IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 September 2023, accepted 14 November 2023, date of publication 16 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333876

RESEARCH ARTICLE

Machine Learning-Based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness

MD. MEHEDI HASSAN[1], (Member, IEEE), MOHAMMAD ABU TAREQ RONY[2], MD. ASIF RAKIB KHAN[3], MD. MAHEDI HASSAN[3], FARHANA YASMIN[4], ANINDYA NAG[1], (Member, IEEE), TAZRIA HELAL ZARIN[5], ANUPAM KUMAR BAIRAGI[1], (Senior Member, IEEE), SAMAH ALSHATHRI[6], AND WALID EL-SHAFAI[7,8], (Senior Member, IEEE)

[1]Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh
[2]Department of Statistics, Noakhali Science and Technology University, Noakhali 3814, Bangladesh
[3]Department of Computer Science and Engineering, Bangladesh University of Business and Technology (BUBT), Dhaka 1216, Bangladesh
[4]Department of Computer Science and Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China
[5]Department of Computer Science and Engineering, University of Development Alternative, Dhaka 1209, Bangladesh
[6]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[7]Security Engineering Laboratory, Computer Science Department, Prince Sultan University, Riyadh 11586, Saudi Arabia
[8]Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt

Corresponding authors: Md. Mehedi Hassan (mehedihassan@ieee.org), Anupam Kumar Bairagi (anupam@cse.ku.ac.bd), Samah Alshathri (ealshathry@pnu.edu.sa), and Walid El-Shafai (eng.waled.elshafai@gmail.com)

This work was supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, through the Princess Nourah bint Abdulrahman University Researchers Supporting Project under Grant PNURSP2023R197.

ABSTRACT Rainfall prediction plays a crucial role in raising awareness about the potential dangers associated with rain and enabling individuals to take proactive measures for their safety. This study aims to utilize machine learning algorithms to accurately predict rainfall, considering the significant impact of scarcity or extreme rainfall on both rural and urban life. The complex nature of rainfall, influenced by various atmospheric, oceanic, and geographical factors, makes it a challenging phenomenon to forecast. This research employs data preprocessing techniques, outlier analysis, correlation analysis, feature selection, and several machine learning algorithms such as Naive Bayes (NB), Decision Tree, Support Vector Machine (SVM), Random Forest, and Logistic Regression. The study focuses on developing the most accurate rainfall prediction model by utilizing machine learning and feature selection techniques. The Artificial Neural Network (ANN) achieves a maximum accuracy of 90% and 91% before and after feature selection, respectively. Furthermore, k-means clustering and Principal Component Analysis (PCA) are applied to examine regional rainfall patterns in Australia. Lastly, to make our proposed machine learning simpler and more usable for general people, we have formulated a web-based application system using Flask in our research paper. Overall, this research demonstrates the effectiveness of different machine-learning techniques in predicting rainfall using Australian weather data.

INDEX TERMS Rainfall, rainfall prediction, machine learning, features selection, PCA, cluster analysis, WebApp.

I. INTRODUCTION

Rainfall is not only essential for the survival of plants and animals but also plays a critical role in maintaining ecological balance by supplying fresh water to the Earth's surface. However, the unpredictable nature of rainfall

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

patterns can give rise to extreme weather events, such as prolonged droughts or devastating floods, which can have far-reaching consequences for ecosystems, agriculture, and human populations [1]. Therefore, accurate and reliable rainfall forecasting is of utmost importance to enhance preparedness, improve resource management, and make informed decisions during severe weather conditions.

According to the National Centers for Environmental Information, the projected global average precipitation for 2021 stands at 2.66 millimeters per day, slightly below the 40-year climatological mean of 2.69 millimeters per day [2]. This highlights the dynamic nature of rainfall patterns and the need for effective forecasting methodologies. In this regard, the field of weather forecasting has witnessed significant advancements with the integration of data analysis and machine learning techniques. Machine learning, a powerful computational approach, harnesses the potential of vast datasets to uncover intricate patterns, correlations, and trends among various meteorological variables. By leveraging this knowledge, machine learning algorithms can make accurate predictions, aiding in better understanding and anticipation of rainfall patterns.

Several well-established rainfall forecasting models are currently employed worldwide. These models include the Weather Research and Forecasting (WRF) model, which combines advanced atmospheric physics with numerical simulations to generate high-resolution weather forecasts. The General Forecasting Model focuses on providing short-term weather predictions, while Seasonal Climate Forecasting aims to anticipate rainfall patterns over longer periods. The Global Data Forecasting Model integrates a wide range of meteorological data from across the globe to produce comprehensive weather forecasts. Although these models offer valuable insights, their computational requirements can be substantial, making them resource-intensive to run and maintain.

In contrast, data mining models offer an alternative approach to rainfall prediction. These models rely on historical data, statistical analysis, and similarity patterns to extract meaningful information and make accurate predictions. Empirical methodologies form an integral part of data mining in rainfall forecasting. Techniques such as regression analysis, artificial neural networks, fuzzy logic, and group data handling have proven to be effective in climate prediction [3]. These empirical models derive insights from past rainfall data, exploring its associations with a wide range of meteorological and oceanic factors on a global scale. By recognizing and understanding these complex relationships, these models can provide valuable predictions of future rainfall patterns.

Machine learning algorithms have been extensively studied for their effectiveness in rainfall prediction. Cramer et al. [4] demonstrated the superiority of machine learning-based intelligent systems, including Genetic Programming, Support Vector Regression, Radial Basis Neural Networks, M5 Rules, M5 Model Trees, and K-Nearest Neighbors, in comparison to existing state-of-the-art methodologies. These algorithms showcased improved predictive accuracy and reduced correlations, enhancing the reliability of rainfall forecasts.

On the other hand, the dynamical approach to rainfall prediction involves constructing physical models based on systems of equations. These models simulate the complex interactions of atmospheric, oceanic, and environmental factors to predict changes in the global climate system in response to initial atmospheric conditions [5]. Integrating machine learning techniques and feature selection methods into this approach can further refine the predictive capabilities and improve the accuracy of rainfall forecasts.

The primary objective of this research is to develop a robust and reliable rainfall prediction model by combining machine learning algorithms and feature selection methods. The ultimate goal is to create an effective rainfall prediction system that can be utilized in remote regions, where accurate forecasts are crucial for efficient water resource management, agricultural planning, and disaster preparedness.A few things set this article apart from others that compare rainfall prediction models.

- We analyze the most well-known prediction models in this work, whereas other studies only consider a small number.
- We employ the most important meteorological factors as input variables to evaluate rainfall prediction models that haven't been examined before.
- Data from weather stations around Australia is used to compare the performance of prediction models.

We have discussed some published work in the related work section based on our rationale. The methodology section describes the overall workflow of this research. We assessed our research's findings and model performance in the outcome section and presented our research's novelty in the discussion section.

II. RELATED WORK

He et al. [6] conducted a comprehensive study on rain forecasting in Australia using an active learning system. They utilized a large dataset from Kaggle, consisting of historical weather data, and employed entropy sampling as the query approach within a pool-based sampling method. The classification model used in their study was logistic regression, which proved to be effective in predicting rainfall. The research not only focused on the accuracy of the predictions but also compared the performance of active learning with random sampling. The experimental results showed that the active learning approach achieved a prediction accuracy of 82.02%, while the random sampling approach achieved 82.0%. These findings highlight the potential of active learning techniques in improving rainfall prediction accuracy. Nolan and Graco [7] aimed to predict weather outcomes, specifically focusing on forecasting rainy and dry days in Sydney for the next day. They employed a decision-tree model with capstone analysis, which allowed for the identification of key features that influenced the weather conditions. The capstone decision-tree model demonstrated the highest accuracy rate of 87.9%,

indicating its effectiveness in accurately predicting weather conditions. Furthermore, the combined-city model achieved an accuracy rate of 75.6%, suggesting the applicability of the approach to different geographical regions. Balamurugan and Manojkumar [8] conducted a comparative study to evaluate the performance of machine learning techniques in rainfall prediction compared to statistical methods. They observed that the percentage of departure of rainfall ranged from 46% to 91% for the month of June 2019, as per the Indian Meteorological Department (IMD). However, based on their study, machine learning techniques outperformed statistical methods in rainfall prediction. Logistic regression achieved an overall accuracy of 84.6%, compared to 72.6% for the ROC statistical technique and 77.6% for the Decision Tree statistical strategy. These results emphasize the superiority of machine learning algorithms in capturing the complex patterns and relationships in rainfall data.

Ejike et al. [9] explored the application of logistic regression modeling to predict rainfall for the following day. They used one year of meteorological data from Canberra, Australia, including temperature, pressure, humidity, sunlight, evaporation, cloud cover, wind direction, and wind speed. The results revealed that, with the inclusion of appropriate meteorological factors, rainfall for the following day can be forecast with an accuracy of 87% using logistic regression. This finding highlights the significance of incorporating relevant features in the modeling process to achieve accurate rainfall predictions. Kumarasiri and Sonnadara, [10] developed neural network models for rainfall prediction at different time scales. They created a one-day-ahead model, which achieved a prediction accuracy of 74.30% for the next day's rainfall. Additionally, they developed a one-year-ahead model for yearly rainfall depth predictions, which achieved an accuracy of 80.0% within a 5% error bound. Moreover, these models were extended to make predictions for multiple time steps into the future. The findings suggest the potential of neural networks in capturing the temporal dynamics of rainfall patterns and making accurate predictions over different time horizons. Ria et al. [11] conducted a rain prediction study using machine learning models based on a dataset acquired from the Bangladesh Jatiyo Tottho Batayon website. They trained and evaluated five different models to predict rainfall. Each model was trained with eight input characteristics related to weather conditions before being used to validate rainfall forecasts. The results of the study indicated that the Random Forest classifier achieved the highest accuracy of 86% compared to other models, demonstrating its effectiveness in capturing the complex relationships between input variables and rainfall patterns. Neelakandan and Paulraj [12] proposed a CA-SVM-based prediction model for rainfall forecasting using real-time data sets. They utilized RMC, Chennai, annual rainfall data for a 12-month period to validate their model. The study employed two learning models, namely the local learning model and the dynamic learning model. The proposed algorithm achieved an accuracy of 89% compared to existing algorithms, demonstrating its potential in accurately predicting rainfall by integrating separate data points without overlap. Sankaranarayanan et al. [13] utilized machine learning methods, including Artificial Neural Networks (ANN) with a single hidden layer, to predict floods based on various factors such as precipitation, temperature, water velocity, water level, and humidity. They developed a deep neural network that incorporated stream flow for flood forecasting. The study collected a large volume of rainfall data, along with other relevant variables. The results of a comparison of the accuracy achieved using four different algorithms indicated that the deep neural network outperformed the benchmark, achieving a higher accuracy of 90%. These findings highlight the potential of deep learning techniques in accurate flood prediction.

Overall, these studies demonstrate the effectiveness of machine learning techniques, including active learning, decision-tree models, logistic regression, neural networks, and ensemble methods, in improving the accuracy of rainfall prediction. By leveraging large datasets and incorporating relevant meteorological factors, these models have the potential to provide valuable insights for weather forecasting and decision-making in various regions and time scales.

III. METHODOLOGY

The following key components are included in the overall architecture described in this article: Data Pre-processing, replacing missing observations, Encoding the string value into numeric, a summary of the dataset using SPSS in Table 2. to find the insight of data, basic visualization to find an idea about different features, Outlier detection, implement classification with the change of parameter before and after feature selection, and comparison outcome of the classification algorithms. In this project, Python 3.8.3 is used for the classification dataset. In addition, SPSS 25 is used for finding details and descriptive statistics [14] of the dataset described in Table 1. Figure 1 illustrates the comprehensive procedure that has been followed throughout the course of this research. This figure encapsulates the sequential steps and methodologies employed to investigate and analyze the phenomenon under study. From data collection and preprocessing to feature extraction, model development, and performance evaluation, each phase is visually depicted in this overarching representation, providing a holistic view of the research approach. In Figure 2, a graphical depiction is presented that delineates the myriad variables exerting influence over the complex process of rainfall. These factors span a wide spectrum, including atmospheric conditions, geographical features, temperature gradients, and more. This figure aims to encapsulate the multifaceted nature of rain formation and the intricate interplay of variables that contribute to its occurrence and intensity. Figure 3 provides a comprehensive visual overview of the entire project. This encompassing image provides an aerial perspective of the research's components, interactions, and outcomes. By integrating the key elements of the project into a singular visual representation, Figure 3 offers a concise yet

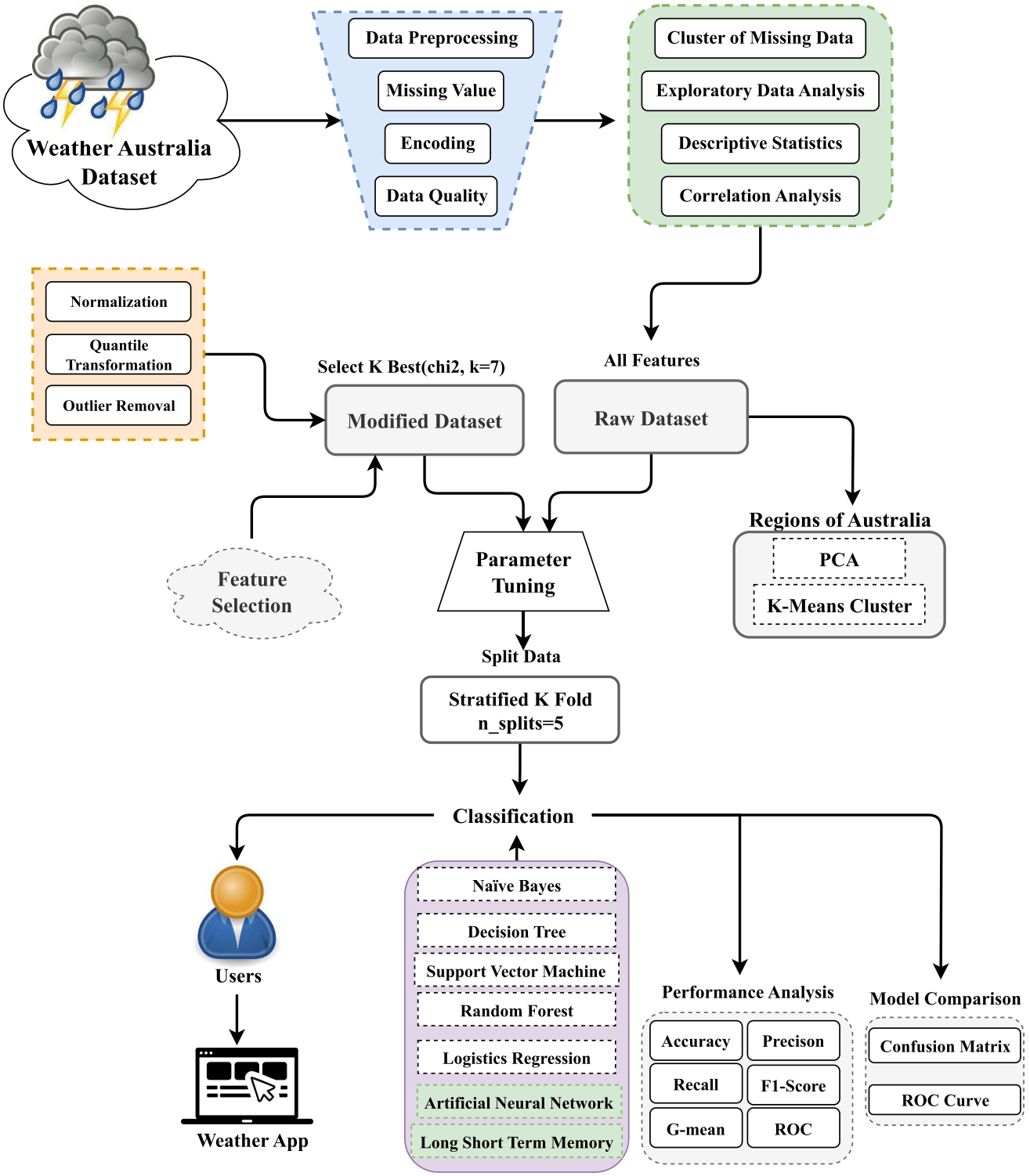


FIGURE 1. The proposed methodology for this study is presented as a sequential workflow diagram, outlining the step-by-step process of the research.

informative snapshot that allows viewers to quickly grasp the project's scope, objectives, and interconnected of various aspects.

A. DATASET DESCRIPTION

The dataset used in this study comprises regular weather observations from various Australian weather stations over

TABLE 1. Features descriptions of the train dataset.

Features	Description
Date	Date of the observation
Location	The location of the weather station's common name
Min Temp	In degrees Celsius, the minimum temperature.
Max Temp	In degrees Celsius, the maximum temperature
Rainfall	The total quantity of rain reported during the day in millimeters
Evaporation	In the 24 hours to 99 am, the so-called Class A pan evaporation (mm)
Sunshine	In the 24 hours leading up to 99 a.m., the so-called Class A pan evaporation (mm)
Wind Gust Dir	In the 24 hours leading up to midnight, the direction of the highest wind gust
Wind Gust Speed	The highest wind gust's speed (km/h) in the 24 hours leading up to midnight
WindDir9am	At 9 a.m., the wind was blowing in the other direction.
WindDir3pm	At 33 p.m., the wind was blowing in the other direction.
WindSpeed9am	Before 9 a.m., the wind speed (km/hr.) averaged over 10 minutes.
WindSpeed3pm	Wind speed (km/hr.) averaged over 10 minutes before 3 pm
Humidity9am	Humidity (percent) at 9 am
Humidity3pm	Humidity (percent) at 3 pm
Pressure9am	Atmospheric pressure (HPA) reduced to mean sea level at 9 am
Pressure9am	Atmospheric pressure (HPA) reduced to mean sea level at 3 pm
Cloud9am	At 9 a.m., a portion of the sky was shrouded by a cloud. The unit of measurement is "oktas," which are eights. It shows how many eights of the sky are covered with clouds. A score of zero denotes a clear sky, while an eight indicates a clouded sky.
Cloud3pm	At 3 p.m., a portion of the sky was shrouded by clouds (in "oktas": eighths). For an explanation of the values, see Cload9am.
Temp9am	Temperature (degrees C) at 9am
Temp3pm	Temperature (degrees C) at 3pm
Rain Today	If the amount of precipitation (mm) in the 24 hours leading up to 9 a.m. exceeds 1 mm, the value is 1. Otherwise, it is 0.
Rain Tomorrow	If the amount of precipitation (mm) in the 24 hours leading up to 9 a.m. is greater than 1 mm, the value is 1; otherwise, the value is 0.

a period of ten years. The goal of the research was to determine the most suitable classification model for predicting whether it will rain tomorrow, based on a range of features extracted from the dataset. By employing effective feature engineering techniques, the researchers aimed to identify the most informative variables that would contribute to the construction of highly accurate prediction models [15].The data collection period spans from 01-12-2008 to 25-06-2017, during which the Australian Bureau of Meteorology automatically collected data from 49 weather stations. The dataset contains a total of 23 distinct features and 145,460 observations. Among these features, 17 are continuous variables, while the remaining six are discrete variables. For instance, the target variable to forecast is whether or not it will rain tomorrow, indicated by a binary value of ''yes'' or ''no.'' In this context, ''yes'' signifies that it will rain the following day if the rainfall for that day is recorded as 1mm or more [8].

To provide a concise overview of the dataset variables, Table 1 presents a summary of their characteristics and descriptions. This information aids in understanding the nature of the dataset and the types of variables considered for rainfall prediction.

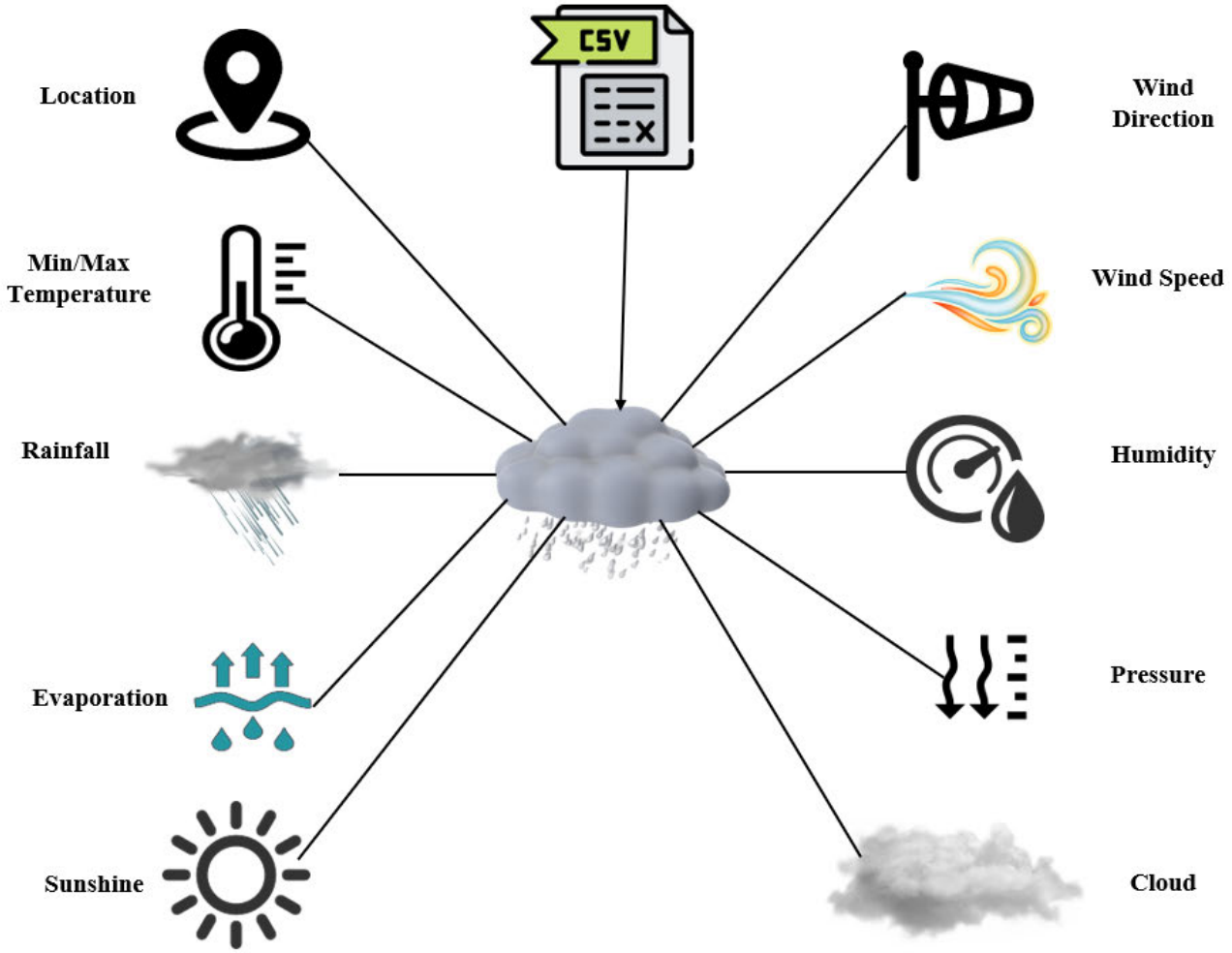


FIGURE 2. The figure illustrates various factors that influence rainfall.

B. DATA PREPROCESSING

Data Preprocessing is often used in the field of machine learning to describe the steps taken to clean, organize, and prepare raw data before it is used to construct machine learning models [16]. Preprocessing methods may be used to get rid of certain abnormalities while keeping others untouched [17]. In our data, the full date attribute is not helpful for machine learning; hence, the Month is extracted from the date. Consequently, the date attribute will be dropped from the dataset. Then Delete the original Date column. Besides, the Encoding of discrete points is performed in this section. Each issue is mapped to a set of numbers defined depending on the number of unique items presented in the attribute. To achieve a uniform or normal distribution for the characteristics, the Quantile Transformation technique is employed, which involves transforming the quantiles. This process effectively distributes the predominant values associated with a specific component. Moreover, it mitigates the influence of (marginal) outliers, establishing a robust preprocessing approach.

C. OUTLIER

An outlier is a value in a random sample of a population that deviates abnormally from the other values [18]. Outliers for each attribute are presented using the boxplot function. These outliers needed to be processed to ensure the model's correct performance since they represent discrepancies in the data instances. In the outlier distribution part, we can see in Figures 4 and 5 that on all numerical digits, we look for outliers. Boxplots will be used. Multiple boxplots make the plot more visible, and we can quickly see outliers. These variables feature many outliers, as seen by the boxplots. Rainfall, evaporation, wind gust speed, and minimum and maximum temperatures Wind speed 9 am, 3 pm; humidity 9 am; pressure 9 am, 3 pm; 9 am temperature 3 pm, 9 am, 3 pm, 3 pm Outliers may be noticed in the Wind Speed 3 pm data. For example, we've presented a boxplot with and without outliers. Examine the distribution of the variables. We will proceed by creating histograms to visually assess whether the distributions exhibit normalcy or skewness. If a variable follows a normal distribution, extreme value

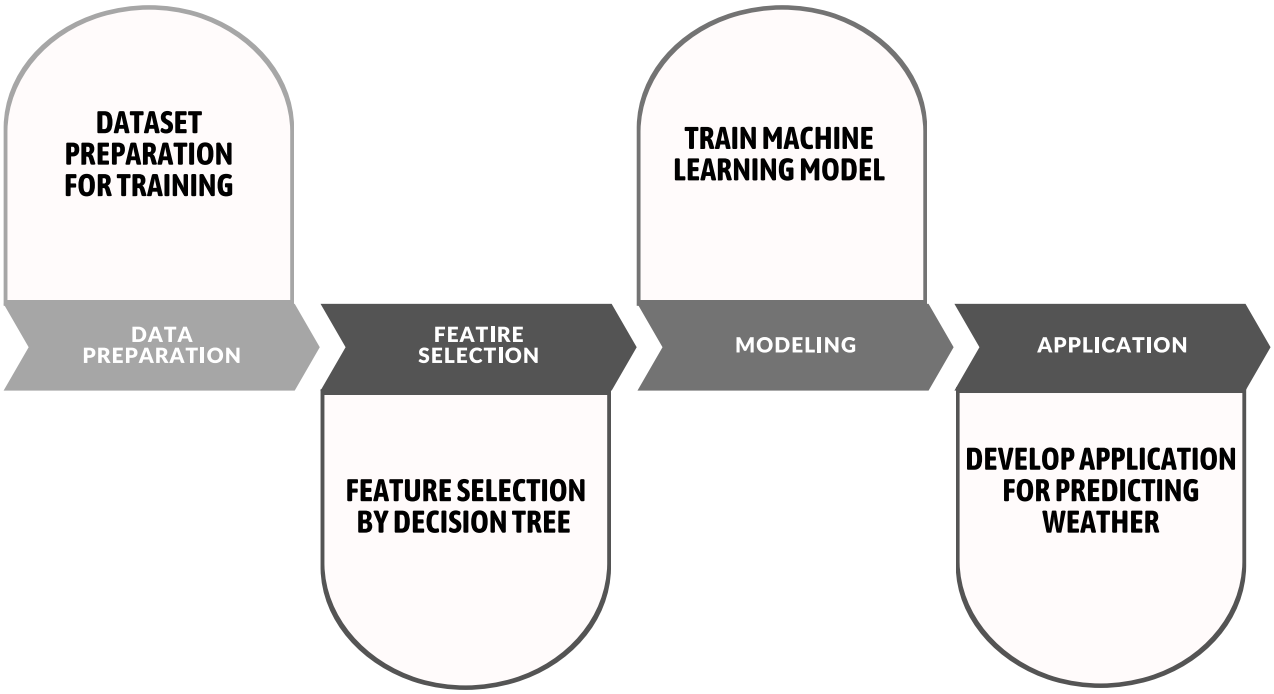


FIGURE 3. Overview of the proposed system of this study.

analysis will be applied. Conversely, if the distribution is skewed, the interquartile range (IQR) approach will be employed. Understanding the concepts of skewness and Gaussian distribution is essential for effectively handling outliers. For this purpose, we can employ the indicators of skewness and Gaussian distribution.

Outliers were removed/dropping outliers from the raw data to find better results from the dataset.

D. NORMALIZATION

A normalization is an approach to reducing the number of inserts, deletes, and changes that happen in a database because of duplicate data that can cause problems [19]. The process of normalization can improve data integrity and reduce dataset redundancy. It aids in the database's data organization. Numerous studies have demonstrated the significance of data normalization for enhancing data quality and, thus, the performance of machine learning algorithms [20]. Setting the data into tabular form and removing duplicate data from relational tables involves several steps.

E. QUANTILE TRANSFORMATION

The variables in this study are converted with the help of quantile transformation techniques which follow a uniform or normal distribution method. So, this transformation technique tends to spread out the most frequent observation for a given variable. Again, it reduces the effect of outliers as raw data has so many outliers making it a robust strategy. Using the Quantile transformer function, quantile transformation is carried out on the training dataset where all characteristics have values between 0 and 1. Figure 6 shows the Quantile Transformation histogram.

F. CLASSIFICATION MODELS

Classification is a supervising technique that categorizes the data into the desired number of classes. Multiple factors make intelligible classification models important. Users must understand a computer-induced model to trust and follow its predictions [21]. This work aims to determine the factors behind burning and predict a person's probability of being dead. In that manner, we can decrease the damage rate and keep people away from this deadly damage. So, we have employed seven classifiers: Naïve Bayes, Decision Tree, SVM, Random Forest ad Logistic Regression. Finally, we compared their performance based on different model evaluation metrics and found the best-fitting algorithm for this problem. In addition, we considered hyperparameter tuning and 10-fold cross-validation to make our model more robust and generic. Table 7 shows the parameter distributions of each classifier.

1) NAÏVE BAYES (NB)

The Bayes Theorem makes the NB classifiers, collecting classification algorithms [22]. It's a collection of algorithms based on the knowledge that every pair of classified qualities is independent of one another. This approach just needs a small quantity of training data to obtain classification parameter estimations [23]. The weather forecast for tomorrow, for example, is included in our dataset. Based on the present circumstances, each tuple determines whether the weather is

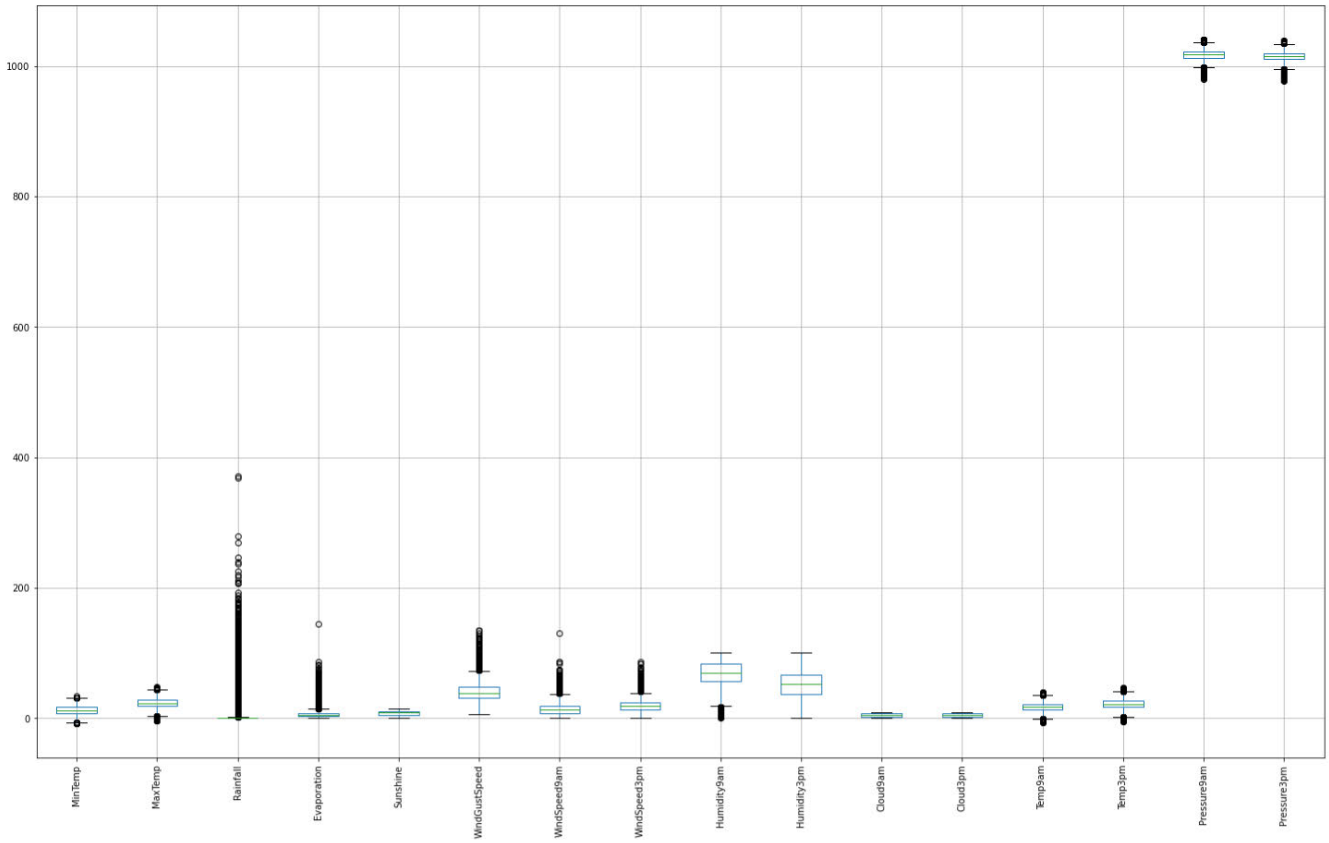


FIGURE 4. The figure presents a visual representation of the Boxplot used for outlier detection. The Boxplot allows for clear visualization of the distribution of data and enables the identification of any potential outliers within the dataset.

suitable (''Yes'') or unsuitable (''No'') for Rain tomorrow. It is the conditional probability that determines this. Bayes' theorem is represented mathematically as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(A)P(B)}, \tag{1}$$

Here, A and B are events, respectively.

The prior probability of A and B is P(A) and P(B), respectively, without consideration of each other. P(A|B), commonly known as posterior probability, is the chance of seeing event A if B is true. P(B|A), commonly known as likelihood, is the likelihood of seeing event B if A is true.

2) DECISION TREE (DT)

The C4.5 method is substantially implemented in J48. The J48 decision tree algorithm detects the most and most minor essential traits, and then separates them into subtree attributes. Unlike the Random Forest decision tree, it yields a binary tree [24]. A decision tree can be done with chi-square, Information gain, or Gini index. The goal behind the decision tree approach is to break the data into smaller pieces and then utilize the knowledge gained from those pieces to choose which characteristic to prioritize. [25]. Figure 7 shows the working flow of the Decision Tree.

When employing a decision tree (DT) as a classifier, two independent processes must be implemented: first, the tree must be built, and then classification must be performed using the top-down technique in DT development. The anticipated value is used at each node to decide what is likely to be the optimal split. The decision rule for each branch in the tree is labeled, and the anticipated value for each terminal node is also labeled. Cross-validation guarantees that the best tree size is chosen, avoiding the problem of overfitting. In each node, the decision tree algorithm always discovers the most significant qualities [26]. As a result, the importance of features is domain knowledge that is useful to the decision tree algorithm. To include this feature relevance score in decision tree learning, our system employs a unique way.

This classification algorithm's penetration level was set to 7 and the desired solution for the used dataset throughout this investigation was achieved by the classifier using this maximum depth value.

ENTROPY

$$H(S) = -(P_{i+}log_2P_{i+} + P_{i-}log_2P_{i-}), \tag{2}$$

where,

$H(S)$ is used to find out the entropy of the current dataset

$H(S)$ = Entropy of the current or main dataset

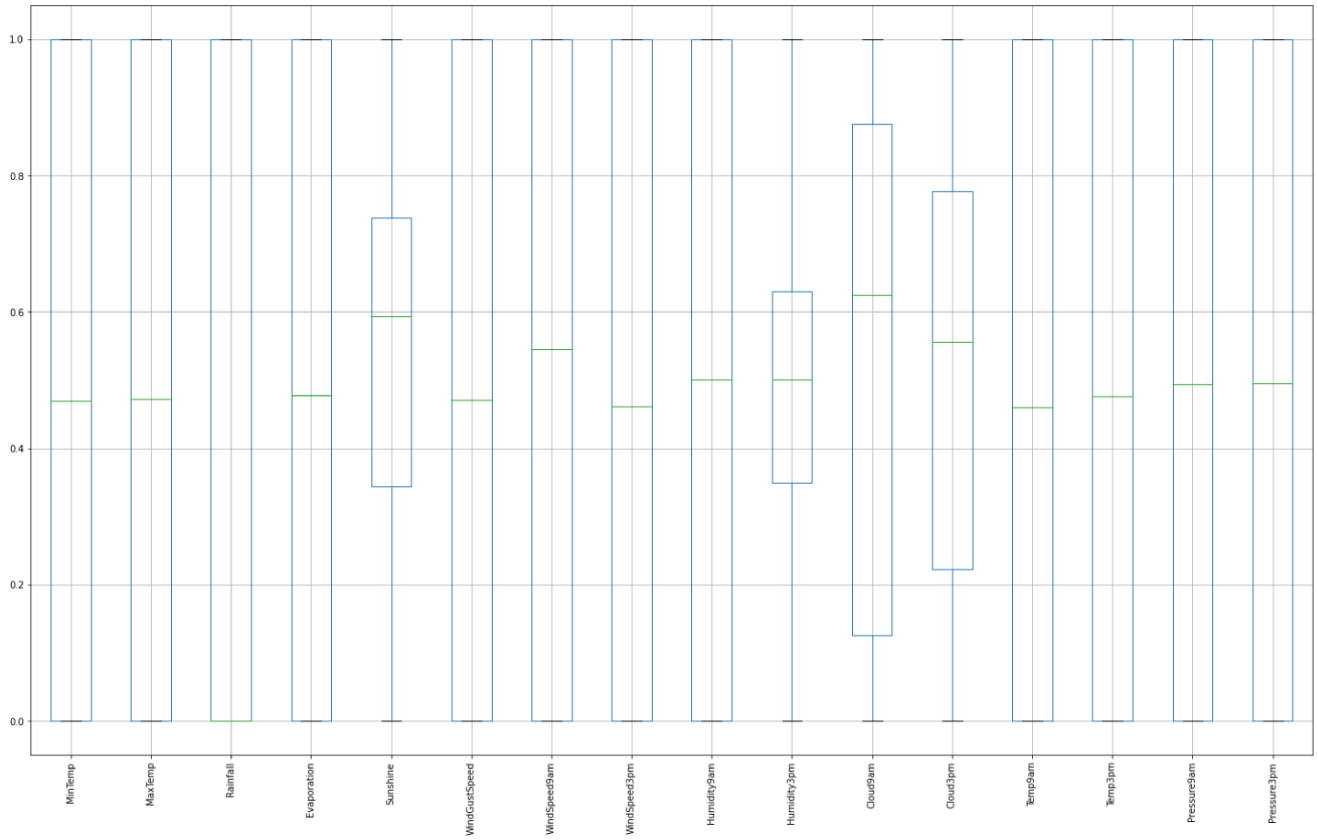


FIGURE 5. The Boxplot visualization is shown after the removal of outliers. It illustrates the distribution of data without the presence of any identified outliers, providing a clearer understanding of the dataset's features.

P_{i+} = Probability of positive class in S
P_{i-} = Probability of negative class in S

INFORMATION GAIN

$$Gain(S, F) = H(S) - \sum_{v \epsilon F} P(v) * H(S_v), \tag{3}$$

where,
S = Target dataset
F = Feature
$v \epsilon F$ = Entropy of selected attribute of feature F
$P(v)$ = Probability of selected attribute of feature F

$$Gini\ Index = 1 - \sum_{i=1}^{n} P_i^2, \tag{4}$$

3) SUPPORT VECTOR MACHINE (SVM)

An SVM is a supervised machine learning model with two groups that utilize classification [27]. It is a collection of linked algorithms that are applied for classifications and regressions [28]. SVM models can categorize data sets after being given labeled training data. Again Support vector machine (SVM) is one of the most popular supervised learning algorithms in Machine Learning. The purpose of the SVM algorithm is to establish the hyperplane or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify new data sets in the future. The term "hyperplane" refers to this optimal decision boundary. SVM selects the extreme vectors and points which help in the creation of the hyperplane. Support vectors, which are used to represent these severe cases, form a foundation for the SVM method.

4) RANDOM FOREST (RF)

The ensemble approach is a powerful technique that improves prediction accuracy by combining the predictions of multiple algorithms [29]. Decision trees, such as classification and regression trees, are known for their high variance, which can lead to overfitting. Random forests, also referred to as random decision forests, are a popular ensemble learning method used for classification, regression, and other tasks. This approach involves constructing a large number of decision trees and combining their results to make predictions. In classification tasks, the final prediction is determined by the majority class among the individual trees, while in regression tasks, it is the mean forecast of the individual trees. Random forests are built using a collection of Classification and Regression Trees (CART) [30]. One of the key advantages of random forests is their ability to mitigate the overfitting tendency of decision trees. By aggregating the predictions of multiple

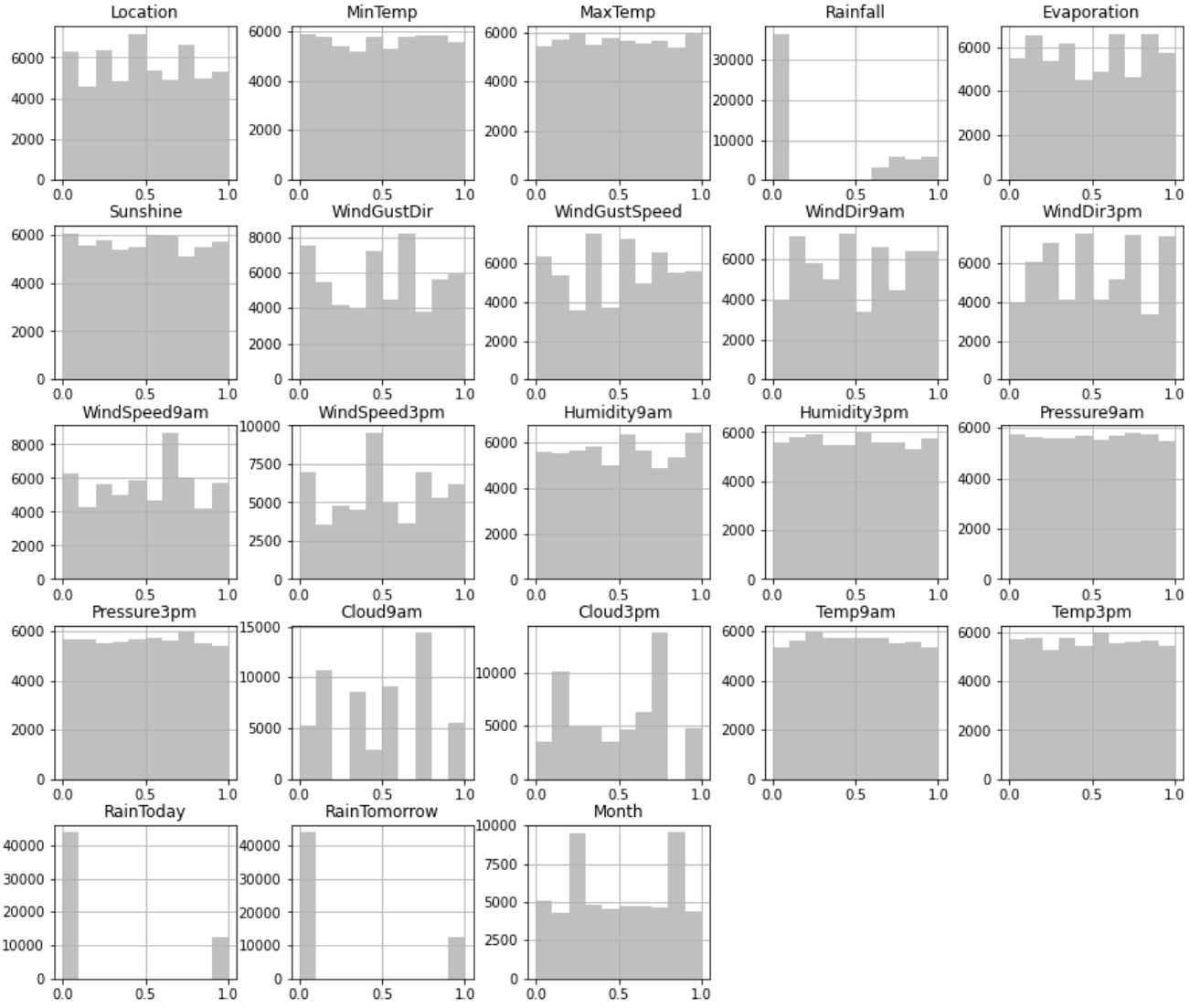


FIGURE 6. Histogram after quantile transformation of the dataset.

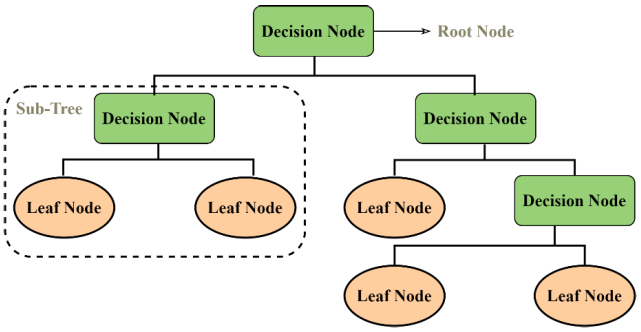


FIGURE 7. The working flow of decision tree.

trees, random forests provide a more robust and accurate prediction. To ensure the effectiveness of random forests, two important conditions should be met:

- All features should contain meaningful information that enables the models built with them to outperform random guessing.
- The predictions of the individual trees, as well as their errors, should have low correlations with each other.

By satisfying these conditions, random forests can effectively capture diverse patterns and make accurate predictions by leveraging the collective knowledge of the ensemble. Random forests are commonly used in supervised learning tasks and have shown great success in various domains. The model architecture of the random forest model is shown in Figure 8.

5) LOGISTICS REGRESSION (LR)

The logistic process is known as logistic regression. The logistic process, also called the sigmoid function, was created by statisticians to characterize the characteristics of

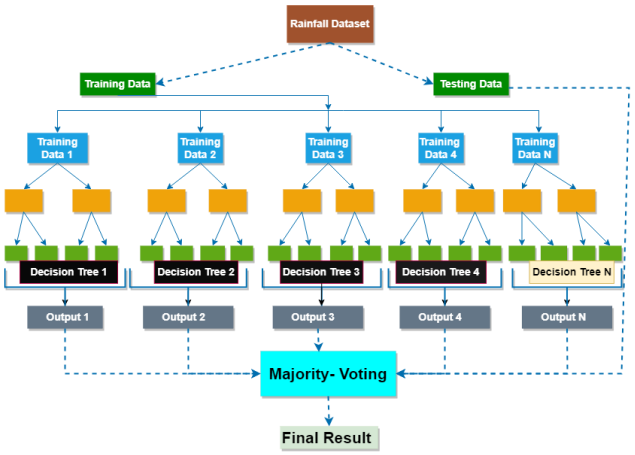


FIGURE 8. Random forest classifier architecture.

population development in ecology. The chance of belonging to one of the two groups in the data set is calculated using a logistic regression model [31]. In the field of machine learning, logistic regression has become an essential technique. It enables machine learning algorithms to categorize incoming data using historical data.

The Linear Regression equation can be applied to calculate the Logistic Regression equation.

For multivariable, straight line equation is,

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_r x_r + \varepsilon, \qquad (5)$$

The ultimate equation of logistic regression is,

$$P(X) = \frac{1}{1 + e^{-f(x)}}, \qquad (6)$$

Which comes under the activation function of sigmoid.

Each and every real number can be transformed into a value between 0 and 1 using the sigmoid function which is a mathematical tool that transforms linear regression to logistic regression.

where,

$$f(x) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_r x_r + \varepsilon, \qquad (7)$$

Also called the logit,

Where the variables $\beta_0, \beta_1, \ldots\ldots, \beta_r$ are the estimator of the regression coefficients, which are also called the predicted weights.

6) ARTIFICIAL NEURAL NETWORK (ANN)

Artificial neural networks (ANNs) are computer programs inspired by the structure and function of the human brain [32]. They aim to mimic the brain's information-processing capabilities. ANNs learn and make predictions by identifying patterns and correlations in data, rather than relying on explicit programming instructions. One of the key advantages of ANNs is their speed, adaptability, and robustness in various settings, including noisy and novel environments. They have found broad applications due to these qualities. ANNs were designed to replicate the behavior of biological systems, specifically the central nervous systems of animals, which consist of interconnected neurons [33]. Neural networks are computational models that consist of simulated neurons, which are analogous to the neurons in the human brain. In the brain, neurons use electrical and chemical signals to transmit and process information. These neurons are connected through structures called synapses, enabling communication between them. Similarly, in artificial neural networks, a large number of simulated neurons are interconnected to form a network. Neural networks are not limited to categorization tasks but can also be used for regression tasks involving continuous target attributes. In the field of data mining, neural networks have numerous applications. They have gained popularity due to their ability to handle large volumes of data and achieve higher accuracy compared to traditional machine learning models. The architecture of ANN model is shown to Figure 9.

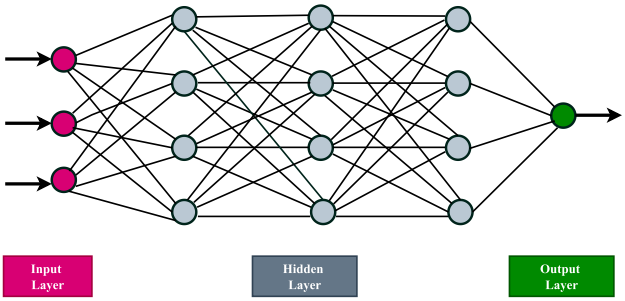


FIGURE 9. The architecture of Artificial Neural Network model.

7) LONG SHORT-TERM MEMORY (LSTM)

The short-term memory neural network is a specific type of recurrent neural network (RNN) that has the ability to retain and utilize previously processed information within the network [34]. Long Short-Term Memory (LSTM) networks, a variant of RNNs, excel in learning order dependencies in sequence prediction tasks, making them well-suited for various complex problem domains such as machine translation and speech recognition. LSTM networks are designed to address the issue of vanishing gradients in traditional RNNs by introducing specialized memory cells that can selectively retain or discard information over time. These memory cells allow LSTMs to effectively capture long-term dependencies in sequential data. The computational complexity of LSTM networks is relatively low, with a per-time-step and per-weight complexity of 0.1 [35]. LSTM networks have demonstrated significant advantages over other recurrent network algorithms, including real-time recurrent learning, backpropagation through time, recurrent cascade correlation, Elman nets, and neural sequence chunking. They tend to achieve higher success rates and learn faster in various tasks. Moreover, LSTM networks excel in handling complex challenges with long-time lags, which previous recurrent network algorithms struggled to solve. Figure 10 illustrates the structural components of the LSTM model. It consists of memory cells, input and output gates, and a forget

gate, which collectively enable the network to effectively process and retain information over time. By leveraging these architectural elements, LSTM networks can capture and utilize the sequential patterns and dependencies present in the data, leading to improved performance in various sequential prediction tasks.

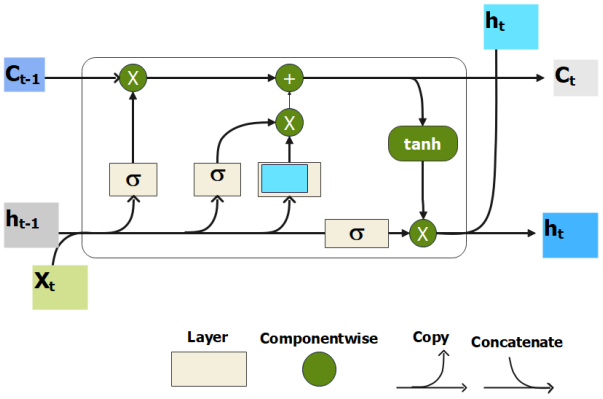


FIGURE 10. LSTM model structure is shown to layer-wise.

G. EVALUATION METRICS

The choice of evaluation metric plays a crucial role in determining the effectiveness of a classifier during classification training [36]. Selecting an appropriate evaluation metric is essential for comparing and identifying the best classifier among different models. There are several evaluation metrics available for assessing the performance of machine learning models, and they provide valuable insights into the classifier's performance. Evaluation metrics based on confusion matrices, such as accuracy, precision, recall, and F1-score, are commonly used to evaluate the performance of individual classifiers. Each metric captures different aspects of the classifier's performance and provides specific insights into its behavior. It is important to understand and interpret these metrics to assess the strengths and weaknesses of the classifier effectively. Table 2 gives a brief explanation of each of them. Besides confusion matrix serves as the foundation for calculating various parameters.

H. IMPLEMENTATION TOOLS AND ENVIRONMENTS

The studies have been carried out on a computer equipped with a 3.2 GHz Intel Core i5-6500 processor with 4 cores. A chipset from Intel was utilized, and the RAM speed was set at 3000 megahertz with 8 gigabytes. All the experiments have been performed in Python. Also, we take help from the cloud platforms like Google Collaboratory and Kaggle Kernel.

IV. RESULT AND ANALYSIS

A. DATASET DESCRIPTIVE STATISTICS

A descriptive statistic is a summary statistic that quantitatively characterizes or summarizes features from a collection of all dataset information. It's a relationship between a group of to-be-defined beings and a set of descriptive values, with

TABLE 2. Evaluation metrics.

Measures	Definitions	Formula
Accuracy	The algorithm's accuracy in predicting variables is calculated by its accuracy.	A = (TP+TN) / (Total no. of samples)
Precision	Precision assess the correctness.	P=TP/(TP+FP)
Recall	The recall is used to evaluate a classifier's completeness or sensitivity.	R=TP/(TP+FN)
F1-score	The average of Precision and Recall is known as F1-score.	F=2*(P*R)/(P+R)
G-mean	Root Product of Class Based Sensitivity.	$G_{mean} = \sqrt{TP * TN}$
ROC	Sensitivity and specificity comparison	Sensitivity Vs. Specificity

the condition that each being is linked to precisely one explanatory value. A summary of data excluding date and location is given in Table 3.

Any prediction model will work best if the data used to train it is accurate and healthy. Weak classifiers are frequently the result of incorrect data values. This step is critical for ensuring valuable data. We count missing values for each attribute except date and location in our preprocessing effort. Missing values are allocated and processed in this section, in which the data instance, including any missing attribute value, is dropped. The dark areas (gaps) in the below heatmap represent the not-available values in the respective attribute. Moreover, in our class (Rain Tomorrow at Figure 11), 22.42% are yes and 77.58% are no. Using data visualisation techniques, researchers use exploratory data analysis (EDA) to analyse and examine data sets and summarise their key features. It can also assist in determining if the statistical approaches considered for data analysis are appropriate.

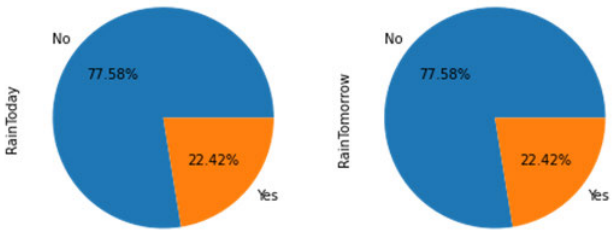


FIGURE 11. Pie chart of today and tomorrow rain prediction.

Figure 12 indicates the changes in Evaporation and Rainfall to Wind Direction.

The dendrogram in Figure 13 presents a comprehensive hierarchical clustering analysis, illuminating how features are grouped based on their missing value patterns through binary distance. The clustering process involves the systematic assembly and division of clusters to optimize the reduction of binary distance within each group. The result is a structured arrangement that highlights relationships among features and their missing data tendencies. Upon closer examination of the dendrogram, a clear distinction emerges between two main clusters. The first cluster encompasses columns characterized by a higher frequency of missing values, while the second cluster comprises features with more complete data. Notably, as we move towards the leaf nodes, the number of missing values decreases, signifying columns with comparatively

TABLE 3. Descriptive statistics of variables.

Variables	Min	Max	Mean	SD	Variance	Skewness	Kurtosis
Min Temp	-8.5	33.9	12.19	6.39	40.94	.02	-.484
Max Temp	-4.8	48.1	23.22	7.11	50.68	.22	-.22
Rainfall	.0	371.0	2.36	8.47	71.87	9.83	178.15
WindSpeed9am	0	130	14.04	8.91	79.48	.77	1.22
WindSpeed3pm	0	87	18.66	8.810	77.61	.628	.764
Humidity9am	0	100	68.88	19.02	362.10	-.48	-.03
Humidity3pm	0	100	51.54	20.79	432.47	.03	-.51
Temp9am	-7.2	40.2	16.99	6.48	42.10	.08	-.34
Temp3pm	-5.4	46.7	21.68	6.93	48.11	.23	-.13
Evaporation	1	359	259.14	124.22	15432.67	-.91	-.67
Sunshine	1	146	108.75	49.03	2404.34	-.94	-.72
Wind Gust Dir	1	17	9.15	4.95	24.59	-.064	-1.24
Wind Gust Speed	1	68	35.18	11.67	136.36	1.52	2.13
WindDir9am	1	17	8.75	4.82	23.28	.05	-1.19
WindDir3pm	1	17	9.39	4.91	24.19	-.14	-1.18
Pressure9am	1	547	220.37	131.85	17385.02	1.43	1.41
Pressure3pm	1	550	203.02	140.03	19610.11	1.51	1.48
Cloud9am	1	11	7.58	3.52	12.42	-.61	-1.07
Cloud3pm	1	11	7.75	3.41	11.66	-.62	-1.04
Rain Today	1	3	2.20	.450	.203	.75	.36
Rain Tomorrow	1	3	2.20	.450	.20	.75	.36

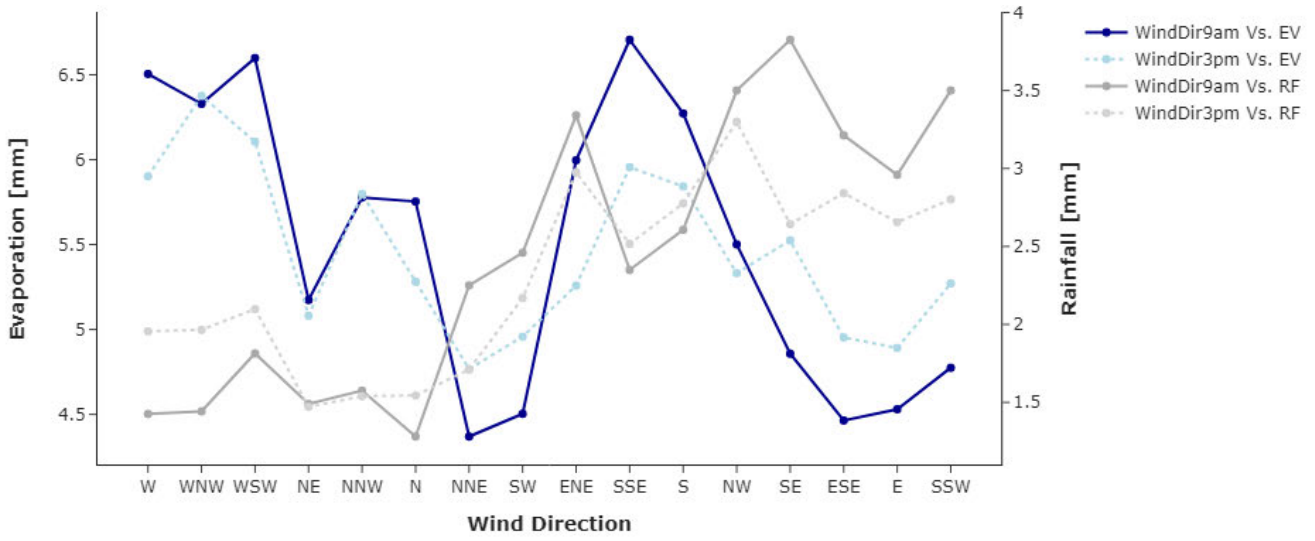


FIGURE 12. Evaporation and rainfall with wind direction.

higher data availability. Analyzing specific features, such as Rain Tomorrow, Rain Today, Rainfall, Date, and Location, reveals a noteworthy pattern they all possess zero missing values, indicating their robust data completeness. Additionally, the proximity of a branch to zero on the dendrogram directly correlates with fewer missing values, underscoring the trend of enhanced data availability. As we shift our focus to predictions, an ensemble of columns, including Humidity 9 am, Wind Speed 9 am, Temp 9 am, Minimum Temperature, Maximum Temperature, Rainfall, Date, Rain Tomorrow, and Rain Today, emerges as particularly coherent and intact for observational predictions. This strategic selection of features reflects a well-informed approach to leveraging data for predictive modeling, enhancing the accuracy and reliability of outcomes.

Within the multivariate analysis section, a comprehensive depiction of the attribute distribution is presented through the utilization of both bar charts and pie charts. Notably, the analysis discerns that a significant portion, amounting to 70%, of the dataset falls under the ''float64'' data type category. In contrast, the remaining 30% of the data is categorized as object type time. This classification of data

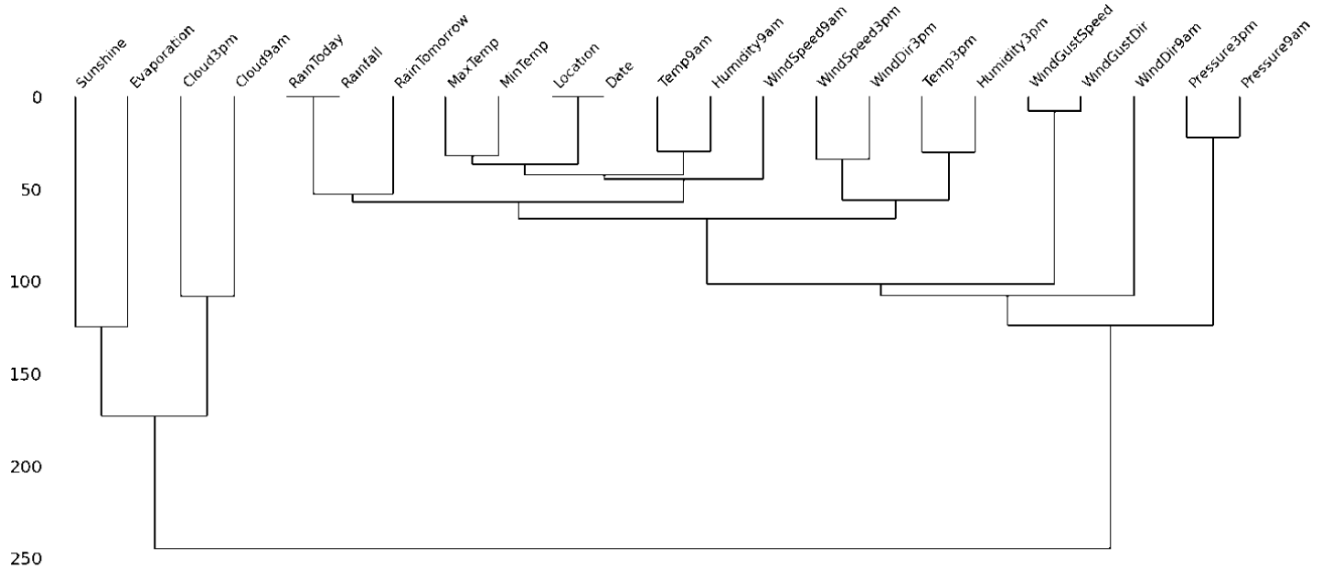


FIGURE 13. Cluster analysis of missing value of the dataset.

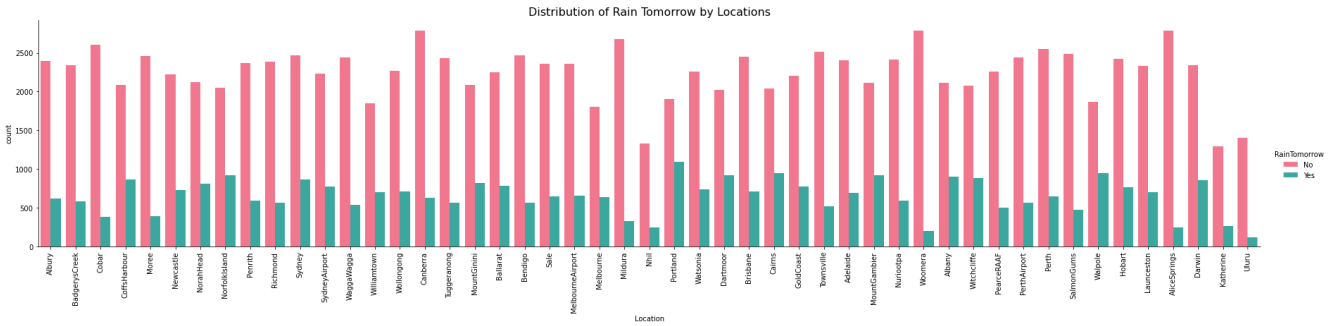


FIGURE 14. Distribution of rain tomorrow concerning location's variables.

types is pivotal, as all attributes are required to be continuous in order to facilitate the seamless execution of the subsequent classification model. This uniformity in data type is also crucial for smooth processing within the subsequent stages of analysis. A visual representation in Figure 14 delves into the Distribution of Rain in relation to Location Variables. This illustrative depiction allows for an insightful examination of how rainfall patterns correlate with specific locations. The histogram serves as a powerful visualization tool, enabling us to uncover potential trends or anomalies within the dataset that might not be readily apparent through numerical analysis alone. This comprehensive approach further enriches our understanding of the dataset, facilitating more informed decision-making as we progress through the analytical stages.

In the provided Figure 15, a comprehensive visualization of the distribution of all numerical attributes is presented. This visualization offers valuable insights into the nature of these attributes, with some features exhibiting a characteristic normal distribution characterized by a mean of zero and a variance of sigma square. This observation carries significant implications for our analysis, as the normal distribution of attributes often serves as an underlying assumption in various statistical methods and machine learning algorithms. Expanding on the insights, Figure 16 offers a pair plot that visualizes relationships and patterns among various features within the training data. This matrix of scatterplots provides a dynamic overview of attribute interactions, revealing potential correlations, trends, and outliers that might exist within the dataset. Such a comprehensive analysis is paramount in uncovering intricate relationships that may not be immediately apparent from individual attribute analyses. By visually exploring these pairwise interactions, we gain a more holistic understanding of the data's underlying structure, allowing us to make more informed decisions when devising our analytical strategies.

Together, the distribution analysis in Figure 15 and the insights from the pair plot in Figure 16 contribute to our ability to glean essential insights from the dataset, shaping our subsequent analytical methodologies and enhancing the accuracy and depth of our findings.

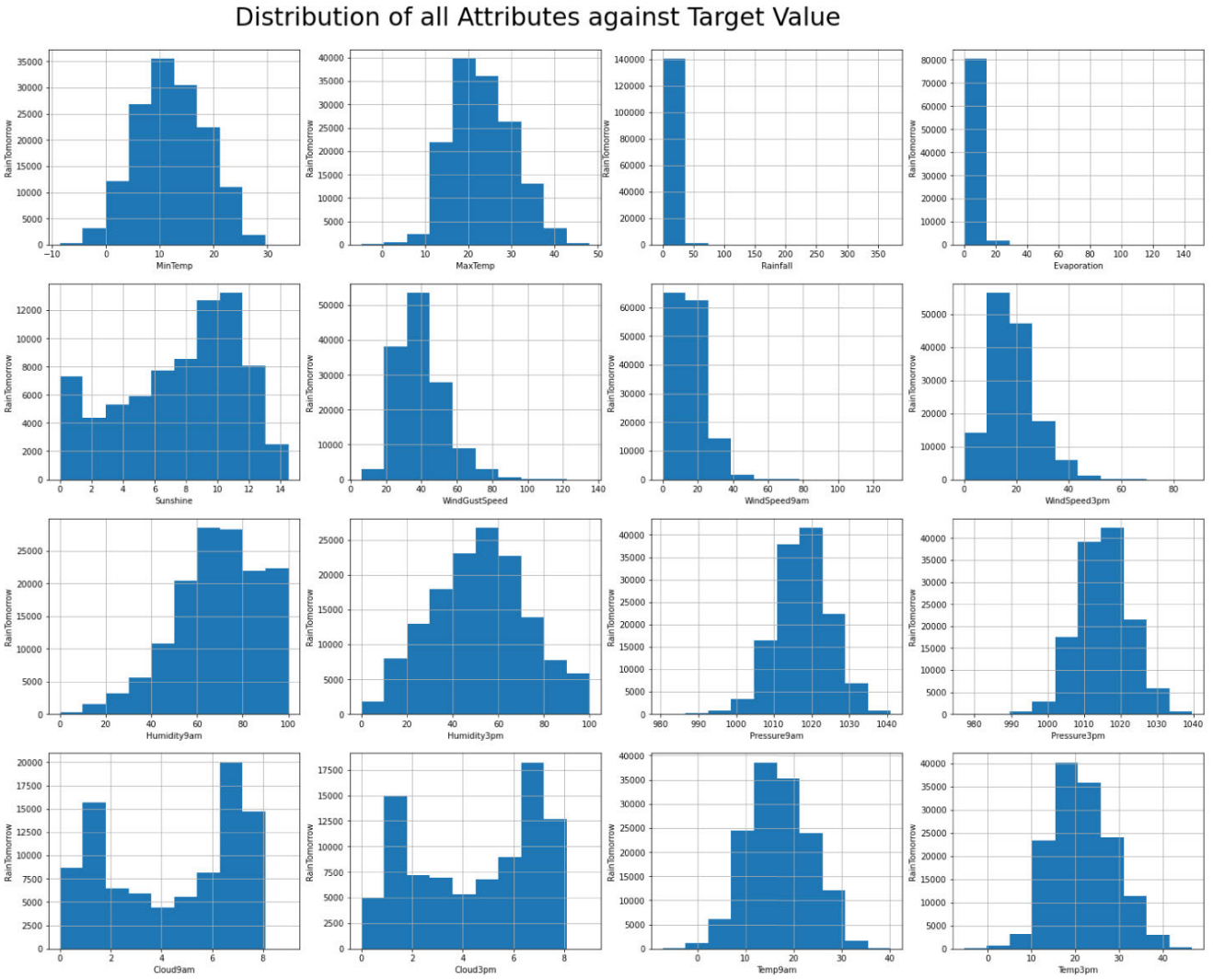


FIGURE 15. Distribution of the variables of the training dataset.

B. CORRELATION ANALYSIS

One notable finding is the strong positive correlation between the variables ''Min Temp'' and ''Temp 3 pm.'' This connection suggests that, in general, the minimum temperature experienced earlier in the day aligns with the temperature at 3 pm, reflecting a consistent trend in temperature variation throughout the day. Similarly, the variables ''Min Temp'' and ''Temp 9 am'' also exhibit a substantial positive correlation, further emphasizing the coherent temperature patterns observed throughout different time intervals. The variables ''Max Temp'' and ''Temp 9 am'' showcase a pronounced positive correlation, indicating that higher temperatures in the morning are often indicative of elevated maximum temperatures later in the day. Similarly, the strong positive correlation between ''Max Temp'' and ''Temp 3 pm'' reinforces the notion of a consistent temperature trend as the day progresses. Among other findings, the heatmap reveals a noteworthy association between ''Wind Gust Speed'' and ''WindSpeed3pm.'' This positive correlation suggests that higher wind gust speeds are accompanied by elevated wind speeds in the afternoon hours, reflecting the interconnected of wind-related attributes. Another correlation that stands out is the significant positive relationship between ''Pressure9am'' and ''Pressure3pm.'' This connection, as indicated by the correlation coefficient, suggests a consistent pressure pattern throughout the day, emphasizing the interplay between morning and afternoon atmospheric pressure levels. The correlations highlighted in the heatmap, such as those between temperature variables and wind-related attributes, unveil underlying patterns and trends within the dataset. These findings provide valuable insights for subsequent analyses, modeling, and decision-making, enhancing the depth and accuracy of our understanding of the dataset's intricate relationships.

The correlation heatmap depicted in Figure 17 provides a visual representation of the relationships among various

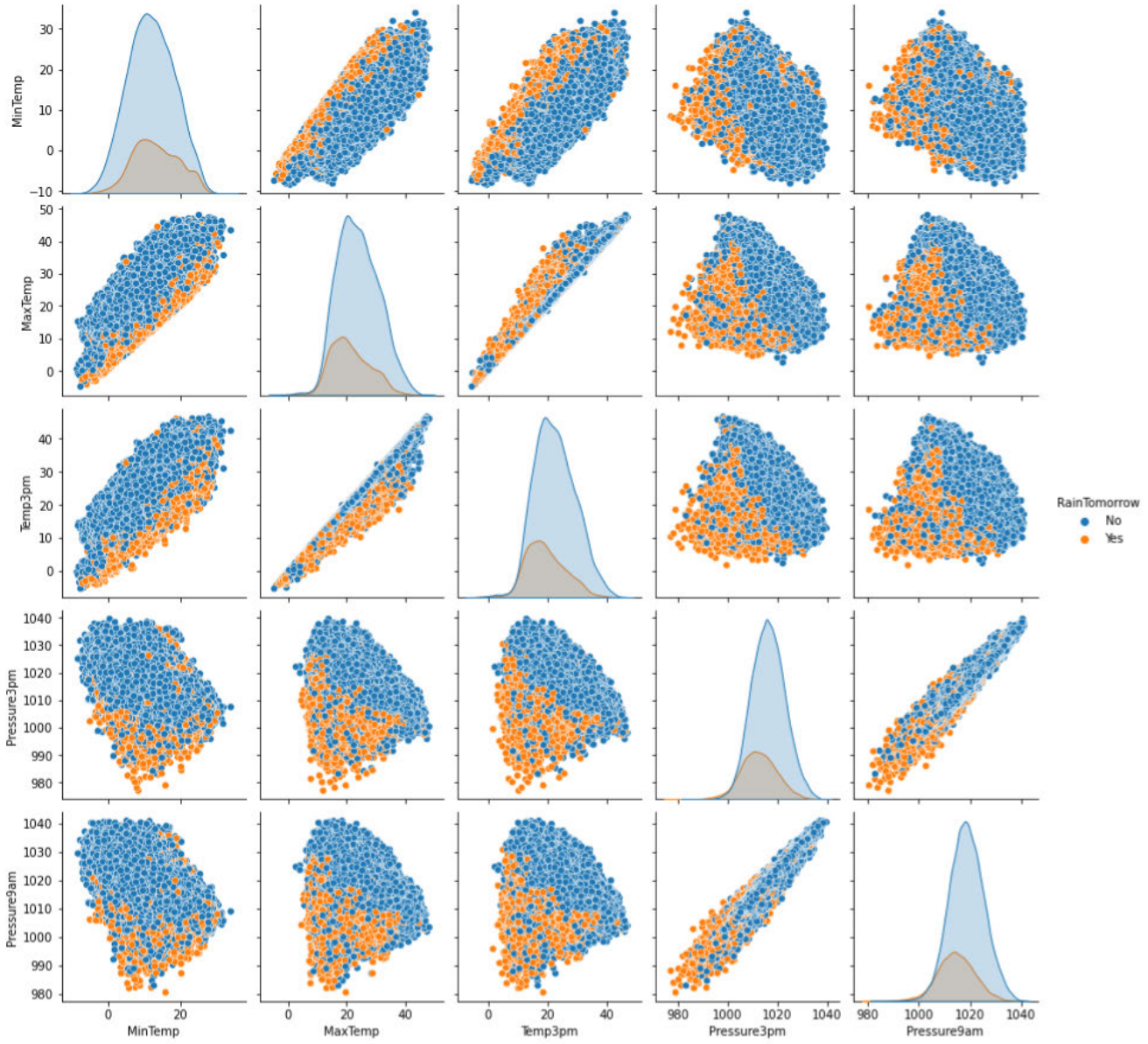


FIGURE 16. Pair plot the training dataset to determine the optimal collection of features for explaining a connection between two variables or for forming the most distinct clusters.

variables within the dataset. Through a careful analysis of this heatmap, several significant positive correlations between different variables emerge, offering insightful findings that contribute to our understanding of the dataset's underlying dynamics.

C. CLUSTER ANALYSIS AND PRINCIPLE COMPONENT ANALYSIS

K-means clustering is a powerful unsupervised machine learning technique used to partition data points into clusters based on their similarity. By applying K-means clustering to the dataset, regions with similar meteorological attributes are grouped together, shedding light on potential geographical patterns and relationships that could influence rainfall. This clustering approach aids in identifying regions that share similar weather characteristics, thereby assisting in understanding the factors contributing to rain occurrence.The application of Principal Component Analysis (PCA) introduces an additional layer of analysis. PCA is a dimensionality reduction technique that identifies key patterns and correlations within the data by transforming the original variables into a new set of orthogonal variables called principal components. In this context, PCA assists in identifying the most influential variables that contribute to the observed patterns of rainfall occurrence. This technique enhances our comprehension of the complex interplay of meteorological factors that impact rainfall. The integration of both K-means clustering and PCA presents a comprehensive approach

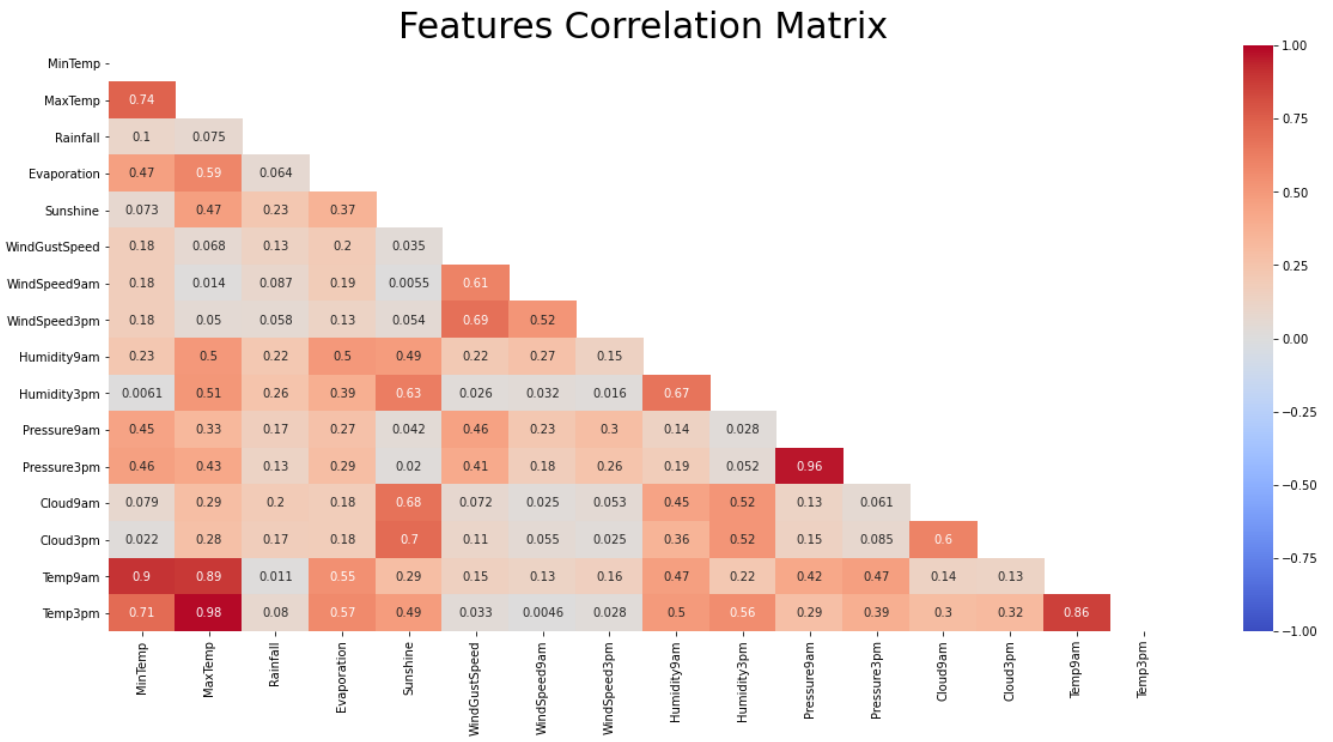


FIGURE 17. Correlation analysis of training dataset.

to discerning rain patterns across Australian regions. By analyzing the clustered regions and the principal components that contribute to these clusters, the study aims to unveil geographical trends, local variations, and meteorological factors that collectively influence the likelihood of rainfall in distinct areas. This holistic approach provides valuable insights into the intricate relationships between weather attributes and rainfall occurrence, contributing to a more nuanced understanding of the factors governing rainfall patterns in Australia.

In Figure 18, a comprehensive analysis is presented wherein the prediction of rainfall in distinct regions across Australia is examined through the application of two distinct techniques: K-means clustering and Principal Component Analysis (PCA). This investigation seeks to discern discernible patterns within the dataset that could potentially provide insights into the likelihood of rainfall occurrence in specific geographic areas.

In the initial segment of Figure 19, a visual representation of the clusters formed based on the relationship between rainfall and temperature across various locations within the dataset is depicted. Subsequently, the application of Principal Component Analysis (PCA) involves the utilization of two principal components to encapsulate the cluster characteristics. Notably, the choice of k=5 clusters aligns with the five distinct places identified in the dataset. Each cluster elucidates distinctive weather patterns, thus enabling the differentiation of meteorological attributes within each cluster.Since the cluster model is employed on the raw dataset, the graph visualization displays outliers in green and purple hues, exemplifying the variances introduced due to the unique characteristics of specific regions. A comprehensive observation reveals that, in a significant portion of the areas, temperature and rainfall exhibit a close correlation. Notably, locations represented by the red and orange clusters are situated in close proximity to one another, indicating similarity in weather attributes.

In the subsequent Figure 19, a distinct cluster analysis is presented, incorporating a broader array of variables. Approximately 80% of the dataset's variance is accounted for, with a focus exclusively on numerical attributes, as PCA operates optimally with such variables. While variables like Wind Direction are excluded from this analysis due to the nature of PCA, the inclusion of attributes like clouds, sunshine, and wind speed in conjunction with rainfall and temperature further enriches the insights gained.The application of PCA introduces dimensionality reduction, highlighting the most significant variables contributing to the clusters. This process effectively condenses the dataset's complexity, allowing a more focused analysis. As a result, the clusters depicted in Figure 19 exhibit greater dispersion, indicating a broader range of weather attributes that collectively contribute to the distinct meteorological patterns.These analyses underscore the power of clustering and PCA in uncovering intricate relationships within the dataset. The visual representations and insights gleaned from these techniques provide a more nuanced understanding of the interplay between various meteorological variables and their

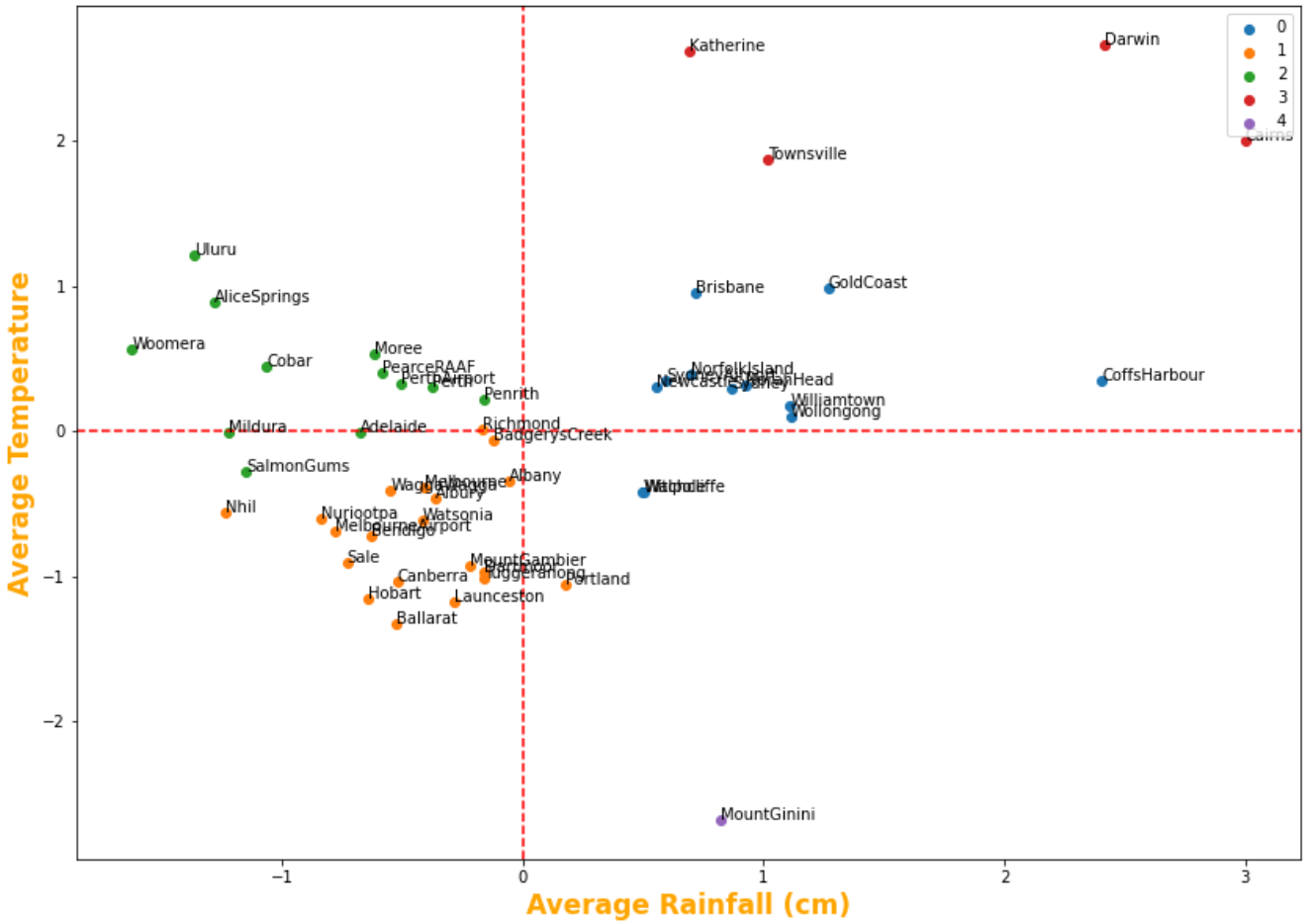


FIGURE 18. Cluster analysis of the dataset.

impact on rainfall and temperature patterns across different locations.

D. FEATURE SELECTION

Feature selection plays a pivotal role in enhancing the accuracy and efficiency of predictive models, such as those used in weather forecasting. By identifying the most influential attributes that contribute significantly to the target variable, Rain Tomorrow in this case, feature selection streamlines computational processes and holds paramount importance in the broader context of machine learning endeavors [37]. Leveraging this feature selection insight, a critical step is undertaken to extract these key attributes from the original dataset. Subsequently, the dataset is primed for the application of diverse machine learning algorithms, aiming to ascertain their performance metrics, including Accuracy, Precision, Recall, and F1-score. These metrics are comprehensively assessed through the analysis of confusion matrices, thereby evaluating the efficacy of the chosen machine learning methods in predicting Rain Tomorrow.

The assessment of feature importance based on scoring reveals that the last five attributes possess remarkable significance, as indicated by their high scores. Accordingly, the optimal features that contribute most notably to the predictive capability are discerned as Rain Today (4199.82), Rainfall (3221.39), Sunshine (1822.64), Humidity 9 am (1510.93), and Cloud 3 pm (1123.83), among others. For instance, the feature importance of the decision tree algorithm is elucidated in Figure 20, showcasing how each attribute contributes to the decision-making process. This analysis empowers us to understand the hierarchical significance of features and their impact on predictions.

E. MODEL PERFORMANCE

Table 4 serves as a comprehensive repository of the performance evaluation metrics for various classification models. It offers a clear snapshot of how each model fares in predicting Rain Tomorrow. However, the true essence of this analysis is unveiled after implementing feature selection, which systematically narrows down the attributes to the most influential ones, refining the model's predictive prowess.Before delving into feature selection, the initial assessment reveals that the Artificial Neural Network (ANN) stands out with a commendable accuracy rate

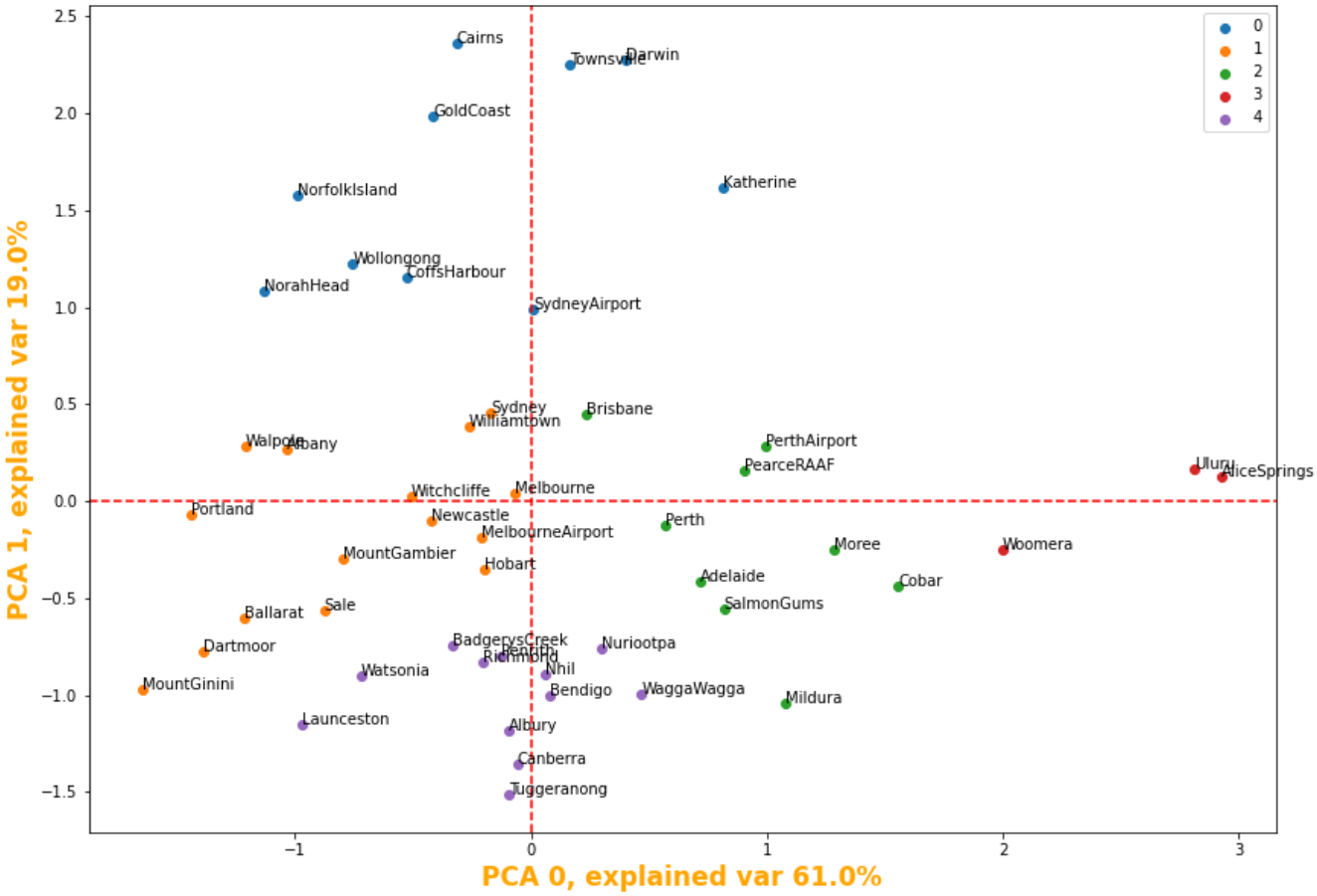


FIGURE 19. The visualization illustrates the cluster analysis conducted on the dataset derived from PCA.

of 90%. This demonstrates its proficiency in deciphering the complex relationships within the dataset and making accurate predictions.

After the feature selection process, the true impact of selected attributes on the model's performance is unveiled. Notably, the ANN maintains its position as the top classification model, reaffirming its supremacy. The precision of 90% signifies the model's ability to accurately predict positive cases out of all cases it labeled as positive. Likewise, the recall rate of 89% highlights the model's capability to identify a significant proportion of actual positive cases. The F1 score, a balanced metric combining precision and recall, also attains a remarkable 89%, further validating the model's overall effectiveness. These findings underscore the pivotal role of feature selection in enhancing the predictive potential of the classification model. By selecting the most influential attributes, the model becomes more attuned to the intricacies of the data, enabling it to make predictions with higher accuracy, precision, recall, and overall performance. The consistent dominance of ANN throughout this evaluation reinforces its suitability for accurate Rain Tomorrow predictions, positioning it as a robust and reliable choice for this forecasting task.

TABLE 4. Comparative performance of classification algorithms on various measures.

Classifier	Accuracy	Precision	Recall	F1-score	G-mean	ROC
Naïve Bayes (NB)	0.84	0.76	0.78	0.77	0.77	0.77
Decision Tree (DT)	0.80	0.70	0.69	0.70	0.68	0.68
SVM	0.86	0.82	0.74	0.76	0.72	0.73
Random Forest (RF)	0.85	0.82	0.71	0.74	0.69	0.70
Logistics Regression (LR)	0.86	0.82	0.74	0.77	0.72	0.73
ANN	0.90	0.90	0.89	0.89	0.78	0.79
LSTM	0.80	0.78	0.80	0.77	0.71	0.50

Upon conducting feature selection to enhance the classifier's effectiveness, Table 5 becomes a comprehensive canvas illustrating the performance metrics of Accuracy, Precision, Recall, and F1-score for the various classification algorithms. This post-feature selection evaluation provides a deeper understanding of each algorithm's ability to predict Rain Tomorrow with improved attribute relevance. Remarkably, the Artificial Neural Network (ANN) emerges as the front runner, showcasing an exceptional accuracy rate of 91%. This achievement reinforces the ANN's competence in extracting and leveraging the most crucial attributes to make highly

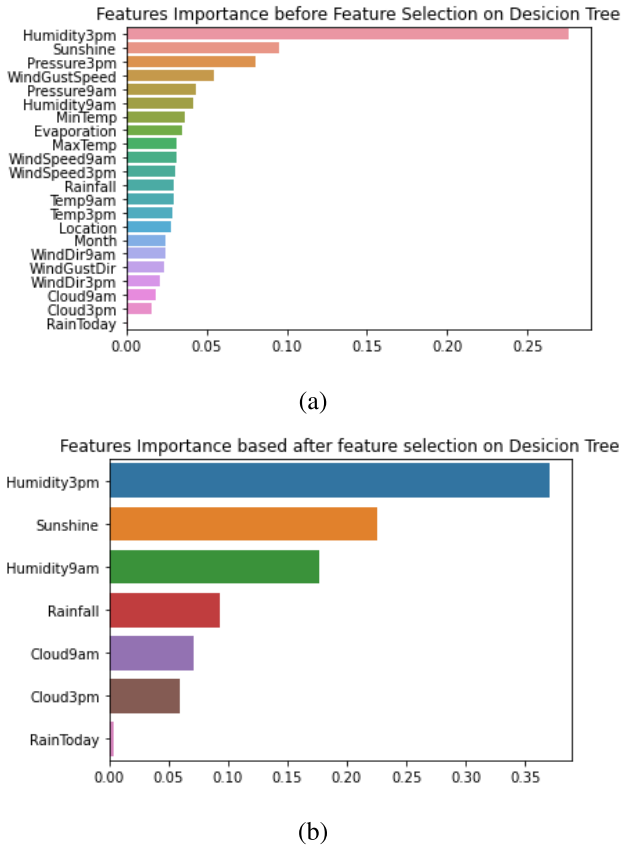


FIGURE 20. The feature importance analysis using a decision tree reveals the significance of different features. In (a) all the feature importances are presented, while in Figure (b) the selected features are ranked based on their importance.

accurate predictions. The precision, recall, and F1-score values associated with ANN further underline its supremacy and its ability to strike a harmonious balance between identifying positive cases and avoiding false positives.

TABLE 5. Performance after feature selection.

Classifier	Accuracy	Precision	Recall	F1-score	G-mean	ROC
NB	0.81	0.79	0.81	0.80	0.77	0.77
DT	0.81	0.80	0.81	0.79	0.68	0.68
SVM	.84	0.82	0.79	0.76	0.72	0.73
RF	.81	0.80	0.81	0.79	0.69	0.70
LR	.83	0.82	0.74	0.65	0.77	0.67
ANN	**.91**	0.88	0.89	0.89	0.87	0.78
LSTM	.83	0.78	0.78	0.81	0.85	0.72

A pivotal tool for evaluating the performance of a classification model is the confusion matrix, which quantifies the accuracy of predictions and any instances of wrong classification. Upon analyzing the performance of the Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) classifiers, the confusion matrices in Figure 22 provide a detailed depiction of their predictive accuracy and potential wrong classifications. The exploration extends to encompass the remaining five models. The Receiver Operating Characteristic (ROC) curve, illustrated in Figure 23, captures the performance of these Machine Learning (ML) models, further revealing their ability to distinguish between positive and negative cases. Additionally, Figure 24 portrays the ROC curve for all the ML models, offering a comprehensive view of their collective ability to balance true positive and false positive rates. Turning our attention to Deep Learning (DL) models, Figure 25 unveils the ROC curve for this subset of algorithms, showcasing their ability to make accurate predictions while minimizing false positives. Moreover, Figure 26 provides a consolidated perspective on the ROC curve for all DL models, offering insights into their collective ability to optimize the trade-off between precision and sensitivity.

F. RESULT ANALYSIS AND COMPARISON

The evaluation of various classifiers based on accuracy and other pertinent evaluation metrics has been meticulously conducted in line with the Australian weather dataset. Each experimental setup adhered to a consistent design, involving the partitioning of data into training and testing sets, along with the essential step of outlier removal to ensure data quality. This methodical approach guarantees robustness and reliability in our findings. The efficacy of the four classification methods was systematically gauged using a spectrum of performance metrics, including Precision, Accuracy, F1-Score, and more. Through these metrics, the models' predictive capabilities were assessed, encompassing both correct and incorrect categorizations. The results from Table 4 underscore the exceptional performance of the Artificial Neural Network (ANN), exhibiting a remarkable accuracy of 90% and an impressive F1-Score of 0.89. These values, approaching the threshold of 1, signify a potent classifier for this specific dataset.Upon delving deeper into the feature selection process, Table 5 reaffirms the prominence of the ANN classifier, as it maintains the highest accuracy post-feature selection. This reinforces the notion that ANN excels in predicting class probabilities, outperforming its counterparts.

Before applying feature selection techniques, we evaluated the model's performance using a confusion matrix, as illustrated in Figure 21. This matrix provided a snapshot of how the model was predicting outcomes based on the entire set of available features. After implementing feature selection, we re-assessed the model's performance, and the resulting confusion matrix is depicted in Figure 22. Notably, the comparison between the two matrices indicates a discernible improvement in the model's performance post-feature selection. This enhanced performance can be attributed to the more focused and informative set of features that were retained. The prevalence of correct predictions, as evidenced by the true positives and true negatives, increased noticeably, contributing to an overall better model performance. This analysis validates the efficacy of the feature selection process in enhancing the model's predictive power. The refined feature set not only led to a more

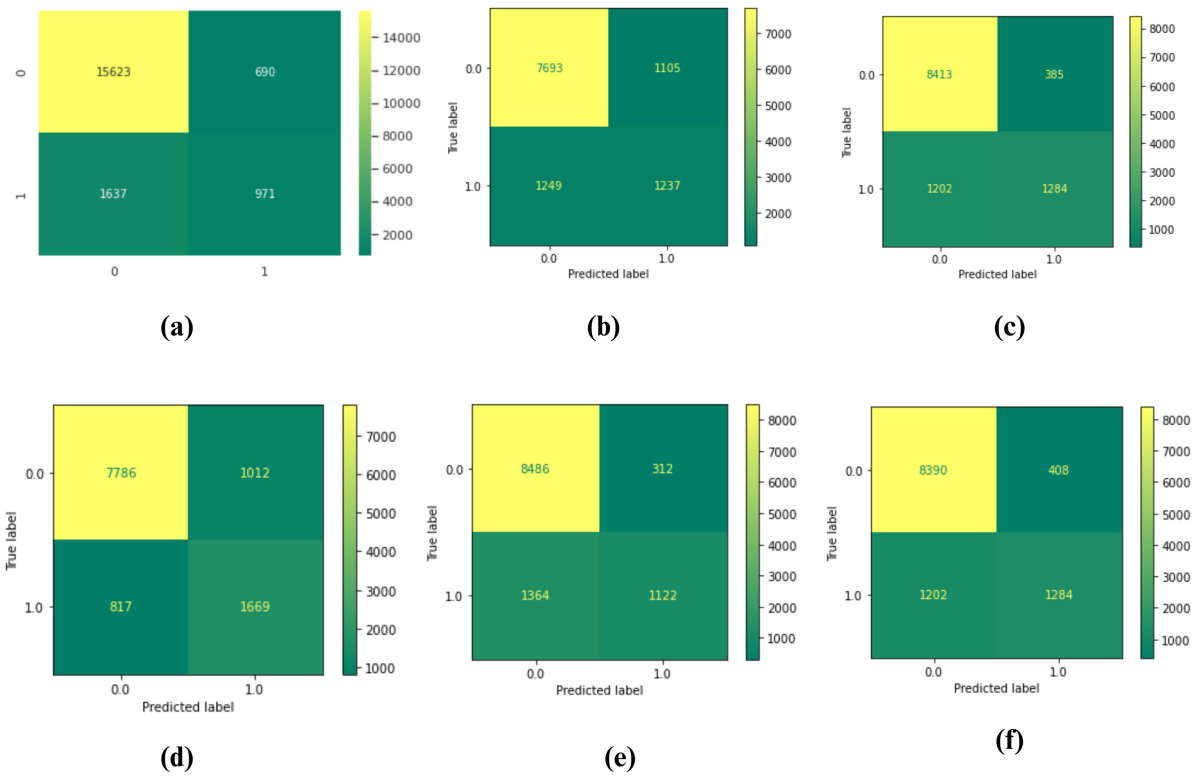


FIGURE 21. The performance evaluation of the different models on the actual dataset is measured using confusion metrics. The confusion matrix is shown to (a) ANN, (b) DT, (c) LR, (d) NB (e) RF, and (f) SVM models.

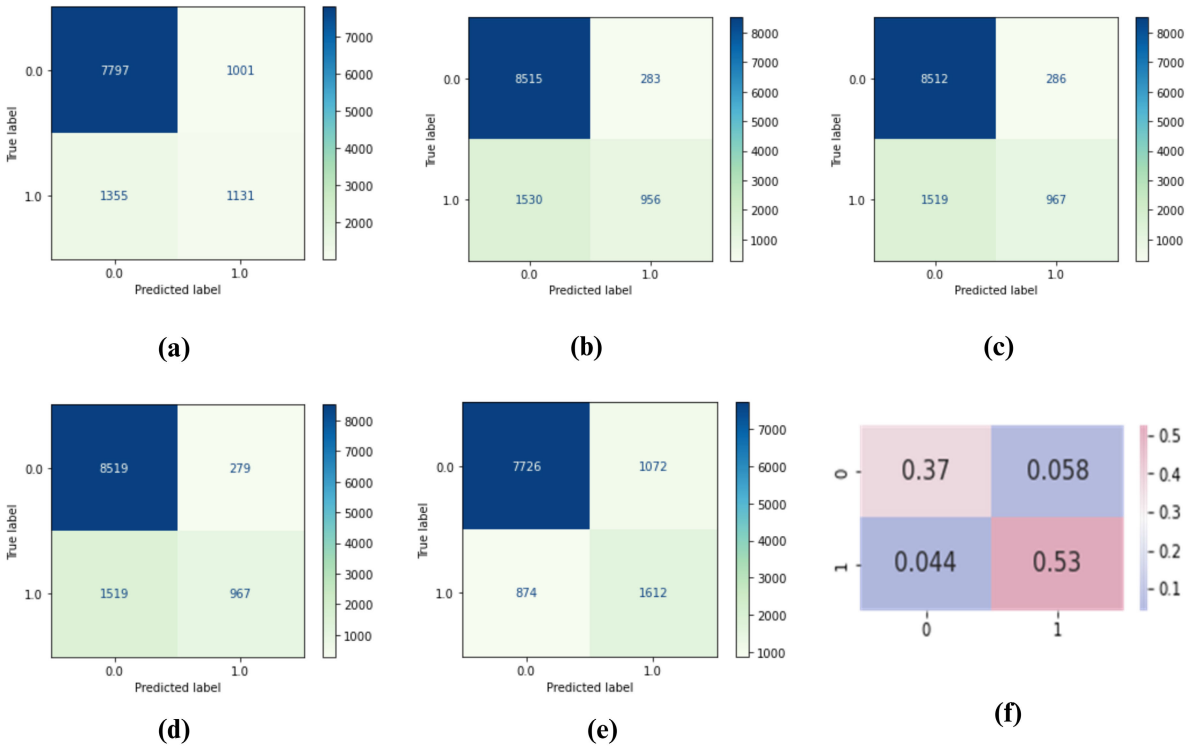


FIGURE 22. The performance evaluation of the different models on the selected features dataset is measured using confusion metrics. The confusion matrix is shown to (a) DT, (b) LR, (c) RF, (d) SVM (e) NB, and (f) ANN models.

concise representation of the data but also resulted in a more accurate and robust predictive model. This finding underscores the importance of feature selection in optimizing machine learning models for better results and underscores

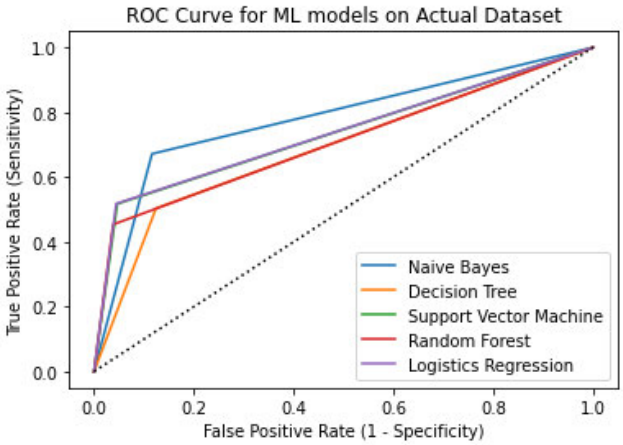


FIGURE 23. ROC analysis of machine learning models on the dataset before feature selection.

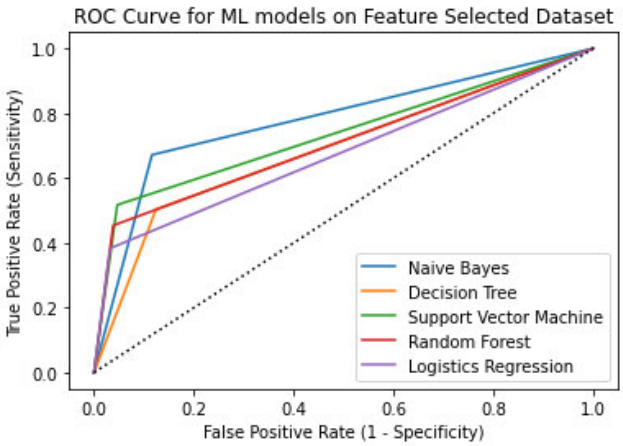


FIGURE 24. ROC analysis of machine learning models on selected faetures.

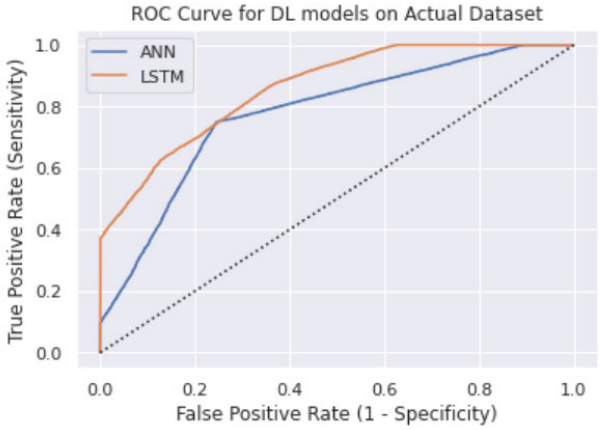


FIGURE 25. ROC of ANN and LSTM model on actual dataset before feature selection.

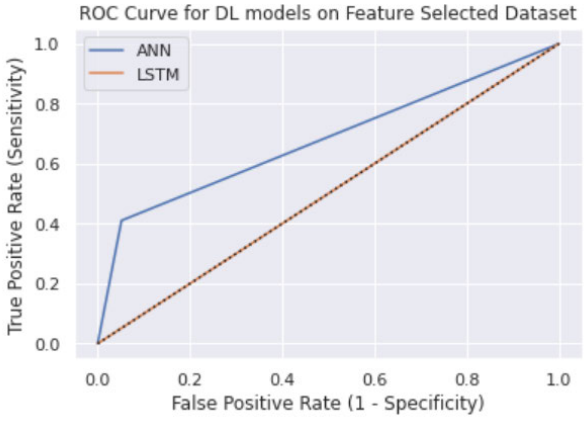


FIGURE 26. ROC of ANN and LSTM model after feature selection.

the value of a systematic approach to model refinement. In the context of Receiver Operating Characteristic (ROC) curves, Figure 23 portrays the performance of various machine learning models against the actual dataset. Comparatively, Figure 24 demonstrates the improved ROC performance after feature selection, highlighting the potency of the selected features in enhancing model discernment. Transitioning to deep learning models, Figure 25 illustrates the ROC curves for both ANN and LSTM models on the original dataset. Remarkably, Figure 26 reflects the enhanced ROC performance achieved by these models after the incorporation of feature selection. These graphs provide a visual representation of the models' ability to balance true positive and false positive rates, demonstrating the superiority of the ANN and LSTM models.

Furthermore, the findings from Table 6 underscore this research's exceptional accuracy when juxtaposed with existing works. Collectively, the presented evaluation metrics and graphical representations establish the dominance of the ANN classifier in predicting weather patterns accurately, especially after feature selection. This empirical exploration substantiates the suitability of the chosen machine learning and deep learning models, positioning our research as a pioneering contribution in weather prediction accuracy.

G. PARAMETER TUNING

The optimization of machine learning algorithms involved the deliberate adjustment of various associated parameters to enhance classification outcomes, as outlined by the reference [38]. This iterative refinement process, commonly known as fine-tuning, aimed to extract the best possible results from the classifiers. The table presents a comprehensive overview of the classifier tuning parameters utilized for various machine learning models in this study. These parameters play a crucial role in determining the behavior and performance of each classifier, thereby influencing the accuracy and reliability of the predictions. For the NB classifier, the parameter settings involve the choice of variance smoothing and the absence of priors. These settings contribute to the handling of categorical data and the absence of assumptions regarding prior probabilities. The DT classifier is configured with parameters such as a random

TABLE 6. Comprehensive comparison of the performance of the proposed rain prediction system with several current systems.

Ref	Methods	Dataset	Best model	Accuracy	Precision	Recall	F1-Score
He et al. [6], (2021)	Active Learning, Random Sampling	Rain in Australia (Kaggle), 21 Attributes	Active Learning	82.02%	-	-	-
Nolan et al. [7], (2017)	Decision Tree, Capital-City, Capstone Analysis	Australian Bureau of Meteorology (BOM)	Decision Tree Model with Capstone Analysis	75.6%	72.30%	91.70%	-
Balamurugan et al. [8], (2019)	LR, DT, RF, SA	India open Data Portal, 123,328 Records	LR	84%	86%	95%	95%
Ejike et al. [9], (2021)	LR, ANN	Open Data, 12 Variables	LR	87.6%	73.7%	61%	67%
Kumarasiri et al. [10], (2008)	NN	Two Combined Datasets	NN	74.3%	64%	-	-
Ria et al. [11], (2021)	DT, KNN, LR, NB, RF	Bangladesh Jatiyo Totho Batayon, 2391 Records	RF	87.68%	90.63%	92.68%	91.54%
Neelakandan et. al. [12], (2021)	Linear SVM, Random Forest, PBIL, CA-SVM W-SVM	Various Source of Real Data	CA-SVM	89%	75%	90%	-
Sankaranarayanan et. al. [13], (2019)	SVM, KNN, Naïve Bayes, DNN	India Water Portal 335 Records	DNN	90%	95%	93%	95%
Proposed Model	NB, DT D SVM, RF, LR ANN, LSTM	145,460 Observations, 23 attributes	ANN	91%	88%	89%	89%

state of 1 to ensure reproducibility, and a minimum impurity decrease of 0.01. These settings are important for controlling the tree growth and preventing overfitting. In the case of the SVM, the choice of parameters includes a maximum iteration of -1, which allows the optimization process to continue until convergence, along with the use of the radial basis function (RBF) kernel and 'auto' gamma value. These settings are pivotal in achieving the optimal separation of classes in higher-dimensional space. The RF classifier incorporates parameters like no maximum depth and a random state of 0, enabling the ensemble model to create diverse decision trees and mitigate overfitting. For Logistic Regression, the 'liblinear' solver is selected, coupled with a maximum iteration of 1000. These settings contribute to the efficient convergence of the logistic regression optimization process. The ANN classifier adopts the 'adam' optimizer and employs 'binary_crossentropy' as the loss function. These settings optimize the ANN's training process to achieve better convergence and prediction accuracy. The LSTM classifier is configured with a learning rate of 0.001, the 'adam' optimizer, and 'binary_crossentropy' loss function. These choices facilitate the effective training of LSTM

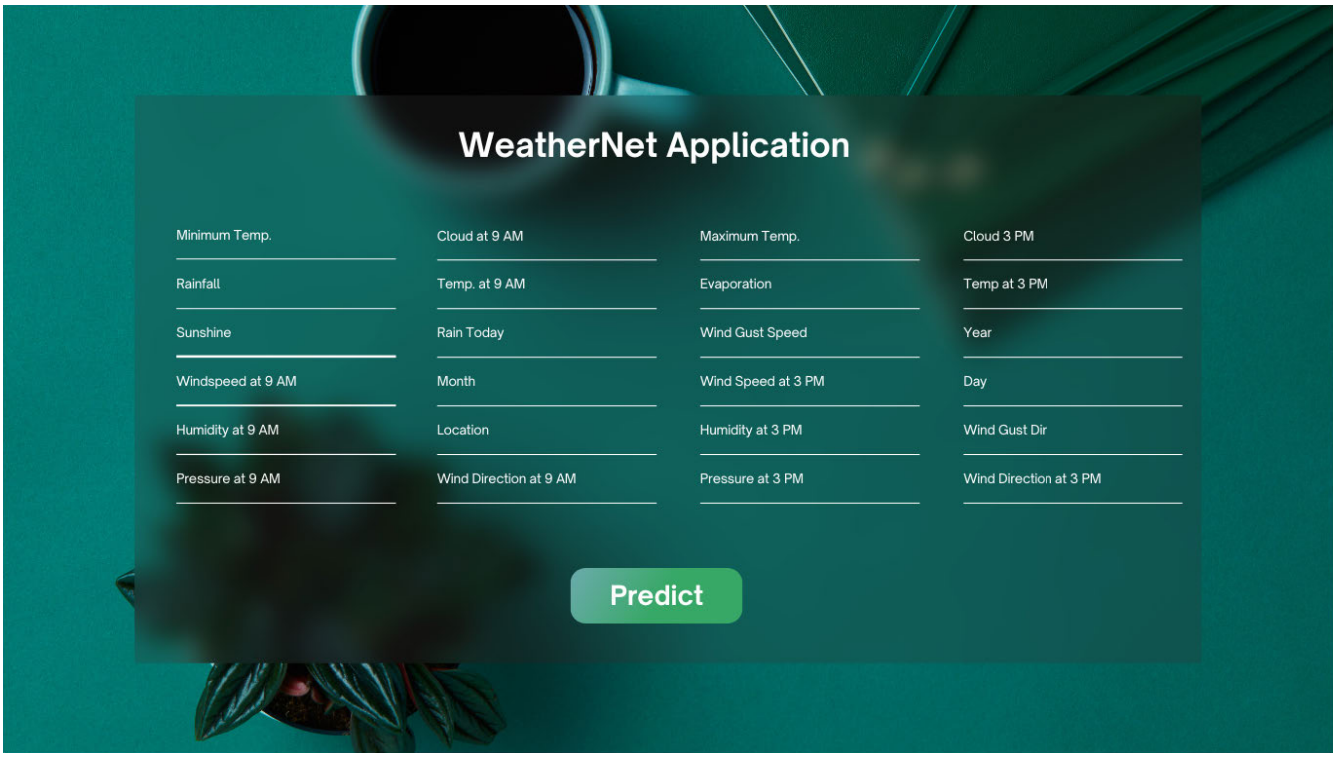


FIGURE 27. The weather application for predicting waterfall.

networks on sequential data. The chosen parameter settings for each classifier are meticulously designed to fine-tune their performance and enhance their ability to handle the dataset. These settings play a pivotal role in achieving accurate and reliable predictions, demonstrating the significance of parameter tuning in machine learning model development. A more granular understanding of the applications of these parameters can be gleaned from the comprehensive breakdown provided in Table 7. This tabulated resource furnishes detailed insights into how each parameter adjustment impacts the behavior and predictive power of the classifiers. This meticulous parameter optimization process, along with strategic pruning, collectively underpins the finesse and reliability of our classification outcomes, substantiating the efficacy of our methodology.

TABLE 7. The table presents an overview of the parameter settings used for the classification models in the study.

Classifier	Tuning Parameters
NB	var smoothing = 1e^-0, 9Priors = none
DT	random state = 1, min impurity decrease=0.01
SVM	max iter = -1, gamma='auto', Kernel = rbf
RF	Max depth = none, random state = 0
LR	Lib linear solver, max iter = 1000
ANN	optimizer='adam', loss='binary_crossentropy'
LSTM	lr = 0.001, optimizer='adam', loss='binary_crossentropy'

V. WEB-BASED WEATHER FORECASTING APPLICATION SYSTEM

In the contemporary landscape, weather forecasting has become an indispensable aspect of modern research. With the integration of a web application system dedicated to Weather Forecasting, the process of checking whether it will rain tomorrow or not has been streamlined for user convenience. This advancement holds paramount importance as it empowers individuals to access accurate weather predictions effortlessly. Our web application system facilitates this process by leveraging the parameters established in the Weather Australia dataset. Upon entering the system, users are prompted to provide pertinent information aligned with these dataset parameters. Through this collaborative interaction, our system harnesses predictive algorithms to determine the likelihood of rain tomorrow. The user interface of our innovative weather prediction web application is depicted in Figure 27. This intuitive interface ensures seamless user engagement and a straightforward experience while obtaining accurate forecasts. This venture embodies a significant contribution to the realm of weather forecasting. By amalgamating real-world datasets, advanced algorithms, and user-friendly interfaces, we empower individuals with actionable insights into imminent weather conditions. The web application not only streamlines information access but also exemplifies the synergy between cutting-edge technology and daily life, fostering informed decision-making and enhancing overall societal resilience. For those interested in delving deeper into the technical aspects, the source code is accessible on GitHub at the following link: https://github.com/tazriahelal/Weather-_Forecasting_App

VI. CONCLUSION

Detecting the occurrence of rainfall tomorrow poses a significant challenge in the realm of data science. This thesis presents a systematic approach to developing a robust classification system for this task. Various machine learning classification techniques are investigated and evaluated at different stages of the research. The suitability of the system is assessed using the Artificial Neural Network (ANN) classification technique, which achieves an accuracy of 91%. Furthermore, this study envisions the potential for using different machine learning methods to predict various outcomes in the future. The research can be extended to address real-world data challenges and enhance automation in analysis by incorporating alternative machine learning algorithms. Instead of analysing datasets based solely on attributes, an intricate algorithm is employed to extract meaningful patterns. Employing superior classification methods can lead to improved predictions and informed decision-making. Future extensions of this research could explore other high-performing classification models and conduct more in-depth descriptive analysis to gain further insights and determine the need for factor analysis. By assessing the predictability of each class and identifying attributes that contribute most to the predictions, the research aims to create an environment where machine learning classification methods can accurately forecast the future. Expanding and refining the work to include a range of machine learning methods and real-world applications of artificial intelligence would enhance analytical automation.Again, learning more about the characteristics that are associated with rainfall in the future can lead to more advanced technology. Rainfall may be predicted using advanced machine learning and deep learning models, and even based on this, one may draw solid, data-driven conclusions that are more effective in determining whether or not it will rain tomorrow. It may also be feasible to create hybrid classifiers by fusing a variety of methodologies. The framework we suggested may be improved upon and updated to create a more sophisticated framework that can be useful for both people and cutting-edge rain prediction systems. While a significant step in the direction of rainfall prediction is represented by our work, certain limitations are acknowledged. One of the primary constraints is the reliance on historical weather data, which may not fully capture the complexity of meteorological patterns. Additionally, the performance of the models may be influenced by the availability of data for specific regions, which can vary significantly. In the future, we are planning to work on big datasets, and we will engage in fedarated learning to improve our application and model performance.

AUTHOR CONTRIBUTIONS

The authors' contributions to this research are as follows: Md. Mehedi Hassan conceived the idea, written full manuscript, conducted and implemented machine learning models, supervised the project, and contributed to writing. Mohammad Abu Tareq Rony assisted in data collection, performed analysis, and contributed to writing. Md. Asif Rakib Khan provided guidance, reviewed results, and helped with writing. Md. Mahedi Hassan assisted with data preprocessing, experiments, and writing. Anindya Nag aided in data analysis, experiments, and writing. Tazria Helal Zarin contributed to data collection, experiments, and literature review. Anupam Kumar Bairagi reviewed the results and aided in writing.

ACKNOWLEDGMENT

This work is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R197), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] P. K. Srivastava, A. Mehta, M. Gupta, S. K. Singh, and T. Islam, "Assessing impact of climate change on Mundra mangrove forest ecosystem, Gulf of Kutch, Western Coast of India: A synergistic evaluation using remote sensing," *Theor. Appl. Climatol.*, vol. 120, nos. nos. 3–4, pp. 685–700, May 2015.

[2] *Annual 2021 Global Climate Report National Centers for Environmental Information (NCEI)*, Nat. Centers Environ. Inf., USA, 2022.

[3] R. Janarthanan, R. Balamurali, A. Annapoorani, and V. Vimala, "Prediction of rainfall using fuzzy logic," *Mater. Today, Proc.*, vol. 37, pp. 959–963, Jan. 2021.

[4] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," *Exp. Syst. Appl.*, vol. 85, pp. 169–181, Nov. 2017.

[5] S.-C. Yang, Z.-M. Huang, C.-Y. Huang, C.-C. Tsai, and T.-K. Yeh, "A case study on the impact of ensemble data assimilation with GNSS-zenith total delay and radar data on heavy rainfall prediction," *Monthly Weather Rev.*, vol. 148, no. 3, pp. 1075–1098, Mar. 2020.

[6] Z. He, "Rain prediction in Australia with active learning algorithm," in *Proc. Int. Conf. Comput. Autom. (CompAuto)*, Sep. 2021, pp. 14–18.

[7] A. G. Nolan and W. J. Graco, "Using the results of capstone analysis to predict a weather outcome," in *Advances in Data Mining. Applications and Theoretical Aspects*. Singapore: Springer, 2017, pp. 269–277.

[8] 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.

[9] O. Ejike, D. L. Ndzi, and A.-H. Al-Hassani, "Logistic regression based next-day rain prediction model," in *Proc. Int. Conf. Commun. Inf. Technol. (ICICT)*, Jun. 2021, pp. 262–267.

[10] A. D. Kumarasiri and U. J. Sonnadara, "Performance of an artificial neural network on forecasting the daily occurrence and annual depth of rainfall at a tropical site," *Hydrological Processes*, vol. 22, no. 17, pp. 3535–3542, Aug. 2008.

[11] N. J. Ria, J. F. Ani, M. Islam, and A. K. M. Masum, "Standardization of rainfall prediction in Bangladesh using machine learning approach," in *Proc. 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2021, pp. 1–5.

[12] S. Neelakandan and D. Paulraj, "RETRACTED ARTICLE: An automated exploring and learning model for data prediction using balanced CA-SVM," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 5, pp. 4979–4990, May 2021.

[13] S. Sankaranarayanan, M. Prabhakar, S. Satish, P. Jain, A. Ramprasad, and A. Krishnan, "Flood prediction based on weather parameters using deep learning," *J. Water Climate Change*, vol. 11, no. 4, pp. 1766–1783, Dec. 2020.

[14] 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.

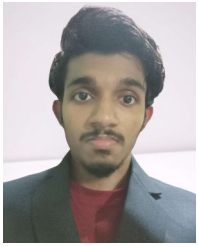
[15] M. M. Hassan, M. A. R. Khan, K. K. Islam, M. M. Hassan, and M. M. F. Rabbi, "Depression detection system with statistical analysis and data mining approaches," in *Proc. Int. Conf. Sci. Contemp. Technol. (ICSCT)*, Aug. 2021, pp. 1–6.

[16] S. A. Azwari, "Predicting myocardial rupture after acute myocardial infarction in hospitalized patients using machine learning," in *Proc. Nat. Comput. Colleges Conf. (NCCC)*, Mar. 2021, pp. 1–6.
[17] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC Trends Anal. Chem.*, vol. 132, Nov. 2020, Art. no. 116045.
[18] L. Klebanov and I. Volchenkova, "Outliers and the ostensibly heavy tails," *Math. Methods Statist.*, vol. 28, no. 1, pp. 74–81, Jan. 2019.
[19] G. Agapito, C. Zucco, and M. Cannataro, "COVID-WAREHOUSE: A data warehouse of Italian COVID-19, pollution, and climate data," *Int. J. Environ. Res. Public Health*, vol. 17, no. 15, p. 5596, Aug. 2020.
[20] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 105524.
[21] A. A. Freitas, "Comprehensible classification models: A position paper," *ACM SIGKDD Explor. Newslett.*, vol. 15, no. 1, pp. 1–10, Mar. 2014.
[22] Q. He, H. Shahabi, A. Shirzadi, S. Li, W. Chen, N. Wang, H. Chai, H. Bian, J. Ma, Y. Chen, X. Wang, K. Chapi, and B. B. Ahmad, "Landslide spatial modelling using novel bivariate statistical based naïve Bayes, RBF classifier, and RBF network machine learning algorithms," *Sci. Total Environ.*, vol. 663, pp. 1–15, May 2019.
[23] A. Ali, A. Khairan, F. Tempola, and A. Fuad, "Application of naïve Bayes to predict the potential of rain in ternate city," in *Proc. ES Web Conf.*, vol. 328, 2021, p. 04011.
[24] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, pp. 20–28, Mar. 2021.
[25] Z. Cheng, J. Lu, Z. Zu, and Y. Li, "Speeding violation type prediction based on decision tree method: A case study in Wujiang, China," *J. Adv. Transp.*, vol. 2019, pp. 1–10, Jun. 2019.
[26] MD. R. A. Iqbal, S. Rahman, S. I. Nabil, and I. U. A. Chowdhury, "Knowledge based decision tree construction with feature importance domain knowledge," in *Proc. 7th Int. Conf. Electr. Comput. Eng.*, Dec. 2012, pp. 659–662.
[27] C. Meng, Y. Hu, Y. Zhang, and F. Guo, "PSBP-SVM: A machine learning-based computational identifier for predicting polystyrene binding peptides," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 245, Mar. 2020.
[28] V. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer, 1999.
[29] I. Reis, D. Baron, and S. Shahaf, "Probabilistic random forest: A machine learning algorithm for noisy data sets," *Astronomical J.*, vol. 157, no. 1, p. 16, Dec. 2018.
[30] G. Rong, S. Alu, K. Li, Y. Su, J. Zhang, Y. Zhang, and T. Li, "Rainfall induced landslide susceptibility mapping based on Bayesian optimized random forest and gradient boosting decision tree models—A case study of Shuicheng County, China," *Water*, vol. 12, no. 11, p. 3066, Nov. 2020.
[31] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Informat.*, vol. 35, nos. 5–6, pp. 352–359, Oct. 2002.
[32] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharmaceutical Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, Jun. 2000.
[33] S. Kumar, T. Roshni, and D. Himayoun, "A comparison of emotional neural network (ENN) and artificial neural network (ANN) approach for rainfall-runoff modelling," *Civil Eng. J.*, vol. 5, no. 10, pp. 2120–2130, Oct. 2019.
[34] Y. Xu, C. Hu, Q. Wu, S. Jian, Z. Li, Y. Chen, G. Zhang, Z. Zhang, and S. Wang, "Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation," *J. Hydrol.*, vol. 608, May 2022, Art. no. 127553.
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[36] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Mining Knowl. Manag. Process*, vol. 5, no. 2, pp. 1–11, Mar. 2015.
[37] M. M. Hassan, M. M. Hassan, F. Yasmin, M. A. R. Khan, S. Zaman, K. K. Islam, and A. K. Bairagi, "A comparative assessment of machine learning algorithms with the least absolute shrinkage and selection operator for breast cancer detection and prediction," *Decis. Anal. J.*, vol. 7, Jun. 2023, Art. no. 100245.
[38] H. Verma and G. Verma, "Prediction model for bollywood movie success: A comparative analysis of performance of supervised machine learning algorithms," *Rev. Socionetwork Strategies*, vol. 14, no. 1, pp. 1–17, Apr. 2020.

MD. MEHEDI HASSAN (Member, IEEE) received the B.Sc. degree in computer science and engineering from North Western University, Khulna, Bangladesh, in 2022, where he excelled in the studies and demonstrated a strong aptitude for research. He is currently pursuing the M.Sc. degree in computer science and engineering with Khulna University, Khulna. As the Founder and the CEO of the Virtual BD IT Firm and the VRD Research Laboratory, Bangladesh, he has established himself as a highly respected leader in the fields of biomedical engineering, data science, and expert systems. He is a dedicated and accomplished Researcher. He is highly skilled in association rule mining, predictive analysis, machine learning, and data analysis, with a particular focus on the biomedical sciences. As a young researcher, he has published 35 articles in various international top journals and conferences, which is a remarkable achievement. His work has been well-received by the research community and has significantly contributed to the advancement of knowledge in the field. Overall, he is a highly motivated and skilled researcher with a strong commitment to improving human health and well-being through cutting-edge scientific research. His accomplishments to date are impressive, and potential for future contributions to the field is very promising. In addition, he serves as a reviewer for 26 prestigious journals. He has filed more than three patents out of which two are granted to his name. His research interests include important human diseases, such as oncology, cancer, and hepatitis, as well as human behavior analysis and mental health.

MOHAMMAD ABU TAREQ RONY received the B.Sc. degree in statistics from Noakhali Science and Technology University, Noakhali, Bangladesh. He is a diligent individual who is self-taught in data analysis, statistics, machine learning, and deep learning. In addition, he has experience in creating an advanced analytics strategy using data. His experience is a mix of 3-year research A.I., machine learning, data science, and statistical analysis.

MD. ASIF RAKIB KHAN received the B.Sc. degree in computer science and engineering from the Bangladesh University of Business and Technology (BUBT). He is currently a Researcher and conducts research in the Discipline of Biomedical Engineering and Environmental Science, BUBT. He has been working on human disease identification and human behavior analysis. As a young researcher, he has had some publications in various reputed journals.



MD. MAHEDI HASSAN received the B.Sc. degree from the Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, in 2021. He is currently a Researcher and a member of the VRD Research Laboratory, Khulna, Bangladesh. He conducts research in the Biomedical Engineering and Environmental Science Disciplines. He collaborates with internal researchers because he believes that without research, innovation is impossible. His research interests include predictive analysis and expert system development.

FARHANA YASMIN received the M.Sc. degree from the Department of Computer Application and Technology, Changzhou University, China. She is currently pursuing the PhD. Degree in computer science and technology with the Nanjing University of Information Science and Technology, China. She is an expert in computer vision, biomedical engineering, and data science. She works with scientists from all around the world to broaden her horizons and increase the depth of her expertise. Her research interests include predictive analysis and the development of expert systems.

ANINDYA NAG (Member, IEEE) received the bachelor's degree from Adamas University, Kolkata, India. He is currently pursuing the master's degree (M.Sc. Engg.) in computer science and engineering from Khulna University, Khulna, Bangladesh. He is an Adjunct Lecturer with the Department of Computer Science and Engineering, North Western University, Khulna. His research interests include NLP, artificial intelligence, the IoT, blockchain, cloud computing, and networking systems.

TAZRIA HELAL ZARIN received the B.Sc. degree in computer science engineering from the University of Development Alternative, Dhaka, Bangladesh. She exemplifies an unwavering commitment to the work. She is also able to think critically as well.

ANUPAM KUMAR BAIRAGI (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science and engineering from Khulna University (KU), Bangladesh, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea. He is currently a Professor with the Computer Science and Engineering Discipline, KU. His research interests include wireless resource management in 5G and beyond, healthcare, the IIoT, cooperative communication, and game theory. He has authored and coauthored around 60 publications, including refereed IEEE/ACM journals and conference papers. He has served as a technical program committee member at different international conferences.

SAMAH ALSHATHRI received the bachelor's degree in computer science and the master's degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, and the Ph.D. degree from the Department of Computer and Mathematics, Plymouth University, Plymouth, U.K. She is currently an Assistant Professor with the Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), Riyadh. Her research interests include wireless networks, cloud computing, fog computing, the IoT, data mining, machine learning, text analytics, image classification, and deep learning. She was the Chair of the Network and Communication Department and participated in organizing many international conferences. She has authored or coauthored many articles published in well-known journals in the research field.

WALID EL-SHAFAI (Senior Member, IEEE) was born in Alexandria, Egypt. He received the B.Sc. degree (Hons.) in electronics and electrical communication engineering from the Faculty of Electronic Engineering (FEE), Menoufia University, Menouf, Egypt, in 2008, the M.Sc. degree from the Egypt-Japan University of Science and Technology (E-JUST), in 2012, and the Ph.D. degree from the FEE, Menoufia University, in 2019. Since January 2021, he has been a Postdoctoral Research Fellow with the Security Engineering Laboratory (SEL), Prince Sultan University (PSU), Riyadh, Saudi Arabia. He is currently a Lecturer and an Assistant Professor with the Department of Electronics and Communication Engineering (ECE), FEE, Menoufia University. His research interests include wireless mobile and multimedia communications systems, image and video signal processing, efficient 2D video/3D multi-view video coding, multi-view video plus depth coding, 3D multi-view video coding and transmission, quality of service and experience, digital communication techniques, cognitive radio networks, adaptive filters design, 3D video watermarking, steganography, encryption, error resilience and concealment algorithms for H.264/AVC, H.264/MVC, and H.265/HEVC video codecs standards, cognitive cryptography, medical image processing, speech processing, security algorithms, software-defined networks, the Internet of Things, medical diagnoses applications, FPGA implementations for signal processing algorithms and communication systems, cancellable biometrics and pattern recognition, image and video magnification, artificial intelligence for signal processing algorithms and communication systems, modulation identification and classification, image and video super-resolution and denoising, cybersecurity applications, malware and ransomware detection and analysis, deep learning in signal processing, and communication systems applications. He also serves as a reviewer for several international journals.