International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)
www.lec.edu.np/journal
Volume 2, Issue no. 2, Page 18-25
Received Date: 10/31/2025
Revised Date: 5/19/2026
Accepted Date: 5/23/2026
ISSN: 3021-999X (Print)
https://doi.org/10.3126/injet-indev.v2i2.95697

Rainfall Prediction in Kathmandu City Using Machine Learning and Deep Learning Techniques

Shaswot Poudyal[1*], Saurav Katwal[2], Rajad Shakya[3]

[1]*IOE Thapathali, Battisputali, Kathmandu, Nepal, shaswot.078bct041@tcioe.edu.np*
[2]*IOE Thapathali, Ekantakuna, Lalitpur, Nepal, saurav.078bct040@tcioe.edu.np*
[3]*IOE Thapathali, Ekantakuna, Lalitpur, Nepal, rshakya.8063@tcioe.edu.np*

Abstract

Precipitation forecasting is a relatively important issue of preparedness to disasters and water resources management and adaptation to climate changes, especially in the geographical location such as Kathmandu Valley in Nepal where precipitation patterns have become erratic and acute with the effect of climate change. The work concerns the depth of rainfall forecasting through the use of machine learning (ML) and deep learning (DL) methods on the meteorological dataset that is 10 years long (2015-2025), with the characteristics that include temperature, humidity, atmospheric pressure, wind direction, and cloud cover. We compare the six predictive models (LightGBM, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Random Forest, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) to find what patterns exist in the historical weather data and predict rainfall instances. We use analysis to show that the LSTM model reaches the best accuracy, which is RMSE of 3.73 and MAE of 2.48, even against the traditional ML models, such as Random Forest (RMSE: 3.89, MAE: 2.63) and SVR (RMSE: 4.28, MAE: 2.57). Its strength is seen in its capacity to reproduce temporal dependencies and nonlinear trends in the seasonal rainfall data, which has made LSTM highly effective. Also, our method minimizes the time of computation of the model, by exploiting optimized hyperparameters and preprocessing mannerisms specific to the zero-inflated nature of precipitation data. The results demonstrate the feasibility of deep learning models in enhancing the accuracy of rainfall predictions which can be used by policymakers and urban planners to help them prepare well ahead of time when it comes to a disaster-prone area.

Keywords: Rainfall prediction, Data mining, Time series forecasting, Machine learning, Weather forecasting, Disaster Preparedness

1. Introduction

Technology has been evolving over the past few years, particularly data mining, Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL), that have been used to tackle meteorological issues. Historical precipitation data is ideal for evaluating the potential of ML methods to unmask patterns that are not readily apparent and are suitable for the modelling of nonlinear and dynamic systems, which is often the case and is the strength of these methods. The main aim of the present study is to evaluate and compare the performance of six models such as Random Forest (RF), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), LightGBM, Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) to find the best model to predict precipitation amounts in the Kathmandu Valley. This study leverages state-of-the-art precipitation data preprocessing and feature engineering techniques specific to these seasons to handle the unique challenges posed by the zero-inflated nature of the data, employing a locally balanced precipitation dataset ranging from 2015-01-01 to 2025-07-06 across the aforementioned seasonal periods. The goal of this piece of work is to establish complex temporal relationships and non-linear trends with the aim of offering valuable information for disaster preparedness and water resources management in a region that is more and more under the impact of climate change, with scenarios characterized by irregular and severe climatic phenomena.

2. Related Work

Prediction of rainfall has been the central topic in meteorology, and the applications of machine learning (ML), deep learning (DL), and time series models present good prospects in this effort. According to Wani et al. (2024), the comparison of ML, DL, and ARIMA models across North-Western Himalayas indicated better results of

* *Corresponding author*

LSTM when compared to ARIMA because of its capacity to capture nonlinear rainfall patterns. Their research however missed out on localized topographies, like the bowl topography of Kathmandu Valley, as well as the urban heat island aspects which pose unusual issues to the precipitation data, such as making them zero inflated.

Pujara and Paudel (2024) used LSTM and GRU to generate forecasts of rainfall, and they proved to be strong in as much as they capture the temporal dependencies. However, their regionally combined data restricted generalization of the model. Compared to previous work, our analysis was conducted based on high-resolution (decade long), localized data to Kathmandu Valley so that modeling can occur with seasonal and localized specifications. We have beyond technical measures, the practical applications such as disaster preparedness and agricultural planning.

The performance indicators of nonlinearity and seasonality issues in traditional methods such as ARIMA are reflected in the low RMSE values provided by Wani et al. (2024). The same can be applied to Bashyal et al. (2025) who paired wavelet transforms with transformers and computed the high cost in computation that could not be used in real-time forecasting. In order to overcome these gaps, we suggest elaborate preprocessing solutions, such as, seasonal subsets (pre-monsoon, monsoon, post-monsoon) and feature engineering to manage sparse rainfall data better. This paper contributes to rainfall forecast in Kathmandu Valley owing to local data, advanced ML/DL techniques, and new preprocessing procedures.

3. Dataset Description

The Kaggle dataset for Kathmandu Weather Data 2015-2025 offers a great amount of daily weather records of Kathmandu Valley area in Nepal. The proposed dataset has a time range of January 1, 2015, to July 7, 2025, which sums up to more than ten years of active weather observation. The data contains key meteorological values with the temperature measurement (maximal, minimal), humidity, and pressure values, rain data, and other weather variables. The positioning of Kathmandu Valley in the geographical system of Himalayan region gives evidence about the monsoons and seasons, climatic changes and long-term climate conditions of this ecologically important region. The data collection also includes a combination of several weather stations across the valley.

Table 1. Daily Weather Observations for Kathmandu 2015–2025

Features	Description	Unit
Date	Record date	YYYY-MM-DD
tempmax	Maximum temperature	Celsius (°C)
tempmin	Minimum temperature	Celsius (°C)
temp	Average temperature	Celsius (°C)
dew	Dew point	Celsius (°C)
humidity	Humidity level	Percentage (%)
pressure	Atmospheric pressure	Pascal (Pa)
windspped	Wind speed	Meters per second (m/s)
windir	Wind direction	Degrees (°)
solarenergy	Solar energy	Joules per square meter (J/m2)
cloudcover	Cloud coverage	Percentage (%)

4. Methodology

The study engaged a clearly defined experimental pipeline overlaying a series of procedural steps for data preparation, feature engineering, model training, and model evaluation.

4.1 Data Preprocessing

- **Zero-Inflated Target**

The fact that the precipitation data of Kathmandu has a zero-inflated distribution of precipitation and because of the distinct wet and dry seasons created problems in forecasting the models, particularly when long dry spells occurred. In order to solve it, this research employed characteristics of Kathmandu with

special seasonal groupings (pre-monsoon, post-monsoon, monsoon), as well as specific practices of feature engineering that take into consideration local climatic conditions such as sudden rains during monsoon.

- **Handling Missing Values**

In case a source presented a missing value, then this value was substituted with the information of that day with another source. When an alternative source did not exist, then the relevant row was deleted.

- **Outlier Detection**

Outliers were identified via the graphical review of the boxplots and by statistical-based review utilizing the technique of interquartile range (IQR). The defective data were fixed with field knowledge or in the cases where the data was beyond the reasonable meteorological limits, data was discarded.

- **Normalization/Scaling**

For tree-based models (e.g. Random Forest, LightGBM), normalization/scaling was not applied since these models are insensitive to feature scaling. For other models (SVR, GRU, LSTM), Yeo-Johnson transformation was applied to ensure optimal performance as:

For $x \geq 0$:

- $If\ \lambda \neq 0, \quad y = \frac{(x+1)^{\lambda}-1}{\lambda}$
- $If\ \lambda = 0, \quad y = ln\ ln\ (x+1)$

(Equation 1)

For $x < 0$:

- $If\ \lambda \neq 2, \quad y = -\frac{(-x+1)^{2-\lambda}-1}{2-\lambda}$
- $If\ \lambda = 2, \quad y = -\ ln\ ln\ (-x+1)$

where y is the transformed value, x is the original value, and λ is the transformation parameter.

4.2 Exploratory Data Analysis (EDA)

In the EDA, there was the need to determine which variables were associated with the target variable (precipitation) using independent variables. The important findings within EDA are as follows:

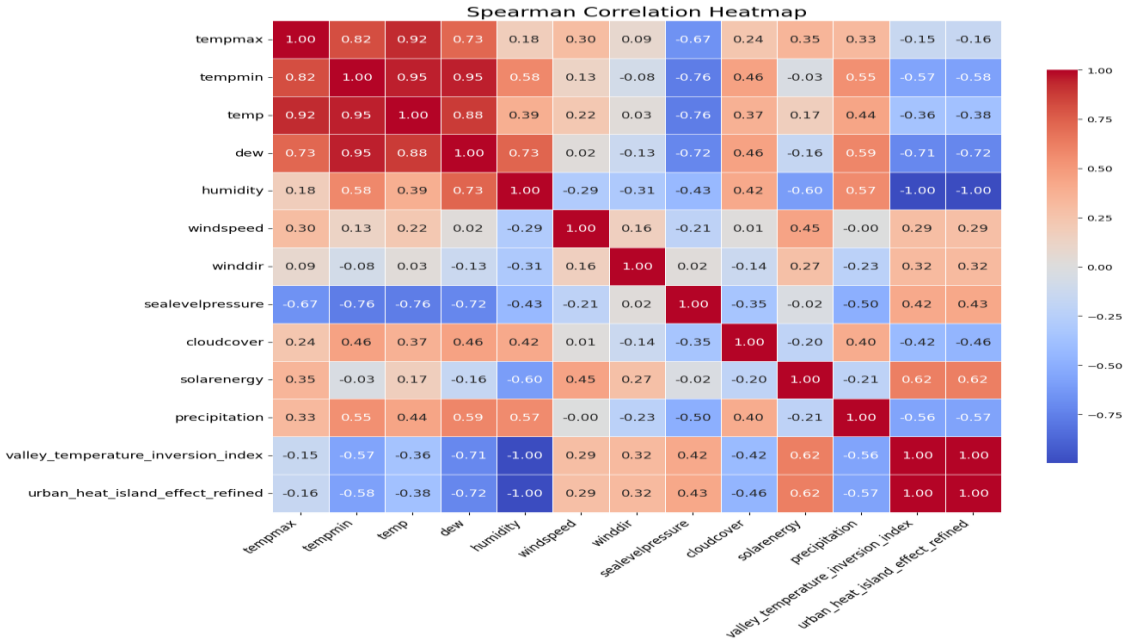

Figure 1: Spearman Correction Heatmap

Correlation Analysis: There were strong correlations between temperature, humidity, and precipitation as shown in the heatmap. These understandings led to the selection of features and generation of the models.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad \text{(Equation 2)}$$

where ρ is the Spearman correlation coefficient, d_i is the difference between the ranks of corresponding variables, and n is the number of observations.

4.3 Feature Engineering

- Features like temperature, dew, windspeed, winddirection, solarenergy had high multicollinearity with retained factors and also possessed poor overall capabilities of predicting precipitation. An extreme value of temperature (Max Temperature / Min Temperature) was found to be more instructive than mean temperature, and the variables humidity and cloudcover rendered superseded dewpoint and solar radiation.
- The cyclical pattern of seasons was created using Fourier variables (season sin, season cos) that were computed from date.

$$\text{season\_sin} = \sin\big((2\pi * \text{season\_number})/\text{total\_seasons}\big) \quad \text{(Equation 3)}$$

$$\text{season\_cos} = \cos\big((2\pi * \text{season\_number})/\text{total\_seasons}\big) \quad \text{(Equation 4)}$$

- Other model parameters unique to the Kathmandu Valley (valley temperature inversion index and urban heat island effect) were also added but eventually removed because of their negligible impact to the model accuracy.

4.4 Model Selection

To determine the amounts of precipitation, six machine learning models were trained and tested. These included both traditional and deep learning approaches. The traditional models were K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Regression (SVR), and Light Gradient Boosting Machine (LightGBM). In addition, deep learning models such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) were also employed to capture the temporal dependencies in the data.

4.5 Model Metrics

All the models were tested based on the following two metrics:

- **Mean Absolute Error (MAE)**: It is the measure of average magnitude of errors between predicted and actual values

$$MAE = \frac{1}{n}\sum |y_i - \hat{y_i}| \quad \text{(Equation 5)}$$

  where y_i is the actual value, $\hat{y_i}$ is the predicted value, and n is the number of observations.

- **Root Mean Square Error (RMSE):** It measures the square root of the average of the squared differences between the predicted values and the actual observed values

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y_i})^2} \quad \text{(Equation 6)}$$

  where y_i is the actual value, $\hat{y_i}$ is the predicted value, and n is the number of observations.

4.6 Model Description

The study uses and tests a variety of machine and deep learning algorithms to forecast the daily rainfall through the meteorological observations. The models target to capture the temporal relations and trends in a series of meteorological parameters (max temperature, min temperature, humidity, pressure, cloud cover) to predict the number of future rainfalls.

- **LSTM (Long Short-Term Memory)**: LSTM model is based on a Long Short-term Memory layer. Such architecture sequentially processes the meteorological data streams forwards and backwards, which enables it to learn dependencies with respect to past and future contexts with respect to every time step in the sequence window. Final hidden states given both directions are flattened together and fed into a fully connected-layer that forecasts the value of rainfall in the next time step. Hyperparameter tuning found the best setup to be 128 hidden units, 2 layers, learning rate of 0.01, and batch size of 16. On retraining it with these parameters, the model got Test RMSE of 3.73, Test MAE of 2.48 and Test MSE of 13.92.
- **GRU (Gated Recurrent Unit)**: GRU utilizes a Gated Recurrent Unit layer to learn temporal dependence. GRU architecture employs the use of update and reset gates to control the flow of information in a manner that controls the amount of information that should be kept in the past and that which should be included. The mechanism offers long-term patterns in the sequence window as it enables the model to learn. The last output of the GRU layer is given to the fully connected layer to generate the forecasted value of rainfall at the next time step. The best configuration found via hyperparameter tuning is the 32 hidden modules, 1 layer, 0.001 learning rate and 32 batch size, attaining validation RMSEof4.73. A test RMSEof4.07 and Test MAEof2.84 was obtained by evaluation on the test set.
- **SVR (Support Vector Regression)**: The support vector regression model takes the form of a kernelized machine learning algorithm. It attempts to identify a function that can differ in value with the actual target values by a maximum of a given margin, and at the same time as flat as possible. Hyperparameter tuning revealed that the Radial Basis Function (RBF) kernel was the best and thus the model was able to capture non-linear relationships in the data. The model had a Test RMSE 4.28 and Test MAE 2.57 indicative of moderate predictive power.
- **KNN (K-Nearest Neighbors)**: KNN regression model makes use of neighbors nearest by in the training data to predict the target value (rainfall) of a new data point as the average value of its K nearest neighbors in training data. A distance measure is usually used to calculate the proximity between pieces of data. The optimal hyperparameter combination n neighbors=105, and Manhattan metric (p=1)(uniform) was found in tuning with the best negative MSE (during the tuning) of-17.53. The Test MAE and Test RMSE on the test set were 2.8649 and 3.9978 respectively. These errors can be considered indicative of acceptable average performance but the comparatively high errors imply a possible weakness in the ability of the method employed to document the intricate rainfall patterns in the context of the dataset.
- **Random Forest**: The Random Forest model is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees. Each tree is trained on a bootstrap sample of the data, and at each node, a random subset of features is considered for splitting. This process helps reduce overfitting and improves generalization. Hyper parameter tuning identified settings such as max depth=20 and max features='sqrt' as optimal. The model demonstrated a Test MAE of 2.63 and a Test RMSE of 3.89, indicating solid performance for this regression task involving inherently noisy weather data.
- **LightGBM**: LightGBM refers to the gradient to boost framework which employs the tree-based learning algorithms to be efficient and performance-oriented, particularly, on large-scale data. It constructs the model in a sequence and every subsequent tree rectifies the mistakes that the earlier trees have committed. On the problem of predicting rainfall, LightGBM was set as a regression model and tuned on Optuna through a time-series cross-validation. The ideal hyperparameters identified entailed a learning rate of 0.047, 59 leaves, and 12 maximum depths among others pertaining to regularization and sampling. The model had an average of cross-validation RMSE of 4.11, MAE 2.92. The analysis of feature importance was also carried out to get the idea about the contribution of the input variables.

5. Results & Discussions

Within the research, several machine learning and deep learning algorithms have been tested to predict rainfall (regression) based on meteorological data. Evaluation of each model on the basis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) was conducted on the test set. The findings are summarized below and then a comparative analysis and discussion is done.

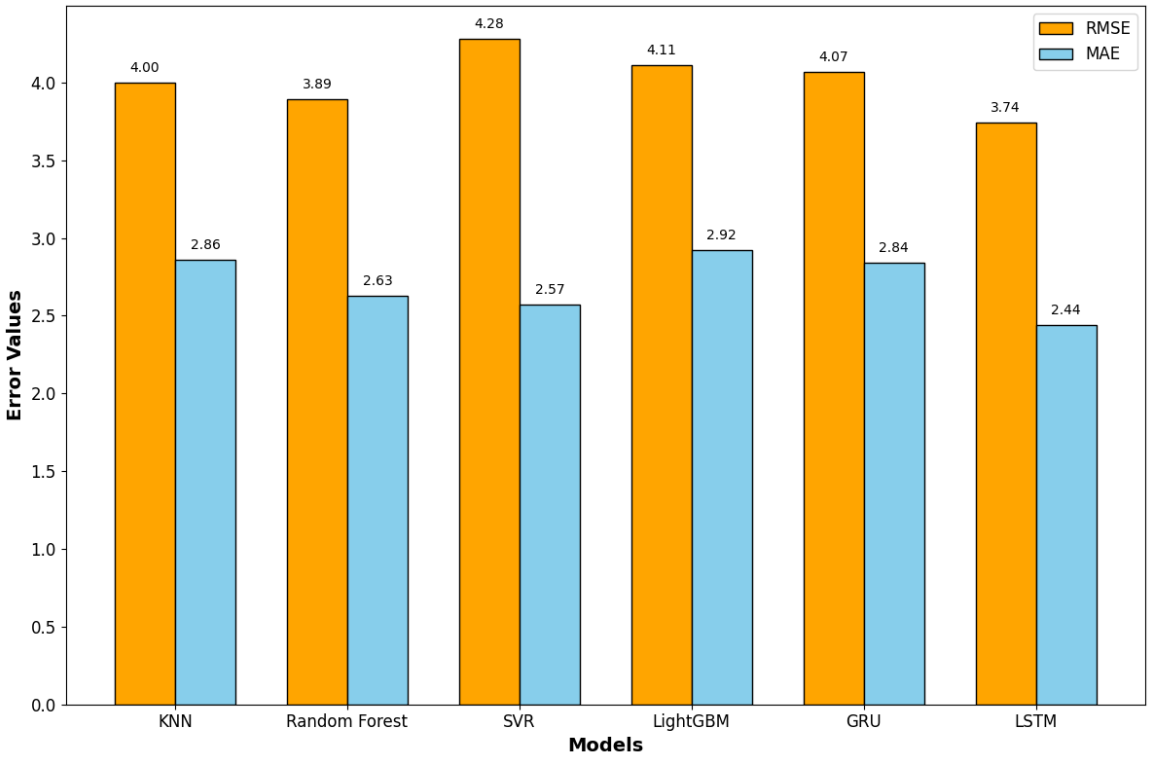

Figure 2: Comparison of MAE and RMSE for Different Models

- **Deep Learning Models (LSTM & GRU)**

  LSTM had the best RMSE (3.73) and MAE (2.48) which means that it was better in picking the temporal patterns in the rainfall data. LSTM has been identified as the best model meaning that the deep learning structures that are equipped with memory are highly applicable in predicting rainfall. GRU did not fare too badly (RMSE: 4.07, MAE 2.84) in comparison with LSTM. The lesser gating structure of the GRUs could have impaired its capacity to capture the long-range dependencies as opposed to LSTM.

- **Traditional Machine Learning Models (SVR, KNN, Random Forest, LightGBM)**

  Random Forest has performed well (RMSE: 3.89, MAE: 2.63) perhaps since its ensemble approach helps to avoid overfitting and vice versa.It was the best-performing traditional ML model, likely due to its robustness against noise and ability to handle non-linear relationships. SVR (RMSE: 4.28, MAE:2.57) had moderate predictive ability and Radial Basis Function (RBF) kernel assisted in taking non-linear associations. Nonetheless, it was beaten by models trained on trees and methods of deep learning.KNN (RMSE: 3.9978, MAE: 2.8649) exhibited moderate level of performance but failed to f it intricate rain patterns, a fact that might have been caused by noise sensitivity and high-dimensional data sensitivity of the model.

Although the numerical outcomes seem to be not that large, their practical implications are considerable. For example: The difference of 4 mm RMSE gives an input to actionable information that can be used on the part of the urban planners and disaster management teams. The stakeholders can use the early warning systems to reduce the effects of flash floods and landslides by storms as periods of heavy rainfall can be reasonably predicted. The advantage is that the model will differentiate between dry periods and high-amount precipitation-related events, although not perfectly, which means that it will contribute to solving the zero-inflated issue of the rainfall records.

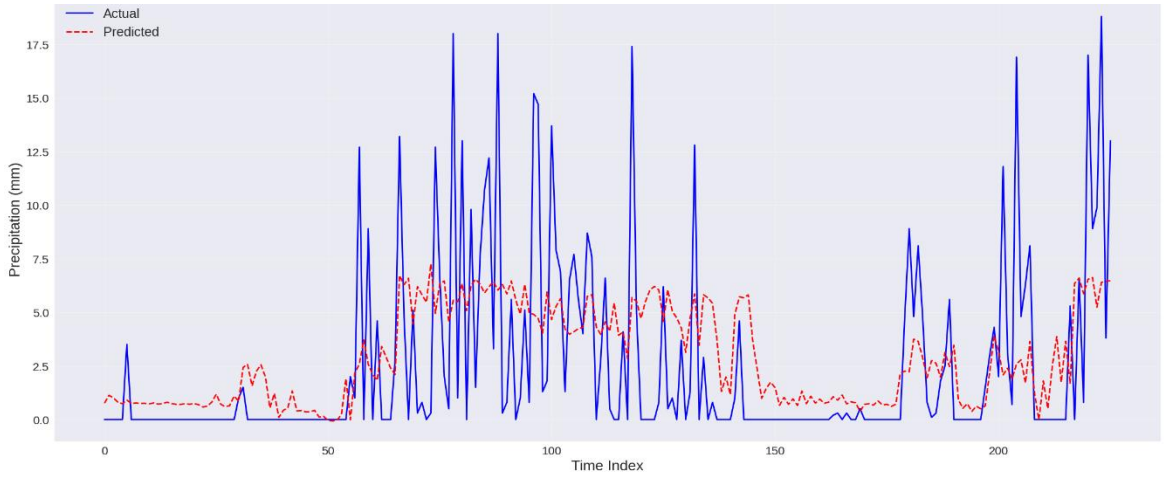

Figure 3: Actual and Predicted Rainfall with LSTM.

6. Contributions and Limitations

6.1 Contributions of the Study

It is a very important contribution to meteorological forecasting, as it gives a decadal benchmark of machine learning and deep learning of the Kathmandu Valley on a localized perspective. The study uses a high-resolution dataset over a time period from 2015 to 2025, providing the possibility of comparing the effectiveness of traditional algorithms such as Random Forest and SVR with more advanced recurrent architectures like LSTM and GRU. A major methodological advancement is the work done on the optimized pre-processing pipeline, which takes into account the nature of precipitation data to be zero-inflated and thus divides the data further into pre-monsoon, monsoon, and post-monsoon data. The identification of LSTM as the best predictor with RMSE of 3.73 and MAE of 2.48 provides a strong technical basis for predicting complex time-dependent climatic patterns in the Himalayas. Finally, the results are presented in usable data that can be used to improve the region's disaster preparedness and urban planning efforts to reduce the effects of unpredictable rainfall and flash floods.

6.2 Limitations of Study

Although the study has proved its predictive ability, there are some limitations that can be improved in further studies. The model is currently based on a combined dataset, and does not currently incorporate the spatial meteorological data from several stations across the valley, and this would be beneficial to include as this would better capture the highly localized weather variation. Moreover, various parameters like the valley temperature inversion index and urban heat island effect were investigated but the insignificant contribution to the accuracy within the experimental setup was the reason for their exclusion from the final model. The study further reveals that while deep learning models perform effectively, some models, such as KNN, do not handle meteorological data's complex noise and dimensionality, resulting in relatively higher errors in capturing the complexity of rainfall patterns than the deep learning models. In addition, the system is currently a historical observation system and does not include a forecasting component that is connected to real-time IoT sensors for real-time monitoring. Last but not least, training deep learning models such as LSTM and GRU can be time-intensive compared to simpler, traditional approaches, which may be a consideration for practical, real-time deployment.

7. Conclusion & Future Work

This paper presents a detailed comparison of different machine learning and deep learning architectures for rainfall forecasting in Kathmandu Valley, using decade-long meteorological data from 2015 to 2025. The research addresses the challenging aspects of zero-inflated precipitation data and complex climatic phenomena dependent on seasonal monsoons and the topographical peculiarities of the Himalayan region. Among the six models evaluated (K-Nearest Neighbors (KNN), Support Vector Regression (SVR), RandomForest (RF), LightGBM, LSTM, and GRU, the Long Short-Term Memory (LSTM) model demonstrated the best performance with the lowest Test RMSE of 3.73 and Test MAE of 2.48. This superior performance can be attributed to LSTM's enhanced capability to learn temporal dependencies and non-linear patterns in historical weather data. Random

Forest also showed competitive results (RMSE: 3.89, MAE: 2.63) compared to other traditional machine learning models, likely due to its inherent resistance to overfitting despite handling non-linear relationships common in meteorological data. Looking ahead, several directions could enhance this work:

- **Addition of Spatial Data**: Incorporating spatial meteorological data from multiple stations throughout the valley, potentially analyzed using Convolutional Neural Networks (CNNs), could enhance forecast accuracy by capturing localized weather patterns.
- **Real-Time Prediction System**: Developing a real-time prediction system using the most suitable model (e.g., LSTM or a hybrid approach) integrated with IoT sensors and live data streams would increase practical applicability for disaster management and urban planning applications.

References

Areru, D. A., Zimale, F. A., & Goshime, D. W. (2026). Integration of machine learning and deep learning models for hydro-meteorological data gap filling and prediction of daily rainfall in the Lake Abaya-Chamo Sub-basin, South Ethiopia. *Journal of Hydrology: Regional Studies*, *65*, 103404.

Talan, T. (2026). Machine learning-based rainfall prediction across temporal scales: model benchmarking and explainability analysis. *Stochastic Environmental Research and Risk Assessment*, *40*(100). https://doi.org/10.1007/s00477-026-03245-8.

Tackling Zero-Inflated Time Series for Precipitation Prediction. (2026). *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(45), 38478–38486.

Bashyal, B., Bhusal, D., Singh, K., & Koju, R. (2025). Rainfall prediction using wavelet transform and transformers. *KEC Journal of Science and Engineering*, *9*(1), 144–153. https://doi.org/10.3126/kjse.v9i1.78379.

Phuyal, S. (2024). Precipitation in Kathmandu Valley: A spatial-temporal study. *International Journal of Environment and Climate Change*, *14*(9), 266–278. https://doi.org/10.9734/ijecc/2024/v14i94410.

Pujara, M., & Paudel, N. (2024). Rainfall prediction using Long Short-Term Memory and Gated Recurrent Unit with various meteorological parameters. *Nepalese Journal of Statistics*, *8*(1), 47–60. https://doi.org/10.3126/njs.v8i1.73165.

Wani, O. A., Mahdi, S. S., Yeasin, M., Kumar, S. S., Gagnon, A. S., Danish, F., Al-Ansari, N., El-Hendawy, S., & Mattar, M. A. (2024). Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. *Scientific Reports*, *14*, Article 27876. https://doi.org/10.1038/s41598-024-77687-x.

Paneru, B., & Paneru, B. (2023). Real-time rainfall prediction in Kathmandu, Kapan area using sensor data with machine learning and linear regression. *Journal of Soft Computing Paradigm*, *5*(3), 266–286. https://doi.org/10.36548/jscp.2023.3.004.