the accuracy and dependability of rainfall forecasting, however, it is noted that due to a limited temporal resolution, and sensor calibration challenges, generalization can be curtailed. The ESN and Deep ESN models of neural networks were implemented in rainfall prediction [19], and Deep ESN performed better than conventional neural networks and SVR algorithms but is too complex and cannot scale up to large volumes of data. A hybrid Multilayer Perceptron-Whale Optimization Algorithm (MLP-WOA) method [20] was proposed, which is more accurate in the predictability of annual rainfall, though it needs much tuning and is computationally costly. Other hybrid approaches that combine projection pursuit, the particle swarm optimization, and support vector regression were proposed in efficient rainfall prediction [21]; but its performance is extremely dependent on the choice of the parameter and can overfit very small datasets. Non-stationarity and high variability of climatic data are a challenge to the process of predicting rainfall occurrence using data mining [22] which is feasible.

The article [23] uses machine learning to select features in order to enhance rainfall prediction, and the ANN accuracy increases between 90% and 91%, but the issue of overfitting exists because of the lack of diversity in the available data. A more refined deep learning architecture that was optimized through particle swarm optimization [24] was not so successful, indicating the sensitivity of the parameters and training instability. An example of an MLP in Ghor (Malaysia) that was hybridized with imperialist, gravitational and CAPSO algorithms, was shown to achieve better monthly rainfall prediction accuracy [26], but at the cost of expensive computation and hyperparameters management. A different AI-based ensemble combining FFNN, LS-SVM, and ANFIS across seven sites in Cyprus increased performance on the basis of linear and nonlinear averaging, but was region-specific and relied on inter-station data consistency. Comparative studies on ANN-based rain forecasting [27 28] also showed good performance in dealing with nonlinearities but poor interpretability and cross-climate performance. Finally, an ANN model developed for daily and monthly rainfall forecasting [29] achieved stable performance but relied solely on local climatic variables, reducing generalizability across diverse environments.

The outcomes of this study show that the ANN model can accurately predict and measure precipitation on a daily and monthly basis utilizing six input parameters: temperature, dew point, humidity, pressure, visibility, and wind speed. The results show that the ANN model is a promising technique for daily rainfall forecasting. The testing phase values for R, RMSE, and MAE for the daily rainfall are 0.8063, 0.2247, and 0.0932, respectively. The performance indicators for R, RMSE, and MAE in the testing part are 0.8012, 0.0731, and 0.0578 for monthly rainfall, correspondingly. Additionally, the results show that monthly rainfall forecasts are more accurate than daily rainfall forecasts. The creation of the finest flood-determining model feasible is the aim of this effort [30]. Decision Tree Model is being evolved and improved as a result. This study compares three machine learning algorithms: To learn Decision Tree (DT), Random Forest (RF), and Gradient Boost (GB). The classification process considers the following characteristics: Place, Era, Month, min_temp, max_temp, clouds, rainfall, the number of days it rained, daily temperature, and times when the river was flooded.

The objective of this study [31] was to evaluate the capability of rainfall forecasting models develop The study [31] evaluates rainfall prediction using LSTM-based architectures compared with traditional machine learning models, including XGBoost, stacked LSTM, and bidirectional LSTM networks. Results show that the Bidirectional-LSTM outperforms the stacked variant with two hidden layers, though it requires more computational time and fine-tuning. In Malaysia, a machine learning system designed to predict rainfall in the Terengganu area [32] uses NNR, DFR, BDTR, and BLR models to predict rainfall in the region to improve agriculture and water management but due to the small regional dataset, cannot be generalized to other areas. The Indian experiment [33] forecasts rainfall till 2030 with ANN-MLP in 34 meteorological subdivisions, and provides information on water resource planning but with uncertainties related to long-term climate variability and with no real time validation.

The research [34] assists the stakeholders including farmers and researchers to comprehend the variability in climate by predicting rainfall with various input variables during monsoon seasons and annual seasons.

Random Forest model gave the best classification accuracy whereas ARIMA and Neural Networks exhibited the top performance with meteorological forecasting but the model has a seasonal emphasis so it is not as adaptable to other climatic regimes. An SVM-HHO hybrid model was created in rainfall time series forecasting in the Assam region of Cachar which was tested using the CC and the RMSE measures and was found to perform well with high sensitivity of the parameters. Subsequent survey [36] summarizes big data analytics methods of rainfall prediction, highlighting their potential but reporting issues with data standardization and generalization of the model. Comparative analyses of hybrid ML frameworks [37] indicate that such systems are more accurate, uncertainty is minimized, and convergence rate is more efficient, but these systems are usually computationally expensive and hard to interpre

A comparison of RF, MARS, SVR, and MLR is done in the study [38] to predict weekly and daily rainfalls at Ranichauri (Uttarakhand) where RF is best calibrated and predictive, but over a single-station dataset. MLR, RF regression and replicated neural networks were used to model daily rain in Semarang, Indonesia, with RF outperforming other models, yet no multi-year testing of model validation was done [39]. A study in Saudi Arabia [40] used different forms of ML with historical meteorological data with data mining techniques classified higher than traditional statistical models, but the methodology had a limitation in scaling features and real-time combination. A hybrid DSP model combining Prophet, SVR, and DWT [41] improved rainfall forecasting by decomposing data into frequency components, yet the model's complexity and dependence on parameter tuning limit operational scalability. Radar-based rainfall estimation using RF and SVM [43] demonstrated effective short-term forecasting with minimal input variables, but accuracy declines over complex terrains due to radar reflectivity uncertainties.

This study [44] contributes towards measurement of important climatic-associated risks within national borders for estimation of impacts in essential sectors of development such as agriculture and water resources. Besides, it wants to provide climate- related data more accessible for the national scale authority where policymaking is more appropriate. Consequently, in situ rain gauge data from CHIRPS scale monthly dataset, which took place from 1998 to 2010, was used whereas the Pearson' correlation coefficient was applied in validation. This paper dealt with the temporal and spatial changes in the key climate parameters in six countries in West Africa including Senegal, Niger, Burkina Faso, Côte d'Ivoire and Benin for the period 1981-2015. The findings further suggest that on balance over the last three decades, precipitation rose significantly in each of the five countries. The study [45] seek to compare analyses of extreme rainfall characteristics; frequency, intensity, seasonality and trends during the period of 1981 and 2015 in the SCWA. Therefore, rainfall estimation products containing in situ observation and satellite rainfall estimation data have been used, together with the daily rainfall at 31 stations distributed uniformly in the southern areas of Côte d'Ivoire, Ghana, Togo, and Benin. Possible topics to look at might include how the distribution of extreme rainfall events over the years would look like, if there are any tendencies apparent in the historical data, and status of potential impacts on regional climate. Such a study is valuable in enhance the existing understanding of climate variability in the western African context, in order to beneficially impact the related risk assessment, disaster response and preparedness, and sustainable development initiatives. Rainfall is an important part of building hydrological mathematical models and performs a definite function of stabilizing the flight of water cycle. These included M5, RF, SVR-poly, SVR-RBF, MLP, and LSTM were explored in this study [46] to predict monthly rainfall at two gauged stations in the Thale Sap Songkhla basin, Thailand. The following study [47] examines the temporal and spatial distribution of precipitation concentration in Pakis tan. This needs to know the pattern of distribution of precipitation with reference to periods within the country's regions. Specifically, concentration of precipitation is explored in terms of spatial (between sites or at different locations) and temporal (at different time points) variations. Aimed at comprehending where within particular intervals of time and space the precipitation is distributed or concentrated.

A comparison of various precipitation databases with ground measurements at 51 Pakistani monitoring stations (19982016) [48] shows spatial discrepancies but the lack of time overlap renders less accuracy. One model is a hybrid forecasting model [49], which combines swarm intelligence optimization and neural networks to make predictions of stable precipitation but is computationally complex in parameter tuning.

Hybrid wavelet neural network (HWNN) is a combination of MI, PSO, and MRA [50] to improve the prediction of monthly rainfall through the use of the long past series but the performance of the model is greatly reliant on proper preprocessing and index selection. ANN models trained by using FFNN and Levenberg Marquardt algorithms [51] were trained using historical data of the northern part of India and gave good results in terms of short-term prediction, but with limited generalization across different climatic regimes. The SVR-PSO hybrid [52] provides a better accuracy in rainfall forecasting in comparison to regression models but its sensitivity to kernel and PSO parameters can make it difficult to be robust. ANN-RBF ensemble model with gamma test and NMR methods [53] is more efficient in prediction, but requires high-quality input features and complicated training. The hybrid model of ACO and neural networks [54] is better than the regular neural networks but is computationally expensive. Hybrid systems that unite NB, C4.5, NN, SVM, and RF [55] enhance short- and long-term predictions of rainfall yet remain challenged by the imbalance in data between rainy and non-rainy days. Pattern recognition is enhanced by Shared Closest Neighbor clustering of Indian monsoon zones [57], but suffers due to regional sensitivity. Research about the variability of rainfall in Saudi Arabia [58] connects the precipitation to the activity of cyclones, but this has not been validated as a predictive. CatBoost, LSTM, PR, and RF are used to analyze urban meteorology data [59], which is highly accurate ($R 2 = 0.76$), but model interpretability is poor. Lastly, more recent work in South Asia [77, 99, 100] integrates logistical regression, SVM, stacking ensembles, and hybrid ML-statistical techniques to predict rainfall in Pakistan and Bangladesh and notes significant improvements but is still limited by the quality of the data and across-regional generalization. The research[101 compared the effect of various activation functions on rainfall forecasting with both ML and DL models, and they discovered that finely tuned activation functions can achieve much higher accuracy. They demonstrated that deep learning models such as LSTM and BiLSTM are more effective than traditional ones in the prediction of multifaceted rainfall behaviors.

2.1. Comparative Analysis

Table 1 includes a critical evaluation of different studies aimed at forecasting rainfall. Every study is thoroughly studied in a number of dimensions. The analysis will start with the determination of the study by serial number and the year in which it was published. It then elaborates on the exact algorithm applied in the process of rainfall prediction and the parameters or features included in the model. These predictions are explained by the results or goals they are meant to achieve and the dataset is described, its origin, and properties. The country in which the study was conducted is mentioned and the results are measured in various metrics to determine the effectiveness of the model. Lastly, an assessment of the limitations of the study has been conducted, which provides information on what may have been improved or what limitations may have influenced the findings of the study. Such an overall approach enables the comparative perception of various means of prediction of rainfall and the effectiveness of their application to various situations.

Table 1. Comparative Analysis of the previous study

Year	Algorithm(s) Used	Parameters / Features	Target / Task	Dataset / Period / Country	Performance (Metric)	Limitations
2023	ML & DL (Polynomial Regression, RF, LSTM)	Temp (Max/Min), Rainfall, Humidity, Wind Speed, Sunshine, Lat/Long, Altitude	Rainfall (mm) Regression	Bangladesh Meteorologica l Dept. (1948–2013)	R^2=0.76 (RF & PR), LSTM Loss=0.09	Small dataset, limited DL models; lacks transfer learning or pre-training
2023	ES-LSTM, ANN, DT	Temp, Rainfall, Wind Speed	Regression & Classificatio n	11 features, 2359 records (Australia)	MAPE=3.17 (ES-LSTM), ANN=96.65% , DT=84%	Needs improved weight tuning and

						DL generalization
2023	MaxEnt (ML)	Rainfall, NDVI, Elevation, TRI, SPI, TWI, Slope, Land Cover	Flood Susceptibility Mapping	1990–2022 (South Africa)	AUC=0.899 (90% accuracy)	Regional generalization limits; requires monsoon-region adaptation
2022	MLR, RF, XGBoost	Temp, Moisture, Humidity, Wind, Sunshine	Daily Rainfall Prediction	1999–2018 (Ethiopia)	XGB: RMSE=7.85, MAE=3.58; RF: RMSE=8.82	Sensor and environmental data missing; limited external validation
2020	ANN	Temp, Dew Point, Pressure, Humidity, Wind Speed	Daily & Monthly Rainfall	Austin, Texas Dataset	R=0.80; RMSE=0.073–0.248	Weak for daily rainfall; no deployment strategy
2024	Logistic Regression, Neural Network	Temp (Min), Humidity, Precipitation, Rain Today	Binary: "Rain Tomorrow"	3525 records (Aligarh, India)	Acc≈82.8%, ROC≈82.4%	Few features; small dataset
2023	MLR, SVR, MARS, RF	Temp, Humidity, Wind, Radiation, Evaporation	Daily & Weekly Rainfall	2000–2017 (Uttarakhand, India)	RF: RMSE=5.70–10.54, R=0.94–0.96	Region-specific tuning needed; lacks long-term trend capture
2023	SVM, RF, FFANN	Evaporation, SST, Pressure, Humidity, Cloud Cover, Wind	Classification (Regular/Heavy Rain)	WMO & IMD Database (Saudi Arabia)	FFANN=96.1%, RF=93.8%, SVM=83.7%	No regional language/alert system integration
2015	Hybrid Wavelet Neural Network (HWNN), ANN, MLR	ENSO, IOD, PDO, SAM, STR	Monthly Rainfall Forecast	255 stations (Australia, 1959–1998)	HWNN improved NSE by 0.17–1.8	Needs integration of multiple climate indices for better results

2.2. Comparison with past studies

Before proceeding any further, it is worth restating some of the key findings of the background investigation: The majority of conventional rainfall forecasting systems along with classical meteorological activities base on human-intensive or even partially-computerized processes of climatic information. The conventional methods are satisfactory in simple forecasting operations but not effective in providing appropriate results when the demands are tied to large datasets or extreme weather events. HEC-HMS, SWAT and MIKE SHE models the movement of water following the occurrences of rainfall using mathematical applications that involve the Manning equations and Navier-Stokes equations respectively. These models require comprehensive data inputs that are comprised of land topographical data and soil properties, the land use pattern and rain data to accomplish accurate outputs. These intricate modeling techniques demonstrate potential accuracy but have computational and data saturation issues limiting their performance at large scales. Multivariate linear regression (MLR) has been used as an analysis tool in combination with Mann-Kendall trend and exponential smoothing and ARIMA time series analysis, which has been applied greatly in analyzing rainfall [84,86].

These analytical methods successfully track broad patterns while maintaining poor capabilities to model climatic data's complex non-linear multivariate characteristics. The application of machine learning approaches for rainfall prediction has experienced growing popularity during recent times. Support Vector Machines (SVM) along with Decision Trees (DT) and k-Nearest Neighbors (k-NN) show enhanced predictive performance through learning from extensive datasets of varied diversity containing minimal data errors [27–32]. The application of ML models surpasses or equals traditional hydrological systems when enough data is available as shown in multiple cases [87]. However, although effective their predictive power becomes weaker for predicting unusual and extreme climate events because of our insufficient understanding of climate complexity.

Beyond the focus covered above, there have been more breakthroughs and advancements in the application of machine learning to flood prediction. However, there are several shortcomings and inadequacies in earlier research [28] [30] [40] [80].

The reliability of projections is impacted by the fact that many research relies on historical rainfall information, which may be erroneous, lacking, or obtained from a small number of weather stations.

While certain machine learning techniques might work well in small areas, they have trouble predicting annual rainfall accurately in broader geographic areas or with a variety of climates. Rainfall patterns can be greatly impacted by changes in terrain, land use, and climate, and models created for smaller, more focused areas might not fully account for these effects. Additionally, many studies that estimate yearly rainfall using machine learning (ML) may not properly evaluate how well their models perform in different regions or check their results against alternative datasets. This could lead to an overestimation of the model's accuracy, which could be problematic in real-world applications where accurate annual rainfall forecasts are crucial for decision-making.

Some models struggle with short-term or near-real-time forecasting due to the limited temporal resolution in historical datasets, which are often averaged over days or months. High-resolution, time-stamped data are crucial for accurate flood prediction, especially for events that develop rapidly.

The challenge related to the combination of ML predictions and standard forecasting used by meteorological organizations is sometimes not addressed in the existing literature, and this fact might limit their applicability and implication in practice. To address these limitations, advances have entailed the utilization of hybrid models, ensemble methods.

The study is unique to the rainfall prediction field especially in reference to Pakistan, owing to several original contributions. This analysis concentrates entirely on the local climatic and meteorological information of Pakistan which was collected at a few stations within the country, unlike in the general studies that use the global or regional information. The projections of the study can be directly translated to Pakistan disaster preparedness and resource management strategies due to its personalized strategy that can facilitate it to manage the unique rainfall patterns and corresponding flood risks of Pakistan. Moreover, the study employs a range of machine learning algorithms as opposed to relying on just one as Random Forest, Support Vector Machines, Gradient Boosting, Adaboost, and Naive Bayes. In this case, the study provides in-depth analysis of the performances of different models through comparing them, thereby showing the most accurate algorithms to use in predicting rainfall. Another issue that the research addresses is the analysis of feature importance, or which meteorological variables, including temperature, humidity, wind speed, and air pressure, produce the most significant impacts on the forecasting of rainfall. The research is a thorough and practical exploration of machine learning to predict rainfall in Pakistan because it does not only enhance the explainability of the model but also offer meaningful information to meteorologists and policymakers.

3. Approaches Used for Research

Basically there are two methods in weather prediction, empirical and dynamical methods [67].

a. Empirical Approach

This approach uses previous data analysis to estimate future situations and searches for correlations between attributes. The most often used techniques in the empirical approach to weather forecasting are classification, regression, Decision trees, Artificial neural networks (ANN), fuzzy logic, ARIMA models, swarm intelligence for outcome optimization, long short term memory (LSTM), and other information processing techniques [67].

b. Dynamical Approach

Instead of focusing solely on the current state of a system, dynamical approaches emphasize understanding how systems evolve and behave dynamically. In a dynamical approach, the aim is indeed to model systems in a way that captures their behavior over time, with the goal of making predictions or understanding future conditions. However, the degree to which the results will match the actual state depends on several factors shown in table 2.

Table 2. Several Factors for prediction

Model Accuracy:	**Data Quality:**	**Parameter Estimation:**	**Initial Conditions**:	**System Complexity**:	**External Factors**:
The models own accuracy is quite important. It is important that the model accurately represents the fundamental dynamics of the system under study.	The forecasting ability of the model can be greatly influenced by the type and amount of data that were utilized to build it. Predictions that are more accurate and thorough can be made with greater confidence.	Many dynamical models involve parameters that need to be estimated from data or theoretical considerations. The accuracy of these parameter estimates can affect the reliability of the model predictions. e.g. such as the mean precipitation rate, variability, and seasonality.	Small initial condition changes can have a big long-term impact on the behavior of many dynamical systems. As a result, precise determination of the initial conditions is essential for forecasting the future state	The complexity of the system being modeled can also influence the accuracy of predictions. Highly complex systems, such as turbulent fluid flows or ecological systems, may exhibit behaviors that are difficult to capture accurately in a model.	External factors, such as environmental changes or external interventions, may influence the system's behavior in ways that are not accounted for in the model, leading to discrepancies between predicted and actual outcomes.

One of the most predictable weather patterns is rain. Precipitation falls to the earth as cloud particles become too heavy to stay suspended in the atmosphere. There are many different types of precipitation, including sleet, snow, freezing rain, and hail. Rain is the term used to describe the precipitation of hydrometeor water particles, usually larger than 0.5 mm in diameter, that fall from clouds and land on Earth's surface. This precipitation occurs when certain atmospheric conditions align, allowing for the formation and release of water droplets from clouds. When a number of factors come together, including humidity, temperature, evaporated water, rising air currents (swirling patterns), ambient air conditions, and the availability of enough moisture, rain occurs in the cloud. The planet's everyday activities are governed by precipitation, which is also crucial for tracking the climate of the water in the earth's reservoirs. Since precipitation is the main element of the hydrological cycle, it is crucial to the study of hydrology. The majority of the variability in terrestrial hydrology can be attributed to precipitation, which is undoubtedly the most significant element of the land-atmospheric system. Precipitation typically shows rapid temporal fluctuations together with considerable and frequent geographical variability. As one of the main objectives of current research efforts in distributed hydrological modeling and land data assimilation systems, precipitation is the most crucial input to generate accurate simulations and forecasts for a suite of hydrological variables (soil moisture, stream flows, and flood levels) [68. This study is empirical, Precipitation observation helps in preparing for and responding to natural disasters such as hurricanes, typhoons, and cyclones. By monitoring rainfall rates and distribution, meteorological agencies can issue early warnings and evacuation orders to protect lives and property from flooding, storm surges, and other hazards associated with heavy rainfall.

The analogy and subjective approaches [70] are used in very brief weather forecasts using various weather models. Table 3 represents the analogical and subjective approach for analyzing and comparing the scenarios of rainfall. With the help of these approaches we can identify the possibilities of unstable weather areas that have a chance of rain.

Table 3. Represent the analogy and subjective approach

Analogy Approach	Subjective Approach
This approach involves comparing current weather conditions to past events with similar characteristics. Meteorologists use historical data to identify patterns or analogs that resemble the current atmospheric conditions. By drawing parallels between past weather events and the present situation, forecasters can make quick predictions about potential weather outcomes. For example, if a specific atmospheric pressure pattern is similar to one observed during a previous storm, meteorologists may anticipate similar weather patterns and issue relevant warnings or advisories [70].	The subjective approach relies on meteorologists' expertise, intuition, and judgment to interpret weather data and make forecasts. This approach involves [70] synthesizing various sources of information, including satellite imagery, radar data, surface observations, and numerical weather prediction models, to assess current and evolving weather patterns. Subjective forecasts often incorporate qualitative assessments of weather trends, such as changes in cloud cover, wind direction, or atmospheric instability, to provide concise and actionable information to the public.

There are frequent daily stable weather trends during the rainy and dry seasons. Dynamic approaches, which forecast future patterns by evaluating past and present meteorological data, can be highly useful during these times. The atmospheric conditions tend to be more stable and predictable, allowing forecasters to make accurate predictions based on past trends and current observations. On the contrary, Transitional seasons, like spring and autumn, are more of a challenge to dynamical methods. These periods are characterized by intense changes in weather conditions, which are characterized by changing temperatures, changing wind patterns and more convective activity. Local weather phenomena, including mountainous terrain and land-sea temperature differences, may be more directly affecting weather patterns, than synoptic-scale mechanisms. Consequently, there is a reduction in predictability of weather in transitional seasons hence making it harder to use dynamical methods to make reliable predictions.

During these transitional periods, forecasters may need to supplement dynamical approaches with other methods, such as analog forecasting or subjective interpretation of atmospheric conditions, to improve the accuracy of short-term weather forecasts.

Data mining [71] is defined as the activity of uncovering valuable and interesting patterns from vast amounts of data. These patterns may include trends, associations, correlations, or anomalies that are not immediately apparent through simple observation. In this work, patterns and associations utilized to anticipate future rainfall were found by analyzing historical meteorological data using data mining techniques. When it comes to supervised learning, the algorithm is rewarded based on past experience that, in turn, are labeled and have input features and target labels that correspond to them. With labeled training data, supervised learning aims towards the building of the mapping function from the input variables to output variables, using the dataset. Here, the mapping function that has been tested and new data can be generated. Therefore, it will be possible to plan things out using this function [60]. Supervised learning tasks may be broadly classified into two categories the first one is regression and second is classification, as shown in fig 2. In this study explanation, categorization will receive a lot of attention.

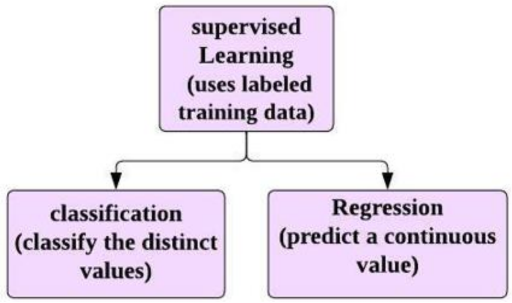


Figure 2. Supervised Learning types

Predicting the categorical class labels of new instances based on input features is the aim of classification problems. There is a limited number of classes or categories that the discrete output variable falls into. The decision function or boundary that divides the various classes in the feature space is learned by the algorithm. Some supervised learning algorithms used in this experimental study e.g. Regression algorithm Support vector machines (SVM), Regress or decision trees, logistic regression. Throughout training, the algorithm discovers how to transfer the input features to the proper sorts of rainfall based on the specified goal.

3.1. Random forest

The Random Forest ensemble learning approach performs their process by creating many of these decision trees during training and deciding on the overall average voting for regression tasks or majority vote for classification tasks as shown in fig 3. Random Forest works well on a range of datasets and is resistant to over fitting. Equation 1 illustrates how it manages missing values and keeps accuracy even when dealing with a large number of features.

Classification**:** Let $\hat{C}_b$(x) be the both random-forest tree's class prediction.

$$\text{Then } \hat{C}_{RF}(x) = \text{ majority vote } \{\hat{C}_b(x)\}_{b=1}^{B} \quad (1)$$

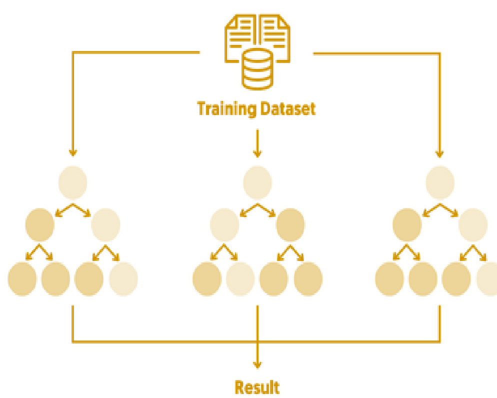


Figure 3. Representation of Random Forest

3.2. Logistic Regressions

In case of binary classification scenarios, an instance's probability of belonging to a certain class is predicted through a statistical model which is referred to as the logistic regression. The logistic function that would transform every real-valued input into a number between 0 and 1. The equation for the logistic function is shown in Equation 2:

$$\text{Sigmoid }(z) = \frac{1}{1+e^{-z}} \quad (2)$$

Where:

- z is a linear combination of the features and their respective coefficients, represented as
- $z = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$, where β_o is the intercept, β_i are the coefficients, and x_i are the feature values.

The logistic regression model calculates the odds of a binary result, such as one of zero or one. The logistic function predicts the likelihood that an occurrence is in the positive class (class 1):

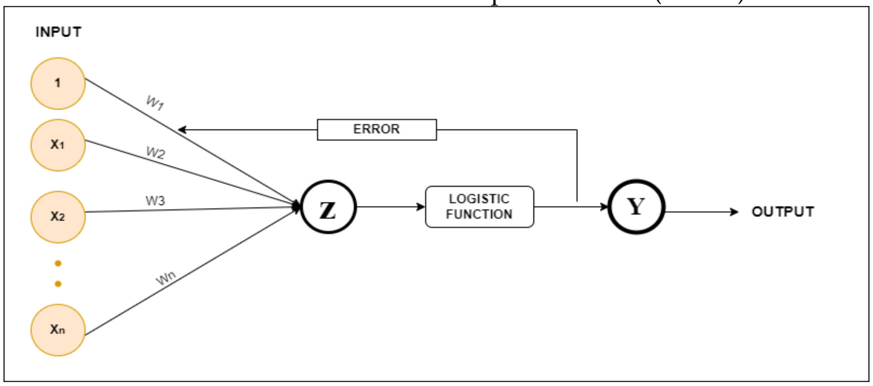


Figure 4. Representation of Logistic Regression

3.3. Naïve Bayes

The technique of the Naïve Bayes has been considered to be the most powerful and successful of the supervised machine learning and the data mining algorithms. Based on its attribute independence property, the Naive Bayes [72] method is a probabilistic machine learning model that incorporates the Bayes' theorem.

In classification tasks, it performs better and better until it reaches a high level of accuracy with its simplicity. The innovative Bayes theorem designates naive to only help one feature for a class, but it has less impact for another feature in the class. Although this is a strong and frequently implausible assumption, it streamlines the computation and increases the computing efficiency of the training data. Equation 3 illustrates the application of Bayes' theorem to determine the likelihood of a class given its attributes, based on this assumption.

$$P(C_k|x_1, x_2, \ldots, x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x_1, x_2, \ldots, x_n)} \quad (3)$$

Where:

$P(C_k\|x_1, x_2, \ldots, x_n)$	Posterior probability of class C_k given the features $x_1, x_2, \ldots, x_n$
$P(C_k)$	Prior probability of class C_k
$P(C_k)$	Likelihood of feature x_i occurring within class C_k
$P(x_1, x_2, \ldots, x_n)$	Evidence pr probability of observing the feature vector x

3.4. AdaBoost

Ada Boost is a type of supervised machine learning algorithm called ensemble learning that means deriving a set of simpler models/decisions. Based on a training set, AdaBoost generates a strong classifier by using weak classifiers, and their combining technique is called decision-stump. Ada Boost or Short for Adaptive Boosting is an ensemble learning method that combines a number of weak learners to create a strong learner. When compared with the other SVM algorithm AdABoost being a collective machine learning approach has the lowest wrong false rate and the highest classification precision for the rainfall prediction ratio. This is done by using a weak ensemble; the usual choice being decision trees to create a strong classifier in the process shown in fig 5 and equation 4.

$$F(x) = \text{sign}(\textstyle\sum_{i=1}^{T} \alpha_t h_t(x)) \quad (4)$$

Where,

- T=iterations (no of weak learners)
- h_t (x)=The tt-th weakest learner's estimate for input x

- α_t =The amount of importance given to the tt-th weak learner in term of weight assign
- sign (.)=The sign function that connects the category identifiers with the sum of the weighting guesses is called sign (.). It anticipates the "positive" class if the total is positive; if not, it predicts the opposite class.

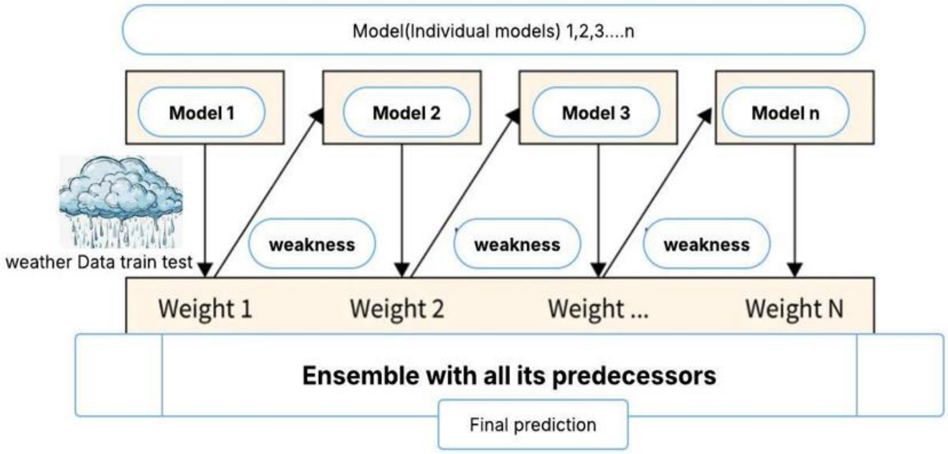


Figure 5. Representation of AdaBoost classifier

3.5. Gradient Boosting

Gradient Boosting is a widely used supervised machine learning algorithm. For classification and regression tasks, gradient boosting algorithms perform well and provide good results, unlike traditional decision tree algorithms like Random Forest, which build multiple trees independently, as shown in Fig. 6. It creates trees sequentially and learns the trees from the errors of its predecessors. It can understand the complex relationship between the attribute and the target result. It works like initializing the model with a simple model, such as a single leaf (constant) value for regression or a constant probability for classification. Then it calculates the residuals or pseudo-residuals for each data point, which represent the errors made by the initial model.

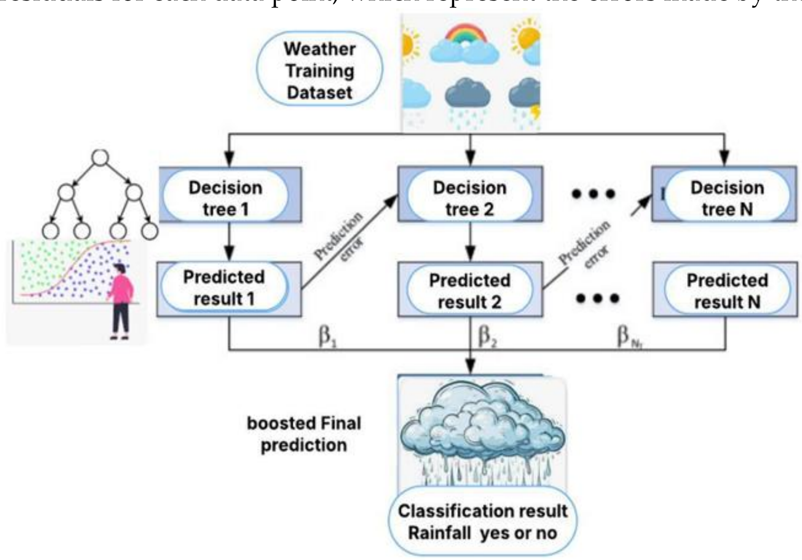


Figure 6. Representation of Gradient Boosted Trees

Its mathematical representation are shown in (equation 5a, 5b and 5c)

- F_m((x) as the current ensemble model (sum of first m weak learners),
- h_m((x) as the m-th weak learner (e.g., decision tree),
- ϱ as the learning rate,
- L as the loss function.

At each iteration, we update the model as follows:

$$F_m(x) = F_{m-1}(x) + \rho \, . \, h_m(x) \quad (5a)$$

Then, the residuals (or pseudo-residuals) are updated:

$$r_{im} = \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \quad (5b)$$

Finally, the prediction at each iteration is given by:

$$\hat{y}(x) = F_M(x) = \sum_{m=1}^{M} \rho . h_m(x) \quad (5c)$$

3.6. K nearest Neighbors

It is a very effective machine learning algorithm used in supervised learning and used in regression and classification. It categorizes new instances based on the most frequent class or highest counting class among the k-Nearest Neighbors in the feature space. KNN is better for small data sets. Derivation for similarity measurement It uses Euclidean distance as illustrated below in equation 6.

$$\hat{y} = \text{majority}\ \{y_i | x_i \in N_k(x)\} \quad (6)$$

Where:

- $\hat{y}$ – Predicted class label for the input instance x_i
- $N_k(x)$ – The set k nearest neighbors of x in the feature space
- y_i – Class label of neighbor x_i
- majority(.) – Returns the most frequent class among the neighbors

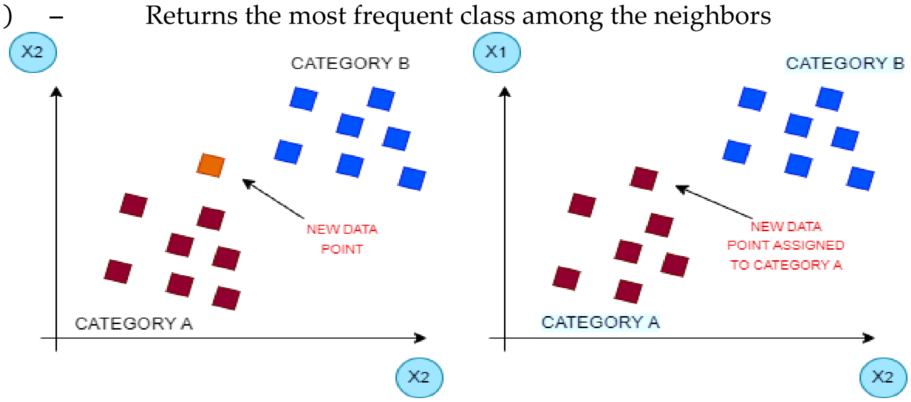


Figure 7. Representation of K nearest neighbors

3.7. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both regression and classification; it gets harder to imagine in feature space or in N dimensions without simulation. The hyperbolic plane in Figure 8 aims at maintaining the greatest possible distance between the closest points pertaining to various categories. The SVM approach finds the good hyper plane that segregates data points into various feature space categories [37, 35]. The dimension of the hyper plane is found by the total number of attributes. It aims to maintain the maximum margin between the closest points of different categories.

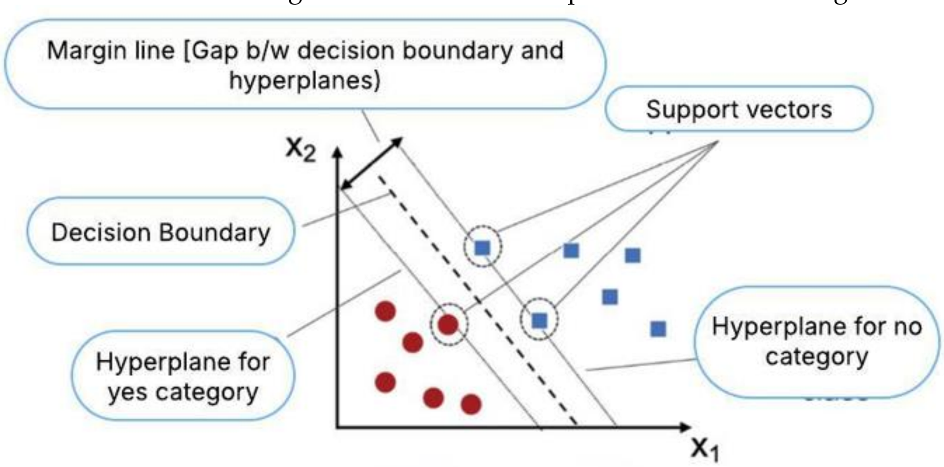


Figure 8. Representation of Support vector machine

The goal function in the optimization problem's dual form gives rise to the main equation. Equation 7 shows the Objective Function (Dual Formulation) in this instance. In its dual form, the objective function of the SVM is stated as follows given a dataset with N sample (x_i, y_i),), where x_i is the input vector and y_i is the class label $y_i \in \{-1,1\}$.

Maximize: $W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$ (7)

Subject: $\sum_{i=1}^{N} \alpha_i y_i = 0, \ \ 0 \le \alpha_i \le C, \ \ i = 1,2,\dots,N$ (8)

- α_i Are the Lagrange multipliers (also called dual variables)?
- C represents the balance of increasing and decreasing margin. it treated as a regularization parameter
- W (α) is the function to be maximized as the target.

Finding the optimum levels of the multipliers of Lagrange αi for maximizing W (α) is the dual problem of the SVM, which is represented by the objective function W (α). Lagrange multipliers are used to identify support vectors and the support vectors are the misclassified or marginal data points.

3.8. Bidirectional Model

In order to capture both past and future historical context, a bidirectional GRU analyses the sequence in both temporal directions. Two GRU cells are stacked as a forward GRU that reads $\{X_1,\dots, X_T\}$ and a reverse GRU that reads $\{X_T,\dots, X_1\}$.Each side's hidden states are concatenated at each time step in accordance with the usual GRU gating in Eqs. (8–11).

Forward and backward recurrences and their concatenation shown in Eq. (8 and 9):

$\overrightarrow{h_t} \ \ GRU_f(X_t, \overrightarrow{h_{t-1}})$, $\overleftarrow{h_t} GRU_b \ (X_t, \overleftarrow{h_t}$-1)

Context fusion by concatenation:

h_t=[$\overrightarrow{h_t}$; $\overleftarrow{h_t}$] ∈ R^{2H} (9)

Sequence labelling prediction (one output per time step) as in Eq. (10):

y_t=g($W_o \ h_t + \ b_o$) (10)

Sequence classification prediction (one output for the entire window) shown in Eq. (11):

y=g(W_o[$\overrightarrow{h_T}$; h_1] +b_o) (11)

Here, g(·) is a task-appropriate activation (e.g., sigmoid for binary classification), HHH is the number of units per direction, and [· ; ·]denotes vector concatenation. BiGRU has computational efficiency similar to a unidirectional.

4. Research Methodology

The purpose of this proposed study is to investigate whether using machine learning that may lead to a reduction in error and an increase in accuracy.

4.1. Dataset Analysis

The dataset comprises daily meteorological observations retrieved from the Visual Crossing weather platform, including temperature (maximum, minimum, and average), humidity, dew point, wind speed, and rainfall-related parameters such as precipitation, precipitation probability, and precipitation coverage. Each record represents one day's data, forming a structured time series suitable for rainfall classification and prediction experiments.

A few missing values were found during the analysis of the dataset. The missing values are shown in Figure 10. Sea level pressure is the only characteristic without a value. Over a range of time periods, it was discovered that the sea-level pressure (SLP) variable frequently and sporadically had missing values. To prevent potential noise amplification and maintain data integrity, this feature was excluded from the model training procedure. None of the other meteorological variables had any missing values after preprocessing. The missing value for other features is 0%.

The dataset size is provided according to the number of rows, columns, type unit, and characteristics in Tables 4 and 5. This comprehension makes it simpler to understand the features of the dataset and the kinds of factors that are considered when making rainfall predictions.

datetime	tempmax	tempmin	temp	feelslikemax	feelslikemin	feelslike	dew	humidity	precip	precipprob	precipcover	preciptype	snow	snowdepth	windgust	windspeed	winddir
2023-12-01	84.5	58.9	71.8	81.9	58.9	71.2	47.8	47.4	0	0	0		0	0	16.3	16.1	50.3
2023-12-02	83.1	62.5	72.8	81	62.5	72.3	47.1	41.8	0	0	0		0	0	17.4	16.1	37.8
2023-12-03	84.2	66.1	72.9	81.6	66.1	72.4	46.4	41.6	0	0	0		0	0	17.2	13.9	40.3
2023-12-04	84.9	57.1	71	81.8	57.1	70.3	42.6	39.8	0	0	0		0	0	12.5	10.3	55.5
2023-12-05	84.2	57.1	71.3	81.5	57.1	70.7	41.1	36.2	0	0	0		0	0	15.2	11.4	53.6
2023-12-06	82.4	55.3	69	80.2	55.3	68.5	39	36.7	0	0	0		0	0	15.9	11.4	58.9
2023-12-07	81.3	58.9	68.5	79.5	58.9	68.3	39.8	36.1	0	0	0		0	0	14.1	15	40.7

Figure 9a. Representation of Pakistan Station's Dataset

pressure	cloudcover	visibility	solarradiation	solarenergy	uvindex	severerisk	sunrise	sunset	moonphase	conditions
1009.4	34.5	5.4	239	20.6	8	5	06:25:01	18:16:27	0.66	Partially cloudy
1008.9	51.3	5	237.3	20.5	8	5	06:25:25	18:15:26	0.69	Partially cloudy
1007.7	69.9	5	231.5	20	8	5	06:25:50	18:14:25	0.75	Partially cloudy
1008.4	85.9	4.9	104.2	9	4	15	06:26:15	18:13:24	0.76	Rain, Partially cloudy
1009.8	62.8	5.2	155.8	13.5	6	15	06:26:41	18:12:24	0.79	Partially cloudy
1010.7	50.5	5.6	225	19.5	8	15	06:27:07	18:11:24	0.82	Partially cloudy
1010.4	76	5.4	188.8	16.1	8	15	06:27:33	18:10:25	0.85	Partially cloudy

Figure 9b. Representation of Pakistan Station's Dataset

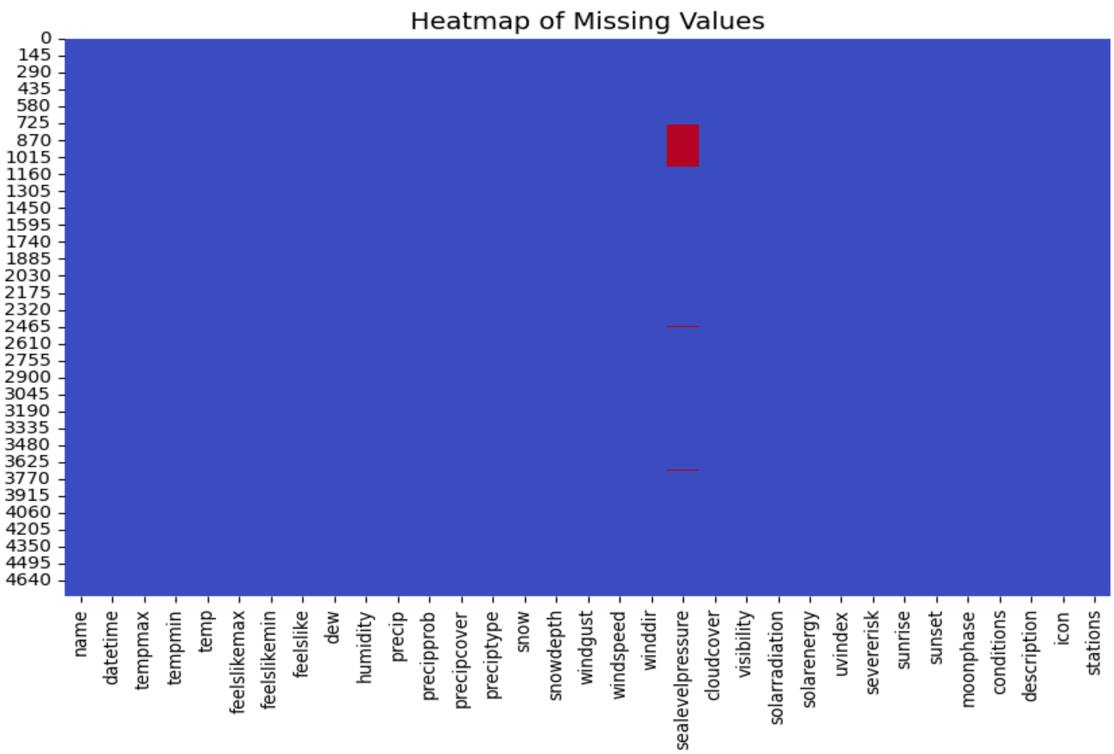


Figure 10. Features Missing values representations

Table 4. Data table set Description

Dataset Description	
Year	**2011 to 2023**
Size	4778 rows, 33 column
Features	7 categorical, 24 numeric
Targets	categorical outcome with 2 classes
Location	Karachi, Pakistan

Meta’s	1 text

Table 5. Features for Deep Learning and Machine Learning Model Training

Sr. No	Feature	Feature Description	Type	Units
1	name	This probably refers to the location's name or address where the weather forecast is given (e.g. Karachi, Lahore)	object	String
2	Date time:	This is the date and time for which the weather conditions are forecasted. It includes information about when the forecast was issued and the period it covers.	object	Time stamp
3	Temp max	Maximum temperature expected during the forecast period.	float	Celsius (°C)
4	Temp min	Minimum temperature expected during the forecast period	float	Celsius (°C)
5	Temp	The general temperature during the forecast period, likely an average or current temperature.	float	Celsius (°C)
6	Feels like max	Maximum "feels-like" temperature, which takes into account factors like humidity and wind to estimate how the temperature might feel to a person.	float	Celsius(°C)
7	Feels like min	Minimum "feels-like" temperature.	float	Celsius (°C)
8	Feels like	The general "feels-like" temperature during the forecast period.	float	Celsius (°C)
9	Dew	The temperature at which moisture saturation of the air occurs and dew formed is known as the dew point.	float	Celsius (°C)
10	Humidity	The relative humidity level, indicating the amount of moisture in the air.	float	Percentage (%)
11	Precip	Precipitation, which could include rain, snow, sleet, etc.	float	Millimeters (mm)
12	Precip prob	Probability of precipitation occurring.	int	Percentage (%
13	Precip cover	The extent or coverage of precipitation in the forecast area	float	Percentage (%)
14	Precip type	The type of precipitation (rain, no rain, etc.).	object	categorical
15	Snow	Amount of snowfall expected.	int64	Centimeters (cm)
16	Snow depth	Depth of snow on the ground.	int64	Centimeters (cm)
17	Wind gust	The maximum wind gust expected during the forecast period.	float	Meters per second (m/s)
18	Wind speed:	The aerage wind speed.	float	Meters per second (m/s)
19	Wind dir	Wind direction characteristic, showing the direction of the wind.	float	Degrees (°)
20	Sea-level pressure:	Atmospheric pressure at sea level.	float	Hectopascal s (hPa)

21	Cloud cover	The proportion of cloud cover in the sky. When humid air rises and cools, water vapor condenses into little water droplets or ice crystals, which is how clouds are formed.	float	Percentage (%)
22	Visibility	The point at which elements are clearly apparent.(Visibility is a crucial parameter in weather forecasting and is often linked to various atmospheric conditions Example fog, rainfall, wind gust)	float	Kilometers (km)
23	Solar radiation	The level of sunlight that penetrates the earth. (Solar radiation is the primary source of energy that drives evaporation from the Earth's surface, particularly from oceans, lakes, and moist land areas. As water vapor rises into the atmosphere, it can condense to form clouds.)	float	Watts per square meter (W/m^2).
24	Solar energy	The solar energy available during the forecast period.(solar energy is a fundamental component of the Earth's climate system, driving the water cycle and influencing the processes that lead to precipitation)	float	Watts per square meter (W/m^2)
25	Uv index	ultraviolet index, which shows how strong the sun's UV radiation is. (Ranges Minimal risk=0–2 , Medium risk 6-7, Extreme risk, 8–10 = Very high hazard 11 +.)	int64	Unit less
26	Severe risk	The risk or severity of severe weather conditions. (The specific feature values associated with severe weather risk can vary depending on the region, the type of severe weather event, and the criteria used by meteorological agencies.) String (e.g., Low, Moderate, High)	int64	categorical
27	Sunrise	Time of sunrise.	object	Date/Time
28	Sunset	Time of sunset.	object	Date/Time
29	Moon phase	the moon phase, such as full or half moon. It can be written as a continuous variable with a range of 0 to 1. This 0 to 1 describe half-moon and full moon	Float	categorical
30	Conditions	General weather conditions, such as clear, cloudy, etc.	object	string
31	Description	Additional descriptive information about the weather. (E.g. partly cloudy etc.)	object	string
32.	Icon	Weather icon, often used for graphical representation in weather apps.	object	string
33.	Stations	Weather stations providing the forecast data.	object	string

One of the ways to predict precipitation; whether it will rain or not in this research is by classifying precipitation through machine learning algorithms that determine yes or no based on the amount of available weather information. Fig 11 shows the total number of rain and no rain.

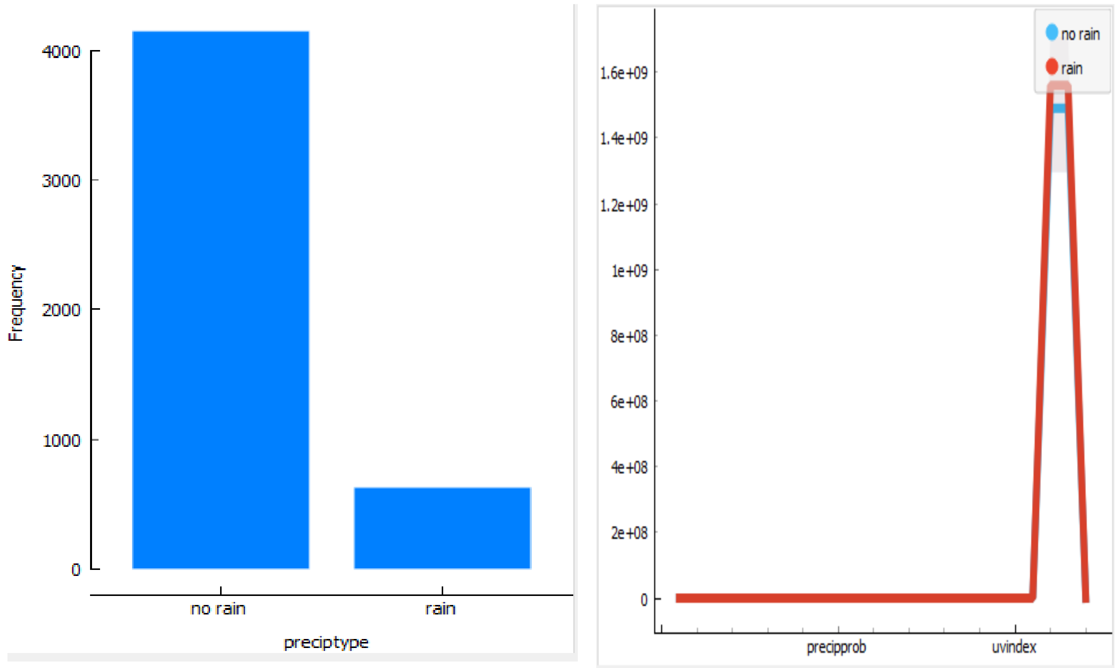


Figure 11. Representation of precipitation in term of rain and no rain

A graphical representation of the several factors influencing the intricate process of rainfall is shown in Fig 12. These variables include a wide range of things, such as temperature gradients, geographical characteristics, and atmospheric conditions. This illustration attempts to capture the complex interactions between several factors that affect the amount and frequency of rain and the multiple nature of rain creation.

Together, the mean, median, mode, and distribution analysis in Table 6 help us extract important insights from the dataset, which in turn helps us shape our following analytical approaches and improves the precision of our conclusions. This table is essential for rainfall prediction in Pakistan as it provides statistical insights into key meteorological factors such as temperature, humidity, wind speed, cloud cover, and precipitation probability. Analyzing the mean, mode, median, and dispersion helps detect patterns, seasonal trends, and extreme weather conditions. The significant variability in humidity, pressure, and wind speed indicates their strong impact on rainfall. Additionally, the high occurrence of zero precipitation values suggests an imbalanced dataset, requiring proper preprocessing. These insights support the selection of suitable machine learning models, enhancing prediction accuracy for better water resource management and disaster preparedness.

For early processing, a data mining approach is employed to convert messy and imprecise input into a structure that the model can easily understand without ambiguity. Large number of data that are collected are inaccurate, imbalanced, and missing, also a lot of extra information contain that are not valuable for experiment.

Data investigation and estimation indicate that the raw model data with the exception of a single missing number of sea level pressure is complete and contains no null, redundant or invalid values. In the preprocessing phase of feature selection, it is only possible to select those features that can be relevant in our rainfall forecasting model. Consequently, the time used in training reduces and accuracy of the model improves. Table 6 and Figure 13 indicate the coefficient of rainfall correlation with various variables. Subsequently, there is feature dropping of the work. In order to facilitate modeling further, a correlation with the both independent and dependent variables is calculated. The columns that followed were omitted. As indicated in table 7, the characteristics related to UV index and solar radiation were the least correlated with the rainfall variable. These attributes were removed before the modeling.

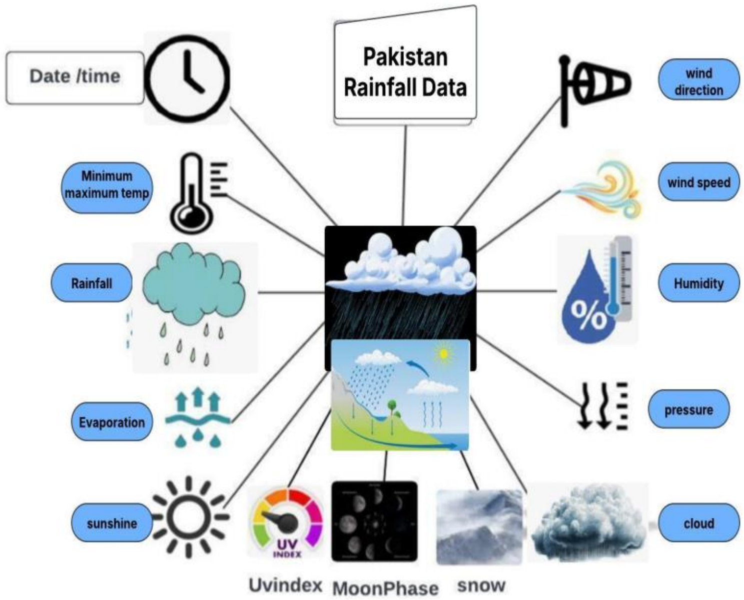
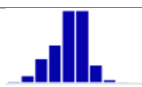
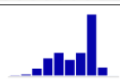
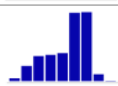
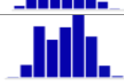
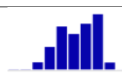
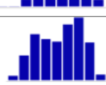
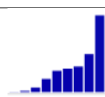
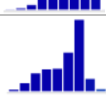
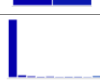
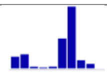
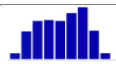
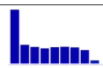
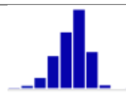
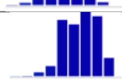
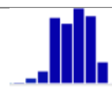
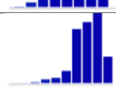
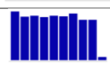
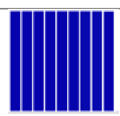
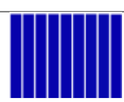
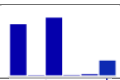
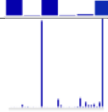
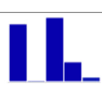
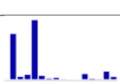


Figure 12. Illustration of various factors that influence precipitation

Table 6. Histogram, and central tendency of dataset

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.
temp maxp		89.6	91.4	91.3	0.080	62.6	116.5
Temp min		71.799	80.5	75.2	0.159	31.9	91.3
temp		79.990	86.1	83.0	0.102	55.8	100.1
Feels like max		93.369	78.8	94.8	0.117	62.6	123.8
Feels like min		74.188	78.7	75.2	0.194	19.0	104.9
Feels like		83.512	95.6	85.4	0.141	54.6	108.8
dew		61.364	74.3	67.0	0.242	4.0	80.1
humidit y		57.601	67.0	62.9	0.258	10.9	91.7
precip		0.0297	0.00	0.00	6.8773	0.00	5.899
Precip prob		13.16	0	0	2.57	0	100
Precip cover		5.1513	0.00	0.00	3.4806	0.00	100.00
Wind gust		3.544	0.0	0.0	2.307	0.0	51.9
Wind speed		17.804	13.9	16.1	0.672	5.8	198.0

Wind dir.		209.42	243.9	242.6	0.422	0.1	360.0
Sea level pressure		1008.835		1009.1	0.007	993.5	1023.3
Cloud cover		34.787	0.0	30.450	0.837	0.0	100.0
visibility		3.453	3.5	3.5	0.153	1.1	5.4
Solar radiation		234.896	283.3	238.9	0.211	18.0	340.2
solar energy		20.288	25.7	20.6	0.212	1.6	29.4
uvindex		8.04	9	8	0.17	1	10
severe risk		3.01	0	0	3.31	0	100
moon phase		0.4828	0.25	0.50	0.5979	0.00	0.98
sunrise		2017-07-16 18:36:45.024278	2011-01-01 07:16:48	2017-07-16 17:52:28	~13 years	2011-01-01 07:16:48	2024-01-31 07:15:16
sunset		2017-07-17 06:45:27.819380	2011-01-01 17:53:57	2017-07-17 07:23:34	~13 years	2011-01-01 17:53:57	2024-01-31 18:15:46
name		-	Karachi	-	0	-	-
snow		-	0	-	0	-	-
Snow depth		-	0	-	0	-	-
conditions		-	Partially cloudy	-	1.06	-	-
description		-	Partly cloudy throughout the day.	-	1.46	-	-
icon		-	partly-cloudy-day	-	1.1	-	-
stations		-	41780099999,OP	-	1.47		-

			KC,remote			
Precip type		-	no rain	-	0.39	-

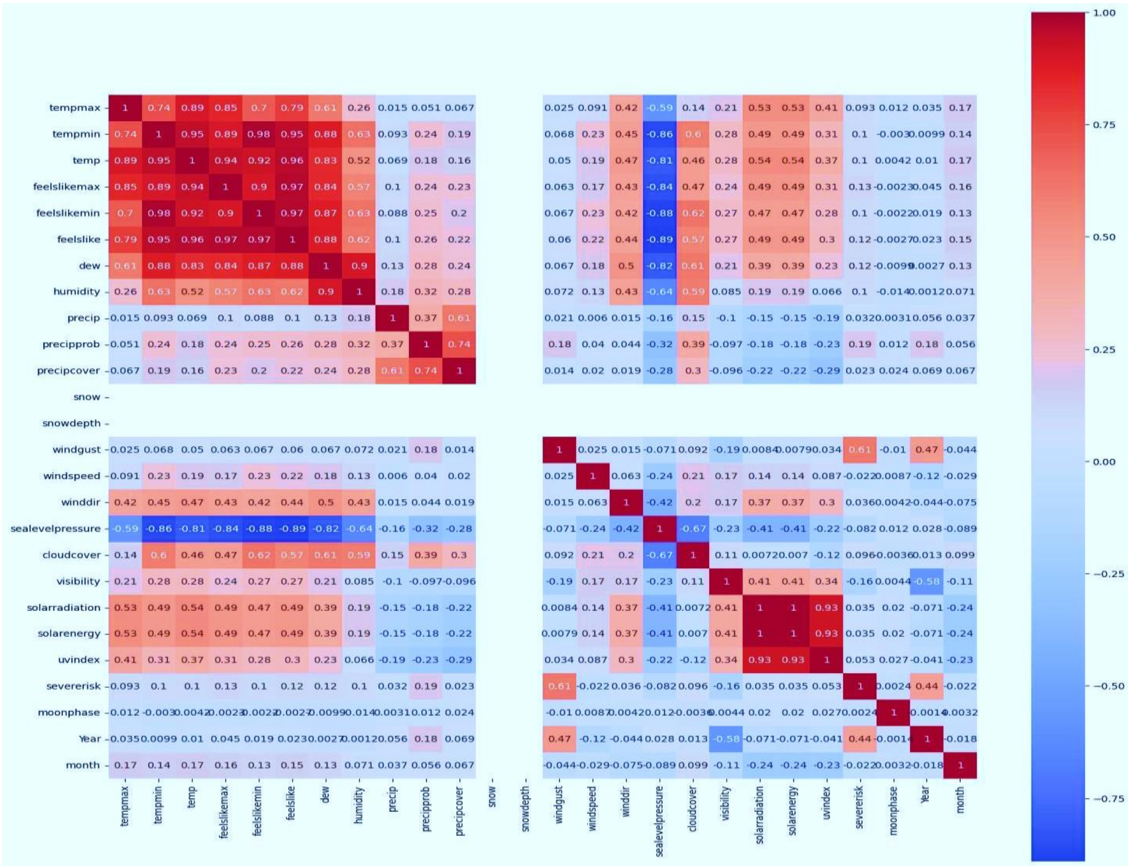


Figure 13. Visual depiction of the relationship between attributes.

Table 7. Correlation coefficients of precipitation with various features

month	year	moon phase	severe risk	Uv index	Solar energy	Solar radiation	visibility	Cloud cover	Sea level pressure	Wind dir.	Wind speed	Wind gust
0.037	0.05	0.003	0.003	-0.19	-0.15	-0.1	0.15	0.16	0.015	0.006	0.015	0.006

Precip cover	Precip probability	precip	humidity	dew	feels like	Feels like min	Feels like max	temp	Temp_min	*Temp_max*
0.61	0.37	1	0.18	0.13	0.1	0.088	0.1	0.069	0.093	0.015

4.2. Proposed Methodology

The study is being conducted of Pakistan weather, and the data used in the analysis came from Visual Crossing, a source of historical and forecast meteorological information. Fig 14 depicts the thinking chart that served as a guide for doing this investigation.

This methodology clearly defines the general modeling workflow, but technical detail is needed to ensure reproducibility, such as clearly defining the hyperparameters of each model, the ratio of training–test split and the validation strategy (K-fold cross-validation), the steps taken for data preprocessing (e.g., handling missing values, feature scaling method, and any approach used for class imbalance), and for hybrid or ensemble models, a clear description of how the weights of the component models were defined and how optimization was applied to the final predictions.

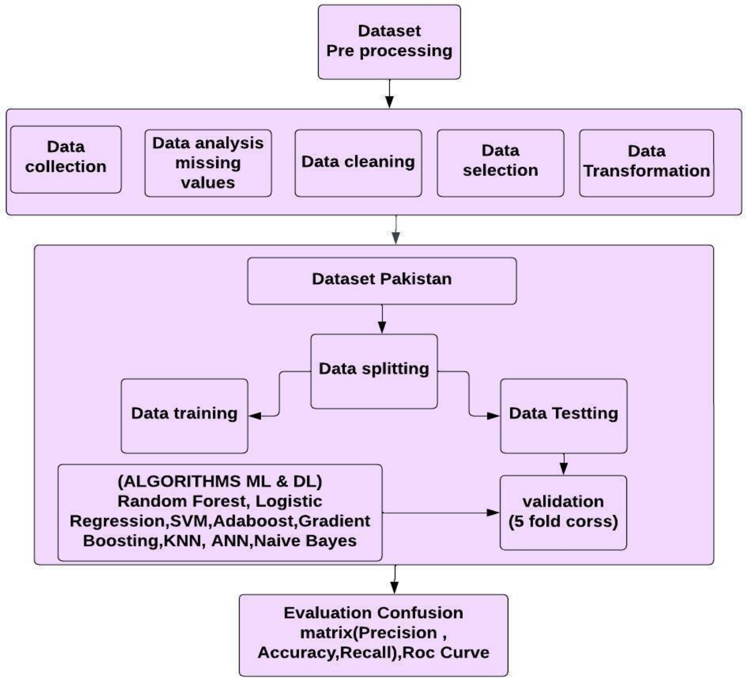


Figure 14. Proposed methodology of the study

4.2.1. Data Collection and Preprocessing

Information gathered from the Visual Crossing website's historical weather data for Pakistan, including variables like humidity, wind speed, temperature, atmospheric pressure, cloud cover, and precipitation.

4.2.2. Feature Selection

Feature selection is a process in machine learning for identifying the most important attributes for rainfall prediction and dropping the unnecessary.

4.2.3. Feature Encoding

In machine learning, feature encoding is the process of transforming non-numeric or categorical data into a numerical format so that algorithms that need numerical input can use it. Some features in this investigation were transformed into numerical

4.2.4. Feature Scaling

Normalization is kind of like pulling some features towards the same level as feature scaling is the process of scaling features. This sought to ensure that the created dataset for the current models was bias free and that the features in the current data set had been standardized utilizing the current feature scaling standards. The equation (12) for feature scaling through the z-score scaling standardization approach is as follows, the feature scaling is done feature by feature.

$$x_{scaled} = \frac{x - \mu}{\sigma} \quad (12)$$

Where

x = original value of the feature,

μ = mean (average) of the feature in the dataset,

σ = standard deviation of the feature in the dataset, and

x_{scaled} = scaled value of the feature

4.2.5. Model Selection and Training

Some of the methodologies used to forecast rainfall were the Naive Bayes, AdaBoost, Logistic Regression, Gradient Boosting Machine (GBM), Decision Tree Classifier (DTC), Random Forest (RF) and Support Vector Machine (SVM). The model performance was checked with the help of cross-validation which separated the dataset into two parts namely the training and the testing dataset. The data was divided into five folds and all the models were trained and validated 5 times- each fold must act as a validation and rest four as a training set. The mean cross-validation results were compared to find out the most effective model in predicting the rainfall.

Each model was manually tuned with the help of a Grid Search approach that was run over a five-fold cross-validation. The number of estimators, learning rate, maximum depth of the tree and the kernel functions (where applicable) were individually optimized in each algorithm to avoid overfitting and fairness. This

4.2.6. Assessment Statistics.

This experimental study utilized some metrics to evaluate the efficacy of the model assessment measures for binary classification tasks, including recall, accuracy, precision, F1-score, and ROC-AUC.

4.2.7. Model Evaluation and Validation

Utilize confusion matrices and ROC curves to visualize model performance and assess classification thresholds.

5. Novelty of Work

This paper introduces a locally optimized, data-intensive rainfall prediction system in Pakistan with a special interest in Karachi, as a representative urban area. Unlike the previous analyses that utilized the models that were generalized or region-agnostic, the research derives and analyses seven state-of-the-art machine learning models, including those of Random Forest, Gradient Boosting, AdaBoost, Decision Tree, Logistic Regression, KNN and SVM, that were trained and tested on high-resolution meteorological data acquired in the Visual Crossing dataset. Multi-parameter tuning of the suggested models minimizes the bias and variance to make the models robust and comparable. The biggest innovation in this is the weaving of 33 different meteorological variables, such as temperature, humidity, dew point, pressure, wind speed, wind direction, solar radiation, and soil moisture, into a single dataset, which allows predicting the occurrence of rainfall with even greater accuracy over time.

Furthermore, incorporation of new variables like solar radiation and soil moisture provide an extension of the feature space typically employed in rainfall prediction across Pakistan, providing a more detailed picture on the behavior of the atmosphere. The statistical reliability and generalization performance of each of the models are validated using rigorous validation measures such as fivefold cross-validation, precision-recall and ROC-AUC measures. The findings reveal that ensemble models especially Random Forest and Gradient Boosting are more stable and interpretable in the prediction than the conventional classifiers (Logistic Regression and Multivariate Linear Regression).

Also, a Bidirectional Long Short-Term Memory (BiLSTM) deep learning model was included, which included sequential temporal dependencies of rainfall data, and the test accuracy was 99.8%. This increases the resilience of deep learning solutions to complement machine learning solutions to meteorological predictions. The study also theoretically constructs a prototype early-warning system that incorporates predictive models with flood preparedness communication systems, which provide a theoretical base on how it can be implemented practically. In general, the analysis adds a scalable analytical framework of climate resilience and adaptive weather forecasts to apply in Pakistan and other semi-arid urban areas.

6. Results and Analysis

6.1. Machine Learning Models Results

Figure 15 illustrates the general steps like selection of column (feature selection) involved then creating a model for prediction. The model predictor utilize in this study are the (RF) Random Forest gives accuracy 0.997,(GB) Gradient Boosting with accuracy 0.994 , AdaBoost(AB) accuracy rate (0.996) ,(SVM)Support Vector Machine with accuracy rate 0.500(50%), (NB)Naïve Bayes provide over fitting in result producing 100%

accuracy rate, and (KNN) k-Nearest Neighbor algorithms with accuracy rate 0.907. best result provided for rainfall prediction Adaboost and random forest

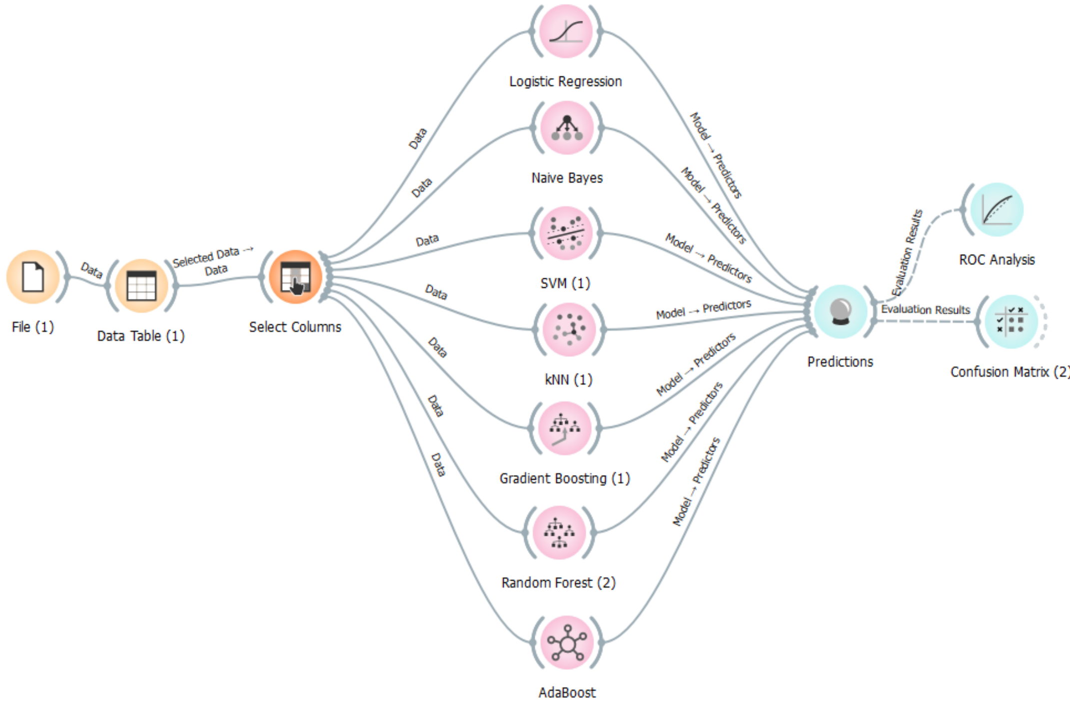


Figure 15. Data Mining Software's Classification Process (Orange Ver. 3.23.1)

To determine the model's reliability, the outcome of the model of each approach is verified using a small amount of input data. To determine which method is the best, the test results are also compared to obtain the highest precision value. In this experiment during the testing phase, the collection of information is divided into two distinct groups: testing data and training data. Eighty percent (80%) of the training data is used in a mining operation to get likelihood values, and the remaining twenty percent (20%) of test data is used to validate the probabilistic values that have been produced.

6.2. Matrix of Confusion

When assessing how good binary classification models perform, such as the one that predicts whether it will rain or not, the ROC curve and confusion matrix are essential tools. The accuracy, precision, and recall of the exam results are assessed using a confusion matrix test. The objective of the test results is to evaluate the accuracy and Area under the Curve (AUC) of the 10-fold Cross Validation procedure. For each algorithm, the test results are displayed in figures 16 through 22 below.

Actual \ Predicted	no rain	rain	Σ
no rain	3567.6	579.4	4147
rain	548.7	82.3	631
Σ	4116	662	4778

Figure 16. Confusion Matrix for Logistic Regression

The findings shown in Fig 16 demonstrate that Logistic regression algorithm yields an accuracy rate of 35%. Of the total datasets analyzed (4778 datasets), 3649 datasets have valid predictions.

The findings shown in Fig 17 demonstrate that Naïve Bayes algorithm yields an accuracy rate of 100%. Of the total datasets analyzed (4778 datasets), 4776 datasets have valid predictions.

The findings shown in Fig 18 demonstrate that the Random algorithm yields an accuracy rate of 99%. Of the total datasets analyzed (4778 datasets), 4772 datasets have valid predictions.

Actual \ Predicted	no rain	rain	Σ
no rain	4147.0	0.0	4147
rain	2.0	629.0	631
Σ	4149	629	4778

Figure 17. Confusion Matrix results for Naïve Bayes

Actual \ Predicted	no rain	rain	Σ
no rain	4144.3	2.7	4147
rain	2.5	628.5	631
Σ	4147	631	4778

Figure 18. Confusion Matrix results for Random forest

Actual \ Predicted	no rain	rain	Σ
no rain	3599.3	547.7	4147
rain	547.7	83.3	631
Σ	4147	631	4778

Figure 19. Confusion Matrix results for Support Vector machine

The findings shown in Fig 19 demonstrate that SVM algorithm yields an accuracy rate of 50%. Of the total datasets analyzed (4778 datasets), 3682 datasets values have valid predictions

Actual \ Predicted	no rain	rain	Σ
no rain	3865.2	281.8	4147
rain	281.0	350.0	631
Σ	4146	632	4778

Figure20. Confusion Matrix results for K nearest neighbors

The findings shown in Fig 20 demonstrate that KNN algorithm yields an accuracy rate of 90%. Of the total datasets analyzed (4778 datasets rows), 4215 datasets values have valid predictions

Actual \ Predicted	no rain	rain	Σ
no rain	4145.4	1.6	4147
rain	2.0	629.0	631
Σ	4147	631	4778

Figure 21. Confusion Matrix results for Gradient Boosting Algorithms

The findings shown in Fig 21 demonstrate that Gradient Boosting algorithm yields an accuracy rate of 99%. Of the total datasets analyzed (4778 datasets rows), 4774 datasets values have true predictions

Actual \ Predicted	no rain	rain	Σ
no rain	4144.0	3.0	4147
rain	2.0	629.0	631
Σ	4146	632	4778

Figure 22. Confusion Matrix results for Ada boost

The findings shown in Fig 22 demonstrate that AdaBoost algorithm yields an accuracy rate of 99%. Of the total datasets analyzed (4778 datasets rows), 4773 datasets values have true predictions.

The classifier accuracy, or classification accuracy, is a key metric for evaluating the performance of a classification model. In fact, accuracy is a measure of the degree to which the predicted and actual values are close. In a binary classification scenario (such as predicting rain or not), accuracy is simply the number of cases in the dataset divided by the number of correctly identified examples (true positives and true negatives). The results highlight the precision, recall, accuracy, and AUC values for each test, as shown in Table 8, which shows the maximum accuracy results for the naïve Bayes, random forest, and AdaBoost tests at 100%, 0.997%, and 0.996, respectively, and the lowest precision values of 0.99, 0.998, and 0.998.

On the validation data, Naïve Bayes attained 100% accuracy, indicating overfitting because of the independence assumption. Its probabilistic assumptions restrict generality for coupled meteorological characteristics, despite its effective performance on straightforward datasets.

The comparative findings show that ensemble classifiers like Random Forest, Gradient Boosting, and AdaBoost are superior to the conventional classifiers because of their innate capacities to embrace nonlinear associations among climatic predictors and to reduce variance by means of aggregation of numerous weak learners. Their tree-like design enables them to capture complex interactions between features, e.g. between temperature, humidity, and wind direction, which cannot be effectively captured by linear models such as Logistic Regression or SVM. This is why the accuracy of ensemble models and their ability to recall and to be used as ROC are always greater, which proves that they are capable of meteorological data that are heterogeneous.

Table 8. Table of test results for accuracy values

S.N	Models Name	AUC	CA	F1	Precision	Recall
1	Logistic Regression	0.357	0.859	0.794	0.738	0.859
2	Random Forest	0.997	0.999	0.999	0.999	0.999
3	Naive Bayes	1.000	0.999	0.999	0.999	0.999
4	Support vector machine	0.500	0.141	0.035	0.020	0.141
5	k Nearest neighbors	0.907	0.909	0.900	0.901	0.909
6	Gradient boosting	0.994	0.998	0.998	0.998	0.998
7	Adaboost Classifier	0.996	0.997	0.997	0.997	0.997

In table 8 Support Vector Machine (SVM) and logistic regression (LR) models performed relatively worse than ensemble methods because their linear decision bounds cannot capture nonlinear correlations, which are prevalent in climatic variables such as temperature–humidity interactions and wind–pressure coupling. Ensemble methods such as Random Forest and AdaBoost, which utilize multiple tree-based learners, can model complex feature dependencies and reduce variance.

The results of this study show that a few models achieved near-perfect levels of accuracy, including Random Forest and Naive Bayes, but the near-perfect accuracy of the Naive Bayes model indicates that it may have over fit. Naive Bayes will be less flexible to accommodate new rainfall patterns that are not like those in the training set, but Random Forest is an ensemble technique that averages predictions over a set of decision trees, and is therefore less prone to overfit even with high levels of accuracy. Moreover, Gradient Boosting and

Adaboost also have high metrics, are more reliable across various measures, and are suitable for prediction jobs where data can be varied a lot, because they do not overfit as much as Naive Bayes does for complex patterns. Lower performing models such as Support Vector Machine and Logistic Regression struggled with predictive accuracy, which suggests that they may not be as well-equipped as ensemble approaches to deal with the complexity and non-linearity of the rainfall data. K Nearest Neighbors performed relatively well compared to the more reliable ensemble approaches, but can be influenced by noise and size of dataset.

6.3. ROC Curve Machine Learning Models

The ROC curve plots the true positive rate (TPR) on the y-axis versus the false positive rate (FPR) on the x-axis across different threshold settings to show the trade-off between the true positive rate (sensitivity) and false positive rate (1 - specificity) of a binary classification model. The ROC curve is a useful tool to evaluate the discriminant power of the machine learning algorithms, and the results are shown in Figures 23 and 24. Therefore, the ROC curve is informative about the performance of the model over the range of thresholds, and it is a two-dimensional graph with the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis.

The Random Forest model outperformed all other algorithms in achieving the most balanced trade-off between sensitivity and specificity, suggesting that it is appropriate for rainfall-related decision-support systems where both erroneous positives and false negatives have operational significance.

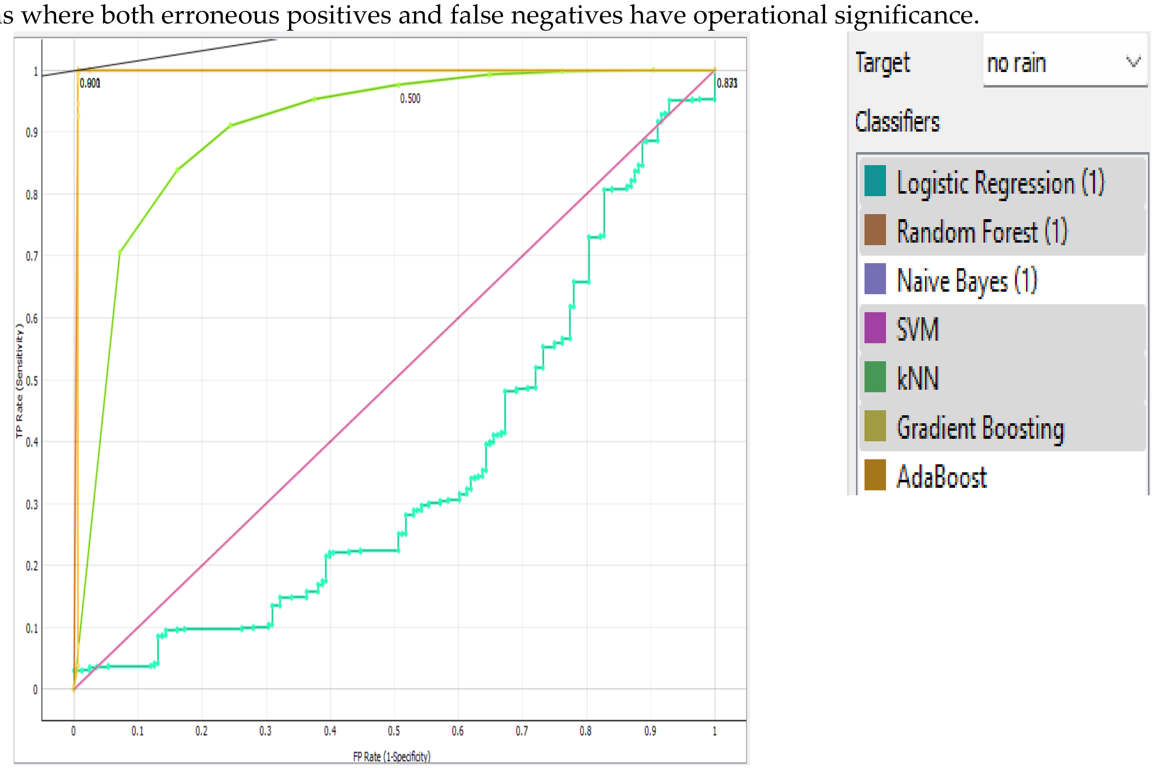


Figure 23. ROC values of Algorithms for target no rain

6.4. Deep Learning Models Results

The Bidirectional Long Short-Term Memory (BiLSTM) was a model that performed exceptionally well in the rainfall classification. The model had a test accuracy of 99.89 with a test loss of 0.0250 after 80 epochs with the use of data balancing and SMOTE alongside potent regularization approaches, including dropout, batch regularization, Gaussian noise, and L2 regularization. The near-perfect accuracy of the classes: rain and no rain prove the strength of the model. The fact that the training and validation curves fit perfectly points to the fact that the training did not overfit, and the distribution of the confidences shows very credible predictions. These results emphasize that the BiLSTM model is quite useful in capturing temporal dependencies and non-linear meteorological patterns and it is more accurate and generalized than traditional machine learning models.

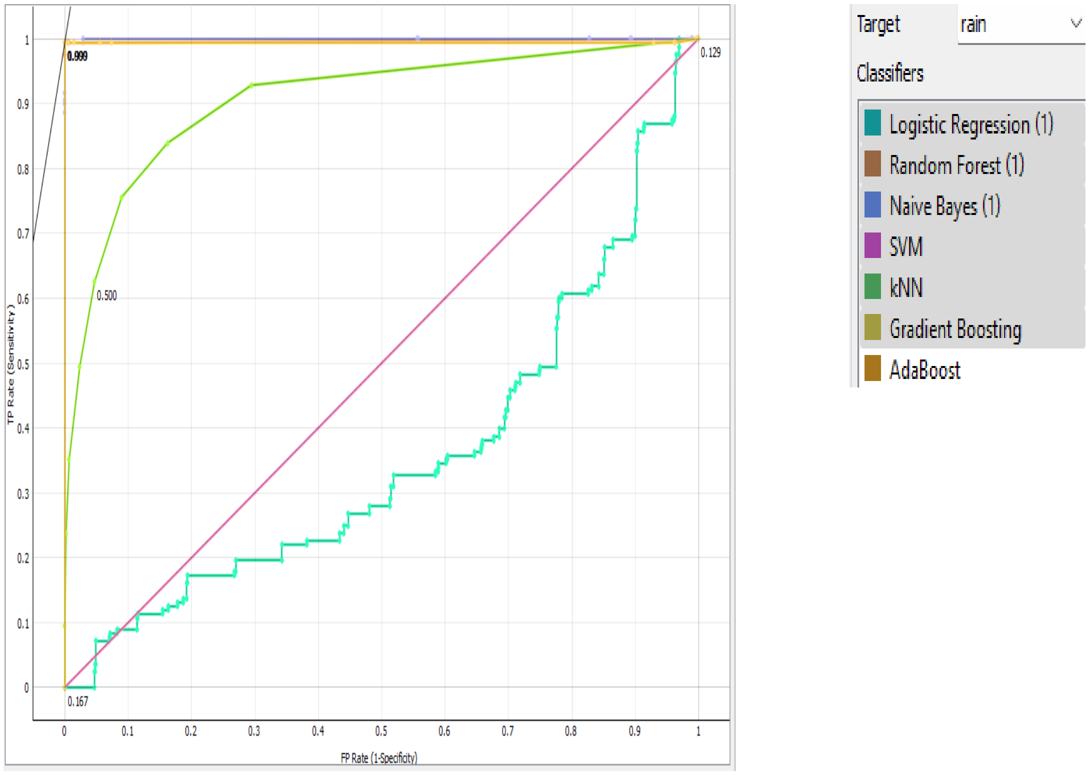


Figure 24. ROC values of Algorithms for target rain

Table 9. Deep learning Models result table

Parameter / Metric	Value
Model Type	BiLSTM
Data Split	80% Train / 20% Test
Balancing Method	SMOTE
Optimizer	Adam (LR = 0.0002)
Loss Function	Categorical Cross-Entropy
Epochs / Batch Size	80 / 32
Regularization	Dropout (0.5–0.3), L2(0.001), BatchNorm
Train Accuracy	99.92%
Validation Accuracy	99.89%
Test Accuracy	99.89%
Test Loss	0.0250

6.5. Criterion for Evaluating Model Confusion Matrix

The most common metrics in machine learning are performance metrics, or measurements that evaluate how well a model is performing some task; this can be as general as clustering and other tasks, or as specific as classification and regression. The measures chosen are heavily influenced by task characteristics. One of the most commonly used techniques for evaluating classification models is a confusion matrix, a two-by-two table that shows the counts of true positives, true negatives, false positives, and false negatives. The situations that our model correctly identified as positive, which means good positive identification is True Positives (TP). The true negative cases, or correctly identified negative outcomes are the negative predictions. Conversely, Type I errors result in False Positives (FP), which is a positive result on a negative event by the model. Finally, Type II errors are

6.5.1. Accuracy

All it measures is the frequency with which the classifier makes accurate predictions. The ratio of the number of accurate forecasts to the total number of predictions (see equation 13) can be used to determine accuracy.

$$\text{Accuracy}=\frac{number\ of\ correct\ prediction}{total\ number\ of\ predictions} \quad (13)$$

The "Number of Correctly Classified Instances" indicates the number of data records that the algorithm correctly classified. The "Total Number of Instances" parameter indicates the total number of data records in the dataset.

6.5.2. Precision (Positive Predictive Value)

This can explain why so many situations which were accurately predicted were indeed positive. Precision is important in the cases described in equation 14 where false positives are a lot worse than false negatives.

$$\text{Precision: } \frac{TP}{TP+FP} \quad (14)$$

6.5.3. Specificity (True Negative Rate):

The specificity of a diagnosis test is explained by Equation 15 as accurate negative predictions divided by total observed negative cases.

$$\text{Specificity: } \frac{TN}{TN+FP} \quad (15)$$

6.5.4. Recall (Sensitivity, True Positive Rate):

This prescribes how many actual positive cases our model was able to predict with absolute certainty. Recall comes is handy measure whenever False Positive is of higher concern than False Negative as shown in equation 16.

$$\text{Sensitivity: } \frac{TP}{TP+FN} \quad (16)$$

6.5.5. Matthews Correlation Coefficient (MCC)

The MCC provides a fair assessment of classification performance, especially when dealing with datasets that are unbalanced and have a large difference in the proportion of positive and negative samples shown in equation 17.

$$\text{MCC}\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)}(TP+FN)(TN+FP)(TN+FN)} \quad (17)$$

7. Limitation of Work

There are some limitations in this study, despite positive performance results observed in the proposed models. The models might not be effective to generalize to other areas in Pakistan because of time and space limits in the dataset. Consistency of prediction can also be different with variation in quality of data, resolution as well as local climatic variability.

Moreover, the models do not use dynamical features and they are not connected with real-time sensor data and large scale meteorological indices like ENSO or NAO. Whilst the models have been successful in the area of training, their applicability in new time patterns and invisible climatic areas is still unclear.

Real-time rainfall forecasting would require automation of the data ingestion process, retraining of the model, and frequent calibration of the model with satellite- and IoT-collected meteorological data streams to be practically deployed. The future research to improve the operational reliability of the model should thus focus on multi-source dataset construction and the implementation of more sophisticated hybrid optimization methods.

8. Conclusion

The main aim of this research was to create a correct and effective machine learning and deep learning -based rain classification model using meteorological data of the Visual Crossing weather portal in relation to Pakistan. The sample size of 4,778 observations and 33 climatic variables (data) was used to train and test seven different sophisticated machine learning models to predict rainfall. The results show that machine learning algorithms could be successfully used to form a meteorological data with clear categorical characteristics that are identified in the form of Receiver Operating Characteristic (ROC) curves and Confusion Matrix. Random Forest was the most accurate with 0.997 followed by AdaBoost (0.996) and Gradient Boosting (0.994), which validated their robust predictor variables of rainfall in the short term. Although it was not as precise, the K-

Nearest Neighbor (0.907) still exhibited potential practicality in rainfall classification. These findings confirm that ensemble learning approaches are better than the classical classifiers, including the Logistic Regression, Support Vector Machine, and Naive Bayes, in solving prediction problems related to weather. Random Forest model, especially, was very reliable and robust and thus it is an appropriate choice in operational weather forecasting. Notably, a deep learning-based Bidirectional LSTM (BiLSTM) model was created and tested as well. It had a test accuracy of 99.89% with the least loss meaning that it was learning at an exemplary level of stability and prediction. It proves that by combining deep learning architectures with machine learning ensembles, the accuracy of rainfall prediction can be further improved, and this approach will become a promising way to improve climate analytics and operational weather forecasting systems in the future. The next direction in work will be the creation of a completely integrated AI-based early warning system that will integrate predictive modeling with real-time communication infrastructure to warn communities in rain-prone areas. This system would improve preparedness against disasters, improve the effects of floods, and promote climate resilience measures throughout Pakistan.

Abbreviations:

KNN: K-Nearest Neighbors
GBM: Gradient Boosting Machine
NB: Naïve Bayes
RF: Random Forest
LR: Logistic Regression
ROC: Receiver Operating Characteristic
FPR: False Positive Rate
TPR: True Positive Rate
SVM: Support vector machines
BiLSTM: Bidirectional Long Short Term memory

References:

1. A. M. Bagirov and A. Mahmood, ''A comparative assessment of models to predict monthly rainfall in Australia,'' Water Resour. Manag., vol. 32, no. 5, pp. 1777–1794, Mar. 2018.
2. Hayder, I. M., Al-Amiedy, T. A., Ghaban, W., Saeed, F., Nasser, M., Al-Ali, G. A., & Younis, H. A. (2023). An Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Advanced Alert System. Processes, 11(2), 481.
3. Mashao, F. M., Mothapo, M. C., Munyai, R. B., Letsoalo, J. M., Mbokodo, I. L., Muofhe, T. P., ... & Chikoore, H. (2023). Extreme rainfall and flood risk prediction over the East Coast of South Africa. Water, 15(1), 50.
4. Pham, B. T., Le, L. M., Le, T. T., Bui, K. T. T., Le, V. M., Ly, H. B., & Prakash, I. (2020). Development of advanced artificial intelligence models for daily rainfall prediction. Atmospheric Research, 237, 104845.
5. Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2018). Rainfall prediction for the Kerala state of India using artificial intelligence approaches. Computers & Electrical Engineering, 70, 66-73.
6. Salehin, I., Talha, I. M., Hasan, M. M., Dip, S. T., Saifuzzaman, M., & Moon, N. N. (2020, December). An artificial intelligence based rainfall prediction using LSTM and neural network. In 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 5-8). IEEE.
7. M. S. Balamurugan and R. Manojkumar, ''Study of short term rain forecasting using machine learning based approach,'' Wireless Netw., vol. 27, no. 8, pp. 5429–5434, Nov. 2021.
8. Gu, J., Liu, S., Zhou, Z., Chalov, S. R., & Zhuang, Q. (2022). A stacking ensemble learning model for monthly rainfall prediction in the Taihu Basin, China. Water, 14(3), 492.
9. Liyew, C. M., & Melese, H. A. (2021). Machine learning techniques to predict daily rainfall amount. Journal of Big Data, 8, 1-11.
10. Hussein, E. A., Ghaziasgar, M., Thron, C., Vaccari, M., & Jafta, Y. (2022). Rainfall Prediction Using Machine Learning Models: Literature Survey. Artificial Intelligence for Data Science in Theory and Practice, 75-108.
11. Chang-Jiang Zhang, Jing Zeng, Hui-Yuan Wang, Lei-Ming Ma, and Hai Chu. Correction model for rainfall forecasts using the Lstm with multiple meteorological factors. Meteorological Applications, 27(1):e1852, 2020.
12. Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review Water, 10(11), 1536.
13. Sankaranarayanan, S., Prabhakar, M., Latish, S., Jain, P., Ramprasad, A., & Krishnan, A. (2020). Flood prediction based on weather parameters using deep learning. Journal of Water and Climate Change, 11(4), 1766-1783.
14. Hammad, M., Shoaib, M., Salah din, H., Baig, M. A. I., Khan, M. M., & Ullah, M. K. (2021). Rainfall forecasting in upper Indus basin using various artificial intelligence techniques. Stochastic Environmental Research and Risk Assessment, 35, 2213-2235.
15. Ali, G., Sajjad, M., Kanwal, S., Xiao, T., Khalid, S., Shoaib, F., & Gul, H. N. (2021). Spatial–temporal characterization of rainfall in Pakistan during the past half-century (1961–2020). Scientific reports, 11(1), 6935.
16. Yasmeen, F., & Hameed, S. (2018). Forecasting of rainfall in pakistan via sliced functional times series (SFTS). World Environment, 8(1), 1-14.
17. Rahman, A. U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., ... & Mosavi, A. (2022). Rainfall prediction system using machine learning fusion for smart cities. Sensors, 22(9), 3504.
18. Latif, S. D., Hazrin, N. A. B., Koo, C. H., Ng, J. L., Chaplot, B., Huang, Y. F., ... & Ahmed, A. N. (2023). Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. Alexandria Engineering Journal, 82, 16-25.
19. Yen, M. H., Liu, D. W., Hsin, Y. C., Lin, C. E., & Chen, C. C. (2019). Application of deep learning for the prediction of rainfall in Southern Taiwan. Scientific reports, 9(1), 12774.
20. Diop, L., Samadianfard, S., Bodian, A., Yaseen, Z. M., Ghorbani, M. A., & Salimi, H. (2020). Annual rainfall forecasting using hybrid artificial intelligence model: integration of multilayer perceptron with whale optimization algorithm. Water Resources Management, 34, 733-746.
21. Wu, J., Liu, M., & Jin, L. (2010). A hybrid support vector regression approach for rainfall forecasting using particle swarm optimization and projection pursuit technology. International Journal of Computational Intelligence and Applications, 9(02), 87-104.

22. Aftab, S., Ahmad, M., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction using data mining techniques: A systematic literature review. International journal of advanced computer science and applications, 9(5).
23. Hassan, M. M., Rony, M. A. T., Khan, M. A. R., Hassan, M. M., Yasmin, F., Nag, A., ... & El-Shafai, W. (2023). Machine Learning-Based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness. IEEE Access, 11, 132196-132222.
24. Cholissodin, I., & Sutrisno, S. (2020). Prediction of rainfall using improved deep learning with particle swarm optimization. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(5), 2498-2504.
25. Beheshti, Z., Firouzi, M., Shamsuddin, S. M., Zibarzani, M., & Yusop, Z. (2016). A new rainfall forecasting model using the CAPSO algorithm and an artificial neural network. Neural Computing and Applications, 27, 2551-2565.
26. Nourani, V., Uzelaltinbulat, S., Sadikoglu, F., & Behfar, N. (2019). Artificial intelligence based ensemble modeling for multi-station prediction of precipitation. Atmosphere, 10(2), 80.
27. Basha, C. Z., Bhavana, N., Bhavya, P., & Sowmya, V. (2020, July). Rainfall prediction using machine learning & deep learning techniques. In 2020 international conference on electronics and sustainable communication systems (ICESC) (pp. 92-97). IEEE.
28. Cramer, S., Kampouridis, M., Freitas, A. A., & Alexandridis, A. K. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. Expert Systems with Applications, 85, 169-181.
29. Nguyen, H. N., Nguyen, T. A., Ly, H. B., Tran, V. Q., Nguyen, L. K., Nguyen, M. V., & Ngo, C. T. (2021). Prediction of daily and monthly rainfall using a backpropagation neural network. Journal of Applied Science and Engineering, 24(3), 367-379.
30. Kunverji, K., Shah, K., & Shah, N. (2021, May). A flood prediction system developed using various machine learning algorithms. In Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021).
31. Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications, 7, 100204.
32. Ridwan, W. M., Sapitang, M., Aziz, A., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A. (2021). Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Engineering Journal, 12(2), 1651-1663.
33. Praveen, B., Talukdar, S., Shah Fahad, Mahato, S., Mondal, J., Sharma, P., ... & Rahman, A. (2020). Analyzing trend and forecasting of rainfall changes in India using non-parametric and machine learning approaches. Scientific reports, 10(1), 10342.
34. Shah, U., Garg, S., Sisodiya, N., Dube, N., & Sharma, S. (2018, December). Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 776-782). IEEE.
35. Moharana, L., Sahoo, A., & Ghose, D. K. (2022, October). Prediction of rainfall using hybrid SVM-HHO model. In IOP Conference Series: Earth and Environmental Science (Vol. 1084, No. 1, p. 012054). IOP Publishing.
36. Cheema, M. A., Saqib, M., & Iqbal, S. Z. (2023). Rainfall Prediction using Big Data Analytics: A Systematic Literature Review. International Journal of Computations, Information and Manufacturing (IJCIM), 3(2), 28-36.
37. Dotse, S. Q., Larbi, I., Limantol, A. M., & De Silva, L. C. (2023). A review of the application of hybrid machine learning models to improve rainfall prediction. Modeling Earth Systems and Environment, 1-26.
38. Markuna, S., Kumar, P., Ali, R., Vishwkarma, D. K., Kushwaha, K. S., Kumar, R., ... & Kuriqi, A. (2023). Application of innovative machine learning techniques for long-term rainfall prediction. Pure and Applied Geophysics, 180(1), 335-363.
39. C. D. Usman, A. P. Widodo, K. Adi, and R. Gernowo, "Rainfall prediction model in Semarang City using machine learning," Indonesian Journal of Electrical Engineering and Computer Science, vol. 30, no. 2, pp. 1224–1231, 2023,
40. Baljon, M., & Sharma, S. K. (2023). Rainfall prediction rate in Saudi Arabia using improved machine learning techniques. Water, 15(4), 826.
41. Li, D., Ma, J., Rao, K., Wang, X., Li, R., Yang, Y., & Zheng, H. (2023). Prediction of rainfall time series using the hybrid DWT-SVR-Prophet model. Water, 15(10), 1935.

42. Hatim, M., Siddiqui, F., & Kumar, R. (2020, January). Addressing challenges and demands of intelligent seasonal rainfall forecasting using artificial intelligence approach. In 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM) (pp. 263-267). IEEE.
43. Yu, P. S., Yang, T. C., Chen, S. Y., Kuo, C. M., & Tseng, H. W. (2017). Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. Journal of hydrology, 552, 92-104.
44. Sacré Regis M, D., Mouhamed, L., Kouakou, K., Adeline, B., Arona, D., Houebagnon Saint. J, C., ... & Issiaka, S. (2020). Using the CHIRPS dataset to investigate historical changes in precipitation extremes in West Africa. Climate, 8(7), 84.
45. Kpanou, M., Laux, P., Brou, T., Vissin, E., Camberlin, P., & Roucou, P. (2021). Spatial patterns and trends of extreme rainfall over the southern coastal belt of West Africa. Theoretical and Applied Climatology, 143, 473-487.
46. Salaeh, N., Ditthakit, P., Pinthong, S., Hasan, M. A., Islam, S., Mohammadi, B., & Linh, N. T. T. (2022). Long-Short Term Memory Technique for Monthly Rainfall Prediction in Thale Sap Songkhla River Basin, Thailand. Symmetry, 14(8), 1599.
47. Zaman, M., Shahid, M. A., Ahmad, I., Shen, Y. J., Usman, M., Khan, M. I., & Saifullah, M. (2023). Spatiotemporal variability of precipitation concentration in Pakistan. International Journal of Climatology, 43(16), 7646-7666.
48. Begum, B., Tajbar, S., Khan, B., & Rafiq, L. (2021). Identification of relationships between climate indices and precipitation fluctuation in Peshawar City-Pakistan. Journal of Research in Environmental Earth Sciences, 264-278.
49. Zhang, W., Kou, M., Lv, M., & Shao, Y. (2022). Improved combined system and application to precipitation forecasting model. Alexandria Engineering Journal, 61(12), 12739-12757.
50. He, X., Guan, H., & Qin, J. (2015). A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall. Journal of Hydrology, 527, 88-100.
51. Mishra, N., Soni, H. K., Sharma, S., & Upadhyay, A. K. (2018). Development and analysis of artificial neural network models for rainfall prediction by using time-series data. International Journal of Intelligent Systems and Applications, 12(1), 16.
52. Yulianto, F., Mahmudy, W. F., & Soebroto, A. A. (2020). Comparison of Regression, Support Vector Regression (SVR), and SVR-Particle Swarm Optimization (PSO) for Rainfall Forecasting. Journal of Information Technology and Computer Science, 5(3), 235-247.
53. Mohamadi, S., Sheikh Khozani, Z., Ehteram, M., Ahmed, A. N., & El-Shafie, A. (2022). Rainfall prediction using multiple inclusive models and large climate indices. Environmental Science and Pollution Research, 29(56), 85312-85349.
54. Rajkumar, K. V., & Subrahmanyam, K. (2021). A novel method for rainfall prediction and classification using neural networks. International Journal of Advanced Computer Science and Applications, 12(7).
55. Sani, N. S., Abd Rahman, A. H., Adam, A., Shlash, I., & Aliff, M. (2020). Ensemble learning for rainfall prediction. International Journal of Advanced Computer Science and Applications, 11(11).
56. Hussein, E. A., Ghaziasgar, M., Thron, C., Vaccari, M., & Bagula, A. (2021). Basic statistical estimation outperforms machine learning in monthly prediction of seasonal climatic parameters. Atmosphere, 12(5), 539.
57. Kakade, S., & Kulkarni, A. (2016). Prediction of summer monsoon rainfall over India and its homogeneous regions. Meteorological Applications, 23(1), 1-13.
58. Hasanean, H., & Almazroui, M. (2015). Rainfall: features and variations over Saudi Arabia, a review. Climate, 3(3), 578-626.
59. Kumar, V., Kedam, N., Sharma, K. V., Khedher, K. M., & Alluqmani, A. E. (2023). A comparison of machine learning models for predicting rainfall in urban metropolitan Cities. Sustainability, 15(18), 13724.
60. Rajab, A., Farman, H., Islam, N., Syed, D., Elmagzoub, M. A., Shaikh, A., & Alrizq, M. (2023). Flood Forecasting by Using Machine Learning: A Study Leveraging Historic Climatic Records of Bangladesh. Water, 15(22), 3970.
61. Xu, Y.; Hu, C.; Wu, Q.; Jian, S.; Li, Z.; Chen, Y.; Zhang, G.; Zhang, Z.; Wang, S. Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation. J. Hydrol. 2022, 608, 127553
62. Dtissibe, F.Y.; Ari, A.A.A.; Titouna, C.; Thiare, O.; Gueroui, A.M. Flood forecasting based on an artificial neural network scheme. Nat. Hazards 2020, 104, 1211–1237
63. Aswad, F.M.; Kareem, A.N.; Khudhur, A.M.; Khalaf, B.A.; Mostafa, S.A. Tree-based machine learning algorithms in the Internet of Things environment for multivariate flood status prediction. J. Intel. Syst. 2021, 31, 1–14.

64. Aydin, M.C.; Sevgi Birincio ˘glu, E. Flood risk analysis using GIs-based analytical hierarchy process: A case study of Bitlis Province. Appl. Water Sci. 2022, 12, 122.
65. B. N. Lakshmi, "A Comparative Study of Classification Algorithms for Risk Prediction in Pregnancy," pp. 0–5, 2015
66. S. S. Bhatkande and R. G. Hubballi, "Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques," vol. 5, no. 5,pp. 483–487, 2016.
67. Prasetya, R., & Ridwan, A. (2019). Data mining application on weather prediction using classification tree, naïve bayes and K-nearest neighbor algorithm with model testing of supervised learning probabilistic brier score, confusion matrix and ROC. J. Appl. Commun. Inf. Technol, 4(2), 25-33
68. Wang, X., H. Xie, H. Sharif, and J. Zeitler, 2008. Validating NEXRDA MPE and Stage III precipitation products for uniform rainfall on the Upper Guadalupe River Basin of the Texas Hill Country. Journal of Hydrology, Vol. 348(1-2): 73-86, doi: 10.1016/j.jhydrol.2007.09.057
69. Richardson, L. F. Lewis Fry Richardson.
70. Fekadu, K. (2015). Ethiopian seasonal rainfall variability and prediction using canonical correlation analysis (CCA). Earth Sci, 4(3), 112-119.
71. Shabib Aftab, M. A., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction in Lahore City using data mining techniques. International journal of advanced computer science and applications, 9(4).
72. M. A. Shadiq, "Keoptimalan Naïve Bayes Dalam Klasifikasi," no. 1, p. 31, 2009.
73. Monego, V. S., Anochi, J. A., & de Campos Velho, H. F. (2022). South America seasonal precipitation prediction by gradient-boosting machine-learning approach. Atmosphere, 13(2), 243.
74. Mishra, N.; Soni, H.K.; Sharma, S.; Upadhyay, A.K. Development and analysis of artificial neural network models for rainfall prediction by using time-series data. Int. J. Intell. Syst. Appl. 2018, 12, 16.
75. Deka, P. C. (2014). Support vector machine applications in the field of hydrology: a review. Applied soft computing, 19, 372-386.
76. Kumar, P. S., & Swathi, M. N. (2023). Rain Fall Prediction using Ada Boost Machine Learning Ensemble Algorithm. JOURNAL OF ADVANCED APPLIED SCIENTIFIC RESEARCH, 5(4), 67-81.
77. Praveena, R., Babu, T. G., Birunda, M., Sudha, G., Sukumar, P., & Gnanasoundharam, J. (2023, March). Prediction of Rainfall Analysis Using Logistic Regression and Support Vector Machine. In Journal of Physics: Conference Series (Vol. 2466, No. 1, p. 012032). IOP Publishing.
78. Javadinejad, S. (2022). Causes and consequences of floods: flash floods, urban floods, river floods and coastal floods. Resources Environment and Information Engineering, 4(1), 173-183.
79. Hajat, S., Ebi, K. L., Kovats, R. S., Menne, B., Edwards, S., & Haines, A. (2005). The human health consequences of flooding in Europe: a review. Extreme weather events and public health responses, 185-196.
80. Wei, L., Hu, K. H., & Hu, X. D. (2018). Rainfall occurrence and its relation to flood damage in China from 2000 to 2015. Journal of Mountain Science, 15(11), 2492-2504.
81. Svetlana, D., Radovan, D., & Ján, D. (2015). The economic impact of floods and their importance in different regions of the world with emphasis on Europe. Procedia Economics and Finance, 34, 649-655.
82. Sholihah, Q., Kuncoro, W., Wahyuni, S., Suwandi, S. P., & Feditasari, E. D. (2020, February). The analysis of the causes of flood disasters and their impacts in the perspective of environmental law. In IOP conference series: earth and environmental science (Vol. 437, No. 1, p. 012056). IOP Publishing.
83. Dun, O. (2011). Migration and displacement triggered by floods in the Mekong Delta. International Migration, 49, e200-e223.
84. Alam, F., Salam, M., Khalil, N. A., khan, O., & Khan, M. (2021). Rainfall trend analysis and weather forecast accuracy in selected parts of Khyber Pakhtunkhwa, Pakistan. SN Applied Sciences, 3, 1-14.
85. Kumar, V.; Azamathulla, H.M.; Sharma, K.V.; Mehta, D.J.; Maharaj, K.T. The state of the art in deep learning applications, challenges, and future prospects: A comprehensive review of flood forecasting and management. Sustainability 2023, 15, 10543.
86. Liyew, C. M., & Melese, H. A. (2021). Machine learning techniques to predict daily rainfall amount. Journal of Big Data, 8, 1-11.

87. Chitwatkulsiri, D.; Miyamoto, H. Real-Time Urban Flood Forecasting Systems for Southeast Asia A Review of Present Modelling and Its Future Prospects. Water 2023, 15, 178.
88. Farman, H., Khan, D., Hassan, S., Hussain, M., & Usmani, S. A. A. (2024). Analyzing Machine Learning Models for Forecasting Precipitation in Australia. Journal of Computing & Biomedical Informatics, 7(01), 439-458.
89. Farman, H., Islam, N., Ali, S. A., Khan, D., Khan, H. A., Ahmed, M., & Farman, A. (2024). Advancing Rainfall Prediction in Pakistan: A Fusion of Machine Learning and Time Series Forecasting Models. International Journal of Emerging Engineering and Technology, 3(1), 17-24.
90. Farman, H., Khan, A. W., Ahmed, S., Khan, D., Imran, M., & Bajaj, P. (2024). An Analysis of Supervised Machine Learning Techniques for Churn Forecasting and Component Identification in the Telecom Sector. Journal of Computing & Biomedical Informatics, 7(01), 264-280.
91. Hayder, I. M., Al-Amiedy, T. A., Ghaban, W., Saeed, F., Nasser, M., Al-Ali, G. A., & Younis, H. A. (2023). An Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Advanced Alert System. Processes, 11(2), 481.
92. Mashao, F. M., Mothapo, M. C., Munyai, R. B., Letsoalo, J. M., Mbokodo, I. L., Muofhe, T. P., ... & Chikoore, H. (2023). Extreme rainfall and flood risk prediction over the East Coast of South Africa. Water, 15(1), 50.
93. Liyew, C. M., & Melese, H. A. (2021). Machine learning techniques to predict daily rainfall amount. Journal of Big Data, 8, 1-11.
94. Nguyen, H. N., Nguyen, T. A., Ly, H. B., Tran, V. Q., Nguyen, L. K., Nguyen, M. V., & Ngo, C. T. (2021). Prediction of daily and monthly rainfall using a backpropagation neural network. Journal of Applied Science and Engineering, 24(3), 367-379.
95. Dotse, S. Q., Larbi, I., Limantol, A. M., & De Silva, L. C. (2023). A review of the application of hybrid machine learning models to improve rainfall prediction. Modeling Earth Systems and Environment, 1-26.
96. Markuna, S., Kumar, P., Ali, R., Vishwkarma, D. K., Kushwaha, K. S., Kumar, R., ... & Kuriqi, A. (2023). Application of innovative machine learning techniques for long-term rainfall prediction. Pure and Applied Geophysics, 180(1), 335-363
97. Baljon, M., & Sharma, S. K. (2023). Rainfall prediction rate in Saudi Arabia using improved machine learning techniques. Water, 15(4), 826.
98. He, X., Guan, H., & Qin, J. (2015). A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall. Journal of Hydrology, 527, 88-100.
99. Farman, H., Islam, N., & Hasany, N. (2024). Harnessing AI for Enhanced Weather Prediction: A Focus on Rainfall. Journal of Independent Studies and Research Computing, 22(2), 124-150.
100. Hasany, N., Salman, H., & Farman, A. (2024). Evaluating Rainfall Prediction Models on Time Series Data from Bangladesh and Pakistan: A Comparative Approach. Journal of Computing & Biomedical Informatics, 7(02).
101. Farman, H., Hussain, M. A., Hassan, S., Shaikh, S., & Ali, K. (2025). Activation function impact on rainfall prediction: comparative insights across ML and DL architectures. Modeling Earth Systems and Environment, 11(6), 451.