Comparative Analysis of Machine Learning and Deep Learning Models for Rainfall Prediction

Syeda Tanjuma Tasnim
Id: 20200204093
Group: B2 (05)
Department of Computer Science and Engineering

Ekramul Huda Chowdhury
Id: 20200204103
Group: B2 (05)
Department of Computer Science and Engineering

Hasan Farabi
Id: 20200204106
Group: B2 (05)
Department of Computer Science and Engineering

***Index Terms*—Rainfall Prediction, Linear Regression, LSTM, Gaussian HMM, Random Forest, SVR, Machine Learning.**

I. Motivation

Rainfall prediction is very important for purposes of agriculture and water management for development and disaster mitigation. Substantial approaches used to predict the occurrence of weather are usually not effective in analyzing massive weather data. An example of using numerical information is when machine learning models allow higher accuracy thanks to statistically trained data.

II. Literature Review

The integration of data mining techniques with weather data analysis has led to significant advancements in various domains. This examines key studies that apply different data mining methods to weather data, focusing on regression algorithms, electric utility load prediction, traffic congestion analysis, and solar energy optimization.

J. Yadav discusses various regression algorithms for weather data analysis, highlighting the use of linear regression, CART, neural networks, and SVM for classification, prediction, and clustering. These methods effectively handle vast amounts of time series data, emphasizing the role of data mining in extracting valuable insights from large datasets [1].

G. E. Godfrey explores the relationship between summer weather and electric utility load, focusing on the impact of air conditioning. The study employs digital regression analysis to develop composite weather variables, enabling precise modeling of temperature-sensitive load variability influenced by both current and antecedent weather conditions [2].

J. Lee et al. investigate the impact of weather conditions on traffic congestion, using machine learning techniques like artificial neural networks and statistical models such as multiple linear regression. The study highlights the significant role of weather parameters like temperature, humidity, and precipitation on traffic flow and accident rates [3].

L. Sunitha et al. apply data mining techniques, particularly linear regression, to analyze weather data and uncover relationships between temperature, humidity, and dew point. They use clustering techniques to manage large datasets, enhancing the accuracy of regression models and allowing for precise weather predictions [4].

In summary, these studies highlight the pivotal role of advanced data mining techniques and preprocessing methods in extracting valuable insights from large datasets. They have wide-ranging applications in weather forecasting, electric utility load prediction, and traffic congestion analysis.

III. Methodology

The dataset used in this study is sourced from the Australian Bureau of Meteorology and contains weather observations from various locations. The following steps outline the methodology used to train and evaluate the models:

A. Data Preprocessing

The dataset was preprocessed by removing rows with missing values to ensure quality. The selected features for prediction were ‘MinTemp‘, ‘MaxTemp‘, ‘Humidity9am‘, ‘Humidity3pm‘, ‘Pressure9am‘, ‘Pressure3pm‘, ‘Temp9am‘, and ‘Temp3pm‘, with ‘Rainfall‘ as the target variable. An 80-20 train-test split was used to train the models on one portion and test them on a separate portion for performance evaluation.

B. Feature Standardization

Features were standardized using ‘StandardScaler‘ from scikit-learn to achieve zero mean and unit variance, enhancing machine learning model performance.

C. Model Implementation

Five different models were implemented for rainfall prediction:

- **Linear Regression:** A basic regression model that fits a linear relationship between the features and the target variable.
- **LSTM:** A neural network model designed for sequential data, capturing long-term dependencies. The model consists of an LSTM layer followed by a Dense layer.
- **HMM:** A probabilistic model used for modeling sequence data. The model was trained using the flattened sequences of weather features.

- **Random Forest:** An ensemble model that builds multiple decision trees and averages their predictions to improve accuracy and prevent overfitting.
- **SVR:** A regression model that uses a Support Vector Machine with a radial basis function kernel to capture non-linear relationships.

D. Model Training and Evaluation

The models were trained on the training dataset and evaluated on the testing dataset using the following error metrics:

- **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted values.
- **Mean Absolute Error (MAE):** Measures the average absolute difference between actual and predicted values.
- **Root Mean Squared Error (RMSE):** Measures the square root of the average squared difference between actual and predicted values.

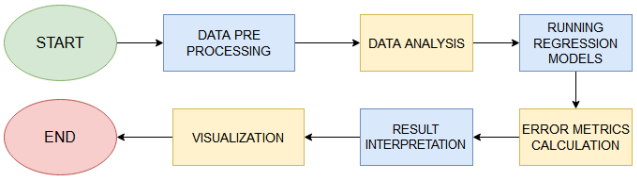

Fig. 1. Working Methodology

IV. Result Analysis

The dataset is from the Australian Bureau of Meteorology, preprocessed by removing missing values and standardizing features. Implemented models include:

1) Linear Regression: The Linear Regression model achieved an MSE of 39.68, MAE of 3.12, and RMSE of 6.30.

2) LSTM: The LSTM model showed an MSE of 52.83, MAE of 3.37, and RMSE of 7.27.

3) Gaussian HMM: The Gaussian HMM model had an MSE of 47.45, MAE of 2.64, and RMSE of 6.89.

4) Random Forest: The Random Forest model resulted in an MSE of 34.68, MAE of 2.46, and RMSE of 5.89.

5) SVR: The SVR model produced an MSE of 37.69, MAE of 1.89, and RMSE of 6.14.

TABLE I
Comparison of Various Error Metrics

Model	MSE	MAE	RMSE
Linear Regression	39.68	3.12	6.30
LSTM	52.83	3.37	7.27
Gaussian HMM	47.45	2.64	6.89
Random Forest	34.68	2.46	5.89
SVR	37.69	1.89	6.14

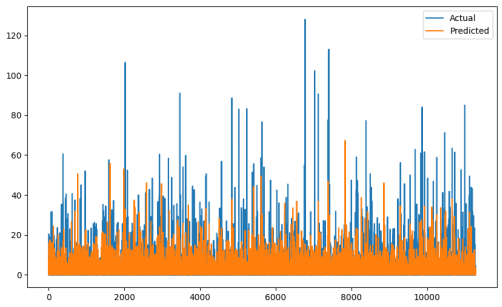

Fig. 2. Actual VS Predicted Values in RF

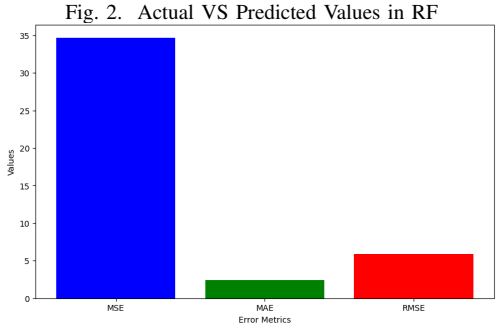

Fig. 3. Comparison of Various Error Metrics

V. Conclusion

We conducted a comparative analysis of various machine learning models for rainfall prediction, including Linear Regression, LSTM, HMM, Random Forest, and SVR. The Random Forest model outperformed others in terms of MSE and RMSE, effectively capturing complex, non-linear weather patterns, while the SVR model exhibited the lowest MAE, demonstrating robustness in regression tasks. These results suggest ensemble methods like Random Forest are well-suited for rainfall prediction, and LSTM models hold potential in time series forecasting. Future work could enhance prediction accuracy by integrating additional meteorological features, experimenting with hybrid models, and applying advanced deep learning techniques, as well as incorporating geographical information and climate indices to improve predictive capabilities.

References

[1] J. Yadav, "Analysis of Weather Data Using Various Regression Algorithms," International Journal of Advanced Research in Computer Science, vol. 9, no. 2, pp. 118-125, 2018.

[2] G. E. Godfrey, "The Relationship of Summer Weather and Electric Utility Load," IEEE Transactions on Power Apparatus and Systems, vol. PAS-85, no. 6, pp. 586-593, June 1966, doi: 10.1109/TPAS.1966.292301.

[3] J. Lee, B. Hong, K. Lee, and Y.-J. Jang, "A Prediction Model of Traffic Congestion Using Weather Data," in 2015 IEEE International Conference on Data Science and Data Intensive Systems, 2015, pp. 81-82.

[4] L. Sunitha, M. Balraju, J. Sasikiran, and B. Anil Kumar, "Finding Relation Between Parameters of Weather Data Using Linear Regression Method," International Journal of Research in Engineering and Technology, vol. 05, Special Issue: 05, pp. 90-93, May 2016.