Weather Forecasting in Bangladesh: A Machine Learning Approach

Niaz Nafi Rahman
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
niaz.nafi.rahman@g.bracu.ac.bd

Zaid Rehman
Computer Science
Brac University
Dhaka, Bangladesh
zaid.rehman@g.bracu.ac.bd

Adib Reza
Computer Science
Brac University
Dhaka, Bangladesh
adib.reza@g.bracu.ac.bd

***Abstract*— Weather forecasting plays a crucial role in safeguarding lives and property. In recent years, machine learning (ML) techniques have emerged as powerful tools for weather prediction, offering advantages over traditional linear approaches. This study explores the application of ML algorithms to predict rainfall occurrence, rainfall amount, and average temperature in Bangladesh. Extensive weather data was collected from multiple regions, preprocessed, and analyzed for patterns. Various supervised ML models, including SVM, KNN, Decision Trees, Random Forest, and ensemble techniques, were employed and evaluated on their accuracy in predicting chosen weather events. Results demonstrate the potential of ML in weather forecasting, particularly using Random Forest and ensemble methods. However, challenges such as class imbalance and the need for model interpretability require further investigation to optimize the use of ML in this field.**

Keywords*— *Weather Forecasting, Rainfall prediction, Average temperature prediction, Machine Learning.

I. Introduction

Weather forecasting, essential for safeguarding life and property, has evolved from manual methods to computer-based models considering numerous atmospheric factors. Traditional linear approaches have given way to nonlinear prediction methods due to the recognition of nonlinearities within weather data attributes. Machine learning (ML) has emerged as a potent tool in this field, capable of unveiling complex patterns within vast datasets. ML's ability to learn from diverse data sources, including satellite imagery and weather radar, enhances the accuracy and timeliness of forecasts across various sectors.

This paper delves into the application of ML techniques in weather forecasting, focusing on predicting rainfall occurrence, amount, and average temperature in several regions of Bangladesh. By training various ML models such as SVM, KNN, decision trees, and random forests on extensive weather data, we aim to explore their efficacy in enhancing weather predictions.

II. Literature review

Singh et al. [5] addressed the need for affordable weather forecasting with a low-cost, sensor-driven system powered by machine learning. Their random forest model, trained on a 20-year Delhi weather dataset, achieved 87.90% accuracy in rain prediction, highlighting humidity as the most significant factor. The system's successful implementation on a Raspberry Pi 3 B board underscores its feasibility. However, the study's focus on rain prediction and reliance on Delhi-specific data may necessitate adjustments for broader weather conditions and other regions.

Rahman et al. [6] developed a machine learning fusion framework for rainfall prediction in Lahore, achieving 94% accuracy with a low miss rate. Their approach combined fuzzy logic with classifiers like Decision Trees and Support Vector Machines, outperforming individual models. While highlighting the potential of ML for weather prediction, the authors acknowledge the importance of data integrity and security measures for real-world implementation.

In the paper [7] (2019), Singh et al. investigated the comparative performance of Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Time Series Recurrent Neural Networks (RNN) for weather forecasting. Their study utilized a dataset of weather conditions collected from airport stations across multiple cities, including parameters like temperature, pressure, humidity, and precipitation. While the dataset's exact size, timeframe, and source remain unspecified, the models were applied for prediction. Key findings indicate that Time Series RNN models achieved superior accuracy compared to SVM and ANN. Additionally, the study highlights the importance of prediction window size, as larger windows correlated with increased errors. The Root Mean Squared Error (RMS error) for each model was reported as follows: SVM: 6.67, ANN: 3.1, Time Series RNN: 1.41.

Liyew and Melese [8] investigated key atmospheric factors influencing rainfall and used machine learning to predict daily rainfall in Ethiopia. Their XGBoost model, trained on 20 years of meteorological data, outperformed MLR and RF. While successful, the study would benefit from correlation analysis between atmospheric attributes and a broader exploration of factors affecting rainfall patterns to enhance model interpretability.

Shafin in [9] (2019) investigated the use of machine learning algorithms to predict average temperatures in Bangladesh. Their study utilized a dataset spanning 1901-2018, containing monthly temperature values, allowing for both yearly and seasonal analysis. The dataset included features such as yearly average, summer, rainy season, and winter season average temperatures. The authors employed Linear Regression, Polynomial Regression, Isotonic Regression, and Support Vector Regressor (SVR) to model the data and forecast future trends. Interestingly, while Isotonic Regression excelled at fitting the training data, Polynomial Regression and SVR yielded more accurate predictions of future temperatures. The study sought to uncover patterns in Bangladeshi temperature fluctuations. Despite limitations stemming from potential dataset quality issues and the inherent assumptions within machine learning models, this research demonstrated the potential of machine learning techniques for understanding and predicting average temperatures in Bangladesh.

Suzuki et al. [10] explored the use of Convolutional Neural Networks (CNN) for rainfall forecasting in Japan. Their model achieved detection ratios between 64% and 76% for 30-minute lead times, demonstrating CNN's ability to learn from spatiotemporal data. However, a high false alarm ratio suggests a need for refinement to improve precision. Further research in diverse environments is needed to determine the model's broader adaptability and potential for operational forecasting.

III. Methodology

A. Data Collection

To analyze the performance of different machine learning techniques, we have collected weather data [3] from different locations across Bangladesh. This dataset includes several attributes essential to weather analysis: daily minimum and maximum temperatures (°C), rainfall (mm), evaporation (mm), sunshine hours, wind gust direction and speed (kmph), wind direction at 9 am and 3pm, humidity at 9 am and 3pm (%), atmospheric pressure at 9 am and 3pm (hPa), cloud cover at 9 am and 3pm, temperatures at 9 am and 3 pm (°C), and whether it rained or not on a given day. This dataset contains 10 years of daily observations from 2013 to 2022. While data for some days may be missing, the substantial size of the dataset helps mitigate this potential limitation.

Feature	**Attribute**	**Description**
Date	Categorical	Date of the observation (DD-MM-YYYY format)
MinTemp	Continuous	Minimum temperature (°C)
MaxTemp	Continuous	Maximum temperature (°C)
Rainfall	Continuous	Amount of rainfall (mm)
Evaporation	Continuous	Amount of evaporation (mm)
Sunshine	Continuous	Hours of bright sunshine
WindGustDir	Categorical	Direction of strongest wind gust (compass points)
WindGustSpeed	Continuous	Speed of strongest wind gust (kmph)
WindDir9am	Categorical	Wind direction at 9 am (compass points)
WindDir3pm	Categorical	Wind direction at 3 pm (compass points)
WindSpeed9am	Continuous	Wind speed at 9 am (kmph)
WindSpeed3pm	Continuous	Wind speed at 3 pm (kmph)
Humidity9am	Continuous	Humidity at 9 am (%)
Humidity3pm	Continuous	Humidity at 3 pm (%)
Pressure9am	Continuous	Atmospheric pressure at 9 am (hPa)
Pressure3pm	Continuous	Atmospheric pressure at 3 pm (hPa)
Cloud9am	Continuous	Cloud cover at 9 am (eighths)
Cloud3pm	Continuous	Cloud cover at 3 pm (eighths)
Temp9am	Continuous	Temperature at 9 am (°C)
Temp3pm	Continuous	Temperature at 3 pm (°C)
RainToday	Categorical	Indicates if it rained on that day (Yes/No)

B. Data Analysis

To understand the characteristics and relationships within the weather dataset, Exploratory Data Analysis (EDA) was performed. This analysis revealed a tendency for higher wind speeds in the afternoon (3 pm), particularly during August and September, while May and June exhibited the highest average wind speeds at 9 am. Humidity levels were consistently higher at 9 am compared to 3 pm, with peak morning humidity observed between January and February and peak afternoon humidity in September and October. Atmospheric pressure readings at 9 am consistently exceeded those at 3 pm, with indications of peak pressure occurring in March and April. Temperature patterns showed expected fluctuations throughout the year, and an initial correlation matrix highlighted potential relationships between various weather features. These EDA findings will guide further analysis, inform subsequent modeling choices, and provide a basis for investigating the dataset's suitability for predicting weather patterns, rainfall, and temperature.

C. Data Preprocessing

There were multiple preprocessing steps used to get the dataset ready for modeling. First, missing values were checked and none were found. Standardizing the features were the next step. This prevented features with larger ranges from controlling distance calculations by ensuring that variables with different scales would have comparable influence during model training. Using one-hot encoding, categorical variables like "WindGustDir," "WindDir9am," and "WindDir3pm" were changed. As a result, each category received new binary features that improved compatibility with machine learning algorithms. A numerical binary format (0, 1) was used to replace the categorical representation ('No', 'Yes') of the target variable 'RainToday'. Two methods were used to address the class imbalance in the target variable "RainToday": undersampling and oversampling. The classification models were then trained using the resampled datasets. The removal of unhelpful elements like "Date" was the last step. Furthermore, in order to produce new derived features that might improve predictive performance, possible feature engineering techniques were investigated.

D. Model Selection and Training

In our weather forecasting research, we employed a diverse range of machine learning models to tackle various prediction tasks. For rainfall occurrence prediction, we utilized Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machine (SVM) with a Linear Kernel, and Random Forest. These models were chosen based on their individual strengths, including interpretability, capacity to capture nonlinear relationships, effectiveness with mixed data types, and robustness in handling high-dimensional datasets. For rainfall amount and average temperature prediction, we employed Linear Regression, Random Forest Regression, and Support Vector Regression (SVR). Each model was selected for its ability to capture different aspects of the complex relationships within weather data.

To maximize prediction performance, we utilized ensemble techniques. The VotingClassifier combined the outputs of individual classification models, while the VotingRegressor combined predictions from multiple regression models. These ensemble methods aimed to improve the accuracy and robustness of our forecasts for rainfall occurrence, amount, and average temperature. By leveraging this diverse set of machine learning techniques, we aimed to develop comprehensive forecasting models capable of addressing the complexities of weather prediction tasks.

IV. Results

A. Rainfall Occurrence Prediction

In our investigation of rainfall occurrence prediction, we employed a range of classification algorithms: Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Random Forest. Initially trained on the original dataset, these models yielded the following performance metrics:

- Logistic Regression achieved an accuracy score of 83.3% and an F1 score of 63.2%.
- KNN demonstrated an accuracy score of 83.7% and an F1 score of 58.5%.
- The Decision Tree classifier exhibited high accuracy, scoring 92.7%, with an F1-score of 86.2%.
- SVM achieved an accuracy score of 83.8% and an F1-score of 63.5%.
- Random Forest outperformed other models with an accuracy score of 94.4% and an F1-score of 88.7%.

However, the presence of class imbalance posed challenges, particularly in minority class prediction. To mitigate this issue, we applied oversampling techniques and re-evaluated the algorithms' performance. While some models demonstrated improvement, others experienced a decrease in performance. For instance, KNN's accuracy dropped to 80.6%, with an F1-score of 50.4%, on the oversampled dataset.

B. Rainfall Amount Prediction

For rainfall amount prediction, we utilized regression algorithms: Linear Regression, Random Forest Regression, and Support Vector Machine (SVM) Regression. The performance of each model was assessed using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics:

- Linear Regression achieved a MAE of 0.499 and an RMSE of 0.948.
- Random Forest Regression demonstrated superior performance with a MAE of 0.381 and an RMSE of 0.887.
- SVM Regression yielded a MAE of 0.365 and an RMSE of 0.972.
- An ensemble regression approach further improved prediction accuracy, achieving a MAE of 0.366 and an RMSE of 0.904.

C. Average Daily Temperature Prediction

In our analysis of average daily temperature prediction, regression models were trained using Linear Regression, Random Forest Regression, and Support Vector Machine (SVM) Regression algorithms. The performance of each model was evaluated based on MAE and RMSE metrics:

- Linear Regression: MAE = 0.471, RMSE = 0.603
- Random Forest Regression: MAE = 0.448, RMSE = 0.569
- SVM Regression: MAE = 0.435, RMSE = 0.560
- Ensemble Regression: MAE = 0.424, RMSE = 0.545

These results highlight the effectiveness of ensemble techniques in capturing complex relationships in temperature data and improving prediction accuracy.

Conclusion

In summary, this study delved into the application of machine learning (ML) in weather forecasting, specifically focusing on predicting rainfall occurrence, amount, and average temperature across Bangladesh. Extensive data collection and preprocessing paved the way for model selection, where a range of classification and regression algorithms were evaluated. Results showcased promising performance, with random forest being the top performer in rainfall occurrence prediction, achieving 94.4% accuracy. Random forest regression proved effective for rainfall amount prediction, while ensemble regression techniques enhanced average daily temperature prediction.

Overall, this study highlights ML's potential to bolster weather forecasting accuracy, offering valuable insights for decision-making across various sectors. However, challenges like class imbalance and model interpretability warrant further exploration. By refining ML approaches, we can advance resilience and preparedness in response to evolving weather patterns.

References

[1] Taylor, K. E., & Leslie, L. M. (2012). The development and early use of ensemble prediction systems. ECMWF Newsletter, 130, 6-12. *(references)*

[2] Sachindra, D. A., Ahmed, K., Rashid, M. M., Shahid, S., & Perera, B. J. C. (2018). Statistical downscaling of precipitation using machine learning techniques. Atmospheric Research, 212, 240-258.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Haupt, S. E., Cowie, J., Linden, S., McCandless, T., Kosovic, B., & Alessandrini, S. (2021). Machine learning for applied weather prediction. IEEE 18th International Conference on Smart Cities, 290-295.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. Nature, 525(7567), 47-55. https://doi.org/10.1038/nature14956

[5] N. Singh, S. Chaturvedi and S. Akhter, "Weather Forecasting Using Machine Learning Algorithm," 2019 International Conference on Signal Processing and Communication (ICSC), NOIDA, India, 2019, pp. 171-174, doi: 10.1109/ICSC45622.2019.8938211.

[6] Rahman, A., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., Khan, M.A., & Mosavi, A.H. (2022). Rainfall Prediction System Using Machine Learning Fusion for Smart Cities. Sensors (Basel, Switzerland), 22.

[7] Singh, Siddharth and Kaushik, Mayank and Gupta, Ambuj and Malviya, Anil Kumar, Weather Forecasting Using Machine Learning Techniques (March 11, 2019). Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019, Available at SSRN: https://ssrn.com/abstract=3350281 or http://dx.doi.org/10.2139/ssrn.3350281

[8] Liyew, C.M., Melese, H.A. Machine learning techniques to predict daily rainfall amount. *J Big Data* **8**, 153 (2021). https://doi.org/10.1186/s40537-021-00545-4

[9] Shafin, Ashfaq. (2019). Machine Learning Approach to Forecast Average Weather Temperature of Bangladesh Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence. 19. 39-48. 10.17406/GJCST

[10] Kim, S., Suzuki, T., & Tachikawa, Y. (2020). RAINFALL OCCURRENCE PREDICTION WITH CONVOLUTIONAL NEURAL NETWORK. *Journal of Japan Society of Civil Engineers, Ser. B1 (Hydraulic Engineering)*.