Journal of Informatics and Web Engineering
Vol. 4 No. 2 (June 2025) eISSN: 2821-370X

A Fundamental Framework for Analysis of Rainfall Prediction Features Significance

Ye Zhian Teoh[1], Yim Ling Loo[2*]
[1,2] Faculty of Computing and Informatics, Universiti Multimedia, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia
**corresponding author: (looyimling@gmail.com; ORCiD: 0000-0001-8217-0166)*

Abstract - Rainfall prediction efforts had been prevalent ever since the impact of climate change on occurrences of natural disasters globally. Implementation of machine and deep learning techniques on features that contribute to rainfall occurrences were conducted with aims of seeking greater prediction accuracy for rainfall occurrences with a lack of study for significance of features in rainfall occurrence prediction. This study presents a framework of rainfall prediction features' significance analysis in the case study of Peninsular Malaysia rainfall occurrences. Features investigated in this study consist of temperature, humidity and wind speed. The designed framework for the investigation includes phases of data collection, data preprocessing, integration of Random Forest (RF) for ensemble classification and Feature Importance (FI) for feature significance calculation and finally model evaluation based on the metrics of precision, recall, F1 score and Receiver Operating Characteristic (ROC) curve. In the preliminary investigation, the prediction model demonstrated accuracy, precision, recall and F1-score of 80.65%, 80%, 81% and 0.80 respectively. Humidity was found to have highest significance to the model's predictive power as compared to temperature and wind speed. Rainfall occurrence correlation with lower temperature and higher humidity and vice versa was identified with further investigation of feature data distribution against rainfall occurrences.

Keywords— Rainfall Predictions, Rainfall Prediction Features Analysis, Features Significance, Features Significance Analysis Framework, Features Selection

Received: 20 November 2024; Accepted: 7 February 2025; Published: 16 June 2025

1. INTRODUCTION

Rainfall or precipitation prediction plays a significant role, particularly for transportation scheduling [1], [2], [3], [4], journey planning [5], [6], [7], [8] and safety measures [9], [10], [11]. Authors in [1] indicated that shifts in temperature and alterations in precipitation patterns significantly impact road environments, escalating risks to road safety in Malaysia. Moreover, alterations in precipitation patterns contribute to diminished adhesion between roads and vehicles, creating hazardous conditions. Excessive rainfall can lead to the accumulation of water on road surfaces, resulting in loss of control for vehicles. The increased frequency and intensity of rainfall also elevate the risk of flash floods and landslides, both of which pose severe threats to road safety, as evidenced by a notable landslide incident in Malaysia

throughout the years [12], [13]. The resulting losses from floods in the same year amounted to approximately RM 622.4 million [14]. To mitigate these risks and losses, accurate weather forecasts especially rainfall prediction become imperative for timely precautions.

Existing research works on rain forecast models reveals a diverse range of methods employed to enhance the accuracy of rainfall predictions, with a predominant focus on machine learning techniques. According to [15], machine learning is a segment of Artificial Intelligence (AI) and computer science which highlight employment of data and algorithms to emulate the learning process of real individuals, with the aim of enhancing its accuracy over time. Predictive models often integrate a machine learning algorithm, and through continuous training, these models can adapt to new data or values over time, providing the desired outcomes. There are many existing machine learning models, such as linear regression, neural network, Ramdom Forest (RF), Support Vector Machines (SVM), Bayesian network models and more for application in wide spectrum of prediction research works. Various researchers have employed distinct machine learning models or statistical analyses for rainfall predictions. Several studies from many different countries' datasets, such as Mexico, China, India, Jordan, Ghana, and Taiwan, have leveraged advanced machine learning algorithms to improve rainfall forecasts [16]-[22].

2. LITERATURE REVIEW

Research work [23] used a dataset of Mexico, suggesting a unique approach utilizing the Clausius–Clapeyron Relation and the CRHUDA model (CRossingHUmidity, Dewpoint, and Atmospheric pressure) successfully predicted rainfall onset based on atmospheric pressure, humidity, and dewpoint. This study contends that the most effective approach for synoptic precipitation onset forecasting involves a blend of humidity, dewpoint, and atmospheric pressure. The study examined two sets of series: the first set focused on humidity, while the second set explored the interplay between atmospheric pressure and dewpoint, and they are proposed from data monitored at every minute. Their method showcases the importance of the features stated in rainfall prediction.

Research work in [7] used the dataset of Indonesia and proposed Multiple Linear Regression and K-Nearest Neighbour (KNN) algorithms to predict rainfall values. Despite dataset limitations leading to higher prediction errors, the study provides insights into the comparative performance of different algorithms. An Attentive Interpretable Tabular Learning neural network (TabNet) was utilised by [24] using dataset from China, incorporating feature engineering and construction to increase the accuracy of satellite observation. The model demonstrated superior performance compared to traditional methods, emphasizing the significance of machine learning in rainfall prediction. It involves many meteorological features, which are Evaporation, Surface temperature. Air pressure, Humidity, Temperature, Windspeed, rainfall, sunshine time.

Studies conducted using dataset from Ghana [16] employed statistical tests to analyze rainfall and temperature data. The research work found that both rainfall and temperature data were investigated within the boundary of normality tests, namely Shapiro–Wilk, Anderson–Darling, Lilliefors, and Jarque–Bera tests. There are also various Homogeneity Test Analysis that have implemented, which are Pettitt's Test, Standard Normal Homogeneity Test (SNHT), Buishand's Test. The Mann–Kendall (MK) nonparametric test was employed to examine the trends present in the data for both rainfall and temperature. Besides, a post hoc test for multiple comparisons using one-way ANOVA was also implemented in the paper to assess variations in the means of temperature and rainfall data. One-way ANOVA was introduced in the study with the deliberate intention of complementing the results obtained from trend tests. Its primary purpose was to facilitate a comprehensive understanding of how ANOVA could effectively elucidate differences in mean scores for rainfall and temperature. The results suggested the suitability of the data for trend analysis, highlighting the importance of understanding the temporal patterns of climatic parameters in rainfall forecasting.

Researchers that use dataset from India [25] have explored various methodologies, including Rough Set Theory (RST), Naïve Bayes, J48, CART, Multi-Layer Perceptron, Bayesian Logistic Regression and RF. The research work introduced a novel Rough Set-based Maximum Frequency Weighted (MFW) feature reduction technique prior to modelling, which resulted in enhanced accuracy, with identified parameters such as relative humidity and solar radiation playing crucial roles in rainfall prediction. Their previous research work [26] employed Rough Set Attribute Reduction Technique (RSART) combined with data mining methods, demonstrating that the RSART-GA approach, coupled with Bayesian Logistics Regression Classifier, outperformed other classifiers after attribute reduction. Bayesian network models proposed by [18] exhibited efficiency in monthly rainfall forecasts, emphasizing the relevance of factors like humidity, cloud cover, temperature, wind speed, and the Southern Oscillation Index (SOI).

The efficiency of the model is found to be above 85 percent for most of the cases. A research work [17] utilised non-parametric tests, including the Pettitt and Mann-Kendall tests, alongside machine learning approaches such as Artificial Neural Network-Multilayer Perceptron (ANN-MLP) to identify influential factors in rainfall changes. The study suggested the impact of moisture divergence, precipitation convective rate and low cloud cover on rainfall variations in India.

Amongst existing research efforts, forest machine learning technique is seen to have been widely used and prevalent in many prediction models [27], [28]. Among research efforts found, research work in [8] implemented RF to predict rainfall and results with high accuracy was generated. Research work in [5] conducted experimental works on Australian rainfall dataset using four main machine learning methodologies of logistic regression, decision trees, SVM and RF. The research concluded that RF is the best prediction model compared to the other classification algorithms.

Research work in [2] explored efficiency of linear discriminant analysis, logistic regression, decision tree and gradient boosted trees, Bernoulli Naïve Bayes, deep learning, KNN for classification and RF on Australian dataset of temperature, rain, evaporation, sunshine and maximum wind gust. The research highlighted that even though RF took slightly longer duration for modelling, the results demonstrated much improved precision. Apart from rainfall prediction, RF classification algorithm was implemented in various prediction models such as business, medical and transportation and natural disaster incidents prediction [29], [30], [31], [32]. It is noteworthy that RF classification algorithm contributes to high accuracy in other prediction models.

Research in [30] applies RF regression to predict daily cases and deaths related to Covid-19. The evaluation of model performance involves metrics such as coefficient of variation, accuracy, relative error and root-mean-square score. The results indicate that RF emerges as the superior choice compare to other models. RF modelling was again found to be outperforming other algorithms such as SVM and traditional logistic regression, achieving good prediction results in landslide susceptibility analysis [32]. The landscape of rain forecasting research is marked by a diverse array of methodologies with many advanced machine learning techniques [22].

From the aforementioned literature reviews, various studies across different countries' dataset, have explored the efficacy of machine learning models like RF, Neural Networks, Logistic Regression and Bayesian approaches in improving rainfall predictions. Notably, each region's unique climatic features and datasets have prompted researchers to tailor their methodologies, emphasizing the importance of specific meteorological parameters like atmospheric pressure, humidity, dewpoint, and solar radiation.

Most of the studies were conducted to find the most effective machine learning model for the specific rainfall datasets in the experimentation [5], [33]. Current research shift to the paradigm of feature engineering for prediction optimization demonstrated the importance of investigating significant features that contributes to rainfall incidents for better prediction [4], [21], [34]. Study in [4] concluded that selecting significant features by feature classification and matching did enhance the predictive performance of flood prediction. Research works in [34] demonstrated the efficiency of wind power prediction with advanced feature engineering and highlighted that the achievement is through revolution of feature synthesis, integration of feature selection with concluded significant features and studies on the complex data patterns.

Integration of Bayesian Networks (BN) with Recursive Feature Elimination (RFE) in [21] identified significant features for prediction of monthly rainfall using ERA 5 Reanalysis Dataset by European Center for Medium Weather Forecasts (ECMWF). The research work concluded that feature selection has improved both short-term and long-term prediction of rainfall for the studied area with relative humidity, total precipitable water and wind found to be significant features contributing to optimised prediction of rainfall. The study further recommended usage of different machine learning algorithms to be accompanied with a feature selection technique to select important or significant features for prediction improvement.

Every machine learning algorithm has a unique method for determining the significance of attributes in performing classification tasks. Feature importance denotes the extent to which each feature contributes to the prediction made by the model. Research work in [35] built model using Random Forest (RF), AdaBoost, and K-Nearest Neighbours (K-NN) algorithms and conducted hyperparameter tuning to enhance their performance. Feature importance analysis was employed to discern the significance of individual features for each model. Through this analysis, they obtained ranked lists of influential features for readmission prediction in each machine learning model. The significance of feature importance analysis, as exemplified in the research work, becomes evident in enhancing the accuracy of predictions by discerning the individual contributions of meteorological features.

While diverse machine learning algorithms have been applied in different regions, there was lack of studies for this problem domain in Malaysia where cases of flood had been rising and the most recent research found was [3]. This brought to the attention of this study to address the gap by answering the following research questions:

RQ1- What is the formulation of dataset for Malaysia rainfall prediction modelling?

RQ2- How various rainfall prediction features can be extracted from formulated Malaysia dataset?

RQ3- Which rainfall prediction feature impact the most prediction model's prediction?

This research aims to address the research questions by leveraging RF ensemble classification in conjunction with feature importance analysis and introduce an innovative methodology for rainfall prediction, offering a unique perspective that holds promise for advancements in the field of meteorological forecasting. This research also focuses on delving into historical temperature, humidity, and wind speed data before rain occurrences for a detailed study of significance between the three features. To that end, the research work curated specific dataset for this work from data collection and analysis as well as implemented validated feature extraction, modelling, feature significance analysis and results validation techniques from previous studies, which will be discussed in the following sections.

3. RESEARCH METHODOLOGY

This section outlines the formulation of rainfall features dataset for this research and framework for analysis of rainfall prediction features significance. Detailed discussions of framework segments follow the outline section.

3.1 Framework Formulation

The meteorological parameters of historical temperature, humidity and wind speed play a crucial role in predicting rainfall and are integral components of weather forecasting systems. Through a detailed examination of complex patterns for temperature, humidity, and wind speed, the research aims to discover important information that can improve the accuracy of rainfall predictions. This newfound knowledge has the potential to make rain forecasting models work better, which, in turn, can help different industries be better prepared and respond more effectively to changing weather conditions. Malaysia climate features are found to have the characteristics of uniform temperature, light wind, high humidity and copious rainfall. Temperature patterns in Malaysia, unlike countries with distinct seasons, remain relatively uniform throughout the year.

Following normal data pattern, there is a slight annual variation, with higher temperatures in April and May and lower temperatures in December and January, which are months associated with maximum rainfall. According to mean annual temperature trend distribution map provided by Malaysia Meteorological Department, n.d., mean temperature for Klang Valley area would be higher as it is an urban conglomeration in Malaysia as illustrated in Figure 1.

Wind patterns in Malaysia have been categorised into four distinct seasons, contributing to the complexity of Malaysia's weather. These include the northeast and southwest monsoons, as well as two shorter inter-monsoon periods. The Northeast Monsoon, prevalent from early November to March, brings steady East or North-East winds ranging from 10 to 30 knots. This season is associated with a wet period, marked by approximately 4-5 monsoon surges that can lead to flooding. Southwest Monsoon dominates from late May or early June to September, featuring generally light Southwest winds below 15 knots. This season is characterised by relatively dry weather, except for Sabah. The Inter-Monsoon seasons occur from late March to early May and October to mid-November, demonstrating light and variable winds, along with the equatorial trough lying over Malaysia. Frequent afternoon thunderstorms are common during these periods.

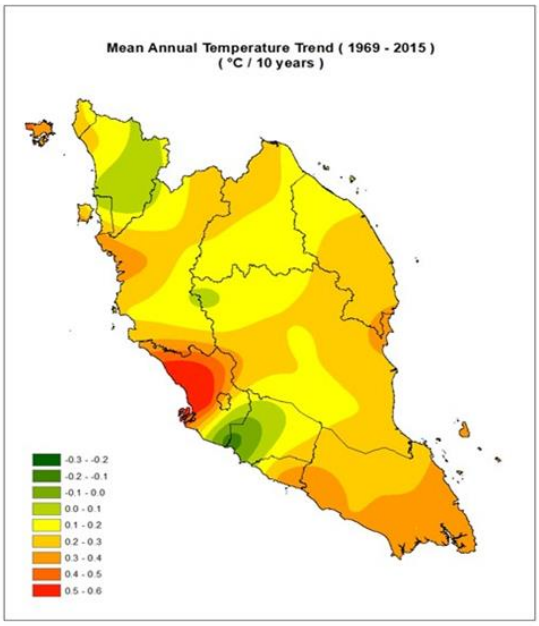


Figure 1. Mean Annual Temperature Trend Distribution Map by Malaysia Meteorological Department

Malaysia experiences variations across different locations and months with humidity ranging from 10% to 90%. In Peninsular Malaysia, the lowest relative humidity typically occurs in January and February, except for the east coast states of Kelantan and Terengganu, where this minimum is usually observed in March and the highest relative humidity is generally experienced in November. Similar to temperature patterns, the daily fluctuations in relative humidity surpass the annual variations. The average daily minimum relative humidity can drop as low as 42% during dry months and rise to around 70% during wet months. Despite regional differences, the maximum daily average relative humidity remains relatively consistent across locations, hovering over 94%, and occasionally reaching as high as 100%. The significant differences in humidity records implicated that there is significant difference between wet and dry seasons. This research work will analyze humidity feature in Klang Valley area for the correlation of the feature to rainfall instances.

RF algorithm which was previously reviewed as widely used machine learning technique is adopted as the main analysis technique for this research work. RF is an ensemble classification technique that compares and outputs the fittest result from multiple decision trees outputs. In detail, RF utilizes technique of bagging and feature randomness to create diverse forest of decision trees in order to address classification and regression problems [8], [29], which is the main investigations of this research. Within RF modelling, feature importance technique scrutinizes how each input feature contributes to the accuracy of model prediction. Such finding is useful to determine the significance of a feature in contribution to the model's performance. As such, we leverage both RF modeling and feature importance technique to analyze historical temperature, humidity and wind speed data within the scope of this research work. This aims to identify the effectiveness of aforementioned threefold input features in predicting rainfall occurrence.

Consequently, this research integrates RF ensemble classification and feature importance analysis technique in the proposed framework to investigate the individual contributions of temperature, humidity, and wind speed in rainfall occurrence prediction. Findings of significance of each input features' contributions to the accuracy of modeling implicate effectiveness of the model in predicting rainfall occurrences and correlation between the input features with rainfall occurrences. Such outcomes lead to improved forecasting and early warning systems. Ultimately, this research aims to mitigate losses caused by precipitation-induced disasters, offering valuable insights for organisations and contributing to overall disaster preparedness in Malaysia. Figure 2 illustrates the rainfall prediction features significance analysis framework.

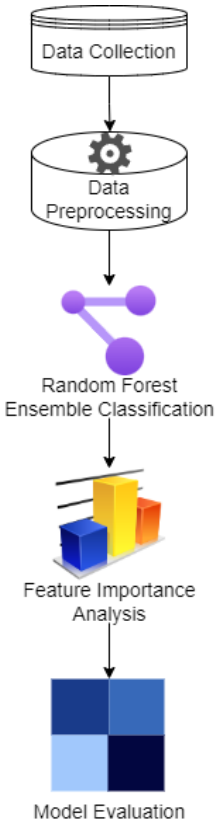


Figure 2. Rainfall Prediction Features Significance Analysis Framework

Proposed framework includes main phases of data collection, data preprocessing, data analysis through RF ensemble classification and feature importance analysis, and model evaluation. This framework includes the essential phases that are in-line with directions from findings of literature. The following sub-sections of this manuscript expound each of the sections in great details.

3.2 Data Collection

For this phase, data is downloaded from Visual Crossing, a weather data and Application Programming Interface (API) website, as the weather data is not readily available through Malaysia official weather websites. The targeted regions include Kuala Lumpur, Putrajaya, and the nine districts within Selangor state, which are Gombak, Hulu Langat, Hulu Selangor, Klang, Kuala Langat, Kuala Selangor, Petaling, Sabak Bernam and Sepang. The states are chosen for feasibility of research to focus on the main state of Malaysia, namely Selangor, Wilayah Persekutuan Kuala Lumpur and Putrajaya to represent the region of Peninsular Malaysia. A four-year dataset spanning from 2020 to 2023 was compiled. A total sample size of 16,071 data was collected using the filter of targeted regions and dateline.

Visual crossing is a website that offer weather data and API which was established in 2003. Visual Crossing stands out as a premier supplier of weather data and sophisticated analytical tools tailored for data scientists, business analysts, professionals, and academics. The website's extensive and comprehensive data assisted the dataset formulation for this research. Figure 3 illustrates the data collection workflows ranging from accessing Visual Crossing API through query builder, setting the filter of location and timeline of data and the extraction of raw data from the platform in this framework.

The dataset of this research is formulated to encompass a total of 33 attributes, including state name, date, tempmax, tempmin, temp, dew, feelslikemax, feelslikemin, feelslike, precip, precipprob, precipcover, preciptype, snow, snowdepth, windspeed, windgust, winddir, visibility, cloudcover, humidity, pressure, solarradiation, solarenergy, uvindex, severerisk, sunrise, sunset, moonphase, icon, conditions, description, and stations. Description of the attributes and sample data are displayed in Table 1.

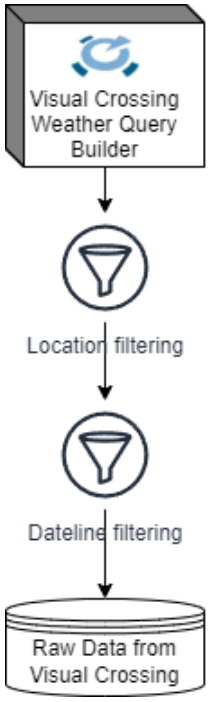


Figure 3. Rainfall Prediction Significance Analysis Framework Features

The wind speed and wind direction display the maximum of the hourly values. Wind speed is typically measured 10m above ground in a location with no nearby obstructions. The wind direction signifies the origin or source from which the wind is blowing. It expressed in degrees from the north, the units of wind direction range from 0 degrees wind (from the North) to 90 degrees (from the East), 180 degrees (from the South), 270 degrees (from the West), and finally returning to 360 degrees. Wind gust is the maximum wind speed measures over a short amount of time.

Precipitation coverage represents the percentage of time during a specified period when measurable precipitation was documented. For example, if there are six hours of measurable rainfall within a 24-hour day, the precipitation coverage would be 25% (calculated as 6/24 multiplied by 100). Relative humidity denotes the quantity of water vapor in the air relative to the maximum amount it could hold at a particular temperature, expressed as a percentage mean. Human comfort levels typically fall within the range of 30-70%. Humidity levels surpassing 70% are categorised as humid, while values below 30% are considered dry.

Ultraviolet (UV) index is a numerical scale that represents the intensity of ultraviolet radiation from the sun at a specific location and time. In this dataset, a value between 0 and 10 is used to represent the maximum level of UV exposure for the day, which 10 represents high level and 0 represents no exposure. There is also a moonphase attribute in this dataset, which record the moon phase of the day. The moon phase is quantified by a decimal value ranging from 0 to 1, where 0 represents the new moon, and 0.5 represents the full moon. This decimal scale shows the moon's illumination throughout its entire cycle. Starting with the new moon at 0, the waxing crescent phase unfolds from 0 to 0.25, leading to the first quarter at precisely 0.25. The waxing gibbous phase follows, extending from 0.25 to 0.5, reaching the full moon at the midpoint, 0.5. Subsequently, the waning gibbous phase spans from 0.5 to 0.75, transitioning to the last quarter at 0.75. The final segment, from 0.75 to 1, encompasses the waning crescent phase, ultimately completing the representation of entire lunar cycle. Subsequent to data collection, the data is then combined into one single excel file to simplify accessibility and enhance efficiency of data management. Further data pre-processing is done to formulate the dataset for the use of this research work.

Table 1. Data Column and Description

Element	Description	Unit (UK)
name	Name of the place	-
date	Date and time	-

tempmax	Maximum Temperature	C
tempmin	Minimum Temperature	C
temp	Temperature (or mean temperature)	C
dew	Dew Point	C
feelslikemax	Feels like (maximum temperature)	C
feelslikemin	Feels like (minimum temperature)	C
feelslike	Feels like	C
precip	Precipitation	mm
precipprob	Precipitation chance	%
precipcover	Precipitation Cover	%
preciptype	Precipitation type	–
snow	Snow	cm
snowdepth	Snow Depth	cm
windspeed	Wind Speed	kph
windgust	Wind Gust	kph
winddir	Wind Direction	degrees
visibility	Visibility	km
cloudcover	Cloud Cover	%
humidity	Relative Humidity	%
pressure	Sea Level Pressure	mb
solarradiation	Solar Radiation	W/m2
solarenergy	Solar Energy	MJ/m2
uvindex	UV Index	–
severerisk	Severe Risk	–
sunrise	Sunrise time	–
sunset	Sunset time	–
moonphase	Moonphase	–
icon	A weather icon	–
conditions	Short text about the weather	–
description	Description of the weather for the day	–

3.3 Data Preprocessing

Subsequent to formulation of the initial dataset from data collection, the initial step involves removing three columns—namely, snow, snowdepth, and severerisk—due to their lack of relevance to Malaysia, a country where snowfall is not applicable. Since all values within these three columns are empty, their exclusion contributes to minimizing data redundancy and streamlining the dataset. Figure 4 illustrates the workflow of data preprocessing from initial stage until formulation of training and testing dataset.

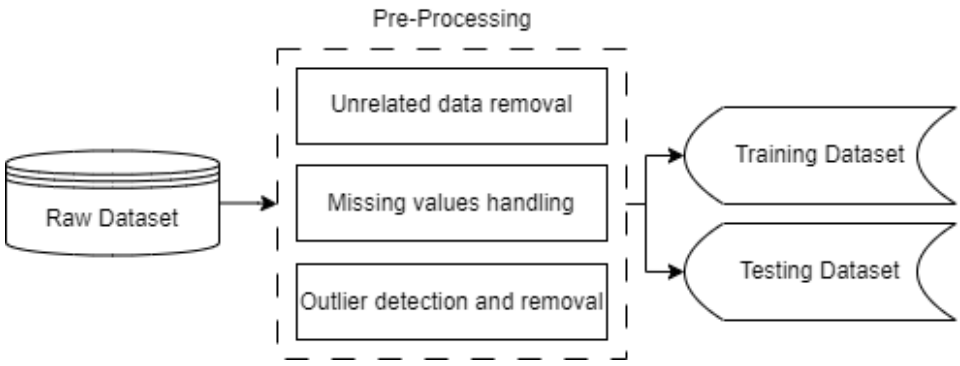


Figure 4. Data Preprocessing Workflow

In next data preprocessing stage, the focus is on cleaning and preparing the data to ensure its suitability for subsequent analysis and model training. Initially, a meticulous process of handling missing values will be executed, addressing any gaps or inaccuracies in the dataset. Pandas is used to identify the missing values within the dataset collected in this research and fill up the missing values with another set values. After addressing missing values in the dataset, outliers within the dataset are addressed by outlier detection and removal. This is done by visualizing the distribution of numerical variables and adopting Z-scores and interquartile range (IQR) statistical measures. These techniques further enhance overall quality and reliability of the dataset.

Additionally, categorised variables will be transformed into numerical representations using one-hot encoding to incorporate the input features data into RF modeling. This step is essential to ensure the categorised variables aspects of the data are appropriately considered during analysis. Binary columns are generated for each categorised variable, a format that is compatible with various machine learning algorithms, for this case, RF modeling algorithm. The binary representation allows RF algorithm to interpret and process categorical variable data effectively, contributing to more accurate model predictions.

After completion of data cleaning and transformation, the dataset is partitioned into training and testing sets. This segmentation is essential to evaluate the model's performance. At this stage, dataset is divided into features and target variables with split ratio of 70-30, following the reviewed anchor research works. The splitting of data is done using functions from scikit-learn libraries. The model is trained on one subset and tested on another to increase the model's ability to produce more precise predictions on data that is not exposed to the model before. This ensures the model to learn from one part of dataset and be tested on another for continuous future machine learning tasks.

Following the partitioning of dataset, data normalization is done to ensure that dataset of input features is normally distributed. Normalization is used for machine learning algorithms that are sensitive to the distributions of the input features. Min-max scaling technique is utilised in this research work to rescale out-of-range dataset to a predetermined normal distribution range. Normalizing the dataset helps in speeding up the convergence of optimisation in machine learning algorithms, enhances the interpretability of the model coefficients and prevents certain features from dominating the learning process due skewed data. This will further imply the overall robustness and efficiency of models during training and evaluation phases.

3.4 RF Ensemble Classification

Preprocessed dataset is to be modeled using RF algorithm. This encompasses a process of constructing an ensemble of classifications of random subsets of features and data samples to train individual classification. Each classification in the RF is trained on distinctive subset of the overall dataset to avoid biased pattern analysis formulation. Randomization concept in RF modeling ensures robustness of analysis output. Training of individual classifications ensure independent learning of patterns and establishment of predictions. This will assist the model to be able to establish predictions on new data, with the training results from all individual classifications that are aggregated to form the final prediction. Through the ensemble approach, a two-fold objectives of prediction model are achieved where accuracy of predictions is enhanced, at the same time, algorithm's ability to recognise new data is increased.

In the context of rainfall prediction, the algorithm will construct an ensemble of classifications by incorporating random subsets of input features and data samples. This technique is advantageous for predicting complex and nonlinear meteorological patterns. For instance, when predicting rainfall, numerous input features such as humidity, temperature and wind speed may influence the outcome. By training individual classification on specific subsets of these input features, the RF Algorithm will capture diverse relationships between the input features. This will enable the analysis to be extended to find correlation between features and subsequent rainfall occurrence. Implementation of the RF Algorithm for rainfall forecasting will be carried out using libraries in Python. Further analysis on significance of each features uncovers the possibility of identifying main features that influence effectiveness of the prediction model.

3.5 Feature Importance Analysis

A prevalent method for feature importance analysis in RFs modeling is based on Gini impurity and information gain metrics. These metrics are intrinsic to the classification within the RF ensemble. Gini impurity measures the degree of disorder in a dataset, while information gain assesses the reduction in uncertainty about the target input feature when a particular feature is used for splitting. RFs algorithm utilizes these metrics across all input features, for the overall calculation of feature importance.

Input features with resultant higher values calculated by the metrics indicate that the input feature made more significant contributions to the model's prediction performance. Likewise, input features with resultant lower values indicate that the input feature made lesser substantial contributions to the model's prediction performance. Feature importance analysis is conducted through integrated feature importance analysis tool in scikit-learn Python framework. Feature importance analysis scores can be accessed and analysed after the completion of model training in RF algorithm.

Comprehensive evaluation of feature importance scores will bring insights for specific features that significantly contribute to a rainfall prediction model's performance for the dataset of Klang Valley, Malaysia. In order to ensure reliability of analysed feature importance results, resultant prediction performance need to be on par with the results of similar research models as reviewed. Thus, model evaluation is essential for a comprehensive measurement of the prediction model's effectiveness.

3.6 Model Evaluation

The final phase in our proposed framework consists of evaluation of the rainfall prediction model's performance. In this research work, metrics such as accuracy (Equation (1)), precision (Equation (2)), recall (Equation (3)), F1 score (Equation (4)), and area under the Receiver Operating Characteristic (ROC) curve will be used for the model evaluation. It is worthy to note that the chosen metrics are used universally as machine learning model performance measurement metrics in the reviewed research works. The evaluation will be carried out on the partitioned test dataset as detailed in Section 3.3.

Precision, recall, F1 score and ROC curve (see Equation (1) to (4)) are the metrics utilised to assess the effectiveness of the prediction model by comparing them to the actual outcomes from the test dataset. Accuracy metric is an overall measure of correct predictions out of the number of predictions performed. Precision metric focuses on the ratio of correctly predicted positive instances or true positives against false positives. Recall metric assesses the proportion of actual true positives against false negatives. F1 score combines precision and recall to provide a balanced measurement of the predictive performance of the model.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

$$Precision = \frac{True\ positive}{(True\ positive + False\ positives)} \quad (2)$$

$$Recall = \frac{True\ positive}{(True\ positive + False\ negatives)} \quad (3)$$

$$F1\ score = \frac{2 \times (Precision + Recall)}{(Precision + Recall)} \quad (4)$$

The ROC curve is utilised to evaluate the model's performance across different thresholds of dataset. The area under the ROC curve (AUC-ROC) summarizes the model's ability to distinguish between positive and negative instances. Resultant higher AUC-ROC indicates the model possesses higher discriminative power. Visualization of the model predictions versus actual outcomes through confusion matrix or ROC curve plot, provides a clear illustration of the model's strengths and weaknesses. Scikit-learn Python library will be used for the computation of metrics and graph visualization.

In addition to the metrics mentioned, Mean Squared Error (MSE) (Equation (5) and R-squared (R2) (Equation (6)) will be used to evaluate the performance of the model in regression tasks. These metrics are particularly relevant to evaluation of models that predict numerical values. MSE calculates the average squared difference between the predicted and actual values, thus providing insights into the accuracy of the model's regressive numerical predictions. A lower MSE indicates better alignment between predicted and true values. R2 evaluates the proportion of variance of predictable dependent variable from the independent variables. It offers a measure of how well the model explains the variability in the data, with higher R2 values indicating a better fit.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (5)$$

Where n = number of observations in the dataset

Y_i = observed value for the i^{th} observation

Yˆi= predicted value for the i^{th} observation

$$R^2 = 1 - \frac{Sum\ of\ Squared\ Residuals\ (SSR)}{Total\ Sum\ of\ Squares\ (SST)} \quad (6)$$

Where SSR = the sum of squared differences between predicted value and actual value

SST = the sum of the squared differences between the actual values and the mean of actual value

The model will be evaluated on the test set to ensure it generalizes well to new, unseen data. Visualization of model predictions vs. actual outcomes using graphs will also be done through libraries such as Matplotlib, Plotly and Seaborn in Python, which can provide a versatile and visually appealing way to represent data through various plots, charts, and graphs. Comprehensive evaluation provides insights into the model's effectiveness in predicting rainfall in the specified regions which proves the reliability of feature significance found. This will also provide some insights on weather patterns of Peninsular Malaysia.

4. RESULTS AND DISCUSSIONS

A preliminary study is done for the framework implementation to analyze significance of three features in causing subsequence rainfall occurrence. The three features are namely temperature, humidity and wind speed which were highlighted by a recent study to be significant features in determining rainfall occurrences [21]. Features data chosen namely temp, humidity and windspeed and the resulting rain occurrence data namely precipprob were extracted from the formulated dataset for the preliminary study. The aim is to determine significance of these features, contributing to subsequent rainfall occurrences and identify the most significant feature out of the experimented three features. Ratio of 7:3 was used to divide the dataset of 16,071 accumulative data into training and testing set. Preliminary result of accuracy, precision, recall and F1-score of 80.65%, 80%, 81% and 0.80 respectively were established. The accuracy, precision and recall are on par with existing related research works [5], [8]. Figure 5 illustrates the confusion matrix of the model.

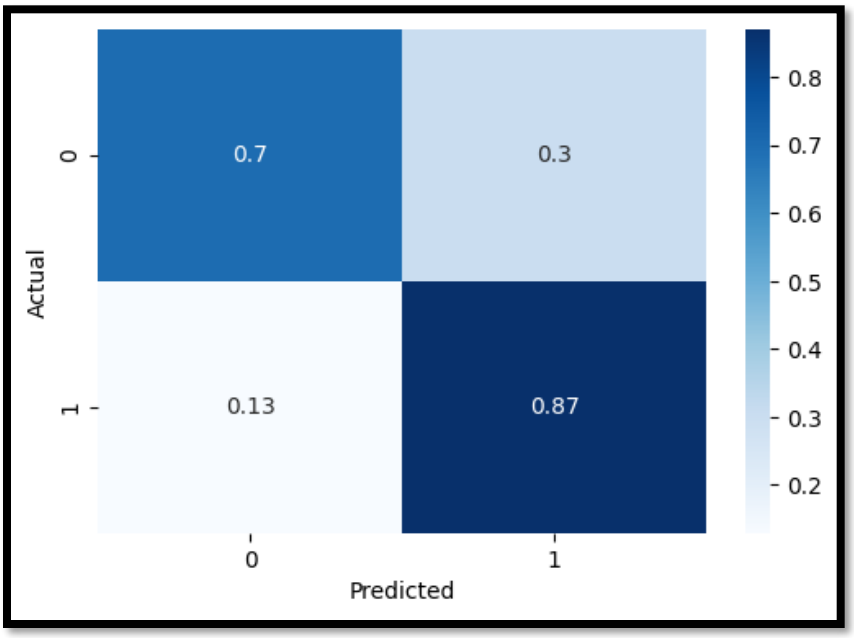


Figure 5. Model's Confusion Matrix

The features used to train the model was found to be effective to predict rainfall based on the results. With the promising result established, feature importance analysis is conducted to identify significance of each experimented features through calculation of importance score. Higher importance score indicates more substantial contributions of the feature to the model's predictive power. Bar graph in Figure 6 illustrated the features with their calculated importance scores.

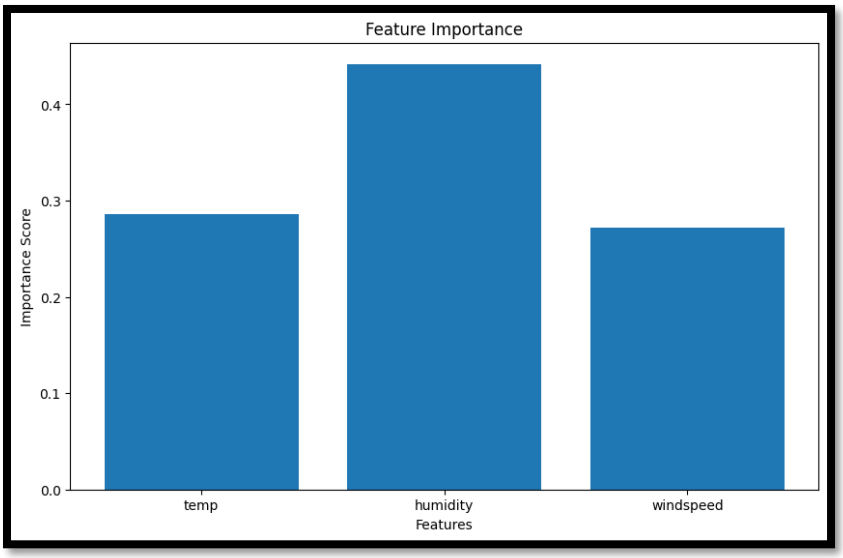


Figure 6. Feature Importance Analysis

Humidity was found to be the most significant feature in determining rainfall occurrences as compared to temperature and wind speed, with the highest importance score. However, it is noteworthy that the result does not imply that the other features are not significant, as both temperature and wind speed do have importance scores that are not too low from humidity feature. Further investigation is conducted to determine distributions of the data in each feature for prediction of rainfall occurrences. Box plot for each of the features against no rainfall occurrence and rainfall occurrence is illustrated in Figure 7.

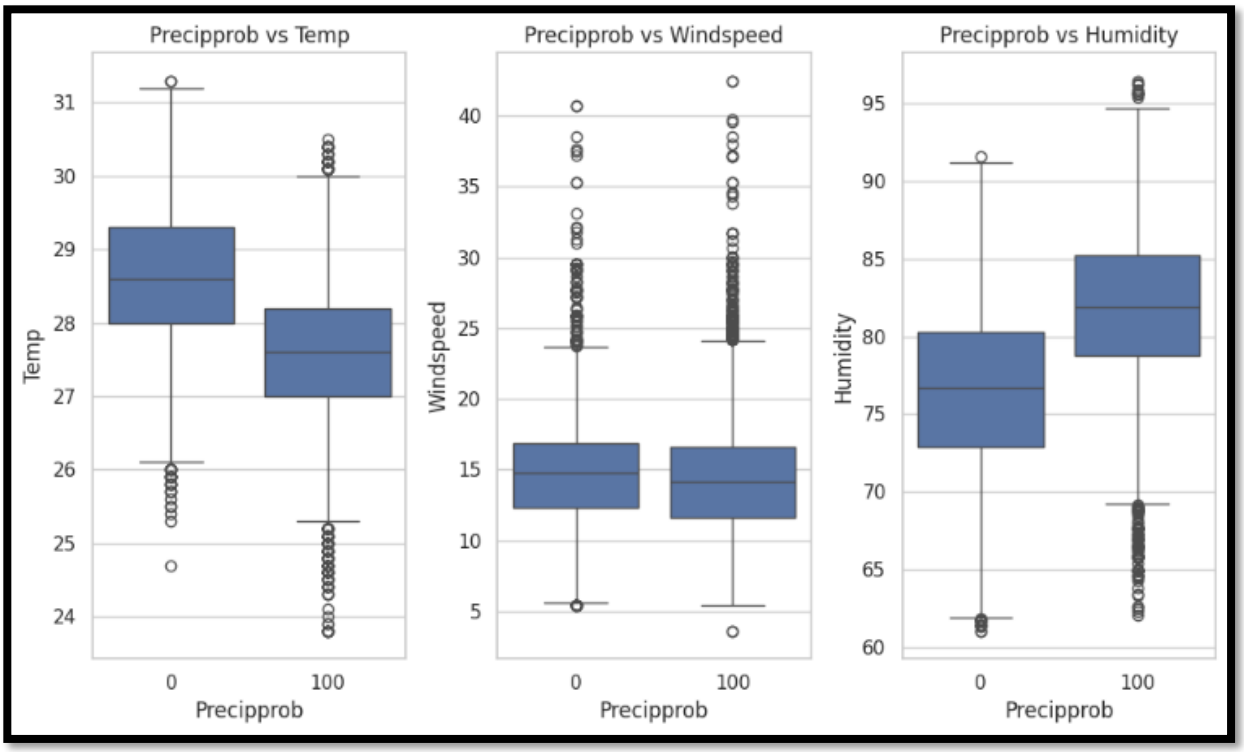


Figure 7. Distribution of Feature Data

The distribution of data illustrated in Figure 7 shown that significant difference of distribution is found for both temperature and humidity for no rainfall occurrence and rainfall occurrence as compared to windspeed. It is noteworthy that windspeed distribution of data did not vary much for no rainfall and rainfall occurrences. On this note, windspeed is found to be not a statistically significant feature in identification of rainfall. According to the investigation, no rainfall occurrence is relative to higher temperature range, while rainfall occurrence is relative to lower temperature range. Lower humidity is found to be relative to no rainfall occurrence while higher humidity range is relative to rainfall occurrence. These findings established a better comprehension of features' relationship with rainfall occurrence for the studied regions.

5. CONCLUSION

Preliminary research work presented in this paper, addressed the three-fold research questions posed in introduction section.

RQ1- What is the formulation of dataset for Malaysia rainfall prediction modelling?

We have presented a detailed data collection segment through formulated framework of rainfall prediction features significance. Manual data collection is conducted from visual crossing API that provided 33 rainfall-related features. Data formulated encompassed Peninsular Malaysia states of Selangor, Wilayah Persekutuan Kuala Lumpur and Putrajaya. The data collected consist of four-year dataset spanning from 2020 to 2023. Detailed data preprocessing segment discussed about removal of irrelevant features such as snow and snow depth. The preprocessed dataset is consisting of 31 rainfall-related features with total amount of 16071 data.

RQ2- How various rainfall prediction features can be extracted from formulated Malaysia dataset?

We have presented a detailed data preprocessing segment within formulated framework of rainfall prediction features significance for extraction of specific features. Element temp or Average Temperature in C UK unit, humidity or Relative Humidity in % UK unit and windspeed or Wind Speed in kph UK unit are specifically extracted for the

experiments in this study. As this study focuses on experimenting the significance of temperature, humidity and wind speed features, the extraction of data from specific three features are conducted from the preprocessed dataset.

RQ3- Which rainfall prediction feature impact the most prediction model's prediction?

We have presented a formulated framework of rainfall prediction features significance analysis which consists of five-fold segments ranging from data collection, data preprocessing, RF ensemble classification, feature significance calculation and finally model evaluation for determination of significant feature(s) that affect rainfall occurrence. Preliminary study conducted using the formulated framework shown that the framework is applicable for analysis of feature significance for rainfall prediction in the investigate regions.

By leveraging historical temperature, humidity, and wind speed data before rain occurrences, this study presented a comprehensive approach to rain forecasting using the RF algorithm and feature importance analysis. This research outlines a structured methodology that involves the application of the RF algorithm and feature importance calculation. The utilization of RF algorithm, helps in determining relationships between meteorological parameters and rainfall effectively, offering a more accurate and adaptable forecasting model, while feature importance helps to determine the significance of each feature towards rain forecast. In the preliminary study, RF model demonstrated high accuracy, precision, recall and F1-score of 80.65%, 80%, 81% and 0.80 respectively. Humidity was found to be having the highest substantial contributions of the feature to the model's predictive power, as compared to temperature and wind speed. Correlation of higher temperature and lower humidity for no rainfall occurrence and vice versa for rainfall occurrence was identified with further investigation of feature data distribution against rainfall occurrences.

Further experiments are recommended to be conducted to further validate the results and determine whether there are other features that demonstrates similar significance in determining rainfall occurrence in Malaysia. This research work contributed valuable insights into the significance of temperature, humidity, and wind speed in predicting rainfall, ultimately enhancing disaster preparedness and reducing the economic impact of precipitation-induced disasters in Malaysia.

ACKNOWLEDGEMENT

This work is sponsored by Multimedia University, Cyberjaya, Malaysia. The authors would like to thank the anonymous reviewers for their valuable comments.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Ye Zhian Teoh: Conceptualization, Data Processing, Methodology, Validation, Writing – Original Draft Preparation; Yim Ling Loo: Project Administration, Supervision, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

REFERENCES

[1] S. Shahid, and A. Minhans, "Climate change and road safety: A review to assess impacts in Malaysia", *J. Teknol.*, vol. 78, no. 4, Mar. 2016, doi: 10.11113/jt.v78.7991.

[2] M. Raval, P. Sivashanmugam, V. Pham, H. Gohel, A. Kaushik, and Y. Wan, "Automated predictive analytics tool for rainfall forecasting," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-95735-8.

[3] W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 1651–1663, Jun. 2021, doi: 10.1016/j.asej.2020.09.011.

[4] N. Chenmin, M. F. Marsani, and F. Pei Shan, "Flood prediction based on feature selection and a hybrid deep learning network," *Journal of Water and Climate Change*, Mar. 2024, doi: 10.2166/wcc.2024.559.

[5] Suvashisa Dash and Answeta Jaiswal, "Machine learning based forecasting model for rainfall prediction," *World Journal of Advanced Research and Reviews*, vol. 21, no. 1, pp. 1678–1686, Jan. 2024, doi: 10.30574/wjarr.2024.21.1.0180.

[6] V.S, G. S, D. M, and K. S, "Rainfall Based Flood Prediction in Kerala Using Machine Learning", *Int J Intell Syst Appl Eng*, vol. 12, no. 16s, pp. 141–144, Feb. 2024.

[7] B. Setya, R. A. Nurhidayatullah, M. B. Hewen, and K. Kusrini, "Comparative Analysis Of Rainfall Value Prediction In Semarang Using Linear And K-Nearest Neighbor Algorithms," in *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICORIS60118.2023.10352274.

[8] P. A. Nandini, B. Meenavalli, A. Puttamreddy, J. Meghana, N. Kataria, and L. Gupta, "Prediction of Rainfall using Random Forest," in *2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/SCEECS54111.2022.9741063.

[9] C. Schutte, M. van der Laan, and B. van der Merwe, "Leveraging historic streamflow and weather data with deep learning for enhanced streamflow predictions," *Journal of Hydroinformatics*, Feb. 2024, doi: 10.2166/hydro.2024.268.

[10] H. Ishibashi, "Framework for risk assessment of economic loss from structures damaged by rainfall-induced landslides using machine learning," *Georisk*, vol. 18, no. 1, pp. 228–243, 2024, doi: 10.1080/17499518.2023.2288606.

[11] R. Sato and Y. Fujimoto, "Rainfall Forecasting with LSTM by Combining Cloud Image Feature Extraction with CNN and Weather Information," *IEEJ Journal of Industry Applications*, vol. 13, no. 1, pp. 24–33, 2024, doi: 10.1541/ieejjia.23002926.

[12] N. A. Majid, "Historical landslide events in Malaysia 1993-2019," *Indian J Sci Technol*, vol. 13, no. 33, pp. 3387–3399, Sep. 2020, doi: 10.17485/IJST/v13i33.884.

[13] G. Singh, M. A. B. H. Abdul Rahman, and M. S. bin Zulkipli, "An Emergency and Mass Casualty Incident Response in the Jalan Batang Kali-Jalan Genting Highlands Malaysia Landslide 2022: A Case Report and Strategies to Improve," *International Journal of Management and Human Sciences*, vol. 07, no. 01, pp. 33–40, 2023, doi: 10.31674/ijmhs.2023.v07i01.005.

[14] Department of Statistics Malaysia, "Jadual statistik." [Online]. Available: https://www.dosm.gov.my/uploads/publications/

[15] I. El Naqa and M. J. Murphy, "*What is machine learning?*," in Machine Learning in Radiation Oncology, Berlin, Germany: Springer, 2015, pp. 3–11, doi: 10.1007/978-3-319-18305-3_1.

[16] Y. Asamoah and K. Ansah-Mensah, "Temporal Description of Annual Temperature and Rainfall in the Bawku Area of Ghana," *Advances in Meteorology*, vol. 2020, 3402178, pp. 1–18, 2020, doi: 10.1155/2020/3402178.

[17] B. Praveen *et al.*, "Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-67228-7.

[18] A. Sharma and M. K. Goyal, "Bayesian Network Model for Monthly Rainfall Forecast." [Online]. Available: https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi

[19] R. He, L. Zhang, and A. W. Z. Chew, "Data-driven multi-step prediction and analysis of monthly rainfall using explainable deep learning," *Expert Syst Appl*, vol. 235, Jan. 2024, doi: 10.1016/j.eswa.2023.121160.

[20] *Proc. IEEE Congress Evolutionary Computation (CEC)*, Vancouver, BC, Canada, 24–29 July 2016. IEEE Press, 2021. doi: 10.5555/978-1-5090-0623-6.

[21] P. Das, D. A. Sachindra, and K. Chanda, "Machine Learning-Based Rainfall Forecasting with Multiple Non-Linear Feature Selection Algorithms," *Water Resources Management*, vol. 36, no. 15, pp. 6043–6071, Dec. 2022, doi: 10.1007/s11269-022-03341-8.

[22] S. Benziane, "Survey: Rainfall Prediction Precipitation, Review of Statistical Methods," *WSEAS Transactions on Systems*, vol. 23, pp. 47–59, Jan. 2024, doi: 10.37394/23202.2024.23.5.

[23] A. Gutierrez-Lopez, I. Cruz-Paz, and M. M. Mandujano, "Algorithm to predict the rainfall starting point as a function of atmospheric pressure, humidity, and dewpoint," *Climate*, vol. 7, no. 11, 2019, doi: 10.3390/cli7110131.

[24] J. Yan, T. Xu, Y. Yu, and H. Xu, "Rainfall forecast model based on the tabnet model," *Water (Switzerland)*, vol. 13, no. 9, May 2021, doi: 10.3390/w13091272.

[25] M. Sudha and V. Balasubramanian, "Identifying effective features and classifiers for short term rainfall forecast using rough sets maximum frequency weighted feature reduction technique," *Journal of Computing and Information Technology*, vol. 24, no. 2, pp. 181–194, 2016, doi: 10.20532/cit.2016.1002715.

[26] M. Sudha and B. Valarmathi, "Rainfall forecast analysis using rough set attribute reduction and data mining methods," *AGRIS on-line Papers in Economics and Informatics*, vol. 6, no. 4, pp. 1-10, Dec. 2014. [Online]. Available: http://ageconsearch.umn.edu

[27] S. M. Wajid, T. Javed, and M. M. Su'ud, "Ensemble Learning-Powered URL Phishing Detection: A Performance Driven Approach," *Journal of Informatics and Web Engineering*, vol. 3, no. 2, pp. 134–145, Jun. 2024, doi: 10.33093/jiwe.2024.3.2.10.

[28] T. Ahmed Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. M. Su'ud, "Sentiment Analysis using Support Vector Machine and Random Forest," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 67–75, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.5.

[29] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[30] F. Ozen, "Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey," *Heliyon*, vol. 10, no. 4, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25746.

[31] Y. Q. Lim and Y. L. Loo, "Characteristics of multiclass suicide risks tweets through feature extraction and machine learning techniques," *International Journal on Informatics Visualization*, vol. 7, no. 4, pp. 2297-2305, 2023, doi: https://dx.doi.org/10.62527/joiv.7.4.2284.

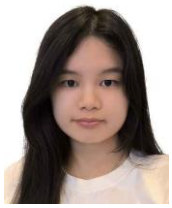
[32] X. Wang, W. Nie, W. Xie, and Y. Zhang, "Incremental learning-random forest model-based landslide susceptibility analysis: A case of Ganzhou City, China," *Earth Sci Inform*, vol. 17, no. 2, pp. 1645–1661, Apr. 2024, doi: 10.1007/s12145-024-01229-2.

[33] M. A. Saleh and H. M. Rasel, "Performance evaluation of Machine Learning based regression models for rainfall forecasting", Research Square, 2024, doi: 10.21203/rs.3.rs-3856741/v1.

[34] M. A. Habib and M. J. Hossain, "Revolutionizing Wind Power Prediction—The Future of Energy Forecasting with Advanced Deep Learning and Strategic Feature Engineering," *Energies (Basel)*, vol. 17, no. 5, Mar. 2024, doi: 10.3390/en17051215.

[35] M. S. A. Magboo and V. P. C. Magboo, "Feature Importance Measures as Explanation for Classification Applied to Hospital Readmission Prediction," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 1388–1397. doi: 10.1016/j.procs.2022.09.195.

BIOGRAPHIES OF AUTHORS

	**Ye Zhian Teoh** graduated from Multemedia University (Cyberjaya), Malaysia with a Bachelor of Computer Science (specialization in Data Science) (Hons.) in 2024. She is currently pursuing a part time Master of Data Science at Sunway University. She can be contacted at email: yezhian020328@gmail.com
	**Yim Ling Loo** graduated from UNITEN (Universiti Tenaga Nasional) with a First Class Honours degree in Bachelor of Computer Science (Software Engineering) (Hons.) in 2008. She pursued and attained Master of Information Technology with full research mode in 2014 and Doctor of Philosophy in Information and Communication Technology in UNITEN in September 2021. Her main research focus is natural language processing, specifically sentiment analysis, as well as feature selection and engineering. She attained her professional technologist recognition from Malaysian Board of Technologist in end of March 2021. She is a certified Train-The-Trainer (TTT) Trainer of HRDCorp Malaysia. She can be contacted at email: looyimling@gmail.com