International Journal of Innovative Research and Scientific Studies, 7(2) 2024, pages: 727-741

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Machine learning classification of rainfall forecasts using Austin weather data

Ting Tin Tin[1*], Enoch Hii Chen Sheng[2], Loo Seng Xian[3], Lee Pei Yee[4], Yeap Sheng Kit[5]

[1]*Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia.*
[2,3,4,5]*Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, 53300 Kuala Lumpur, Malaysia.*

Corresponding author: Ting Tin Tin (*Email: tintin.ting@newinti.edu.my*)

Abstract

The paper examines the machine learning classification of rainfall forecasts using Austin weather data. Rain is a natural phenomenon that is essential for the Earth's water cycle. Rain brings benefits to daily lives and also causes disasters, such as floods, which will endanger lives in addition to causing great losses. Due to this, many methods have been studied and experimented with to find a solution to predict rainfall and prevent tragedies from happening. In this research, the Austin weather dataset is applied to make predictions of rainfall through the implementation of machine learning models. The models used to predict rainfall based on the data set were Extreme Gradient Boosting, Support Vector Machine, Long Short-Term Memory, and Random Forest models. 21 variables with 1319 records were present in the dataset, but the variables used for the modelling were 18 variables from the original data, and 1 variable, "Precipitation Sum," was converted to the variable "Precipitation Range," which contained the classes "no rain," "small rain," "moderate rain," and "heavy rain" based on specific value ranges. After training and predicting the data on the models, it was shown that Extreme Gradient Boosting gave the best results of 85.17% accuracy, 83.19% F1 score, 85.17% recall score, and 82.14% precision score, and was able to give predictions on all 4 classes of rainfall. This study and the way to implement machine learning models for rainfall prediction have the potential to provide new insights and methodologies for future studies and pave the way for finding a high-accuracy rainfall prediction method to avoid disaster.

Keywords: Disaster risk reduction, Extreme gradient boosting, Long short-term memory, Machine learning, Rainfall prediction, Random forest, Support vector machine.

DOI: 10.53894/ijirss.v7i2.2881
Funding: This research is supported by INTI International University (Grant number: T&E3422).
History: Received: 9 October 2023/**Revised:** 10 January 2024/**Accepted:** 21 February 2024/**Published:** 11 March 2024

Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Institutional Review Board Statement: The Ethical Committee of the INTI International University, Malaysia has granted approval for this study (Ref. No. T&E3422).
Publisher: Innovative Research Publishing

1. Introduction

Rainfall is an essential aspect of the planet's water cycle and plays an important role in sustaining the ecosystem. Almost every country experiences rainfall at some point throughout the year, and the amount and frequency of rainfall can have significant impacts on the environment and human activities. On the one hand, rainfall is essential to maintain healthy ecosystems, support agriculture and food production, and replenish water resources. Adequate rainfall is necessary for plants to grow and survive, as they form the basis of the food chain and help replenish rivers, lakes, and groundwater reserves.

However, rainfall can also have negative impacts, especially when it occurs in excess or in areas that are not equipped to handle large amounts of water. Heavy rain can cause flooding, landslides, and soil erosion, which can damage infrastructure, destroy crops, and lead to the loss of life. Floods have been one of the worst natural disasters faced by many countries, especially Malaysia. Almost every year, Malaysian citizens face this destructive disaster and face great losses. The Department of Statistics of Malaysia estimates that floods in 2022 and 2021 caused losses of RM622.4 million and RM6.1 billion [1]. According to the statistics report, the losses consisted of different types of damage, which were living quarters, vehicles, business premises, agriculture, manufacturing, and public assets and infrastructure, where the losses accordingly were RM157.4 million, RM18.8 million, RM50.3 million, RM154.5 million, RM8.7 million, and RM232.7 million in 2022 [1]. In 2021, the loss amount was RM1,622 million, RM982.8 million, RM525.8 million, RM90.6 million, RM891.4 million, and RM2,000 million [1].

For years, weather prediction has been a challenge that meteorologists have intended to master. With the help of weather prediction, many problems in our daily lives could be avoided. However, weather forecasting cannot be easily mastered as it faces many challenges. Lewis Fry Richardson introduced Numerical Weather Prediction (NWP) to the field of weather forecasting in1922, but he was unable to produce an accurate forecast result [2]. It was found that his failure was due to a lack of consideration of computational facilities [2]. Additionally, upper air sounding and his unclear understanding of the hydrodynamic properties of atmospheric circulation contributed to his failure [3], which made time integration accuracy difficult. Many years later, the NWP was practically used for weather prediction on the IBM 701, which was installed in 1955 [2]. Since then, the application of NWP has become an important tool in the field of weather prediction and forecasting. However, there are many problems that cause an inaccurate prediction result.

In addition to the implementation of NWP, machine learning models for weather prediction were applied to find a better solution to predict weather more accurately than NWP. Throughout this research, 4 machine learning models-Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Random Forest (RF)-were introduced and compared to find a model suitable for predicting rain precipitation range. With the implementation and comparison of these 4 machine learning models, the most accurate one can be selected and considered as a choice to replace NWP. With the help of the selected machine learning model, the weather station can benefit in both aspects of accuracy and efficiency in rain forecasting, which can help reduce the flood losses experienced by Malaysians every year.

2. Literature Review

2.1. Extreme Gradient Boosting (XGBoost)

XG Boost is a scalable machine learning model with the use of ensemble learning algorithms for tree boosting [4]. XGBoost is a powerful model with a higher accuracy rate that is widely implemented in different fields such as business, human disease, man-made disasters, etc. Due to its ability to operate faster than current effective solutions, this model is what makes it successful in any circumstance [4]. This model can also handle regression and classification problems, and it has the ability to handle missing data [5]. In a study by Srinivas, et al. [6] on rainfall prediction, they compared XGBoost with other machine learning methods, including LSTM and RF, and discovered that XGBoost performed better than them in terms of accuracy and efficiency. The performance of XGBoost is 99% accuracy, while RF gives 92% accuracy, and LSTM gives an accuracy of 42% [6]. In addition, Meihong, et al. [7] used XGBoost to predict daily rainfall in the Yunnan province of China. They found that XGBoost outperformed other algorithms for machine learning, including RF and Support Vector Regression (SVR), as XGBoost only got aRoot Mean Squared Error (RMSE) of 4.08 mm/day compared to other models. When it comes to forecasting the risk of flash floods, XGBoost is more reliable (84%), while Yunnan's southeast and southwest are predicted to be high-risk areas [7]. Furthermore, Liyew and Melese [8] examined the results of different algorithmic models: XGBoost, Multi-Linear Regression (MLR), and RF. XGBoost shows the best results among other models by using the high correlation coefficient for environmental characteristics. The Mean Absolute Error (MAE) and RMSE values of the XGBoost algorithms were 3.58 and 7.85, respectively [8]. However, this model can be time consuming when it processes large amounts of data [7]. Therefore, in order to choose the optimal model with the highest accuracy in rainfall prediction, XGBoost is suggested as one of the models to be tested in this study.

2.2. Support Vector Machine (SVM)

SVM is a powerful supervised machine learning technique that has been widely used for various prediction tasks, such as classification and regression tasks [9]. It maps the input data to high-dimensional feature spaces and creates a binary linear classifier to assign new data points to one of two classes [10]. In the context of rainfall prediction, various studies have compared the performance of SVM models with other machine learning techniques such as Seasonal Autoregressive Integrated Moving Average (SARIMA), K-Nearest Neighbours (KNN), Extreme Learning Machines (ELM), Particle Swarm Optimisation and Adaptive Neuro Fuzzy Inference System (PSOANFIS), and also neural network approaches like Artificial Neural Network (ANN). For example, a study by Abdullah, et al. [11] compared the performance of SVM and a

variation of moving average models (SARIMA)to forecast the rainfall value (mm) every day in Indonesia [11]. Based on lower rate in MAE, RMSE, and Mean Absolute Percentage Error (MAPE) anda stronger coefficient value of 0.655, the authors came to the conclusion that the SVM model did better than the SARIMA model. Besides, the authors have also found that SVM can perform better with seasonal constraints by implementing fixed interval patterns such as daily, weekly, or monthly. Similarly, a study conducted by Kaushik, et al. [12] compared the performance of SVM with other supervised machine learning algorithms such as KNN and ELM for the prediction of annual rainfall values in India [12]. As a result, the SVM model was the best model with the smallest number of error values and was well matched with the observed output. Interestingly, the authors also found that SVM has better computational efficiency compared to KNN and ELM, even though it has a limited number of data points. In another study by Pham, et al. [13], the authors have suggested three different models, such as SVM, a hybrid model that combines fuzzy inference and swarm optimisation techniques for improved performance, and PSOANFIS and ANN. The authors compared the performance of each model and evaluated its effectiveness in predicting rainfall in Vietnam [13]. The result showed that SVM was the best performing-model, with the lowest MAE value while maintaining the highest correlation coefficient (R) value of 0.829. It is also highlighted that SVM performed better on daily prediction and was able to track the rain behaviour between a non-rain period and a rain period. With that being said, all three studies indicate that the SVM model holds great potential as one of the methods to predict rainfall, especially for daily prediction tasks that also work well with interval and period constraints.

2.3. Long Short-Term Memory (LSTM)

LSTM, a part of the Recurrent Neural Network (RNN) family, is very powerful in deep learning as it can be used for time series prediction [14] by capturing and memorising values from historical data, which are then used for future prediction [15]. The introduction of LSTM was to overcome the problems faced by standard RNN, where system training had difficulty capturing long-term dependencies if vanishing gradients were present [16]. The LSTM model is a powerful and famous model as it can also be used to solve sequential problems other than time-series prediction. According to the research on precipitation prediction carried out by Salehin, et al. [17] LSTM and the neural network were applied [17]. In their research, data were obtained from 6 regions of Bangladesh. The parameters used for the LSTM model to predict the amount of rainfall (mm) were temperature, humidity, dew point, wind pressure, wind speed, and wind direction, which achieved a result of 76% accuracy. In the research, the memory block of the LSTM model played an important role in rainfall time calculation, where 3 gates were applied, which were the input, output, and forget/hidden gate. The Input Gate was used for the storage of new information in the cell state; the Hidden Gate was used to know what information needed to be kept; and the Output Gate was used to allow the result of the block to be activated at timestamp. In another study by Ouma, et al. [18], LSTM was compared with the Wavelet Neural Network (WNN) in the analysis of the trend of rainfall and runoff time series. The result of the predicted rainfall amount in the research found that LSTM, $R^2 = 0.8610$, gave a better predictive result than WNN, $R^2 = 0.7825$. These 2 models were trained with meteorological data where the parameters were precipitation, mean temperature, relative humidity, wind speed, and solar radiation. The topological structure was formed by 4 hidden layers, each consisting of 30 neurons. The performance of both models was shown to increase as the number of hidden layers and neurons corresponding to them increased. It was also found that, with the help of the increase in neurons, the WNN prediction error could be minimised compared to the LSTM model. At the end of the research, it was shown that the LSTM performed better than the WNN as it obtained a higher R^2 accuracy than the WNN model throughout the investigation.

2.4. Random Forest (RF)

RF is made up of several basic classifiers that are not dependent on each other, and how it works is that a test sample will be input into the new classifier, and then the class label of the sample will be decided based on the voting results from every classification, respectively [19]. Besides that, RF is especially popular as it can increase the forecast accuracy and can also prove effective in solving the overfitting problem. This has made RF very useful in various fields, such as text and image classification [19]. Research by Zamani Joharestani, et al. [20] has used RF as part of their machine learning algorithm to predict the level of pollution; they included 23 features in the model and found that RF does not perform as well as XGBoost with an R^2 value of a slight variation between 0.66 and 0.78 [20]. Moving on, another recent study done by Yao, et al. [21] applied RF to hail forecasting in the Shandong Peninsula region with the help of observation data from 41 meteorological stations from 1998-2013 with a focus on thermal factors for prediction. The team has used cross-validation to select an optimal probability for the forecast, as it has great simulation accuracy, minimal average error, and provides a very stable fit [21]. In addition to that, Hill, et al. [22] have experimented with RF together with other models in the context of severe weather predictions for 3 days. They have trained the models, including RF, for different regions such as the Western, Central, and Eastern Continental US (CONUS). From the results of their findings, they have found that on day 1, RF slightly underperformed when compared with the other models, but outperformed the rest by a large margin on days 2 and 3. Therefore, they concluded that with the help of RF, they would be able to improve the operational severe weather forecast over the 3-days period Hill, et al. [22]. Ali, et al. [23] have also tried to incorporate the use of RF to design a hybrid model that includes Complete Ensemble Empirical Mode Decomposition (CEEMD)and Kernel Ridge Regression (KRR) to test whether there are improvements in terms of monthly rainfall forecasts in the area of Parachinar, Gilgit, and Muzaffarabad in Pakistan. With the RF model alone, they managed to get an average benchmark efficiency of 0.585, which is the fifth in the leaderboard among seven other model configurations. But with the combination of RF, CEEMD and KRR, they managed to achieve the highest average benchmark efficiency of 0.892 Ali, et al. [23]. Zainudin, et al. [24] have also analysed RF together with other models in terms of Malaysian rainfall prediction. Among the 5 classifier models

compared (Naïve Bayes (NB), SVM, Decision Tree (DT), Neural Network (NN), and RF), RF and DT yield the best scores for the F measure of 71.9% and 73.7%, respectively at a data split of 30-70 [24]. Last but not least, Mohan and Gupta [25] used RF to produce a low-cost, portable, yet effective solution. They used a 25-75 data split on a dataset with 7339 records and managed to obtain an accuracy of 87.90% [25].

2.5. Multi-Linear Regression (MLR)

Research showed that MLR was also applied for weather forecasting. The purpose of using MLR was to find and analyse the relationship between the variables X and Y, which are independent and dependent, respectively [26]. When compared with Simple Linear Regression (SLR), SLR can analyse only one X variable, while MLR can analyse multiple X variables [27]. Therefore, MLR uses various independent variables X to predict the sum of precipitation, not to mention the outliers found [28]. In the field of weather forecasting, using old techniques such as Numerical Weather Prediction (NWP) is not accurate as it is not capable of taking into account erratic changes in weather conditions, thus providing results that are not reliable and, at times, are not reliable.

Inaccurate when MLR is able to capture multiple changes in weather conditions like the temperature, wind speed, wind direction, etc. Therefore, by comparing MLR with the others, this model has proven to be the most accurate when compared with models like the SVM, Bayesian Enhanced Modified Approach (BEMA), Ranys method, Broyden–Fletcher–Goldfarb–Shanno (BFGS), and Multi-Variant method when used to predict weather by Anusha, et al. [29]. Besides weather forecasting, MLR has been used for rainfall prediction to detect floods that could cause terrible damage to crops or even fatalities; they took into account variables like time, min/max temperature, wdf, Cirrocumulus, precipitation, and vapour pressure. Comparisons between SLR and MLR have also been made, but they have a higher error rate compared to MLR carried out by Sreehari and Srivastava [27]. In addition to that, MLR has been used to estimate the time taken for the rice cultivation process in order to accelerate the exchange rate for their local farmers, as MLR can use more than one explanatory variable. They proceeded with their experimentation to achieve an RMSE value as low as possible; this was achieved with the help of a 2016-2017 weather dataset for the purpose of training and testing of data. Luminto and Harlili [28] choose regions with one of the highest crop productions. Gandhi, et al. [30]conducted an experiment to predict rice crop yield, which is also closely related to the weather in India, using SVM but with a precision of 78.76% [30]. In a study by Gupta, et al. [31], they used MLR to perform a temperature prediction for a day and obtained an absolute mean error of 2.8, which translates to the predicted mean temperature being only off by 2.8 degrees Celsius. This means that, with a deviation of 2.8 degrees Celsius, their model can be used in a simulation to get a rough idea of what the temperature would be [31].

3. Methodology

3.1. Data Description

The data set applied in this investigation was a data set on Austin weather data, which was obtained from Kaggle and originally sourced from WeatherUnderground.com at the Austin KATT Station. The data set contained 21 variables, which included date, temperature, dew point, humidity, sea level pressure, visibility, high miles, wind speed, precipitation sum, and events. However, only 19 variables were used in this investigation, and another variable, precipitate range, was added, as shown in Table 1. Data in the dataset were recorded daily. The raw data recorded in total were 1319 records, which were collected from December 21, 2013 to July 31, 2017.

Table 1.
Description of variables.

Variables	Description
High, low, and avg (F)	The temperature of the atmosphere is measured at the highest point, the lowest point, and the average.
Dew point high, low, avg (F)	The Dew Point is the temperature to which air must be cooled for water vapour in it to condense into dew, the highest and lowest point is recorded together with their average.
Humidity high, low, avg (%)	This measures how close the air temperature is to the dew point, which means that high humidity will have a higher percentage of water vapour in the air. The highest, lowest, and average are recorded.
Sea level pressure high, low, and avg (Inches)	This is the atmospheric pressure measured at sea level; the pressure will vary slightly according to the weather patterns. The highest, lowest, and average are recorded.
Visibility high, low, avg (Miles)	A measure of a distance at which objects can be clearly seen in the atmosphere, this can be recorded with a visibility sensor. Events such as fog, smoke, precipitation, or even have will have an effect.
Wind high, low, avg (miles per hour - mph)	This variable records the wind movement of air in the atmosphere in terms of speed. The highest, lowest, and average point is recorded.
Wind gust (mph)	WindGust records the brief increases in wind speed that occur over a short period of time, typically less than 20 seconds. This can be associated with rapidly changing weather conditions.
Precipitate range (Categorical)	PrecipitateRange records the labels that correspond to the values in Precipitation Sum Inches, which are also within the range provided.

3.2. Data Preprocessing

Figure 1 shows the process of cleaning and transforming the original data set. In order to apply machine learning models, among others, the value of some variables will be substituted with alternative considerations. Missing values in different variables were replaced with various considerations. In cases with more than 30% missing value, the missing value of the variable will be removed; otherwise, the missing value of the variable will be replaced by "0", or the mean value that is not adopted. Once all missing values were handled, the data types of all variables were converted to numeric formats such as float and int. To create a suitable variable for a classification problem, the variable PrecipitateSumInches is replaced with a new variable called PrecipitateRange. The new column will check the values in "PrecipitateSumInches." If the values match the if-else rules, the corresponding label will be assigned to the new variable, such as "small," which indicates values <0.020, "Moderate," which indicates values between 0.021 and 0.157 and "Heavy," which indicates values >= 0.158[32]. Upon finishing the label encoding, unwanted variables such as "date," "events," and "precipitationSumInches" were removed as they do not contribute to the model's performance. The total number of variables and records after data cleaning was 19 and 1315, respectively.

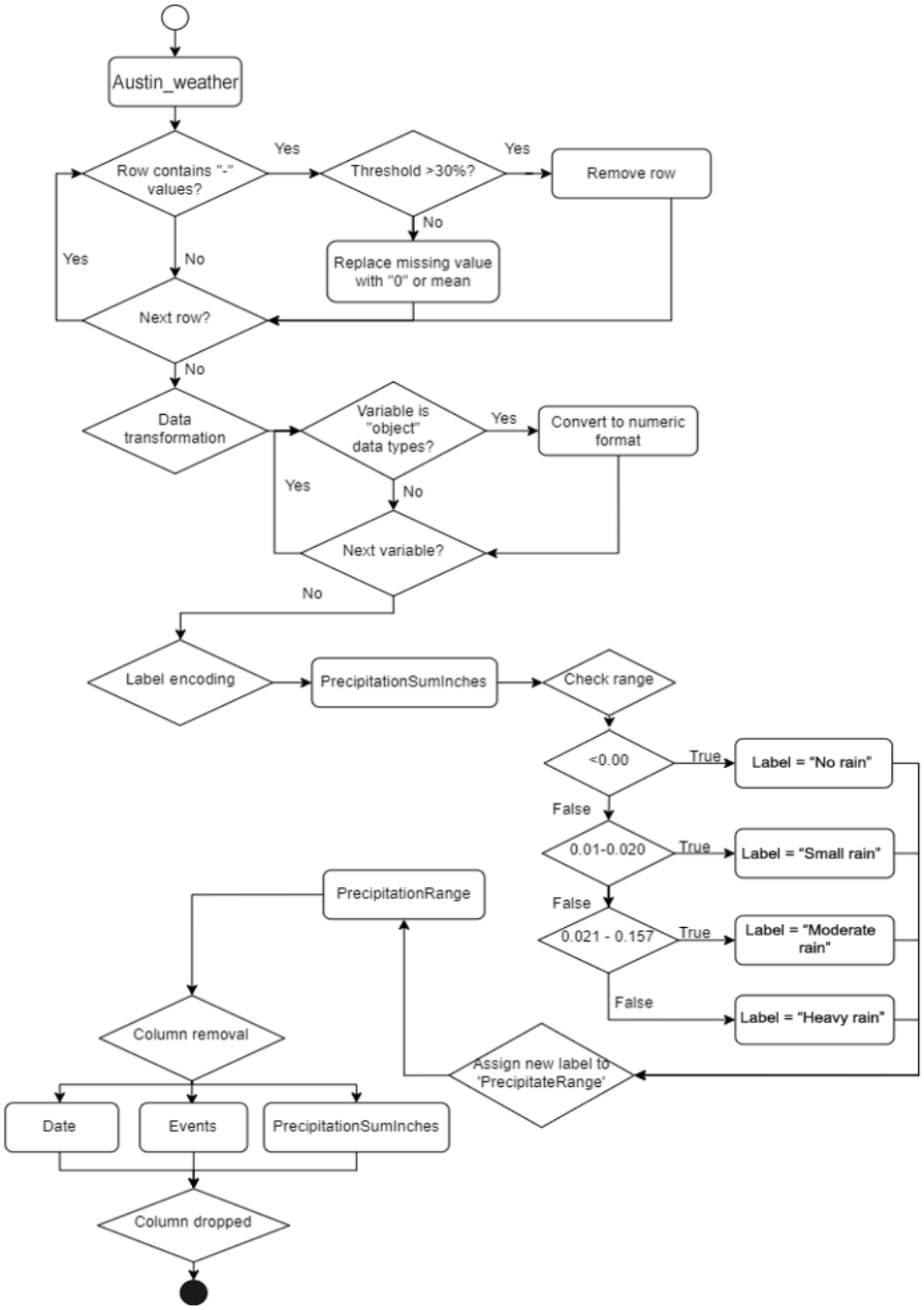

Figure 1.
Data preprocessing process.

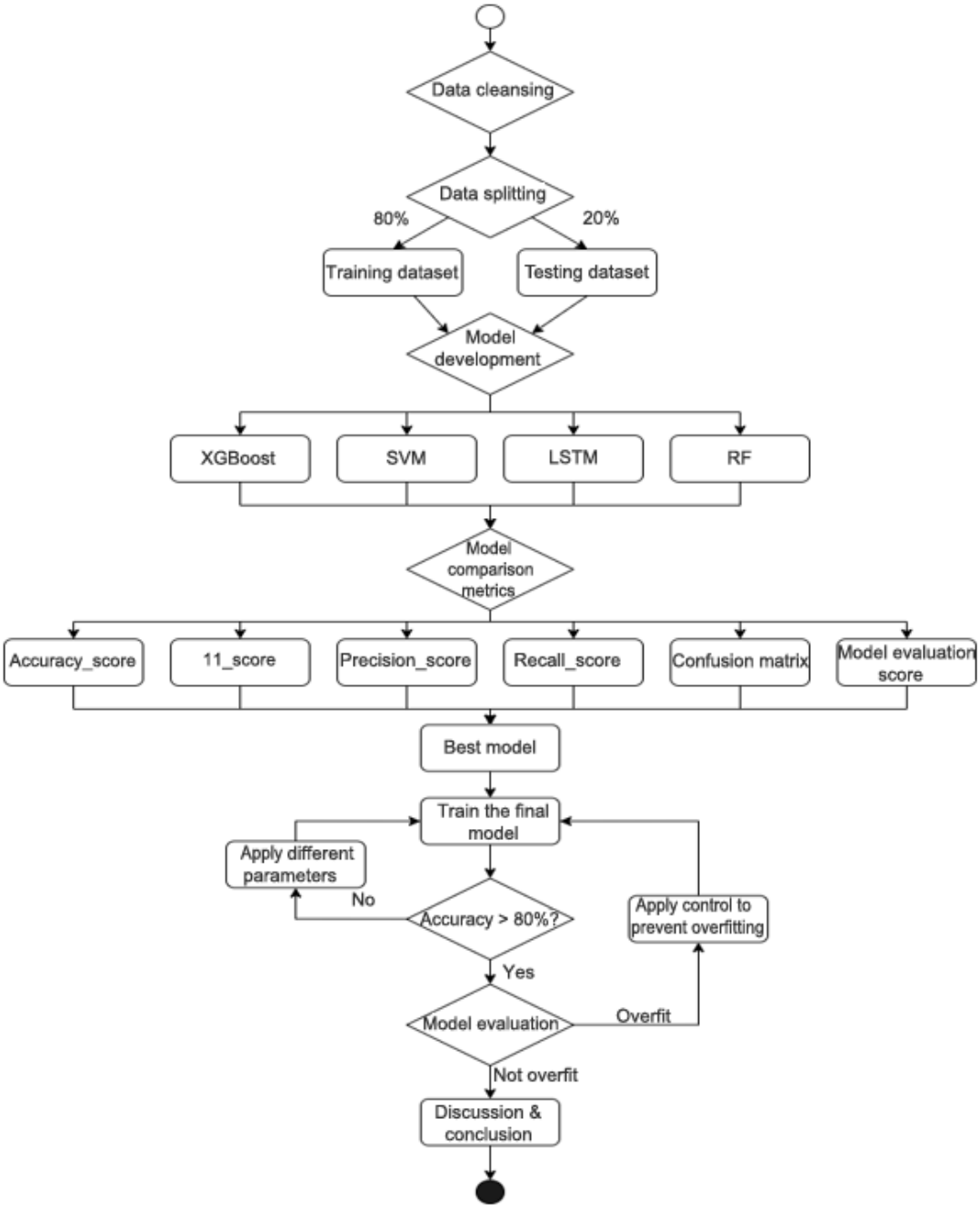

Figure 2.
Prediction model construction methodology.

3.3. Model Comparison Workflow

In this study, multiple predictive models for rainfall prediction tasks were constructed, and their precision was compared as shown in Figure 2. First, the data set was cleaned as shown in Figure 1 before being split into 80% of the training data and 20% of the test data. There were 4 models involved in the comparison: Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), Long Short-term Memory Network (LSTM), and Random Forest (RF). The models were evaluated using common metrics such as the accuracy score, the f1 score, the precision score, and the recall score. Furthermore, the 4 models were also evaluated using model evaluation scores such as "cross-value," and the overall result was visualised using a confusion matrix. Based on the evaluation, the best-performing model will be selected and trained to achieve an accuracy of at least 80%. To ensure that the model is not overtrained, the model evaluation will be applied to prevent any overfitting. This workflow demonstrates the process for determining the most suitable model for this study and improving its performance through training.

4. Results

4.1. Performance of the Prediction Models

The prediction models XGBoost, SVM, LSTM, and RF were applied for the prediction of rainfall, and performance based on the results predicted by each model was calculated. The performance of each model was measured and tabulated as shown in Table 2 for comparison purposes. The results shown in the table were measured in terms of metrics on how well

the machine learning model was able to classify data into its correct class. The performance metrics used to measure the prediction result were accuracy, F1 score, recall, and precision. Among the 4 models, it could be shown that XGBoost achieved the best score of metrics among the other 3 models in predicting rainfall, as its accuracy (R^2), F1 score, recall score and precision score were the highest, which were 0.8517, 0.8319, 0.8517 and 0.8214, respectively. For the SVM model, the metrics scores for accuracy (R^2), F1 score, recall score, and precision score were 0.8137, 0.7590, 0.8137 and 0.7146, while LSTM scored 0.8289, 0.7770, 0.8289 and 0.7323 for its metrics score. Accuracy (R2), F1 score, recall score, and precision score for RF were 0.8403, 0.7996, 0.8403, and 0.7701.

Table 2.
Metrics performance for prediction models.

Model	Accuracy	F1	Recall the following	Precision
XGBoost	0.8517	0.8319	0.8517	0.8214
SVM	0.8137	0.7590	0.8137	0.7146
LSTM	0.8289	0.7770	0.8289	0.7323
RF	0.8403	0.7996	0.8403	0.7701

4.2. Assessment of Overfitting in the Models

An evaluation was carried out on the accuracy of the model on train data. The purpose of the evaluation of the train data was to compare its accuracy with the prediction performance so that the model could be determined if there was a presence of overfitting, where the model learned data too well, causing it to memorise the training data rather than understanding the patterns and relationships between the data. Table 3 shows the evaluation performance for each model by measuring the accuracy. There was a small difference in accuracy between the evaluation performance in the train data and the prediction performance in the test data. Therefore, it can be shown that all 4 models were not overfitting and predicted the data according to their understanding of the pattern and relationship of the data.

Table 3.
Evaluation Vs. prediction.

Model	Accuracy of the evaluation (Train)	Prediction accuracy (Test)
XGBoost	0.8146	0.8517
SVM	0.7966	0.8137
LSTM	0.8380	0.8289
RF	0.8194	0.8403

4.3. Confusion Metrics

As shown in Figure 3, Figure 4, Figure 5, and Figure 6, the confusion matrix of the predicted results for each model. The confusion matrix gives information on the summary of each predicted class and finds the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class. According to these numbers, the evaluation metrics such as precisionrecall and F1 score are calculated to understand the performance of the model for each class.

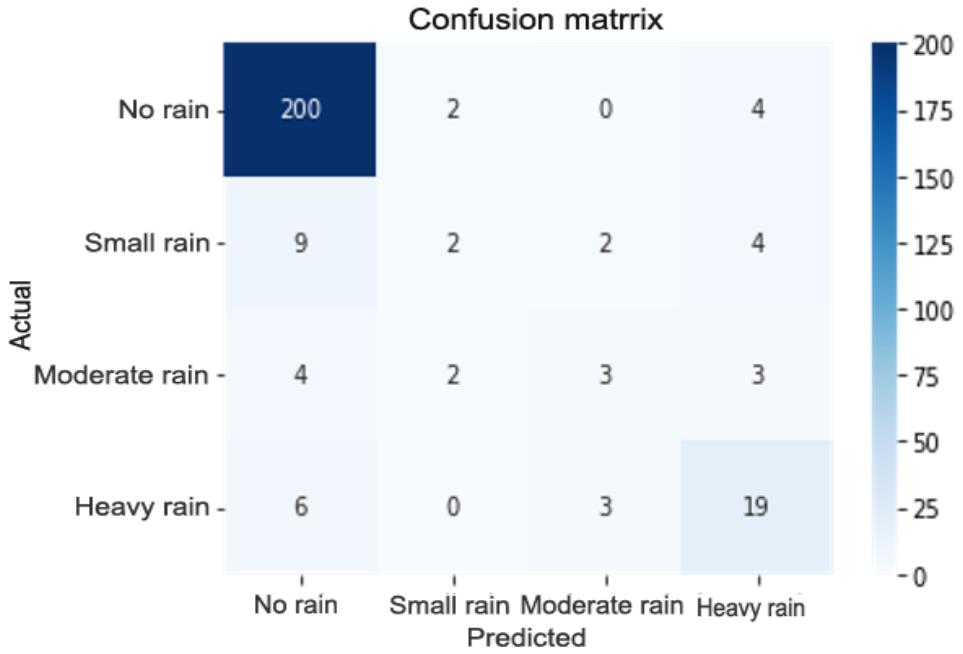

Figure 3.
Confusion matrix for XGBoost.

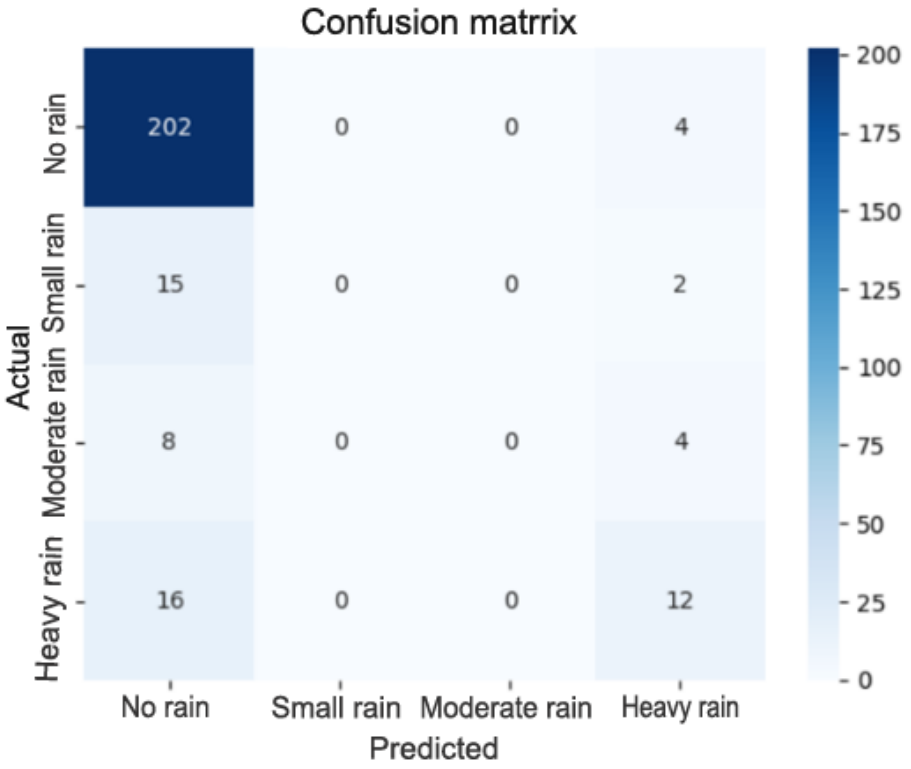

Figure 4.
Confusion matrix for SVM.

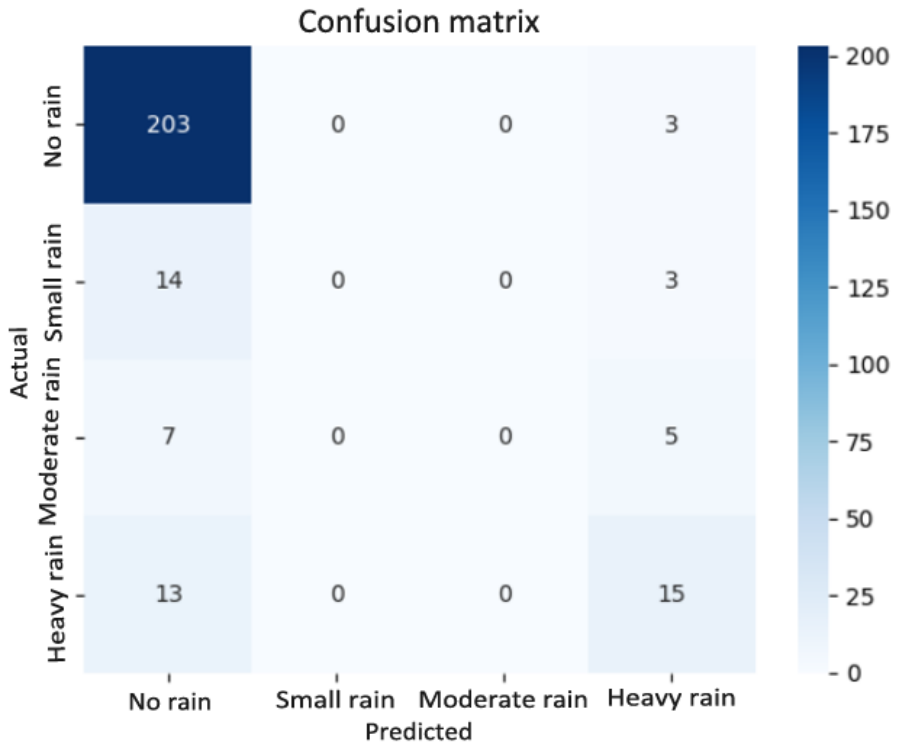

Figure 5.
Confusion matrix for LSTM.

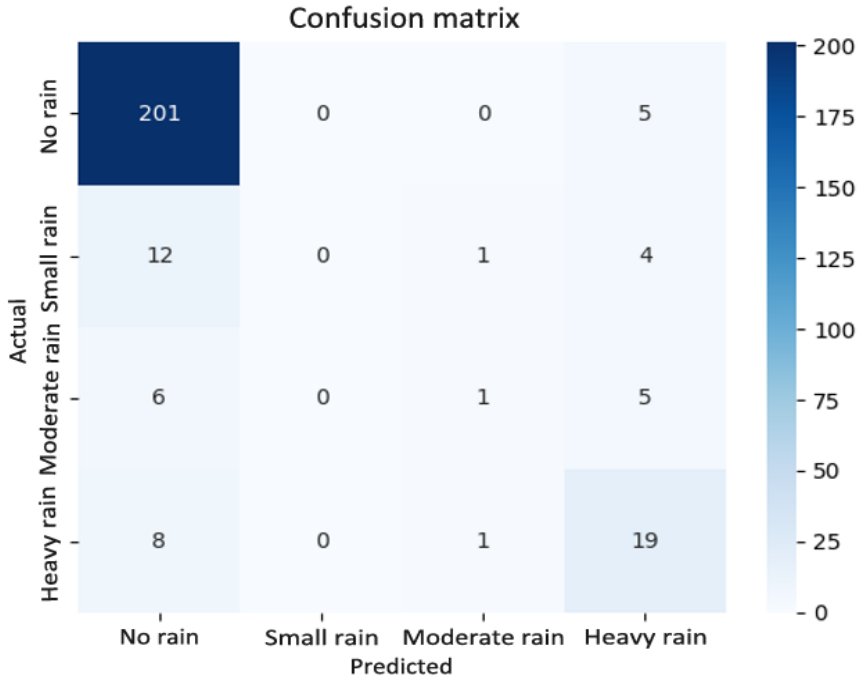

Figure 6.
Confusion matrix for RF.

According to Table 4, it shows the summary of the XGBoost model, where for each class, no rain, small rain, moderate rain, and heavy rain were given predicted values as the precision, recall, and F1-score were stated for each class. It was shown that XGBoost performed well in weather prediction as it was not biased, even though the train data given for small rain and moderate rain were little, which was 17 and 12, respectively.

Table 4.
Classification summary for XGBoost.

Class	Precision	Recall	F1	Support
No rain.	0.91	0.97	0.94	206
Small rain	0.33	0.12	0.17	17
Moderate rain	0.38	0.25	0.30	12
Heavy rain	0.63	0.68	0.66	28

According to Table 5, it shows the summary of the SVM model where only 2 classes, no rain and heavy rain, were given predicted values as the precision, recall, and F1-score were stated for only both classes. It could be shown that the SVM model was biased as it only gave a predictive value for the only classes of no rain and heavy rain, as it had little data to train for small rain and moderate rain.

Table 5.
Classification summary for SVM.

Class	Precision	Recall	F1	Support
No rain	0.84	0.98	0.90	206
Small rain	0.00	0.00	0.00	17
Moderate rain	0.00	0.00	0.00	12
Heavy rain	0.55	0.43	0.48	28

According to Table 6, it shows a summary of the LSTM model. Like SVM, LSTM predicted 2 classes: no rain, and heavy rain where the precision, recall, and F1 score were stated and calculated for both classes. It could be shown that the LSTM model was also biased due to the same reasons as SVM, where only little data leading to small and moderate rain was given for the model to train.

Table 6.
Classification summary for LSTM.

Class	Precision	Recall	F1	Support
No rain	0.86	0.99	0.92	206
Small rain	0.00	0.00	0.00	17
Moderate rain	0.00	0.00	0.00	12
Heavy rain	0.58	0.54	0.56	28

Table 7shows the summary of the RF model. From the table, it is shown that RF was able to predict 3 classes, which were no rain, medium rain, and heavy rain, as precision, recall, and F1-score were calculated for these classes. It could be shown that RF was able to learn the pattern for the moderate rain class even though it was given little data; however, it was unable to find out the predicted values for the small rain class.

Table 7.
Classification summary for RF.

Class	Precision	Recall	F1	Support
No rain	0.89	0.98	0.93	206
Small rain	0.00	0.00	0.00	17
Moderate rain	0.33	0.08	0.13	12
Heavy rain	0.58	0.68	0.62	28

5. Discussion

An accurate rainfall prediction method can be a great change in history, as humans are able to overcome many of the problems caused by rain. This study involved the implementation of four machine learning models: XGBoost, SVM, LSTM, and RF, to predict rainfall based on Austin weather data. In previous studies, these four models were also used for rainfall prediction. The studies and results are summarised and shown in Table 8.

In many previous studies, time-series and regression modelling were applied for research on finding a way to predict rainfall or weather forecasting. In terms of this research, the classification method was applied to the precipitation values, where each specific precipitation range was grouped into different classes of rainfall level. After implementing all four of the models, it was found that not all the models were able to predict all four classes. It was found that only XGBoost was able to predict the 4 classes and achieved the highest accuracy of 0.8517: no rain, small rain, moderate rain, and heavy rain. By building a weighted sum of the predictions of decision trees to make a final prediction, XGBoost takes advantage of high-dimensional data by handling it more effectively using regularisation techniques. As for RF, it outperformed SVM and LSTM with an accuracy of 0.8403; however, like SVM and LSTM, it was unable to give predictions for all 4 classes. It can be shown that all 3 models were biased, and one of the possible reasons was the imbalance of classes in the dataset. This was because the “no rain” class had a significantly higher number of instances than the other classes, as shown in Figure 7. Another possible reason could be the choice of hyperparameters used in the models, which could not be optimal for the dataset.

To resolve the bias in the models, future studies can fine-tune the hyperparameters, such as the penalty parameter “C” and the kernel function. Other than that, a larger data set with a balanced count of classes should be prepared to allow models to learn patterns for each class and increase the accuracy of the models. Normalisation or scaling of the data set can also be applied to all models, as in this research only the data set for the LSTM model was normalized, which may also cause an imbalance in the result. By applying this, the models can then receive a consistent and comparable range of input features, thus allowing the pattern in the data to be captured and making more accurate predictions.

Table 8.
Summary of previous studies.

Ref.	Methods	Variables	Location	Metrics	Value
Srinivas, et al. [6]	XGBoost,RF,LSTM	Minutes past, radar distance, radar reflectivity, maximum reflectivity, correlation coefficient(RhoHV), differential reflectivity(ZDR), specific differential phase(KDP)	Midwestern corn-growing states.	MAE MSE RMSE Accuracy	MAEXGBoost=0.02 degrees,MAERF=0.18 degrees，MAELSTM=0.51 degrees MSEXGBoost=0.0008,MSERF=0.679,MSELSTM=0.536 RMSEXGBoost=0.999,RMSERF=0.927,RMSELSTM=0.422, AccuracyXGBoost=99%,AccuracyRF=92%,AccuracyLSTM=42%
Meihong, et al. [7]	XGBoost, least squares SVM and radial basis function (LSSVM_RBF)	Meteorological: Annual maximum 3 hours precipitation(3-H-P),annual maximum 24 hours precipitation(24-H-P), annual precipitation Topographical: Digital elevation model, slope, river density(RD), vegetation coverage Hydrological: Curve number,topographic wetness index(TWI), soil moisture Anthropological:population, gross domestic product	Yunnan Province,China	Accuracy Precision Recall F-Score Kappa	AccuracyXGBoost=0.84,AccuracyLSSVM_RBF=0.79 PrecisionXGBoost=0.85,PrecisionLSSVM_RBF=0.82 RecallXGBoost=0.83,RecallLSSVM_RBF=0.77 F-ScoreXGBoost=0.83,F-ScoreLSSVM_RBF=0.79 KappaXGBoost=0.68,KappaLSSVM_RBF=0.59
Liyew and Melese [8]	XGBoost,RF,MLR	Year, month, date, evaporation, sunshine, maximum temperature, minimum temperature, humidity, wind speed, and rainfall	Bahir Dar city, Ethiopia	MAE RMSE	MAEXGBoost=4.49,MAERF=4.97, MAEMLR=3.58, RMSEXGBoost=7.85,RMSERF=8.82,RMSEMLR=8.61
Abdullah, et al. [11]	SVM, SARIMA	Monthly rainfall data over the period of January 2011 – June 2020	Bogor city, Indonesia	MAE RMSE MAPE r-Pearson	MAESARIMA(1,0,0,1,0,0) = 128.366, RMSESARIMA(1,0,0,1,0,0) = 156.767, MAPESARIMA(1,0,0,1,0,0) =93.480, r-PearsonSARIMA(1,0,0,1,0,0) = 0.204 MAESARIMA(1,0,1,1,0,1) = 128.173, RMSESARIMA(1,0,1,1,0,1) = 155.401, MAPESARIMA(1,0,1,1,0,1) = 93.835, r-PearsonSARIMA(1,0,1,1,0,1) = 0.257 MAESVM(non-seasonal constraint) = 131.074, RMSESVM(non-seasonal constraint) = 158.749, MAPESVM(non-seasonal constraint) = 94.954, r-PearsonSVM(non-seasonal constraint) = 0.122 MAESVM(seasonal constraint) = 97.16, RMSESVM(seasonal constraint) = 121.62, MAPESVM(seasonal constraint) = 63.28, r-PearsonSVM(Seasonal constraint) = 0.655
Kaushik,	SVM, KNN, ELM	min temperature (MINT), max	Punjab, India	MAE	MAEKNN=7.9,

et al. [12]		temperature (Max T), average wind speed (WIND) and average humidity (HUMD) (Dataset is time series data from year 1973 to 2008)		RMSE Standard deviation (SD) Performance parameter (PP) Elapsed time (ET)	RMSEKNN=38.3,DKNN=101.9,PPKNN=0.62,ETKNN =8 MAEELM=1.15, RMSEELM=24.5,SDELM=96.7,PPELM=0.75,ETELM= 19 MAESVM=1.7, RMSESVM=7.6,SDSVM=100.4,PPSVM=0.92,ETSVM =13
Pham, et al. [13]	SVM, PSOANFIS, ANN	Min temperature, max temperature, solarradiation, wind speed, relative humidity	Vietnam	R MAE Probability of detection (POD) Critical success index (CSI) False alarm ratio (FAR)	RSVM = 0.849 MAESVM = 2.846 mm PODSVM(2 mm) = 0.89 CSISVM(2 mm) = 0.78 FARSVM(2 mm) = 0.14 RPSOANFIS = 0.844 MAEPSOANFIS = 3.155 mm PODPSOANFIS(2 mm) = 0.94 CSIPSOANFIS(2 mm) = 0.75 FARPSOANFIS(2 mm) = 0.21 RANN = 0.829 MAEANN = 3.155 mm PODANN(2 mm) = 0.91 CSIANN(2 mm) = 0.76 FARANN(2 mm) = 0.18
Salehin, et al. [17]	LSTM	Temperature, humidity, dew point, wind pressure, wind speed and wind direction	Bangladesh	Accuracy	Accuracy = 76%
Ouma, et al. [18]	LSTM,WNN	Precipitation, mean temperature, relative humidity, wind speed and solar radiation	Nzoia River	R^2	R^2LSTM = 0.8610 R^2WNN = 0.7825
Zamani Joharestani , et al. [20]	RF, deep learning, XGBoost	Dew point, air temp, max/min air temp, the relative humidity, daily rainfall, visibility, speed of wind, air pressure, sustained wind speed	Mehrabad	R MAE RMSE	RF (Include AODs) R^2 =0.66, RMSE = 15.30, MAE = 11.15 RF (Include AOD10) R^2 =0.78, RMSE = 14.54, MAE = 10.8 RF (NoAODs) R^2 =0.78, RMSE = 14.47, MAE = 10.78 XGBoost(Include AODs) R^2 = 0.67, RMSE = 15.15, MAE = 10.94 XGBoost(Include AOD10) R^2 = 0.80, RMSE = 13.62, MAE = 10.0 XGBoost(No AODs) R^2 = 0.8, RMSE = 13.66, MAE = 10.0

					DeepLearning(Include AODs) R^2 = 0.63, RMSE = 15.89, MAE = 11.66 DeepLearning(Include AOD10) R^2 = 0.77, RMSE = 14.65, MAE = 10.88 DeepLearning(No AODs) R^2 = 0.76, RMSE = 15.11, MAE = 11.2
Ali, et al. [23]	RF, KRR, CEEMD	Longitude, latitude, elevation	Pakistan	RMSE	**All average value** RF Gilgit = 0.5373, Muzaffarabad = 0.5797, Parachinar= 0.634 KRR Linear Gilgit = 0.0203, Muzaffarabad = -0.0373, Parachinar = 0.1967 KRR Gaussian Gilgit = -0.1863, Muzaffarabad = -1.5057, Parachinar = -0.2683 KRR Gaussian Gilgit = 0.0800, Muzaffarabad = 0.1180, Parachinar = 0.2263 CEEMD-RF Gilgit = 0.858, Muzaffarabad = 0.8103, Parachinar = 0.845 CEEMD-RF-KRR Linear Gilgit = 0.7653, Muzaffarabad = 0.6157, Parachinar = 0.3860 CEEMD-RF-KRR Polynomial Gilgit = 0.916, Muzaffarabad = 0.8727, Parachinar = 0.8863 CEEMD-RF-KRR Gaussian Gilgit = 0.7990, Muzaffarabad = 0.807, Parachinar = 0.845
Mohan and Gupta [25]	RF	Datetime, temp_avg, hum_avg, pressure_avg, rain	Dehli	Accuracy	Accuracy = 0.879

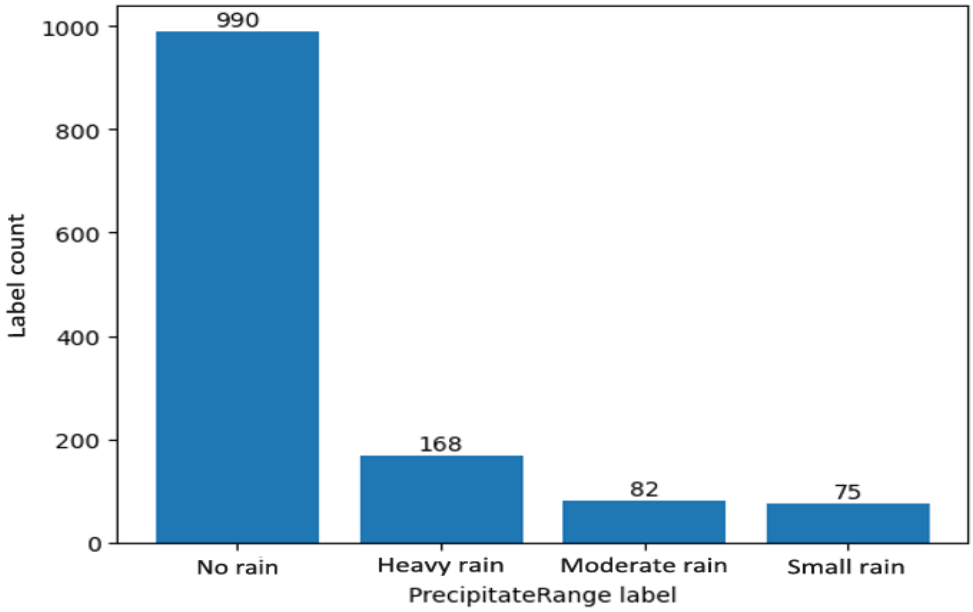

Figure 7.
Count for each classes in the dataset.

6. Conclusion

In this present study, XGBoost, SVM, LSTM, and RF models were applied to predict rainfall in Austin. The variables used for the models in this study were highest temperature, lowest temperature, average temperature, highest point of dew, lowest point of dew, average point of dew, highest humidity, lowest humidity, average humidity, highest sea level pressure, lowest sea level pressure, average sea level pressure, high visibility, low visibility, high wind speed, low wind speed, and precipitation range. The results predicted by the models were measured using metrics such as accuracy; precision score, F1 score, and recall score. The accuracy of all the models was more than 0.80, which means that the model had a high accuracy in predicting the rain. Among these models, XGBoost was shown to be the best, with an accuracy of 0.8517. In addition to that, it achieved a higher precision score, F1 score, and recall score compared to SVM, LSTM, and RF, which were 0.8214, 0.8319, and 0.8517, respectively. The XGBoost model was also efficient and capable of learning the pattern of rainfall class data provided. It could be shown that 4 of the rainfall classes could be predicted by XGBoost despite the insufficient amount of data for the small and medium rain classes, which caused SVM, LSTM, and RF to be biased in their predictions.

This study could provide a new way of predicting rainfall, in addition to using NWP for prediction. This study shows that there is a high accuracy of more than 80% in predicting rainfall using XGBoost, SVM, LSTM, and RF for the classification method. In future work, fine-tuning the parameters of the models and using a larger amount of data for modelling purposes will make rainfall prediction possible using machine learning models. On the other hand, research could include Internet of Things (IoT) technology in collecting data and enhancing flood management in the country [33]. Nevertheless, top management also plays an important role in the sustainable environmental research and implementation of disaster plans [34].

References

[1] Department of Statistics Malaysia, "Ministry of economy department of statistics malaysia special report on impact of floods in Malaysia," Retrieved: https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=496&bul_id=RDVmbnlKUk1rdzRaZmhpK1F6SEZCUT09&menu_id=WjJGK0Z5bTk1ZElVT09yUW1tRG41Zz09#. [Accessed 3-31-2023], 2022.

[2] F. G. Shuman, "History of NWP," *Weather Forecast*, vol. 4, no. 3, pp. 286–296, 1988.

[3] R. Kimura, "Numerical weather prediction," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 90, no. 12-15, pp. 1403-1414, 2002.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[5] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain Informatics*, vol. 4, no. 3, pp. 159-169, 2017. https://doi.org/10.1007/s40708-017-0065-7

[6] A. S. T. Srinivas, R. Somula, K. Govinda, A. Saxena, and P. A. Reddy, "Estimating rainfall using machine learning strategies based on weather radar data," *International Journal of Communication Systems*, vol. 33, no. 13, 2020. https://doi.org/10.1002/dac.3999

[7] M. Meihong *et al.*, "XGBoost-based method for flash flood risk assessment," *Journal of Hydrology*, vol. 598, p. 126382, 2021. https://doi.org/10.1016/j.jhydrol.2021.126382

[8] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," *Journal of Big Data*, vol. 8, pp. 1-11, 2021. https://doi.org/10.1186/s40537-021-00545-4

[9] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988-999, 1999. https://doi.org/10.1109/72.788640

[10] A. Garg and H. Pandey, "Rainfall prediction using machine learning," *International Journal of Innovative Science and Research Technology*, vol. 4, no. 5, pp. 56-58, 2019.

[11] A. S. Abdullah, B. N. Ruchjana, and I. G. N. M. Jaya, "Comparison of SARIMA and SVM model for rainfall forecasting in Bogor city, Indonesia," *In Journal of Physics: Conference Series*, vol. 1722, no. 1, p. 012061, 2021. https://doi.org/10.1088/1742-6596/1722/1/012061

[12] S. Kaushik, A. Bhardwaj, and L. Sapra, "Predicting annual rainfall for the Indian state of Punjab using machine learning techniques," presented at the In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, 2020.

[13] B. T. Pham *et al.*, "Development of advanced artificial intelligence models for daily rainfall prediction," *Atmospheric Research*, vol. 237, p. 104845, 2020. https://doi.org/10.1016/j.atmosres.2020.104845

[14] Z. Karevan and J. A. Suykens, "Transductive LSTM for time-series prediction: An application to weather forecasting," *Neural Networks*, vol. 125, pp. 1-9, 2020. https://doi.org/10.1016/j.neunet.2019.12.030

[15] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, "A Comparison of ARIMA and LSTM in forecasting time series," in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, Institute of Electrical and Electronics Engineers Inc*, 2018, pp. 1394–1401.

[16] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157-166, 1994. https://doi.org/10.1109/72.279181

[17] I. Salehin, I. M. Talha, M. M. Hasan, S. T. Dip, M. Saifuzzaman, and N. N. Moon, "An artificial intelligence based rainfall prediction using LSTM and neural network," presented at the In 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, 2020.

[18] Y. O. Ouma, R. Cheruyot, and A. N. Wachera, "Rainfall and runoff time-series trend analysis using LSTM recurrent neural network and wavelet neural network with satellite-based meteorological data: case study of Nzoia hydrologic basin," *Complex and Intelligent Systems*, vol. 8, no. 1, pp. 213–236, 2022. https://doi.org/10.1007/s40747-021-00365-2

[19] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," presented at the In International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018 Springer International Publishing, 2019.

[20] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data," *Atmosphere*, vol. 10, no. 7, p. 373, 2019. https://doi.org/10.3390/atmos10070373

[21] H. Yao, X. Li, H. Pang, L. Sheng, and W. Wang, "Application of random forest algorithm in hail forecasting over Shandong Peninsula," *Atmospheric Research*, vol. 244, p. 105093, 2020. https://doi.org/10.1016/j.atmosres.2020.105093

[22] A. J. Hill, G. R. Herman, and R. S. Schumacher, "Forecasting severe weather with random forests," *Monthly Weather Review*, vol. 148, no. 5, pp. 2135-2161, 2020. https://doi.org/10.1175/mwr-d-19-0344.1

[23] M. Ali, R. Prasad, Y. Xiang, and Z. M. Yaseen, "Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts," *Journal of Hydrology*, vol. 584, p. 124647, 2020. https://doi.org/10.1016/j.jhydrol.2020.124647

[24] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for Malaysian rainfall prediction," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 6, pp. 1148-1153, 2016. https://doi.org/10.18517/ijaseit.6.6.1487

[25] J. Mohan and A. Gupta, "Jaypee institute of information technology university, institute of electrical and electronics engineers uttar pradesh section. sp/cs joint chapter, and institute of electrical and electronics engineers," presented at the International Conference on Signal Processing and Communication (ICSC) : 07-09 March 2019, Jaypee Institute of Information Technology, NOIDA, 2019.

[26] Z. Zhang, Y. Li, L. Li, Z. Li, and S. Liu, "Multiple linear regression for high efficiency video intra coding," presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.

[27] E. Sreehari and S. Srivastava, "Prediction of climate variable using multiple linear regression," presented at the International Conference on Computing Communication and Automation, 2018.

[28] Luminto and Harlili, "Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate," presented at the International Conference on Advanced Informatics, Concepts, Theory, and Applications, 2017.

[29] N. Anusha, M. Sai Chaithanya, and G. Jithendranath Reddy, "Weather prediction using multi linear regression algorithm," presented at the In IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, 2019.

[30] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," presented at the In 2016 13th International Joint Conference on Computer Science and Software Engineering, 2016.

[31] I. Gupta, H. Mittal, D. Rikhari, and A. K. Singh, "MLRM: A multiple linear regression based model for average temperature prediction of a day," *arXiV*, 2022. https://doi.org/10.48550/arXiv.2203.05835

[32] Water in the Atmosphere, "The national meteorological library and archive," vol. 12. United Kingdom: National Meteorological Library, 2012.

[33] S. S. Maidin, B. Lau Simon, M. Othman, A. A. Latif, and M. F. M. Saad, "Data governance conceptual model for IoT (DGCMIoT) in flood management: The Malaysia perspective," *Journal of Pharmaceutical Negative Results*, pp. 271–276, 2022. https://doi.org/10.47750/pnr.2022.13.S10.028

[34] S. B. Memon, A. Rasli, A. S. Dahri, and I. Hermilinda Abas, "Importance of top management commitment to organizational citizenship behaviour towards the environment, green training and environmental performance in Pakistani industries," *Sustainability*, vol. 14, no. 17, p. 11059, 2022. https://doi.org/10.3390/su141711059