International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 3, Number: 1, Pages: 17 - 24, Year: 2024

Advancing Rainfall Prediction in Pakistan: A Fusion of Machine Learning and Time Series Forecasting Models

Hira Farman[1], Noman Islam[2], Syed Akhmas Ali[2], Dodo Khan[3], Hassan Ali Khan[2], Moiz Ahmed[2], Alisha Farman[2]

[1] Department of Computer Science, IQRA University, Karachi Institute of Economics and Technology, Pakistan

[2] Department of Computer Science, IQRA University, Karachi, Pakistan

[3] Thar Institute of Engineering Science and Technology, Mithi, Pakistan

*Corresponding Author: Hira Farman. (Email: hira.farman@iqra.edu.pk; 14307@kiet.edu.pk)

***Abstract*— This study brings about an innovative approach as rainfall forecast predominantly Boolean calculation is reviewed overall for 6 major cities of Pakistan for the time span of the last quarter century. The study aims to improve the accuracy and reliability of rainfall forecasting by using the capabilities of Artificial Intelligence (AI). This research investigates the efficacy of various machine learning models, including Naive Bayes, Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbors (KNN), Gradient Boosting, and time series forecasting model ARIMA (Autoregressive Integrated Moving Average), for rainfall prediction. The model gets trained using 20 years of Pakistan's historical data performance, where improving precision is the objective. With a highly scientific evaluation against real-world datasets, the new approach shows remarkable improvements in the accuracy of rainfall prediction compared to other conventional methods. Machine learning model KNN and time series forecasting model ARIM provide good results regarding higher accuracy and lower RMSE. This results in combining environmental data science for the innovation of meteorological forecasting in this polarized area, where the judgments are inaccurate.**

Keywords*— *rain forecast; predictive modelling; machine learning; random forest; time series forecasting

I. Introduction

Rainfall prediction is among the most significant and intriguing tasks in the 21st century. Climate and precipitation are typically highly complex, non-linear phenomena that require advanced computer mathematical modelling and simulation to be effectively forecasted. This is because precipitation shows both temporal and spatial relationships. Weather forecasting is becoming an increasingly essential field of research due to the increase in rainfall-related flood disasters in recent decades. The researcher often sought to establish a linear relationship between the required data and the available meteorological data [1]. The exacerbation by the impact of climate change on rainfall patterns worldwide, which are diversified and hardly predictable, alarmingly increases the importance of reasonable and corrective weather predictions. There needs to be more than the classical methods, such as empirical equations and basic linear models, to understand the complex dynamics of hydrological systems. These restraints are often the reason for making forecasting mistakes, which hinder such activities as promoting water management, agriculture planning, and commuting, among other things. Spatial and temporal rainfall [4] variations in Bangladesh are examined by making Make Sense and multilayer perceptron neural networks to predict the national rainfall accurately. Using machine learning algorithms and spatial models, this study is expected to enhance the efficiency and reliability of rainfall forecasting at the national level, thereby improving the preparedness and management of their resources in sectors such as agriculture, disaster preparation, and water resources.

The hydrological prediction niche soon became an arena for the arrival of machine learning and artificial intelligence based on data. These cutting-edge methods open up a new field that will help us get more accurate and reliable rainfall forecasts. As demonstrated by studies, algorithms that meet the criteria, such as LSTM networks and Random Forest, exercise effectiveness in several hydrological forecasting tasks. This research used advanced techniques to design a rainfall prediction model for Pakistan. This is a big problem because Guyana is very sensitive to the negative consequences of climate change on rainfall variability.

Our study confronts this void by formulating and assessing an extremely advanced and great pair of hybrid ML-AI rainfall forecast models well-matched to Pakistan's existing climate and geographical conditions. Since our model consists of LSTM networks built into a well-designed system, Random Forest algorithms and advanced time series forecasting techniques are combined. This method, a combination of all three techniques, is intended to strengthen these techniques and make the accuracy, robustness, and applicability of rainfall predictions in six major cities of Pakistan over 24 years. The research intends to revamp rainfall forecasting in Pakistan through the advantages of machine learning and artificial intelligence. The high forecast accuracy that could be achieved with this system is revolutionary. It allows water managers to develop more advanced strategies and farmers to better use water. Thus, the model serves to inform future efforts for urban planning. In addition, the study is a part of the process of building a knowledge base that can be used by policymakers, researchers, and stakeholders to make decisions to mitigate the impacts of climate change on water and agriculture in Pakistan. This study [26] employed a deep neural network to predict the probability of flooding based on temperature and rainfall. To reliably predict rainfall using just historical rainfall records, a unique Wavelet-coupled Multi-order Time Lagged Neural Network (WMTLNN) model is built [27]. This study's main objective is to notify individuals to take preventive actions, evacuate when necessary, and safeguard lives and property. Artificial Intelligence (AI) enters the picture in this situation. Artificial Intelligence (AI) has emerged as a powerful tool in enhancing the accuracy and efficiency of rainfall forecasting for flood awareness. To increase their application to critically important live real-time early warning systems, such as those for flood or bathing water quality risks and a wider range of predictive modelling scenarios, the overall goal is to contribute to understanding artificial intelligence as a predictive technique. Predicting when it will rain is difficult because the outcomes need to be precise. Numerous hardware tools are available for predicting rainfall based on weather circumstances, such as maximum and minimum temperatures, relative humidity at 7 a.m. and 2 p.m., bright sunshine, rainfall, etc. [42].

Unlike other circumstances, the extreme floods used to represent to the public a great danger to their lives, property and the very means of their sustenance, which have grown undeniably sensitive to climate change. Since rainfall is almost unpredictable and only happens in certain periods during the year, the system is designed to reduce its disastrous potential and provide early warnings for residents so that they can exercise their timely evacuation and preparation.

The high rainfall, which is above the norms, can harm the ecosystem. Thus, this accounts for soil depletion, species loss, and ruined aquatic habitats. The community should plan on having valid rainfall forecasts often that help maintain natural balance and establish agreement between population and nature.

Infrastructure like houses, roads, and communications is the most vulnerable to runoff caused by more rain. Utilizing the rainfall prediction models, which are developed based on the historical rainfall data, this study plans to assist the local communities in enhancing their infrastructure resilience theory to do away with expenditure incurred as a result of repair works and other social resilience issues that can be caused by high rainfall.

II. MATERIALS AND METHODS

To sharpen precipitation forecasting further, Elman neural network (ENN) models are joined with tangent wavelet transform (TWT), Ridgelet expansion approach (REM), curvelet expansion, and wavelet transformation (WT) algorithm are applied in the decomposition as preprocessing steps. This approach [1] forecasting, with the Himawari-8 and GPM IMERG data, is used to generate high-resolution rainfall forecasts. These techniques built on top of the LSTM networks and the multivariate time series analysis, for example, enable achieving more precise and trustworthy prediction of the rainfall patterns at both high spatial and temporal resolutions. They bring valuable insights about weather forecasting and risk management applications [3].With this, Seasonal Decomposition-Discrete Wavelet Transform- Artificial Neural Network- Seasonal Artificial Neural Network hybrid rainfall prediction methods are generated. Using multiple modelling techniques and decomposition methods, these hybrid models combine the best of both worlds, thus enhancing the predictive performance and robustness; in particular, they can capture the seasonal variations and long-term trends in rainfall patterns. Spatial and temporal rainfall [4] variations in Bangladesh are examined by making Make Sense and multilayer perceptron neural networks to predict the national rainfall accurately. Using machine learning algorithms and spatial models, this study is expected to enhance the efficiency and reliability of rainfall forecasting at the national level, thereby improving the preparedness and management of their resources in sectors such as agriculture, disaster preparation, and water resources. The prediction framework integrates [5] fuzzy logic and machine learning algorithms—decision tree, Naïve Bayes, K-nearest neighbours and support vector machines for rainfall prediction in Lahore. This framework comprises fuzzy logic-based inference combined with machine learning algorithms. The interpolation provides more accurate and interpretable forecasting for rainfall in urban areas, contributing significantly to effective risk management and disaster preparedness. M5P and support vector regression are utilized to tackle the problem of extreme rainfall in northern Bangladesh through machine learning. The study seeks [6]to enhance the accuracy and reliability of rainfall forecasts in the region through state-of-the-art regression techniques and machine learning algorithms, thus supporting various applications such as agricultural, water resource management, and disaster risk reduction. Among the six machine learning approaches utilized for long-term multiple-month ahead forecasts of monthly rainfall are the M5 model tree, random forest, support vector regression (SVR), and multilayer perceptron (MLP). The current study [6] aims to compare the efficiency of these procedures, design appropriate methods for predicting rainfall accurately for a long period of time and subsequently support these applications, which include agriculture, water resources management and disaster risk reduction. Classical linear time series models fail to properly capture and forecast complex geophysical phenomena like rainfall. Sharper approaches, such as Artificial Neural Networks and Hidden Markov Models, are examined as a way of overcoming these constraints, ensuring better accuracy and performance in representing the complicated spatial and temporal rainfall features, which, in turn, support a wide range of applications, including weather forecasting, climate modelling, disaster risk reduction, and so on. WF tenancy features changes in rainfall variability in Pakistan, which is more visible in the Baluchistan, Sindh, and Punjab regions. Through the analysis of historical rainfall data, this study [9] intends to seek and quantify the temporal and spatial variations in rainfall patterns, which will be an important factor in climate modelling, water resource management, and disaster risk reduction projects in that region. Trends in both (10) total and (11) monsoon rains have decreased over Pakistan's North, North, West, and Coastal regions. By examining the long-term rainfall data, this research attempts to detect the trends in rainfall occurrences and the amount of rain, making the description of rain in various parts proper for climate modelling, water resource management, and disaster risk reduction more accurate. A sliced functional time series model [11] is used to predict the average rainfall in Pakistan, a monthly forecast for the next ten years. The model will be designed to perform advanced time series analysis techniques and will be expected to limit the long-term rainfall forecast to weather accuracy and reliability enough to be applied in vital areas, including grassland and rivers and disaster risk reduction. Monthly and seasonal climate indices are evaluated for their apparent relationship for precipitation in Peshawar City, Pakistan, and evidence correlation for the initial.

The study [12] of the role that climate indices play in forming precipitation patterns is a way of getting insight into the upstream causes of rainfall variability, which contributes to making better decisions on climate adaptation and disaster risk reduction in the region. The models [13] of rainfall prediction are built around the meteorological factors and statistical downscaling techniques applied for the summer monsoon rainfall over the monsoon region of Pakistan. More accurate and reliable rainfall predictions are obtained using these models, incorporating main weather data and statistical downscaling processes vital for agriculture, water resource management, and disaster risk reduction. Factor and cluster analysis techniques are employed to identify rainfall areas in Pakistan that are needed for agriculture and food security. By studying past rainfall data and finding a specific rainfall pattern and clusters, this study focuses on delivering generally recognized trends for region diversity and providing practical findings in the decision-making process of agriculture and food security.

The structures of Malaysia [15] weather forecasting stations are modelled using the hybrid soft computing algorithm known as the MLP - HGSO method, which beats the MLP method alone. The HGSO (Hybrid Genetic Algorithm), combined with machine learning techniques, is a hybrid model that offers better accuracy and reliability in rainfall prediction and supports different applications such as agriculture, water resource management, and disaster risk reduction. Climatological data for Hyderabad and Nawabshah in the Sindh province of Pakistan is used to find the impacts of climate change on precipitation patterns. The study's purpose [16] is to use historical rainfall data to quantify the changing trends in precipitation patterns. From this analysis, the study can provide many useful insights that could be applied to the fight against the effects of climate change on regional rainfall variability, helping inform decision-making in the field of climate adaptation and disaster risk reduction in the region.

The relationship of climate indices is explored with the rainfall in Peshawar City, Pakistan, to understand the climate change adaptation strategies. Through the examination of the role of climate indices [17] in the precipitation trends in the region, this research will provide important evidence for designing the strategies and policies which would deal with the consequences of climatic changes and would make it possible to reach the sustainable development goals and to undertake the work on the reduction of disasters risk. Using wavelets, quadratic variation in rainfall across Pakistan is analyzed and is characterized by huge variation, too. Through historical rainfall data analysis and wavelet transformation techniques application, this study [18] tries to figure out and quantify regional variations in rainfall patterns, thus getting valuable information on the underlying drivers of rainfall variability and using it to support informed decision-making for water resource management and disaster risk reduction in different regions of Pakistan.

A variety of multi-day rainfall events observed in Punjab, Pakistan, utilizes probability distributions, which is required in flood mitigation policies. The study will [19] concentrate on the multi-day occurrence and classification of rainfall events by using statistical distributions to provide a necessary basis for identifying risks and reducing the exposure of floods, which support disaster risk reduction measures. The raining dynamics in hydrology and meteorology are studied precisely, and machine learning techniques are applied to rainfall prediction. This study [20] is expected to attain the purpose by using techniques such as rain dynamics analysis and machine learning for rainfall prediction, thereby improving the accuracy and reliability of the rain forecast that helps various sectors like agriculture, distribution of water resources and disaster risk reduction. The algorithm has a remarkable generalization capability [21], yielding precipitation predictions in most meteorological stations with a reasonable correspondence to climatic zones. Yet, some stations show a lower estimation of annual total precipitation; this indicates that the prediction model needs to be further improved and validated to solve the discrepancies and improve the accuracy.

A DL architecture [22] designed to estimate the precipitation accumulation for the next day is being introduced. This architecture performs better than those employed in previous studies. Deep learning algorithms can harness the true power of this approach to generate an increased level of preciseness and accuracy in short-term precipitation forecasts, and this technology is expected to find its way into all sectors of agriculture, water resources management, and disaster preparedness. This work [23] presents a study of the methods of landslide prediction, namely, the random forest (RF) and the logistic regression (LR), based on the rainfall datasets and the antecedent cumulative rainfall data.

These methods use rainfall information about where a landslide happened to forecast early warning systems for landslide risk reduction, which are the basis of disaster resilience and public well-being. The work [24] investigates whether machine learning algorithms may serve for precipitation prediction implemented on the data on the rainfall from major urban areas in Australia during the last decade. Several machine learning algorithms, such as k nearest neighbours (KNN), decision tree(DT), random forest(RF), and neural networks(NN), are tested, with neural networks showing to be the most effective model for rainfall prediction, which is a very important knowledge for the improvement of weather forecasting abilities. The main aim of this study [25-27] is to develop a flexible computing system for the same decimal point rainfall forecast in Upstate New York. It is a kind of algorithm that combines classics and trends and involves algorithms such as K-Nearest Neighbors, Support Vector Machine, Deep Neural

Network, Wide Neural Network, Reservoir Computing, and Long short-term memory with the technology which provides accurate and reliable rainfall forecasts that are tailored to local geographic conditions and climate variations. This research investigates the efficacy of various machine learning models, including Naive Bayes, Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbors (KNN), Gradient Boosting, and time series forecasting model ARIMA (Autoregressive Integrated Moving Average), for rainfall prediction.

III. Proposed System

A. *Dataset*

The website, 'Visual Crossing,' was targeted by employing a web-scraping methodology to capture the numbers and statistics for this study. Web pages were systematically pulled out to capture different weather parameters using a web scraping mechanism from variable parameters available on the platform. The site has a detailed weather history from which one can easily extract data, such as temperature, wind speed, precipitation, etc. It had a very well-organized approach and focused on the data collection procedure to minimize the extraction of irrelevant data all at once for diversity. Considering the formatting, the organized dataset was handled with the quantity parameter represented in integer or decimal within the numerical term, and some of the data was a textual unstructured type. This unique numeric dataset has been just incorporated with real-world weather information data for space environmental research and data science. It becomes the primary source for conducting weather forecasts, analysis and induction learning.

A series of steps were required to make it a reality, such as web scraping skills, data pre-processing, and numerical dataset creation. This prevented the right to use them in research and analysis. This extensive dataset captures a comprehensive weather history spanning over two decades (2000 to 2023) for six major cities in Pakistan. The dataset, characterized by 35 rows and 52,597 columns, offers an intricate exploration of meteorological conditions and patterns in the region. Table 1 describes the dataset information regarding data types and missing values.

Table 1: Dataset Description

DATA	DESCRIPTION
Datatype	Numeric (integer, float) & String
Features/Attributes (Rows)	35
Numeric Attributes	27
String Attributes	6
Days Count (Columns)	52, 597
Missing Values	133271
Data Unit	Fahrenheit(°F)
Country	Pakistan
Cities	Karachi, Islamabad Lahore, Peshawar, Rawalpindi & Faisalabad
Year Range	2000 to 2023

B. *System Model*

This research focuses on modernizing weather data collection and prediction methodologies for cities in Pakistan. By integrating advanced weather monitoring systems and leveraging cutting-edge research methodologies, this study aims to surpass existing limits in weather prediction capabilities.

The research involves collecting comprehensive weather data from various sources, including ground-based weather stations. The study explores applying advanced prediction methods, including machine learning algorithms and time series forecasting techniques, to improve the accuracy of weather predictions for Pakistan's cities. By leveraging the vast amount of collected data and employing sophisticated modelling approaches, this research seeks to enhance the understanding of local weather patterns and phenomena. The outcomes of this research are expected to provide valuable insights and tools for decision-makers in various sectors, including urban planning, agriculture, and disaster management. By harnessing the power of modern technology and research methodologies, this study aims to contribute to the development of more resilient and adaptive strategies for addressing weather-related challenges in Pakistan's urban areas. This study's proposed system is a great move in weather surveillance and prediction for Pakistan's cities. Frontline technology and methodologies refined with real-time data assimilation and produced by this system, the forecasts thus generated are said to be valid and meaningful. Users are meant to get trustworthy data they can choose from so they can make decisions instantly to lessen the harm. Fig 1 describes the overall flow of work proposed in this study.

The proposed procedure for enhancing rainfall prediction accuracy comprises several sequential steps. Firstly, it involves accumulating requisite data for analysis, such as historical precipitation records, which serve as a foundation for predicting future precipitation patterns. Following data collection, the next step involves data cleaning and preprocessing to rectify flaws in the raw data and address missing values, errors, and inconsistencies to ensure data integrity and reliability. Subsequently, feature selection techniques are employed to identify the most significant data criteria for prediction. Machine learning modelling techniques, including time series forecasting and random forests, are then applied to develop predictive models based on the curated dataset. Once the models are constructed, they undergo training using historical data, allowing them to learn intricate patterns and relationships necessary for accurate predictions. Following training, the models are evaluated using current data to assess their performance, pinpointing areas requiring enhancement for improved accuracy. Ensemble modelling techniques are then utilized to amalgamate forecasts from multiple models, thereby raising the overall performance, particularly in predicting comprehensive data. Finally, the results of the rainfall prediction models are analyzed and presented, showcasing metrics such as model accuracy, prediction rates, and output quality. This comprehensive procedure enables informed decision-making in various sectors reliant on accurate rainfall forecasts, such as water resource management, agriculture, and disaster preparedness.

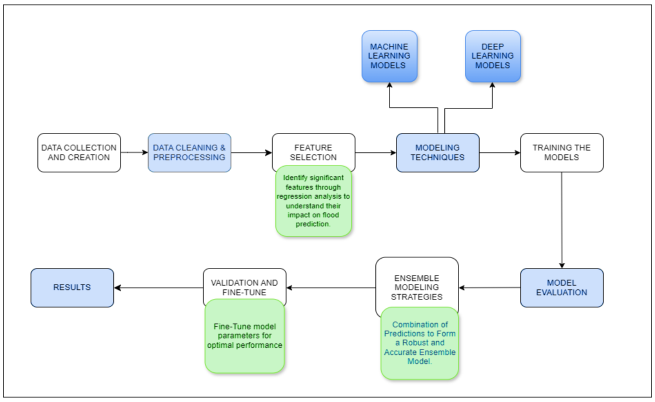

MACHINE LEARNING MODELS
DEEP LEARNING MODELS
DATA COLLECTION AND CREATION
DATA CLEANING & PREPROCESSING
FEATURE SELECTION
Identify significant features through regression analysis to understand their impact on flood prediction.
MODELING TECHNIQUES
TRAINING THE MODELS
RESULTS
VALIDATION AND FINE-TUNE
Fine-Tune model parameters for optimal performance
ENSEMBLE MODELING STRATEGIES
Combination of Predictions to Form a Robust and Accurate Ensemble Model.
MODEL EVALUATION

Figure 1: Framework of the system.

C. ML Algorithm

In this research endeavour, a suite of machine learning algorithms has been harnessed to bolster the accuracy of rainfall prediction. These algorithms encompass Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Gradient Boosting, Random Forest, and K Nearest Neighbors (KNN). The figure denoted as 5a presumably displays the result of one of these models, potentially the outcome of the "orange" model, providing insights into its predictive performance.

Logistic regression is a statistical method used for binary classification tasks. In rainfall prediction, logistic regression models are trained to predict whether rainfall will occur (binary outcome) based on input features such as temperature, humidity, and atmospheric pressure.

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. In rainfall prediction, Naive Bayes classifiers are used to estimate the probability of rainfall given a set of observed weather conditions.

SVM is a supervised machine learning algorithm used for classification and regression tasks. In the context of rainfall prediction, SVM models are trained to classify weather conditions, such as rainy or non-rainy, based on input features.

Gradient boosting is an ensemble learning technique that sequentially builds a strong predictive model by combining multiple weak models. In rainfall prediction, gradient boosting algorithms such as XGBoost or Light GBM can improve prediction accuracy by iteratively fitting new models to the residuals of previous models.

Random forest is another ensemble learning technique that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the individual trees' mean prediction (regression). In rainfall prediction, random forest models can capture complex interactions between weather variables and make accurate predictions. KNN is a non-parametric method used for classification and regression tasks. In the context of rainfall prediction, KNN algorithms classify new data points based on the majority class of their nearest neighbours in the feature space.

ARIMA model amalgamates the Box-Jenkins approach, which involves trend, seasonality, and an autoregressive component to extract the underlying series from data (AR, MA, I). (p, d, q) represents the orders of the autoregressive, integrated, and moving average components, respectively.

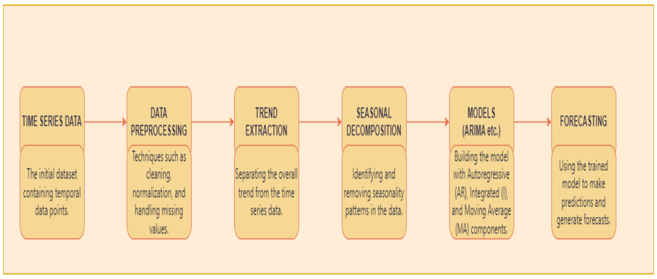

TIME SERIES DATA
The initial dataset containing temporal data points.
DATA PREPROCESSING
Techniques such as cleaning, normalization, and handling missing values.
TREND EXTRACTION
Separating the overall trend from the time series data.
SEASONAL DECOMPOSITION
Identifying and removing seasonality patterns in the data.
MODELS (ARIMA etc.)
Building the model with Autoregressive (AR), Integrated (I), and Moving Average (MA) components.
FORECASTING
Using the trained model to make predictions and generate forecasts.

Figure 2. Flow Diagram of Time Series Forecasting TSF

On the other hand, VAR models are appropriate for analyzing and forecasting multivariate time series data, where multiple variables are interrelated and influence each other. VAR models capture the dynamic interactions among the variables by estimating a system of equations, where each variable is regressed on its own lagged values and other variables in the system. VAR models are advantageous when understanding the relationships and interdependencies among multiple time series variables. Figure 3 shows the flow of TSF. The choice between ARIMA (Autoregressive Integrated Moving Average) and VAR (Vector Auto Regression) depends on the type of data and the intricacy of the relationships involved when determining the best modelling strategy for rainfall prediction. When concentrating on a particular location or variable, ARIMA models are useful for rainfall prediction because they are skilled at capturing the temporal dynamics of univariate time series data. These models are particularly good at identifying and predicting common patterns in rainfall data, such as trends and seasonality.

VAR models are excellent for multivariate rainfall prediction tasks that analyze the relationships between rainfall patterns across various locations or factors influencing precipitation. They are well-suited for analyzing the interdependencies among multiple variables over time. Figure 3b presumably displays the result of one of these models, potentially the outcome of the "orange" model, providing insights into its predictive performance.

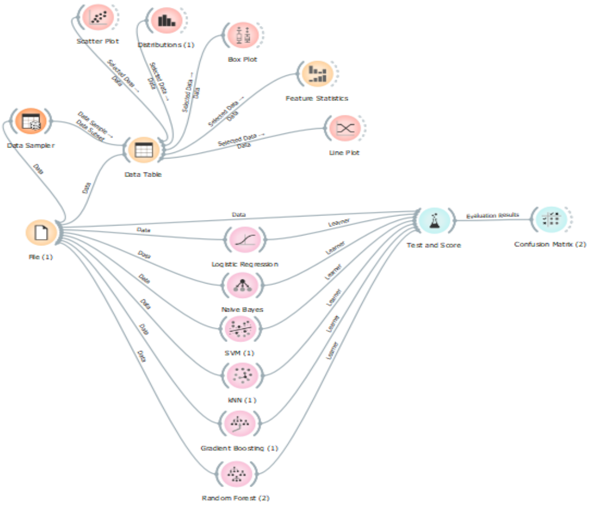


Fig 3a. Orange Tool Workflow, Test and Score of Different Models

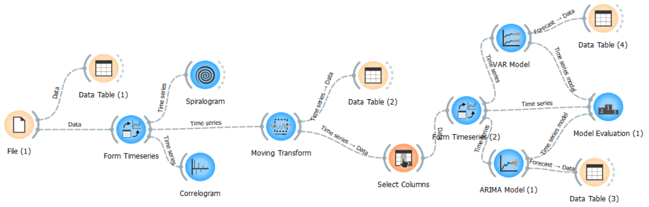
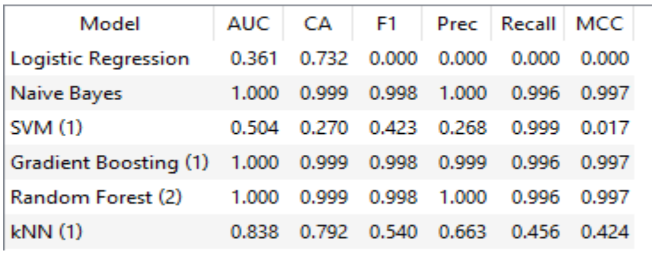


Fig 3b. Orange Tool Workflow (ARIMA vs. VAR)

IV. Results

The percentage of data records that an algorithm successfully classifies after assessing the classification outcomes is known as the metric or classification accuracy metric. It's a key performance indicator for evaluating a classification model's effectiveness. Indeed, the degree of agreement between the expected and actual values can be used to determine accuracy. Accuracy in classification tasks quantifies the frequency with which the classifier generates accurate predictions. Accuracy in a binary classification situation (for example, forecasting rain or not) is determined by dividing the total number of cases in the dataset by the proportion of correctly identified examples (true positives and true negatives). The test results highlight the performance metrics for each trial, including precision, recall, accuracy, and AUC values. As depicted in Figure 4, the maximum accuracy results for the different algorithms are as follows: Naive Bayes (100%), Logistic Regression (0.361), Support Vector Machine (SVM) (0.504), Gradient Boosting (100%), Random Forest (100%), and K Nearest Neighbors (KNN) (0.83). Notably, the KNN algorithm demonstrates a commendable precision of 83%, indicating its ability to classify rainfall instances accurately. Conversely, Logistic Regression and SVM exhibit lower accuracy rates, suggesting potential limitations in their predictive capabilities for this dataset. The detailed performance metrics provided offer valuable insights into the strengths and weaknesses of each algorithm, aiding in informed decision-making regarding their suitability for rainfall prediction tasks. In terms of tp, fp, fn, and tn, Figs. 5a through 5f define the confusion matrix.



Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.361	0.732	0.000	0.000	0.000	0.000
Naive Bayes	1.000	0.999	0.998	1.000	0.996	0.997
SVM (1)	0.504	0.270	0.423	0.268	0.999	0.017
Gradient Boosting (1)	1.000	0.999	0.998	0.999	0.996	0.997
Random Forest (2)	1.000	0.999	0.998	1.000	0.996	0.997
kNN (1)	0.838	0.792	0.540	0.663	0.456	0.424

Fig 4. Classification Metrics of Different Models

In machine learning, performance metrics indicate how successfully a model operates on a particular job. The type of task—classification, regression, clustering, or other particular goals—determines which metrics apply. A confusion matrix is a table shown in Fig. 5a to 5f that displays the percentages of true positives, true negatives, false positives, and false negatives to evaluate the effectiveness of a categorization system.

True Positives (TP) are cases correctly predicted to be positive.

True Negatives (TN) are cases that were correctly predicted to be negative.

False Positive (FP): Situations in which a Type I error resulted in a false positive prediction.

False Negative (FN): Instances in which the negative prediction (Type II mistake) was made incorrectly.

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 can be used to determine accuracy.

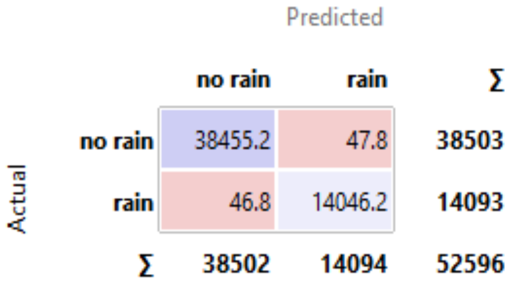
Accuracy=(number of correct prediction)/(total number of predictions)

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 dataset's total number of data records.

Precision (Positive Predictive Value): It explains why many correctly predicted cases turned out to be positive. Precision is useful in scenarios where False Positives are more problematic than False Negatives.

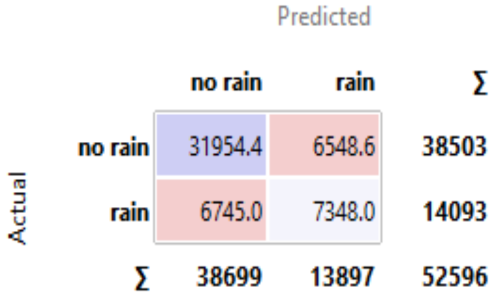
Specificity (True Negative Rate): Specificity as the ratio of accurate negative predictions to the total number of actual negatives observed.

Recall (Sensitivity, True Positive Rate): It makes clear how many actual positive cases our model correctly anticipated. Recall is a useful metric when a False Positive is more worrying than a False Negative.



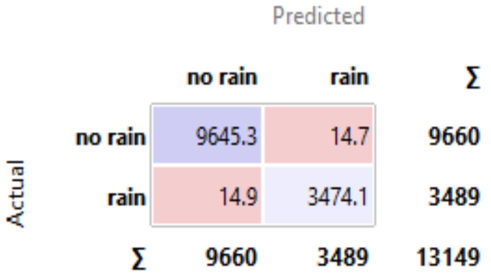
		Predicted		
		no rain	rain	Σ
Actual	no rain	38455.2	47.8	38503
	rain	46.8	14046.2	14093
	Σ	38502	14094	52596

Fig 5a. Confusion Matrix of Random Forest



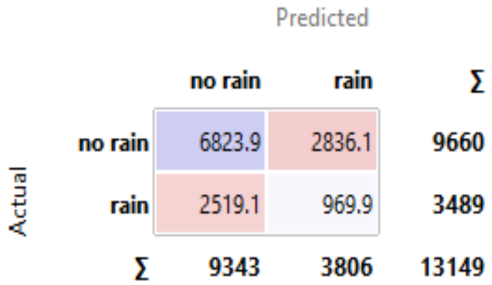
		Predicted		
		no rain	rain	Σ
Actual	no rain	31954.4	6548.6	38503
	rain	6745.0	7348.0	14093
	Σ	38699	13897	52596

Fig 5b. Confusion Matrix of KNN



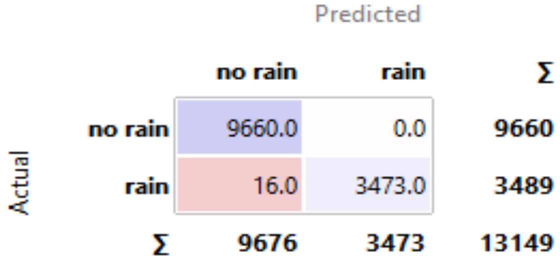
		Predicted		
		no rain	rain	Σ
Actual	no rain	9645.3	14.7	9660
	rain	14.9	3474.1	3489
	Σ	9660	3489	13149

Fig 5c. Confusion Matrix of Gradient Boosting



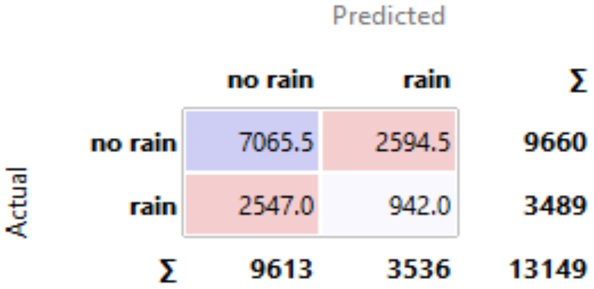
		Predicted		
		no rain	rain	Σ
Actual	no rain	6823.9	2836.1	9660
	rain	2519.1	969.9	3489
	Σ	9343	3806	13149

Fig 5d. Confusion Matrix of Logistic Regression



		Predicted		
		no rain	rain	Σ
Actual	no rain	9660.0	0.0	9660
	rain	16.0	3473.0	3489
	Σ	9676	3473	13149

Fig 5e. Confusion Matrix of Naïve Bayes



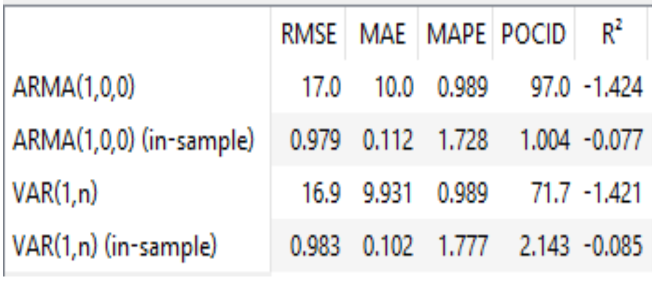
		Predicted		
		no rain	rain	Σ
Actual	no rain	7065.5	2594.5	9660
	rain	2547.0	942.0	3489
	Σ	9613	3536	13149

Fig 5f. Confusion Matrix of SVM

K nearest neighbour (KNN) provided good results. It correctly predicted rain 73480 and no rain 31954 from out of 529596 datasets shown in Fig. 5b. It provides 84% accuracy, better than the other models. As depicted in Figure 6, the maximum accuracy results for the different algorithms are as follows: Naive Bayes (100%), Logistic Regression (0.361), Support Vector Machine (SVM) (0.504), Gradient Boosting (100%), Random Forest (100%), and K Nearest Neighbors (KNN) (0.83). Notably, the KNN algorithm demonstrates a commendable precision of 83%, indicating its ability to classify instances of rainfall accurately. Conversely, Gradient boosting, random forest overfit the data by producing 100% accuracy.

V. Evaluation Metrics for Predictive Model Time Series Forecasting

Comparing the performance of models like ARIMA and VAR (Vector Autoregression) can depend on various factors, including the nature of the data, the specific characteristics of the time series, and how well the models are tuned for the particular dataset.RMSE (Root Mean Square Error) is a common metric used to evaluate the accuracy of forecasting models. A lower RMSE indicates better performance in terms of forecasting accuracy. Figure 8 demonstrates that ARIMA provides a lower RMSE than VAR, suggesting that ARIMA might be a better choice regarding predictive accuracy for the specific dataset and forecasting horizon considered.



	RMSE	MAE	MAPE	POCID	R^2
ARMA(1,0,0)	17.0	10.0	0.989	97.0	-1.424
ARMA(1,0,0) (in-sample)	0.979	0.112	1.728	1.004	-0.077
VAR(1,n)	16.9	9.931	0.989	71.7	-1.421
VAR(1,n) (in-sample)	0.983	0.102	1.777	2.143	-0.085

Fig 6. Model Comparison of VAR AND ARIMA

VI. CONCLUSIONS

The introduction of artificial intelligence into rainfall prediction marks a new beginning in meteorological forecasting, with great innovative discoveries and opportunities in the field. Integration of machine learning techniques into traditional forecasting models shows that AI can improve the accuracy and reliability of the predictions. This research explores how predictive algorithms powered by AI can analyze such complex relationships in the data associated with such extreme weather conditions to produce rainfall forecasts with a high degree of accuracy. Using data from historical weather conditions over several years and utilizing state-of-the-art algorithms, such as Long Short-Term Memory Networks, Random Forests, and time series forecasting models, the predictive model has continuously produced more accurate results than the traditional models. K nearest neighbour (KNN) provides good results. It correctly predicted rain 73480 and no rain 31954 from out of 529596 datasets. It provides 84% accuracy, better than the other models.

The conclusions from this study make fundamental contributions to the literature in environmental data science, highlighting the important findings regarding localized meteorological forecasting. This implies that using rainfall prediction models with the aid of AI will have numerous applications in agriculture, water resource programs, disaster response, and urban development. The next steps in this research field should focus on consistently refining prediction models, increasing the number of datasets, and facilitating the larger applicability of AI prediction models in forecast processes. In addition, policymakers and meteorologists should work with scientists to ensure that the technology can effectively integrate AI into regional operational forecasting. In conclusion, integrating AI in terms of rainfall prediction is a huge and revolutionary milestone for meteorological science in providing a much safer approach to extreme weather event threats. To combat climate vulnerability and offer the knowledge that people could use to minimize their risk to development and risk reduction.

Acknowledgment

The authors greatly appreciate the technical support of the National Centre for Physics Islamabad, Pakistan.

References

[1] Song, C., & Chen, X. (2021). Performance comparison of machine learning models for annual precipitation prediction using different decomposition methods. Remote Sensing, 13(5), 1018.

[2] Simanjuntak, F., Jamaluddin, I., Lin, T. H., Siahaan, H. A. W., & Chen, Y. N. (2022). Rainfall Forecast Using Machine Learning with High

Spatiotemporal Satellite Imagery Every 10 Minutes. Remote Sensing, 14(23), 5950.

[3] Sheikhi, Y., Ashrafi, S. M., Nikoo, M. R., & Haghighi, A. (2023). Enhancing daily rainfall prediction in urban areas: a comparative study of hybrid artificial intelligence models with optimization algorithms. Applied Water Science, 13(12), 232.

[4] 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.

[5] Di Nunno, F., Granata, F., Pham, Q. B., & de Marinis, G. (2022). Precipitation forecasting in Northern Bangladesh using a hybrid machine learning model. Sustainability, 14(5), 2663.

[6] Wu, G., Zhang, J., & Xue, H. (2023). Long-Term Prediction of Hydrometeorological Time Series Using a PSO-Based Combined Model Composed of EEMD and LSTM. Sustainability, 15(17), 13209.

[7] Hu, C., Wu, Q., Li, H., Jian, S., Li, N., & Lou, Z. (2018). Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. Water, 10(11), 1543.

[8] Hussain, M. S., & Lee, S. (2009). A classification of rainfall regions in Pakistan., 44(5), 605-623.

[9] Salma, S., Rehman, S., & Shah, M. A. (2012). Rainfall trends in different climate zones of Pakistan. Pakistan Journal of Meteorology, 9(17).

[10] Yasmeen, F., & Hameed, S. (2018). Forecasting of rainfall in Pakistan via sliced functional times series (SFTS). World Environment, 8(1), 1-14.

[11] 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.

[12] Adnan, M., Rehman, N., Ali, S., Mehmood, S., Mir, K. A., Khan, A. A., & Khalid, B. (2017). Prediction of summer rainfall in Pakistan from global sea-surface temperature and sea-level pressure. Weather, 72(3), 76-84.

[13] Zhang, X., Zhao, D., Wang, T., Wu, X., & Duan, B. (2022). A novel rainfall prediction model based on CEEMDAN-PSO-ELM coupled model. Water Supply, 22(4), 4531-4543.

[14] Sammen, S. S., Kisi, O., Ehteram, M., El-Shafie, A., Al-Ansari, N., Ghorbani, M. A., ... & Shahid, S. (2023). Rainfall modeling using two different neural networks improved by metaheuristic algorithms. Environmental Sciences Europe, 35(1), 112.

[15] Mahessar, A. A., Qureshi, A. L., Sadiqui, B., Kori, S. M., Mukwana, K. C., Qureshi, A. S., & Leghari, K. Q. (2020). Rainfall Analysis for Hyderabad and Nawabshah, Sindh, Pakistan. Engineering, Technology & Applied Science Research, 10(6), 6597-6602.

[16] 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.

[17] Akhter, M. F., & Abbas, S. (2021). Variability of provincial capital rainfall in Pakistan using wavelet transformation. Pure and Applied Geophysics, 178(10), 4147-4157.

[18] Iqbal, M. J., & Ali, M. (2013). A probabilistic approach for estimating return period of extreme annual rainfall in different cities of Punjab. Arabian Journal of Geosciences, 6, 2599-2606.

[19] Umamaheswari, P., & Ramaswamy, V. (2022). Optimized preprocessing using time variant particle swarm optimization (TVPSO) and deep learning on rainfall data. Journal of Scientific and Industrial Research, 1317-1325.

[20] Abdul-Kader, H., & Mohamed, M. (2021). Hybrid machine learning model for rainfall forecasting. Journal of Intelligent Systems and Internet of Things, 1(1), 5-12.

[21] Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016). Rainfall prediction: A deep learning approach. In Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016, Proceedings 11 (pp. 151-162). Springer International Publishing.

[22] Kuradusenge, M., Kumaran, S., & Zennaro, M. (2020). Rainfall-induced landslide prediction using machine learning models: The case of Ngororero District, Rwanda. International journal of environmental research and public health, 17(11), 4147.

[23] Sarasa-Cabezuelo, A. (2022). Prediction of rainfall in Australia using machine learning. Information, 13(4), 163.

[24] Yu, N., & Haskins, T. (2021, July). Bagging machine learning algorithms: A generic computing framework based on machine-learning methods for regional rainfall forecasting in upstate New York. In Informatics (Vol. 8, No. 3, p. 47). MDPI.

[25] 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.

[26] Sankaranarayanan, S., Prabhakar, M., Satish, 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.

[27] Hammad, M., Shoaib, M., Salahudin, 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.