4.1 Dataset

In this study, a weather dataset comprising 22 columns, including temperature, rainfall, wind speed and direction, humidity, pressure, cloud cover, and rainfall indicators, was utilized. The dataset includes attributes such as `MinTemp`, `MaxTemp`,`Rainfall`,`Evaporation`,`Sunshine`, `WindGustDir`,`WindGustSpeed`,`WindDir9am`,`WindDir3pm`,`WindSpeed9am`,`WindSpeed3pm`,`Humidity9am`,`Humidity3pm`,pressure9am`,`Pressure3pm`,`Cloud9am`,`Cloud3pm`,`Temp9am`,`Temp3pm`,`RainToday`,and`RainTomorrow`. Preprocessing involved label encoding categorical variables, adding random noise to numerical features for data augmentation, and standardizing the data using StandardScaler. The dataset was then split into training and testing sets for predicting `RainToday` and `RainTomorrow`, with 80% of the data used for training and 20% for testing. This comprehensive approach aimed to enhance model robustness and accuracy in predicting daily rainfall.

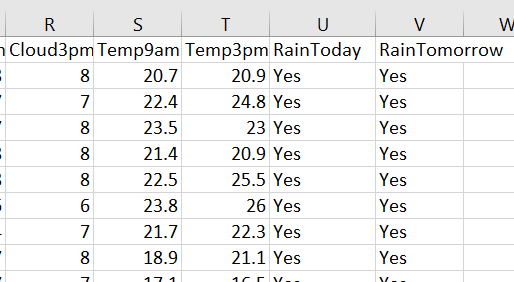


Figure 1. Some examples of the dataset

4.2 Implementation details

In this study, various machine learning models were applied to predict daily rainfall ("RainToday") and next-day rainfall ("RainTomorrow") using a comprehensive weather dataset. Initially, the dataset underwent preprocessing, which included label encoding of categorical features and data augmentation by adding random noise to numerical features. The data was then standardized using StandardScaler. For prediction of "RainToday," logistic regression, random forest classifier, and XGBoost models were trained and evaluated, achieving high accuracy and recall scores.

Furthermore, a hyperparameter tuning process was conducted using GridSearchCV for the RandomForestClassifier to identify the optimal parameters, resulting in the best parameters being `{'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}`. Additionally, an LSTM neural network was implemented for predicting "RainTomorrow." The data was normalized, reshaped, and split into training and testing sets, and an LSTM model with 50 units was trained for 100 epochs. The LSTM model achieved an accuracy of 97.2% on the test set. This comprehensive approach of using different machine learning models and hyperparameter tuning ensured robust and accurate predictions for rainfall.

This table summarizes the hyperparameter's optimal values for each machine learning model used in the study. Table 2 hyperparameter table

|  |  |  |
| --- | --- | --- |
| Model | Hyperparameter | Best value |
| Random Forest | n\_estimators | 100 |
|  | Min\_sample\_split | 2 |
|  | Min\_sample\_leaf | 1 |
| LSTM | Units | 50 |
|  | Epochs | 100 |
|  | Batch size | 1 |

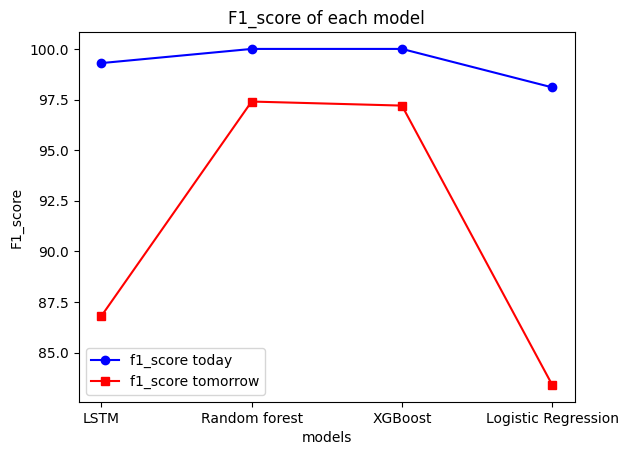
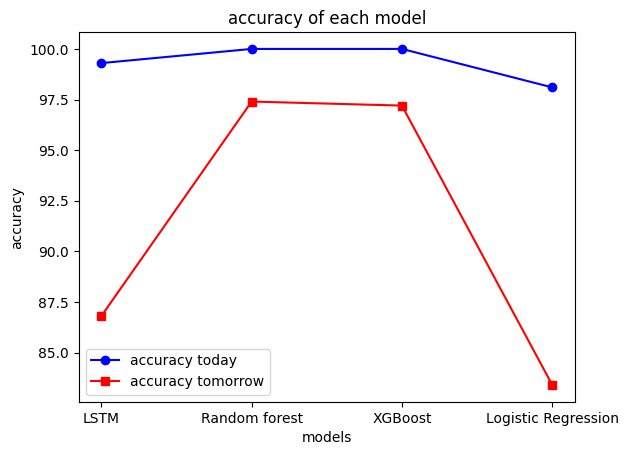
4.3 Results and Analysis

In this study, various machine learning models were applied to predict daily rainfall ("RainToday") and next-day rainfall ("RainTomorrow") using a comprehensive weather dataset. The models evaluated include Logistic Regression, Random Forest Classifier, XGBoost, and an LSTM neural network. The Random Forest and XGBoost models achieved perfect accuracy (100%) for "RainToday" and high accuracies for "RainTomorrow" at 97.4% and 97.2%, respectively, indicating their strong predictive capabilities but also suggesting potential overfitting. The LSTM neural network demonstrated high accuracy for "RainToday" at 99.3% but showed a notable drop for "RainTomorrow" at 86.8%, likely due to its difficulty in capturing longer-term dependencies. Logistic Regression, while simpler, achieved 98.1% accuracy for "RainToday" and 83.4% for "RainTomorrow," showing reasonable performance as a baseline model. Overall, the Random Forest model exhibited the best balance of high accuracy for both predictions, highlighting its robustness in handling the complexities of weather data. The results underscore that while simpler models provide good baseline performance, advanced machine learning techniques significantly enhance predictive accuracy in weather forecasting tasks.

See table 3 and chart 1 for proper visualization.

| **Model Name** | **Accuracy today(%)** | **Accuracy tomorrow(%)** |
| --- | --- | --- |
| Lstm | 99.3 | 86.8 |
| Random forest | 100 | 97.4 |
| XGBoost | 100 | 97.2 |
| Logistic regression | 98.1 | 83.4 |

Table 3 Accuracy of proposed mode



**CHART 1**. containing f1 score and accuracy for all models