International Journal of Applied Engineering & Technology

A MACHINE LEARNING-BASED APPROACH TO CAPTURE EXTREME RAINFALL EVENTS

Willy Mbenza[1] and Kenjiro Sho[2]
[1, 2] Nagoya Institute of Technology, Nagoya, Japan
[1] w.mbenza.608@stn.nitech.ac.jp and [2] show@nitech.ac.jp

ABSTRACT

Increasing efforts are directed towards a better understanding and foreknowledge of extreme precipitation probability, given the adverse effects associated with their occurrence. This knowledge plays a crucial role in long-term planning and the formulation of effective emergency response. However, predicting extreme events reliably presents a challenge to conventional empirical/statistical methods due to the involvement of numerous variables spanning different time and space scales.

In the recent time, Machine Learning (ML) has emerged as a promising tool for predicting the dynamics of extreme precipitations. ML techniques enables the consideration of both local and regional physical variables that have a strong influence on the likelihood of extreme precipitations. These variables encompass factors such as air temperature, wind speed, specific humidity, atmospheric pressure, among others. In this study, we develop an ML model that incorporates both local and regional variables, while establishing a robust relationship between physical variables (features) and precipitation during the downscaling process. For capturing extreme rainfall events, the model adeptly handles both regression and classification tasks, delivering commendable performance in each. In the regression task, it overall achieved good results with a MAE and RMSE of 1.85 mm and 4.81 mm, respectively. For the classification task, the model demonstrated an impressive accuracy rate of 92%.

Keywords: Features, Machine Learning (ML), Prediction, Rainfall, XGBoost.

INTRODUCTION

Rainfall forecasting stands as a crucial subject, pivotal for comprehending the meteorological processes and making dependable predictions. Despite considerable research efforts aimed at enhancing our understanding of rainfall dynamics, there remains a continual pursuit for refining forecasting accuracy. Although contemporary weather forecasting services are generally deemed reliable in both qualitative and quantitative predictions of rainfall probability, there is still room for improvement. Consequently, this topic continues to attract attention, focusing on the development of comprehensive methodologies to minimize forecasting errors.

Rainfall plays a central role in the hydrological cycle, serving as an indispensable component. Its consistent and measured presence nourishes rivers, recharges groundwater tables, and fundamentally shapes the entire hydrological system. Moreover, rainfall plays an essential role in the natural ecosystem, serving as a vital climatic variable essential for the survival of humans, animals, and plants. Its influence extends deeply into multiple sectors, directly impacting crucial elements like irrigation, hydropower generation, domestic water supply, and industrial processes [12]. Recognized as a quintessential factor, rainfall stands at the core of human development, shaping the course of societies and their sustainable growth [1].

Accurate rainfall predictions contribute significantly to the prevention of natural disasters such as floods, landslides, mass movements, and avalanches. Notably, in 2019, floods emerged as the most prevalent form of natural disasters, coupled with typhoons, causing the highest impact in terms of human casualties and economic losses, according to the Emergency Events Database [2]. Given that floods are often triggered by heavy rainfall, precise prediction becomes imperative for safeguarding lives and properties. Conversely, rainfall scarcity leads to droughts, adversely affecting agricultural activities and exerting a substantial impact on the economy. This underscores the importance of reliable predictions for efficient management of available water resources.

In the realm of research, rainfall prediction remains a perpetual concern, particularly in drought-prone countries where the efficient utilization of rainfall water is a pressing research area. Traditional models, ranging from deterministic to stochastic approaches, have struggled to accurately predict rainfall. Empirical models, driven by historical rainfall data, address some limitations but are most reliable in localized regions [13]. The advent of artificial intelligence and machine learning in the twenty-first century has emerged as an efficient tool for estimating and predicting rainfall with notable accuracy. Machine learning techniques surpass traditional deterministic methods, offering enhanced capabilities [11]. Scholars have employed machine learning to identify atmospheric features influencing rainfall, predict intensity, and estimate threshold values to prevent floods [8, 9]. This transformative approach enables both short-term and long-term predictions, providing crucial early warnings for floods, cloud bursts, and landslides in disaster-prone areas [1].

STUDY AREA AND DATA

A. Study Area

Nagoya, one of the major cities in Japan, is situated in the Chubu region's central expanse, has a substantial land area spanning approximately 326.45 km². It experiences a humid temperate climate characterized by mild temperatures and consistent moisture throughout the year, eliminating the presence of a distinct dry season. The city encounters hot and muggy summers accompanied by frequent thunderstorms [3]. Geographically, Nagoya can be precisely located at coordinates 35°10'N, 136°57'E, with an elevation of 56 meters.

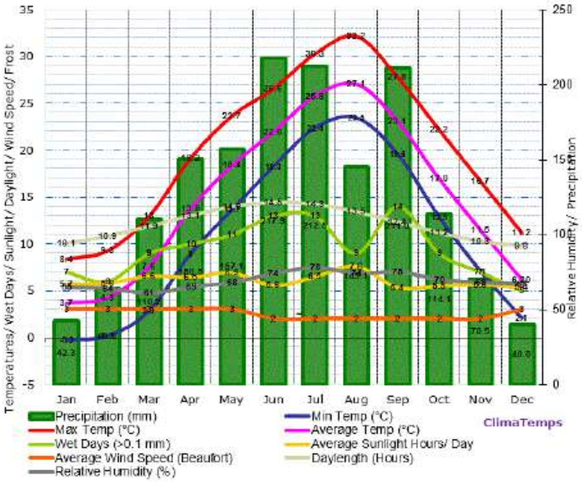

Fig. 1 Nagoya Climate Graph [3]

The climatic dynamics in Nagoya exhibit a distinct seasonal pattern. The monsoon season prevails from June to July, contributing to increased rainfall during June. The typhoon season follows in August to September, resulting in heightened precipitation during September (Fig. 1). Conversely, December emerges as the driest month of the year. The annual average rainfall for Nagoya is recorded at approximately 1573 mm, underlining the city's climatic variability shaped by monsoons and typhoons.

B. Data

The Japan Meteorological Agency (JMA) served as the exclusive source of meteorological data for the experimental framework [4]. Collaboratively, Nagoya City and JMA oversee an array of rain gauges deployed strategically for the systematic collection of precipitation-related information. The dataset under consideration spans an extensive temporal scope, encompassing 52 years of observations (daily data from 1972 to 2023). This dataset manifests as a comprehensive table, comprising 18,899 rows and 18 columns.

The meteorological variables encapsulated in the raw dataset include Mean Cloud Cover, Snowfall, Rainfall, Humidity, Atmospheric Pressure, Mean Sea Level Pressure, Vapor Pressure, Sunlight Hours, Solar Radiation, Temperature, and Wind. Notably, variables such as Temperature, Wind, and Pressure exhibit granularity, being further stratified into maximum, minimum, and mean values, or in some instances, solely maximum and minimum values.

To render the dataset amenable to our experimental objectives, preliminary data processing was imperative. This involved not only categorizing features but also restructuring the dataset into a coherent and interpretable format. Specifically, the original data obtained from JMA's website were organized annually, transitioning from one variable to the next. Consequently, meticulous organization and restructuring were undertaken to align the dataset with the requisites of our experiment.

Table I furnishes an initial snapshot of the dataset before further data processing, providing a foundational overview for subsequent analysis.

METHODOLOGY

A. Data pre-processing

Pre-processing of raw data was essential to facilitate its utilization. This involved a conversion into the CSV format, enabling compatibility with the Pandas library in Python. Addressing missing values was a pivotal step, considering the likelihood of inaccuracies and omissions in the raw data. Fortunately, the dataset exhibited a minimal number of missing values: 250 out of 18,899 (1.32%) for Solar Radiation, 5 out of 18,899 (0.03%) for Sunlight Hours, 1 out of 18,899 (0.005%) for Average Temperature, 1 out of 18,899 (0.005%) for Maximum Temperature, 5 out of 18,899 (0.03%) for Minimum Wind Speed, and 3 (0.02%) for Maximum Instantaneous Wind Speed. Imputation was employed to address these missing values. It consisted of replacing the missing value by the mean value [5].

B. Independent Features

The subsequent phase involved the identification of pertinent features for rainfall predictions. This encompassed a correlation analysis to discern highly correlated features. Features that merely contribute to increased dimensionality without enhancing model performance were subsequently removed [5]. Fig. 2 shows the highly correlated features, following which Maximum Temperature, Minimum Temperature, Minimum Humidity, and Minimum Wind Speed were excluded. Snowfall was also eliminated as it constitutes another form of precipitation without influencing rainfall likelihood.

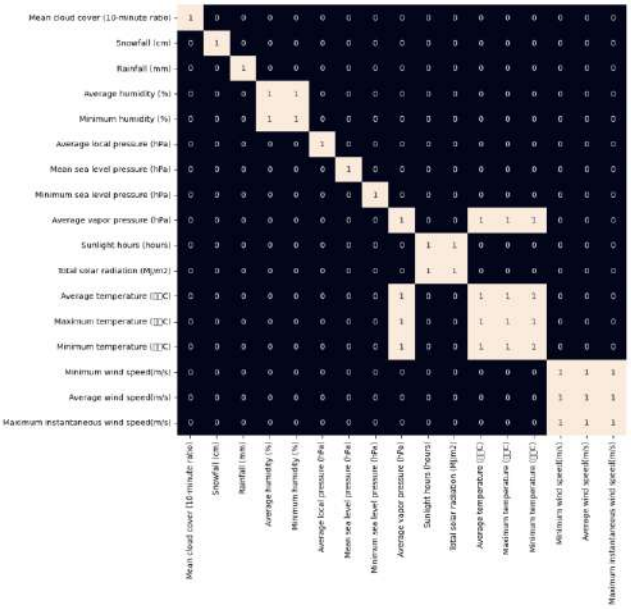

Fig. 2 Assessment of highly correlated features

Table I. RAW Daily Data Obtained from JMA

Feature	count	mean	std	min	25%	50%	75%	max
Mean cloud cover	18899	6.36	3.11	0	4	7	9.5	10
Snowfall (cm)	18899	0.03	0.55	0	0	0	0	23
Rainfall (mm)	18899	4.32	12.26	0	0	0	1.5	428
Average humidity (%)	18899	66.48	12.92	24	58	66	75	100
Minimum humidity (%)	18899	45.18	15.51	5	34	43	54	99
Average local pressure (hPa)	18899	1008.02	6.39	980.6	1003.5	1007.9	1012.6	1027.8
Mean sea level pressure (hPa)	18899	1014.50	14.10	20.5	1010	1014.5	1019.4	1034.8
Minimum sea level pressure (hPa)	18899	456.40	503.40	0	0	0	1010.7	1032.2
Average vapor pressure (hPa)	18899	13.79	8.00	2.3	6.5	11.9	20.5	33.1
Sunlight hours (hours)	18894	5.89	4.02	0	1.9	6.5	9.3	16.93
Total solar radiation (MJ/m2)	18649	13.71	6.93	0	8.5	13.1	19	31.17
Average temperature (°C)	18898	15.94	8.48	-2.9	8.1	16.3	23.2	33.3
Maximum temperature (°C)	18898	20.77	8.69	-0.2	13.1	21.2	28	40.3
Minimum temperature (°C)	18899	11.98	8.74	-6.2	3.9	12.1	19.9	28.8
Minimum wind speed(m/s)	18894	2.95	1.18	0.4	2.1	2.7	3.6	10.2
Average wind speed(m/s)	18899	6.22	2.10	1.5	4.7	5.9	7.5	26.3
Maximum instantaneous wind speed(m/s)	18896	10.51	3.86	3	7.7	9.9	12.8	42.6

C. Rainfall Threshold

To capture extreme events, a rainfall threshold was defined to categorize the type of rainfall. Following the World Meteorological Organization's guidelines, rainfall is deemed heavy if its depth exceeds 50 mm within the past 24 hours [6]. For our experiment, rainfall was categorized into three events: heavy rain if the depth exceeded 50 mm, rain if the depth ranged between 50 mm and 0, and no rain if the rainfall's cell indicated 0. Fig. 3 illustrates different types of rainfall and their frequency in the dataset. Notably, over 52 years of observation, a mere 1.4% of rainfall events were classified as heavy.

D. Feature Engineering

Feature engineering stands as a crucial step designed to elevate the model's performance, emphasizing that superior feature engineering correlates with enhanced machine learning model efficacy: better features make better models [7]. In our experiment, we addressed various feature categories, namely temporal, categorical, weather-related, and cumulative rainfall features.

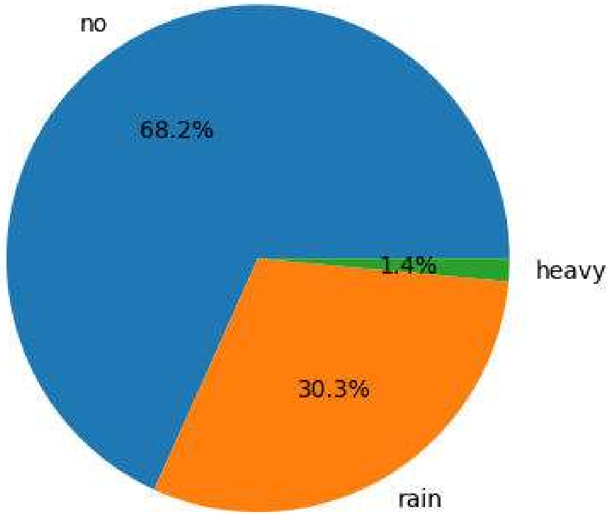

Fig. 3 Rainfall category

D.1. Temporal Features:

To capture temporal patterns, we incorporated lag features, spanning from day one to day seven (a week), acknowledging the significance of long-term prediction. Additionally, statistical features, including rolling mean and rolling standard deviation, were created to mitigate noise in the data and emphasize trends.

D.2. Categorical Features:

Incorporating 'day of the week,' 'month,' and 'season' as categorical features enhances our model's understanding by extracting valuable information from the timestamp. Noteworthy variations in rainfall patterns based on these temporal factors are considered. However, the inclusion of public holidays was deemed irrelevant for the context of this experiment and, thus, omitted.

D.3. Weather-Related Features:

Handling weather-related features such as Temperature, Humidity, and Wind Speed enables the model to account for environmental conditions influencing rainfall. Although the initial dataset cleaning effectively managed all missing values, this step ensures uniformity by bringing weather-related features to a comparable scale through normalization or standardization, mitigating potential model sensitivity.

D.4. Cumulative Rainfall:

Finally, accounting for cumulative rainfall up to day seven plays a pivotal role. Introducing a feature related to previous rainy days proves valuable in capturing the cumulative impact of rainfall over time. This involves calculating the cumulative sum of rainfall over the preceding seven days.

E. Data Splitting

Before delving into the training, predicting, and capturing of extreme rainfall, a crucial initial step involved data splitting. This facilitated the model's learning and prediction based on previously unseen data—a fundamental practice in building and evaluating machine learning models. The systematic partitioning of the dataset into subsets dedicated to training (80%) and validation (20%) proved essential to prevent overfitting issues [8].

F. Model Selection

While conventional practice involves training multiple models and selecting the best-performing one, empirical evidence indicates that XGBoost consistently outperforms other machine learning models in various scenarios [8], [9]. Consequently, for this study, we opted for XGBoost—a decision tree ensembles algorithm widely acknowledged for its efficiency, default splitting criteria, and regularization capabilities.

XGBoost's mathematical intricacies involve constructing a regularized objective function, as illustrated below:

$$\mathcal{L}(\phi) = \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{1}$$

Where ℓ represents a differentiable convex loss function, measuring differences between predictions $\hat{y}_i$ and target y_i, and f_k is an additive function used for predictions. The regularization term Ω penalizes model complexity [10].

For this study, we implemented a hybrid approach, addressing both regression and classification tasks. The regression task predicts actual rainfall amounts, while the classification task uses predefined thresholds to categorize predicted amounts into heavy rain (extreme rainfall events), rain, and no rain.

G. Model performance

G.1. Evaluation of Regression Task

After implementing XGBoost as a regressor, we assessed model performance using statistical metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

- Mean Absolute Error (MAE): average absolute difference between the observed actual values and the predictions. A lower MAE indicates better performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

- Mean Squared Error (MSE): average of the squared differences between the observed actual values and the predictions. A lower MSE indicates better performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

- Root Mean Squared Error (RMSE): the square root of the MSE. Like MAE and MSE, a lower RMSE indicates better performance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

- R-squared (R^2): proportion of the variance in the target variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{SS_{tot}}{SS_{res}} \tag{5}$$

With $SS_{tot} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $SS_{res} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

G.2. Evaluation of Classification Task

The evaluation of the classification task involves assessing accuracy, precision, recall, F-1 score, and support.

- Accuracy: Proportion of correctly predicted examples out of the total examples.
- Precision: Ratio of correctly predicted positive observations to the total predicted positives.
- Recall (Sensitivity): Ratio of correctly predicted positive observations to all observations in the actual class.
- F-1 Score: Weighted average of precision and recall, ranging between 0 to 1.
- Support: Number of actual occurrences of the class in the dataset.

RESULTS

A. Regression Task

Table II provides a comprehensive summary of all statistical metrics obtained after training and validating the model. The collective evidence from these metrics affirms the commendable performance of the XGBoost regression model.

Table II. Metrics' Summary

Metrics	Score
MAE (mm)	1.85
MSE (mm²)	23.14
RMSE (mm)	4.81
R-squared	0.84

Upon careful examination of Table II, the statistical metrics collectively advocate for the efficacy of the XGBoost regression model.

The lower values for Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) strongly suggest that the model's predictions closely align with the actual values. For instance, a MAE of 1.85 means that, on average, the model's predictions deviate by approximately 1.85 mm from the rainfall actual values. Notably, the higher R² value (0.83) underscores that a substantial proportion of the variability in the target variable is successfully captured by the model.

To provide a visual representation of the regression model's capabilities, Fig. 6 encapsulates a scatter plot illustrating the alignment between the model's predictions and the actual values. This graphical representation offers an intuitive insight into the model's precision.

Moreover, Fig. 4 intricately delineates the importance of various features in predicting the amount of rainfall. This comprehensive overview serves to elucidate the intricate relationships and contributions of different variables in the model's forecasting capabilities.

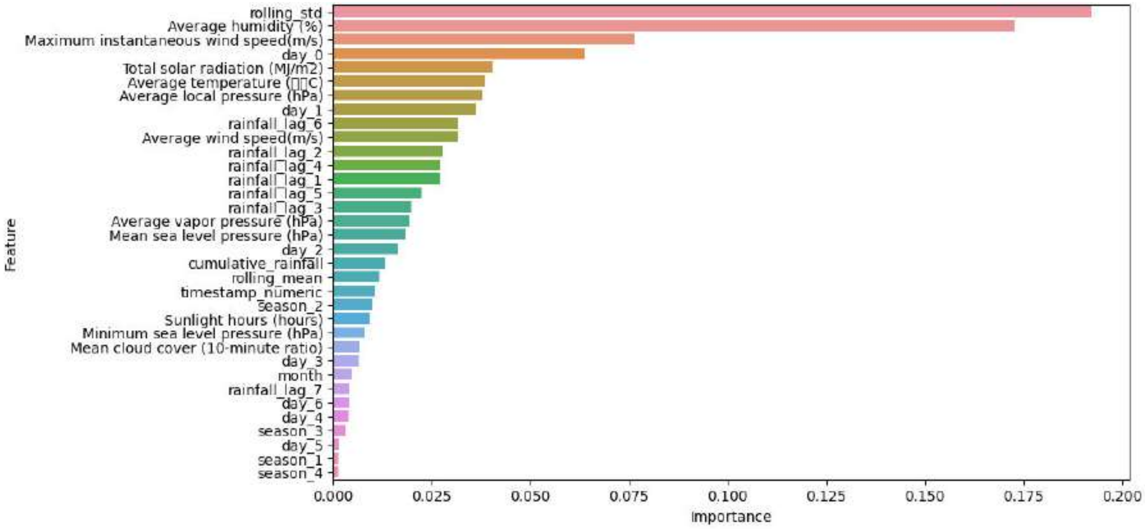

Fig. 4 Feature importance for regression task

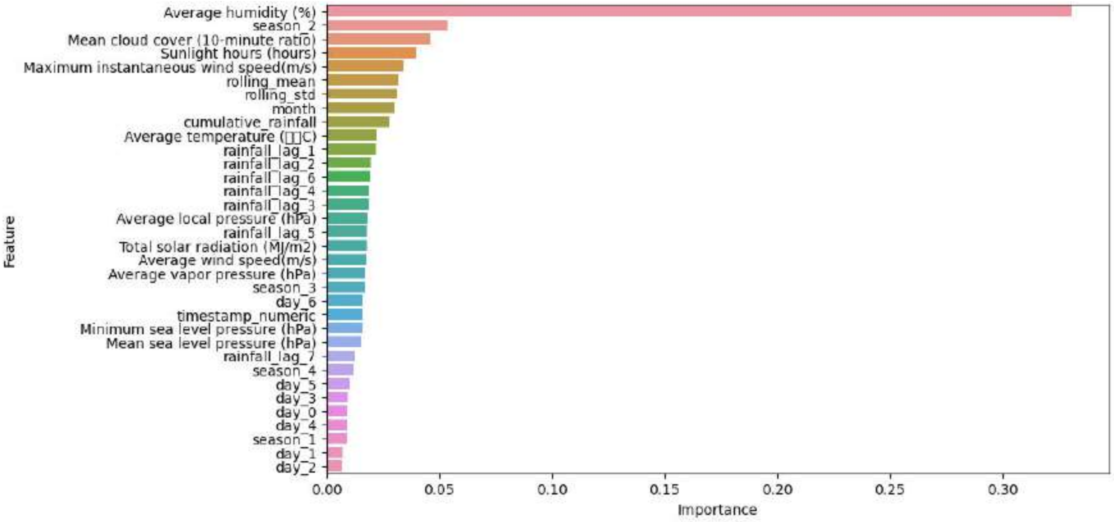

Fig. 5 Feature importance for Classification task.

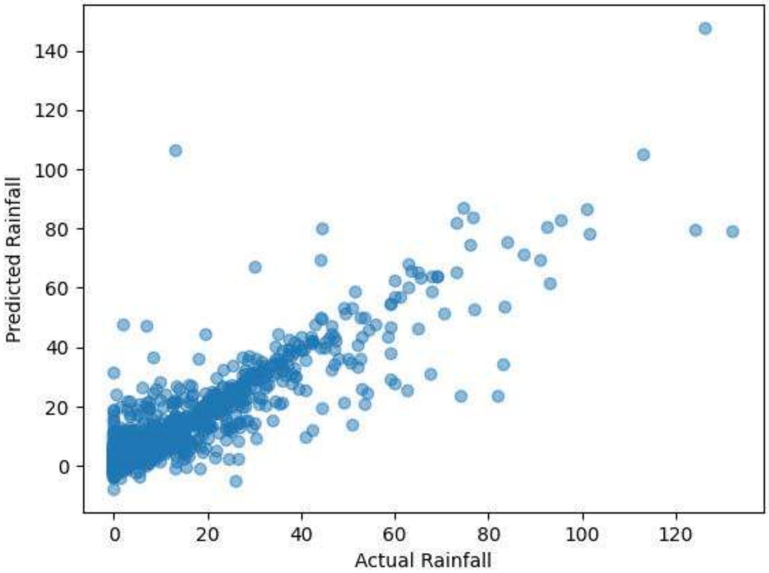

Fig. 6 Actual vs Predicted rainfall

B. Classification Task

In Table III, a comprehensive classification report is provided, revealing the robust performance of the classification model. Notably proficient in accurately identifying instances of heavy rain, the model demonstrates strong overall efficacy.

Much like its counterpart in the regression task, this classification model is accompanied by a feature importance breakdown and a revealing confusion matrix (Fig.5 and Fig. 7). These additional insights further enrich our understanding of the model's dynamics, shedding light on the significant variables influencing its classification capabilities.

Table III. Classification Report

Class	Precision	Recall	F1-score	Support
No rain (0)	0.94	0.96	0.95	2595

Rain (1)	0.89	0.85	0.87	1122
Heavy rain (2)	0.92	0.77	0.84	61
Accuracy			0.92	3778
Macro average	0.92	0.86	0.88	3778
Weighted average	0.92	0.92	0.92	3778

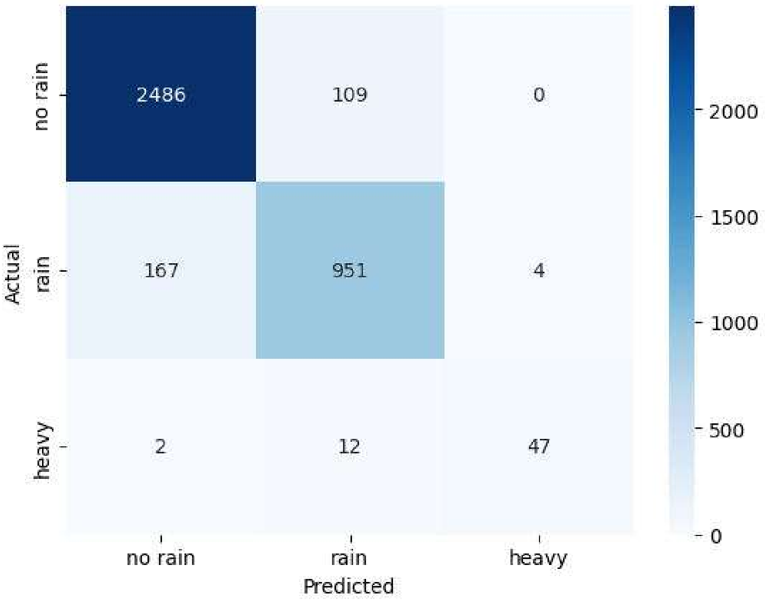

Fig. 7 Confusion matrix

DISCUSSION

In the field of rainfall predictions, various scholars diverge in their approaches. Some concentrate on identifying decisive features with significant influence on rainfall likelihood, while others seek the most performant algorithm based on rigorous model evaluations [1, 8, 9]. Our study, while yielding promising results, acknowledges areas for potential improvement.

Initially experimenting with Support Vector Machine (SVM) and Logistic Regression before settling on XGBoost, our findings indicated that even without hyperparameter tuning, XGBoost outperformed other models. Subsequent fine-tuning enhanced its performance, surpassing related studies [1, 8, 9]. While smaller metric values suggest good performance, additional insights or considerations may further refine the model.

Notably, average humidity emerged as the standout feature for our hybrid model. Incorporating features like lag features for temporal patterns, rolling statistics for trends, and monitoring previous rainy days proved crucial for optimal model learning. This prompts further exploration, such as examining the impact of varying rolling statistics and the relevance of additional features for our specific case.

The selection of a threshold yielding a tertiary classification task aligns with our research focus on capturing extreme rainfall events. While categorizing heavy rainfall as opposed to normal conditions was essential for our needs, the discussion opens avenues for considering binary classification in scenarios where a simple prediction of rain or no rain suffices [5].

In terms of real-world applications, the potential deployment of our research lies in early warning systems and urban planning. Continuous monitoring and improvement are crucial, as accurate rainfall predictions can safeguard lives in disaster-prone areas and optimize outdoor activities by enabling timely actions.

However, it is essential to acknowledge that, despite our model's promising performance, challenges may arise due to data constraints, model assumptions, or external factors such as climate change and urbanization. Future improvements in this study could involve further feature engineering to enhance predictive capabilities, exploring more advanced algorithms like Neural Networks, or integrating comprehensive forecasting models like Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM) among others.

CONCLUSION

The primary objective of this study was to accurately predict heavy rainfall on a daily basis, while also addressing the challenge of long-term predictions extending up to one week. The implementation of careful feature engineering was imperative to discern patterns in the data, thereby enhancing the model's performance. Both the regression and classification models exhibited promising capabilities in capturing extreme rainfall events. The regression model provided accurate predictions of rainfall amounts, yielding a MAE and RMSE of 1.85 mm and 4.81 mm, respectively. Simultaneously, the classification model demonstrated high precision in identifying heavy rainfall events, signifying its reliability in predicting these extreme events.

This hybrid model can serve as a strong foundation for further considerations, particularly regarding its potential deployment for operational use in applications related to predicting and managing extreme rainfall events. Fine-tuning the model or incorporating additional contextual information may further amplify its effectiveness in capturing extreme rainfall events.

ACKNOWLEDGMENT

We extend our sincere gratitude to Iizuka Hiroto for his invaluable support in acquiring data from JMA’s website. Special thanks are due to Reuben Sinha and Herman Musumari for their expert guidance in machine learning, providing valuable recommendations, and engaging in fruitful and insightful discussions about this research. Their contributions have significantly enriched the quality and depth of our study.

REFERENCES

[1] S. Markuna et al. *Application of Innovative Machine Learning Techniques for Long-Term Rainfall Prediction*, Pure and Applied Geophysics 180 (2023), pp. 335-363.

[2] Asian Disaster Reduction Center. *Natural Disaster Data Book 2019 an Analytical Overview,* 1st ed.; Kobe, Japan, 2019; pp. 2-20.

[3] https://www.nagoya.climatemps.com/

[4] Japan Meteorological Agency, https://www.jma.go.jp/bosai/map.html#5/34.5/137/&elem=temp&contents=amedas&lang=en&interval=60

[5] https://www.geeksforgeeks.org/rainfall-prediction-using-machine-learning-python/

[6] World Meteorological Organization, https://severeweather.wmo.int/rain/

[7] Kaggle, feature engineering, https://www.kaggle.com/code/ryanholbrook/what-is-feature-engineering/tutorial

[8] Liyew and Melese, *Machine learning techniques to predict daily rainfall amount*, Journal of Big Data (2021) https://doi.org/10.1186/s40537-021-00545-4.

[9] Chu at al. *Estimation of Threshold Rainfall in Ungauged Areas Using Machine Learning*, Water, MPDI, 2022.

[10] T. Chen and C. Guestrin, *XGBoost: A scalable Tree Boosting System*, KDD, San Francisco, CA, USA.

[11] K. Namitha et al. *Rainfall prediction using artificial neural network on map-reduce framework*. ACM. 2015. https://doi.org/10.1145/2791405.2791468.

[12] Kumar, R., et al. *Assessment of climate change impact on snowmelt runoff in Himalayan region*. Sustainability, 14(3), 1150 (2022). https://doi.org/10.3390/su14031150.

[13] Vitart, F., & Robertson, A. W. (2019). Chapter 1-Introduction: *Why sub-seasonal to seasonal prediction (S2S)?* In A. W. Robertson & F. Vitart (Eds.), *Sub-seasonal to Seasonal Prediction* (pp. 3–15). Elsevier. https://doi.org/10.1016/B978-0-12-811714-9.00001-2.