2023

A Multidimensionality Reduction Approach to Rainfall Prediction

Menatallah Abdel Azeem
University College Dublin, Ireland

Prasanjit Dey
Technological University Dublin, Ireland, prasanjit.dey@tudublin.ie

Soumyabrata Dev
University College Dublin, Ireland

Funder: his research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin

A Multidimensionality Reduction Approach to Rainfall Prediction

Menatallah Abdel Azeem[1], **Prasanjit Dey**[2], **and Soumyabrata Dev**[1,3]

[1]School of Computer Science, University College Dublin, Ireland
[2]School of Computer Science, Technological University Dublin, Dublin, Ireland
[3]ADAPT SFI Research Centre, Dublin, Ireland

Abstract— The rainfall has an impact on various fields and industries, including transportation, construction, tourism, health, and wildlife preservation. Accurate rainfall prediction is essential for mitigating the negative impact of rainfall on these sectors. However, previous studies on rainfall prediction have been mainly based on datasets from North America, Europe, Australia, and Central Asia, covering different periods. This study proposes using weather datasets covering the past 5 to 10 years to capture recent patterns in weather data. Additionally, the curse of dimensionality can impact model performance and lead to overfitting. Therefore, this study proposes utilizing dimensionality reduction techniques to ensure that only the significant features are used for rainfall prediction. Multiple Linear Regression (MLR) with dimensionality reduction is applied to improve the accuracy of rainfall prediction. The experimental result shows that UMAP+MLR and t-SNE+MLR have lower MSEs of 57.27 and 56.74 and higher r^2 scores of 0.130 and 0.138, respectively. The proposed approach can be valuable in optimizing resource utilization and mitigating the impacts of rainfall on various fields and industries. The source code for our research is available on GitHub repository: https://github.com/Prasanjit-Dey/Dimension_Reduction.

1. INTRODUCTION

Rainfall is an essential component of the Earth's natural environment, and it has a profound impact on various fields and industries [1]. Transportation, construction, tourism, health, and wildlife preservation are among the many areas that can be significantly affected by rainfall conditions. In the transportation industry, rainfall can lead to disruptions in road, air, and sea traffic. Heavy rainfall can cause flooding, which can damage roads and bridges, leading to traffic congestion and transportation delays [2]. Accurate rainfall prediction can help transportation authorities prepare for these events and mitigate their impact on traffic flow. The construction industry is also highly susceptible to rainfall conditions. Heavy rain can delay construction work and lead to damage of construction sites, causing significant delays and cost overruns [3]. By accurately predicting rainfall, construction companies can plan their activities accordingly, reducing the impact of rainfall on their projects. Tourism is another field that can be impacted by rainfall conditions [4]. Excessive rainfall can make it difficult for tourists to access popular destinations, leading to a decrease in tourism revenue. Additionally, damage to tourist infrastructure caused by rainfall can be costly to repair, impacting the industry's profitability. Rainfall can also have a significant impact on human health [5]. Excessive rainfall can lead to waterlogging, creating ideal breeding conditions for disease-carrying insects like mosquitoes. This can lead to an increase in the incidence of vector-borne diseases such as malaria, dengue, and chikungunya [6]. Accurate rainfall prediction can help health authorities take preventive measures to reduce the spread of these diseases.

Previous work has been done on rainfall prediction on datasets from North America, Europe, Australia, and Central Asia covering a huge range of periods [7–9]. We are currently observing unexpected changes in weather conditions all over the globe. Thus, the weather patterns have taken a new nature, therefore the study of rainfall needs to take into consideration more recent data sets. In the past five to ten years, climate change has gone through a faster rate of change which led to an increase in the frequency and intensity of extreme weather events and other related impacts on the climate. We have also observed recent severe flooding in many countries all over the world. As part of this paper, we propose utilizing weather datasets covering the past 5 to 10 years to ensure capturing the recent patterns that have taken place in weather data [10].

In addition, since meteorological data has numerous dimensions, the curse of dimensionality is likely to occur. This can significantly impact model performance and lead to overfitting. Therefore, we propose using dimensionality reduction techniques on the datasets to ensure that only the significant features are used for rainfall prediction. There are multiple methods for dimensionality reduction [11], including Principal Component Analysis (PCA), Multidimensional Scaling (MDS), t-Distributed Stochastic Neighbour Embedding (t-SNE), and Uniform Manifold Approximation and

2023 Photonics & Electromagnetics Research Symposium (PIERS) | 979-8-3503-1284-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/PIERS59004.2023.10221498

Projection (UMAP). In this paper, we apply Multiple Linear Regression (MLR) with dimensionality reduction to improve the accuracy of rainfall prediction.

2. RELATED WORKS

This section provides an overview of research utilizing time-series data to train Machine Learning (ML) models for rainfall forecasting.

Kumar et al. [12] proposed a back-propagation feedforward neural network to predict the rainfall of the Udupi district in India. The proposed model used 70% of data from model training and the remaining 30% for testing. Qiu et al. [13] presented a Multi-task CNN network for short-term weather prediction. Through multitasking, this network automatically collects characteristics from several sites for weather prediction. The suggested approach outperforms the public weather forecasting model in terms of prediction accuracy. Aswin et al. [14] introduced a deep neural network based on LSTM and ConvNet to forecast rainfall. The suggested framework can forecast worldwide monthly average rainfall for 10368 geographical regions for 468 months. The LSTM network obtained an RMSE of 2.55 in the suggested framework, whereas the ConvNet network achieved an RMSE of 2.44. Basha et al. [15] introduced various machine learning and deep learning models to predict seasonal rainfall, including Auto-Regressive Integrated Moving Average (ARIMA), Artificial Neural Network, Logistic Regression, Support Vector Machine, and Self Organization Map. The Artificial Neural network predicted rains using backpropagation and cascade Neural Networks. Cramer et al. [16] developed Markov-Chain Extended with Rainfall Prediction (MCRP) and Genetic Programming (GP) to predict rainfall. GP and MCRP were evaluated using 21 distinct datasets from various European locations. In this study, 10-year datasets are used to train the model, while 1-year datasets are used to validate it. The experimental result indicates that GP outperforms MCRP in terms of prediction accuracy. Similarly, Darji et al. [17] have presented a survey of Artificial Neural Networks (ANN) for rainfall forecasting. Feedforward Neural Networks, Recurrent Neural Networks (RNN), and Time-Delay Neural Networks (TDNN) are primarily discussed here for predicting rainfall data. This study also addressed the application of neural networks for yearly, monthly, and daily rainfall forecasting. Mohapatra et al. [18] proposed a prediction model for forecasting rainfall in the Indian city of Bangalore using data mining and linear regression. In this study, data are collected from Bangalore, India, between 1901 and 2002. Pandas and Scikit Learn were utilized for testing and generating numerical results. K fold has been employed to forecast the rainfall for various times of the year. The monsoon season forecast had been more precise than the summer season forecast. Finally, Chatterjee et al. [19] introduced the Hybrid Neural Network for Rainfall forecasting. This study utilized Dumdum, West Bengal meteorological data ranging from the years 1989 to 1995. The Hybrid Neural Network achieved a prediction accuracy of 89.54%

3. METHODS

The MLR with dimensionality reduction-based forecasting system is depicted in Fig. 1. During the data collection phase, publicly available rainfall and time-series data are collected from West Australia [10]. In the data processing phase, unnecessary data is eliminated, and absent values are replaced. The preprocessed data is then fed into the model for dimension reduction. In the dimension reduction phase, different techniques are used to eliminate unwanted features from the preprocessed data. Then, these data are fed into the MLR model for predicting the rainfall. Finally, different evaluation metrics are used to determine the effectiveness of the model.

3.1. Data Collection and Processing

We've also seen recent catastrophic floods in a number of areas throughout the globe. As part of this work, we use meteorological datasets from West Australia over the last 5 to 10 years in order to ensure that we capture the most current trends in weather data. However, some data might have gone missing as a consequence of system failure, device collapse, or other unexpected situations. The existence of missing data has a direct impact on the time-series forecasting model's efficiency. As a result, before the training phase, we replaced the values that were missing using the subsequent information approach. Furthermore, we apply min-max normalization in order to preserve data consistency and improve the train machine learning model's forecasting precision. The following equation describes the min-max normalization:

$$Normalize\_value = \frac{Input - \min(Input)}{\max(Input) - \min(Input)} \tag{1}$$

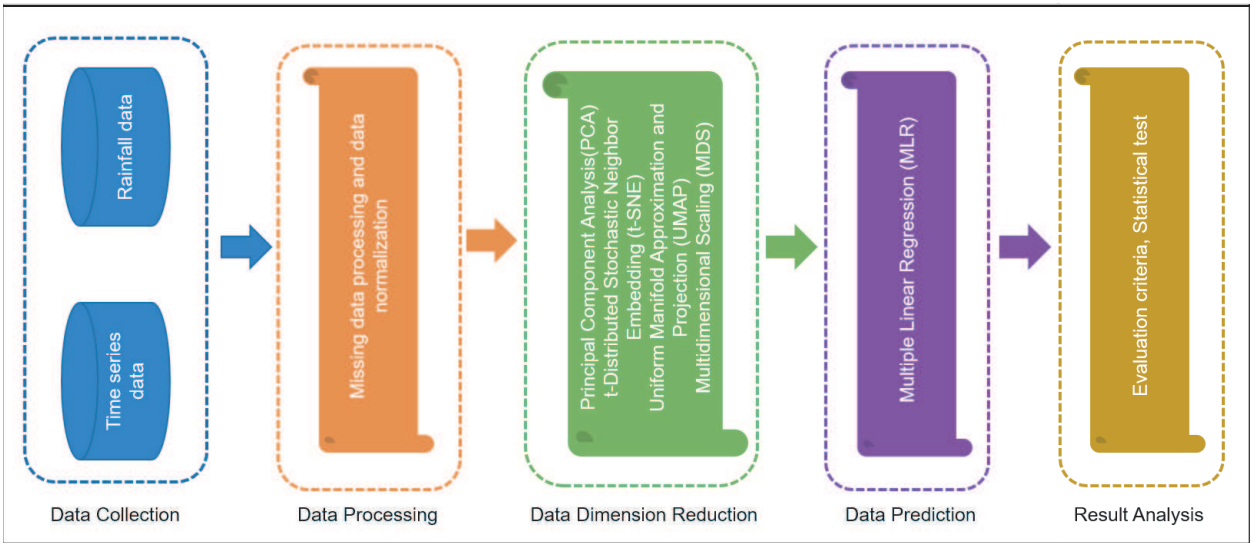


Figure 1: The proposed architecture for predicting rainfall. This framework is divided into five parts: data collection, data processing, data dimension reduction, data prediction, and result analysis.

After normalization, we divided the data into two parts: the training set, which included 80% of the data, and the test set, which contained 20% of the data. The training set is used to train the MLR model, whereas the test set is used to evaluate the trained model's performance.

3.2. Data Dimension Reduction

We have examined the impact of meteorological data's various elements on the performance of our model. The curse of dimensionality may result in overfitting, which can have a major impact on model accuracy. As a result, we propose dimensionality reduction methods to extract only the important information from the rainfall data. In this study, we used four dimension reduction approaches, namely, PCA, MDS, t-SNE, and UMAP, to evaluate the impact of the prediction model. The lower dimensionality of the rainfall data can reduce the chance of overfitting. We found that these approaches considerably enhance the performance of our model and enable us to accurately estimate rainfall in various locations.

3.2.1. Principal Component Analysis (PCA)

It is a popular dimension reduction approach in data analysis, particularly time-series data. PCA's purpose is to convert the data into a lower-dimensional space while maintaining as much of the data's variation as feasible. Given a time-series rainfall dataset X of dimension $(n \times p)$, where n is the number of time points and p is the number of variables, the steps for performing PCA are:

Standardize the data: Subtract the mean value of the variable from each time point and divide by the standard deviation. This guarantees that each variable has the same scale and enables PCA to detect patterns in the data's variance. The following equation describes this process:

$$Z = \frac{(X - \mu)}{\sigma} \tag{2}$$

where, μ is denoted mean vector of dimension $(1 \times p)$, σ is denoted standard deviation vector of dimension $(1 \times p)$, and Z is denoted standardized time-series rainfall data of dimension $(n \times p)$.

Calculate the covariance matrix: Compute the standardized time-series rainfall data covariance matrix. The covariance matrix describes the correlations between variables, and its eigenvectors and eigenvalues may be used to determine the data's main components. The process is described below:

$$C = \frac{Z^T Z}{n - 1} \tag{3}$$

where, C represents the covariance matrix of dimension $(p \times p)$, Z^T represents the transpose of Z, and $(n - 1)$ represents the degrees of freedom correction.

Calculate the eigenvectors and eigenvalues: The eigenvectors are the directions in the data that have the largest variation, and the eigenvalues are the amount of variance explained by

each eigenvector. The process is described below:

$$V, \lambda = eig(C) \tag{4}$$

where, V is denoted matrix of dimension $(p \times p)$ containing the eigenvectors of C as its columns, λ is denoted a vector of dimension $(p \times 1)$ containing the eigenvalues of C in descending order.

Select the principal components: Sort the eigenvectors in decreasing order by their associated eigenvalues. The top k eigenvectors with the greatest eigenvalues reflect the data's main components. The equation is described below:

$$V_k = [v_1, v_2, \ldots, v_k] \tag{5}$$

where V_k represents the matrix of the first k eigenvectors of V, and v_i represents the ith eigenvector.

Transform the data: Project the standardized time-series rainfall data onto the principal components to obtain a new set of variables with lower dimensionality. The is described below:

$$Y = Zv_k \tag{6}$$

where Y is the transformed dataset of dimension $(n \times k)$ containing the new variables, and V_k is the matrix of the first k eigenvectors.

3.2.2. Multidimensional Scaling (MDS)

It is a technique for dimension reduction that aims to preserve the pairwise distances between data points in a lower-dimensional space. MDS can be applied to time-series rainfall data to extract a reduced set of dimensions that capture the similarity structure between the time points. The steps for performing MDS are:

Compute the pairwise distances: Calculate the pairwise distances between the initial high-dimensional time points. The distance metric used will be determined by the nature of the data, but popular distance measures include Euclidean distance and correlation distance. The process is described as follows:

$$D = dist(X) \tag{7}$$

where D is a matrix of dimension $(n \times n)$ containing the pairwise distances between the time points, and dist() is a function that calculates the distance measure.

Transform the distance matrix: Transform the pairwise distance matrix into a new matrix that satisfies certain criteria, such as preserving the original distances as much as possible. The main process describes below:

$$B = -0.5HD^2H \tag{8}$$

where, B represents a matrix of dimension $(n \times n)$ containing the transformed distances, D^2 is denoted the matrix of squared pairwise distances, H is denoted the centering matrix defined as $H = I - \frac{1}{n}11^T$, where I is denoted the identity matrix and 1 is a vector of dimension $(n \times 1)$ containing only ones.

Find the low-dimensional coordinates: Find a set of low-dimensional coordinates for each time point that best approximates the transformed pairwise distance matrix. The process is described below:

$$V, \lambda = eig(B) \tag{9}$$

where, V represents a matrix of dimension $(n \times k)$ containing the low-dimensional coordinates of the time points as its columns, λ represents a vector of dimension $(k \times 1)$ containing the eigenvalues in descending order, and k is the desired number of dimensions.

Select the number of dimensions: Based on the amount of variation described by each eigenvalue, determine the number of dimensions k to keep in the reduced set of coordinates. A popular strategy is to choose the least k so that the sum of the k biggest eigenvalues exceeds a specified proportion of the overall sum of eigenvalues, such as 90% or 95%.

3.2.3. t-Distributed Stochastic Neighbour Embedding (t-SNE)

The t-SNE approach is a well-known non-linear dimensionality reduction technique for displaying high-dimensional data in a low-dimensional environment. Here's how to use t-SNE to reduce the dimension of time-series data:

Choose hyperparameters: t-SNE contains various hyperparameters that must be properly selected. The number of dimensions in the low-dimensional space (typically 2 or 3), the perplexity

(a measure of the effective number of neighbors), and the learning rate are among them. (the step size used in the optimization process).

Compute pairwise similarities: Next, compute pairwise similarities between data points. One frequent strategy for time-series data is to employ dynamic time warping (DTW) to determine the similarity between two-time series. The t-SNE equation for the pairwise similarity between two data points i and j is given by:

$$p_{j|i} = \frac{\exp\left(-|x_i - x_j|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-|x_i - x_k|^2/2\sigma_i^2\right)} \tag{10}$$

where x_i and x_j are the feature vectors for data points i and j, respectively, and σ_i is the bandwidth parameter chosen for point i.

Compute conditional probabilities: Using the pairwise similarities, conditional probabilities are computed for each point. These probabilities reflect the likelihood that a point would choose another point as its neighbor, given their similarities. The conditional probability of choosing point j as a neighbor of point i is then given by:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{11}$$

where, n is the total number of data points.

Optimize the embedding: Finally, an optimization algorithm is used to find a low-dimensional embedding of the data that preserves the pairwise similarities as closely as possible. This involves minimizing the Kullback-Leibler divergence between the conditional probabilities in the high-dimensional and low-dimensional spaces:

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{12}$$

where, P is the high-dimensional probability distribution and Q is the low-dimensional probability distribution. The optimization is typically done using gradient descent, with the learning rate adjusted dynamically during the optimization process.

3.2.4. Uniform Manifold Approximation and Projection (UMAP)

UMAP is a popular non-linear dimensionality reduction technique that can be used for time-series data. Here is a method for using UMAP for dimension reduction of time-series data:

Compute pairwise distances: It is necessary to compute the pairwise distances between the data points. A common method for computing the distance between two-time series for time-series data is to use dynamic time warping (DTW).

Compute fuzzy simplicial sets: Using the pairwise distances, fuzzy simplicial sets are constructed. These sets represent the local structure of the data and reflect the degree of similarity between data points. The UMAP equation for the construction of the fuzzy simplicial set is given by:

$$p_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma_i \sigma_j}\right) \tag{13}$$

where d_{ij} is the distance between data points i and j, and σ_i and σ_j are the bandwidth parameters chosen for points i and j, respectively. The fuzzy simplicial set is then constructed using a modified version of the UMAP algorithm that includes an additional fuzzy factor:

$$q_{ij} = \frac{\left(1 + a \cdot (p_{ij})^b\right)}{\sum_{k\neq i}\left(1 + a \cdot (p_{ik})^b\right)} \tag{14}$$

where a and b are parameters that control the strength of the fuzzy factor.

Optimize the embedding: Finally, an optimization algorithm is used to find a low-dimensional embedding of the data that preserves the local structure as closely as possible. This involves minimizing the cross-entropy between the fuzzy simplicial sets in the high-dimensional and low-dimensional spaces. The low-dimensional embedding is obtained by minimizing the cross-entropy between the high-dimensional and low-dimensional fuzzy simplicial sets:

$$H(P, Q) = -\sum_i \sum_j p_{ij} \log(q_{ij}) \tag{15}$$

where P is the high-dimensional fuzzy simplicial set and Q is the low-dimensional fuzzy simplicial set. The optimization is typically done using stochastic gradient descent, with the learning rate adjusted dynamically during the optimization process.

3.3. Data Prediction and Evaluation Metric

Multiple Linear Regression (MLR) is a statistical method used to analyze the relationship between multiple independent variables and a single dependent variable. When applied to time-series data, MLR can be used to make predictions about future values of the dependent variable based on past values of both the dependent and independent variables. The equation for Multiple Linear Regression is:

$$y = b_0 + b_{1x1} + b_{2x2} + \ldots + b_{nxn} \tag{16}$$

where, y is the dependent variable; $x_1, x_2, \ldots, x_n$ are the independent variables and $b_0, b_1, b_2, \ldots, b_n$ are the coefficients.

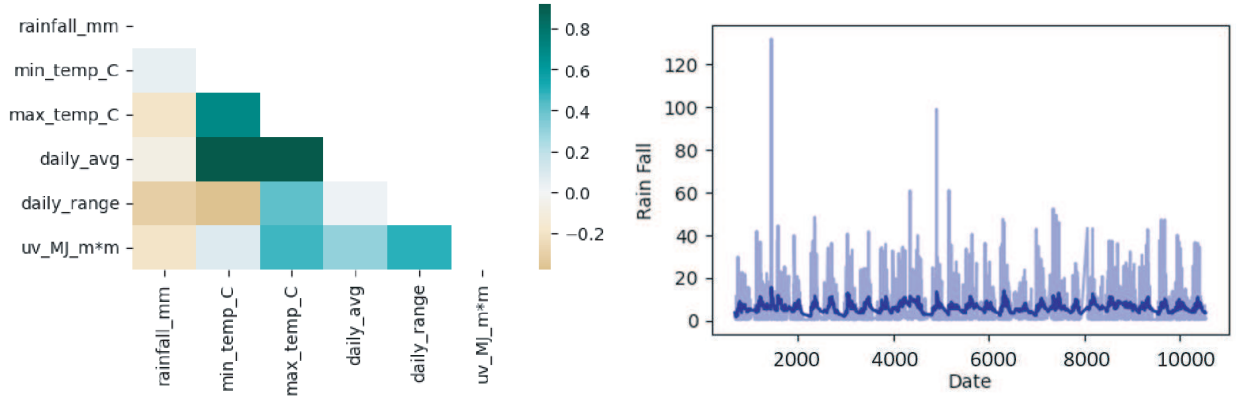
At first, low dimension rainfall are into training and testing sets: Split the data into two sets: a training set and a testing set. The training set is used to fit the MLR model, while the testing set is used to evaluate the performance of the model. Then Fit the MLR model using the training set. This involves estimating the coefficients $(b_0, b_1, b_2, \ldots, b_n)$ of the model by minimizing the sum of squared errors between the predicted values of y and the actual values of y in the training set. Next, Use the testing set to evaluate the performance of the MLR model. Calculate metrics such as the mean squared error (MSE), and coefficient of determination (R^2) to assess how well the model predicts the dependent variable. Finally, the model has been fit and evaluated, use it to make predictions on new data. Given the values of the independent variables $(x_1, x_2, \ldots, x_n)$ for a future time period, fed them into the MLR equation to predict the value of the rainfall.

4. RESULT AND DISCUSSION

A series of experiments were conducted to evaluate the effectiveness of the MLR model. The test dataset was subjected to dimensionality reduction techniques such as PCA, MDS, UMAP, and t-SNE in combination with MLR to determine the accuracy of predictions of rainfall. The accuracy of MLR was then compared with the combination of MLR and dimensionality reduction techniques to measure performance.

4.1. Data Analysis

As a component of our project, we have incorporated meteorological records from Western Australia from the past 5 to 10 years, to ensure that we have gathered the most up-to-date information on weather patterns. This research measures the correlation coefficient between rainfall data and other parameters. Fig. 2(a) illustrates the correlation coefficient of the data. As shown in the figure, it can be observed that rainfall does not have a strong correlation with other parameters. Therefore, different dimensionality reduction techniques have been employed to eliminate unwanted features, which can increase the prediction accuracy of the model.



(a) (b)

Figure 2: illustrated the (a) correlated coefficient and (b) data distribution of rainfall data from West Australia.

Similarly, Fig. 2(b) displays the distribution of rainfall during different time periods. The deep blue line represents the 7-day exponential moving average of rainfall data, indicating that the change in rainfall data over time is consistent, which suggests the presence of a hidden pattern in historical rainfall data. This pattern can be utilized in the prediction model to accurately forecast rainfall.

4.2. Performance Analysis of Short and Long Terms Prediction

The results of this study are presented in terms of short-term and long-term predictions of rainfall using different techniques. Fig. 3 shows the short-term prediction accuracy of rainfall for five different techniques, namely, Multiple Linear Regression (MLR), Multidimensional Scaling (MDS)+MLR, Principal Component Analysis (PCA)+MLR, Uniform Manifold Approximation and Projection (UMAP)+MLR, and t-Distributed Stochastic Neighbour Embedding (t-SNE)+MLR. The x-axis of the figure represents the first 20 hours of time periods of the test data, while the y-axis represents the predicted and actual rainfall over different times. The blue curve represents the actual rainfall, and the orange curve represents the predicted rainfall. The findings indicate that the UMAP+MLR and t-SNE+MLR methods produce better prediction results than other methods, as their orange curves are closer to the blue curve. This result indicates a higher degree of accuracy in the short-term prediction of rainfall using these techniques. The UMAP+MLR and t-SNE+MLR methods have the potential to perform better in predicting rainfall for a short duration of time.

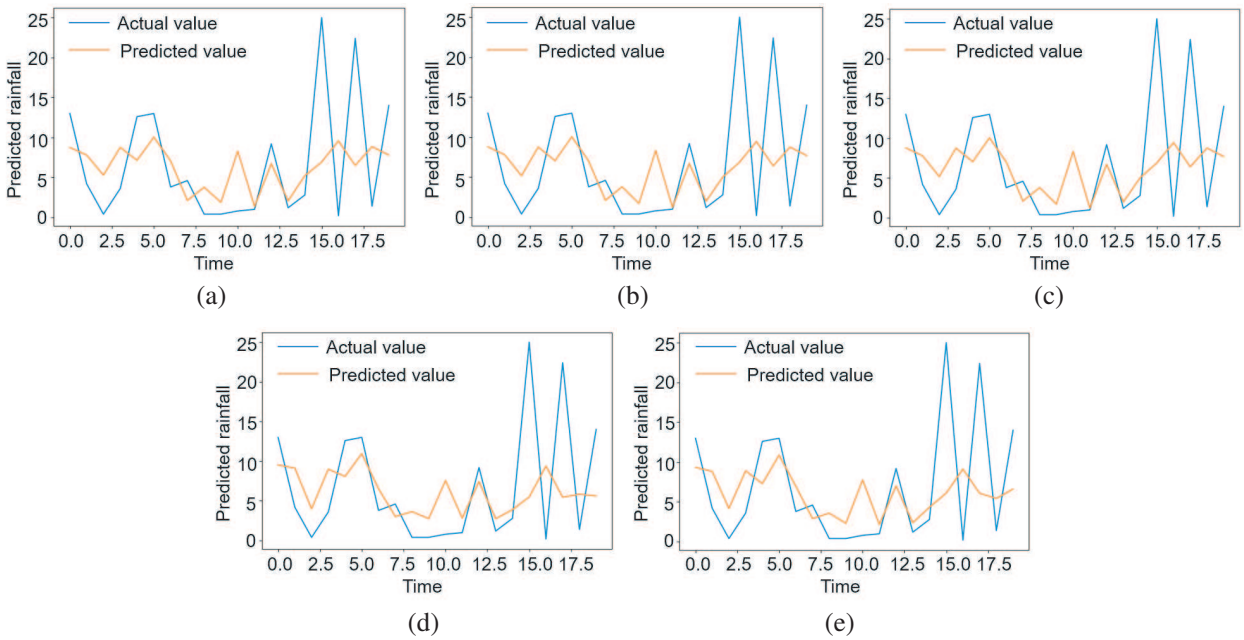


Figure 3: Short-term prediction using: (a) MLR, (b) MDS+MLR, (c) PCA+MLR, (d) UMAP+MLR, and (e) t-SNE+MLR.

Similarly, Fig. 4 shows the long-term prediction accuracy of rainfall for the same five techniques used in the short-term prediction. The x-axis represents more than 500 hours of time periods of the test data, and the y-axis represents the predicted and actual rainfall over different times. The figure shows that the UMAP+MLR and t-SNE+MLR methods produce better prediction results than other methods, as their orange curves are closer to the blue curve. This result indicates that the UMAP+MLR and t-SNE+MLR methods have better generalization ability for unseen data compared to other methods. The UMAP+MLR and t-SNE+MLR methods have the potential to perform better in predicting rainfall for a long duration of time.

Overall, the results of this study suggest that the UMAP+MLR and t-SNE+MLR methods have a higher degree of accuracy and better generalization ability than other techniques in predicting rainfall for both short-term and long-term periods.

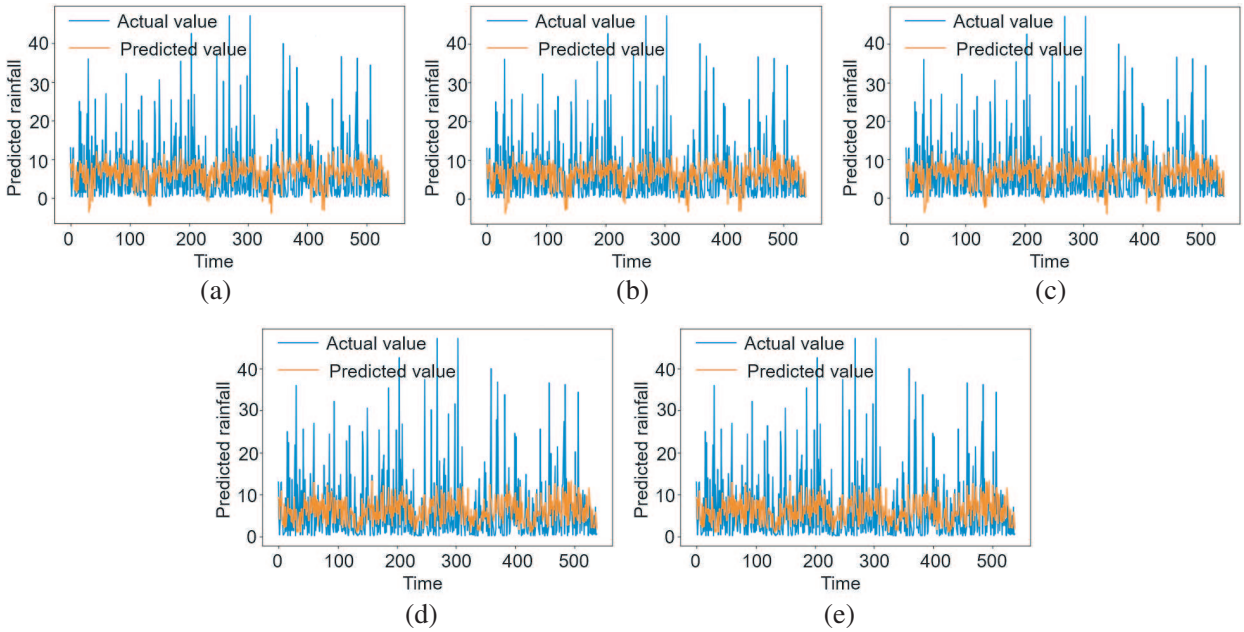


Figure 4: Long-term prediction using: (a) MLR, (b) MDS+MLR, (c) PCA+MLR, (d) UMAP+MLR, and (e) t-SNE+MLR.

4.3. Comparative Analysis

To validate the performance analysis of the MLR, MDS+MLR, PCA+MLR, UMAP+MLR, and t-SNE+MLR methods, this study employed statistical analysis, including mean square error (MSE) and correlated coefficient (R^2). Table 1 presents a numerical comparison of the different techniques in terms of MSE and R^2 for rainfall prediction. The results show that the UMAP+MLR and t-SNE+MLR methods have lower MSE values of 57.27 and 56.74, respectively, compared to the other methods. Furthermore, these methods have higher R^2 scores of 0.130 and 0.138, respectively. These findings indicate that the UMAP+MLR and t-SNE+MLR methods have better prediction performance than the other methods.

Table 1: Comparative analysis of the MLR and MDS+MLR, PCA+MLR, UMAP+MLR and t-SNE+MLR.

Methods	**Metric**	
	MSE	**R^2**
MLR	59.20	0.100
MDS+MLR	59.21	0.100
PCA+MLR	59.26	0.099
UMAP+MLR	57.27	0.130
t-SNE+MLR	56.74	0.138

As shown in Fig. 5, the UMAP+MLR and t-SNE+MLR methods had a better fit than the other methods. This indicates that these methods were better at capturing the underlying patterns in the data and producing more accurate predictions.

The MSE values, R^2 scores, and fitting curves provide a statistical measure of the accuracy of the prediction results. The lower the MSE value, the closer the predicted values are to the actual values. Similarly, the higher the R^2 score, the better the fit of the predicted values to the actual values. The statistical analysis results presented in Table 1 further support the findings from the figures presented in the previous section. The UMAP+MLR and t-SNE+MLR methods demonstrate better prediction accuracy and generalization ability compared to other methods for rainfall prediction. Therefore, the results of our study demonstrate that the combination of dimensional reduction techniques and MLR models can significantly increase the accuracy of rainfall predictions. Specifically, our findings suggest that the use of UMAP+MLR and t-SNE+MLR methods results in better fits and more accurate predictions than other methods.

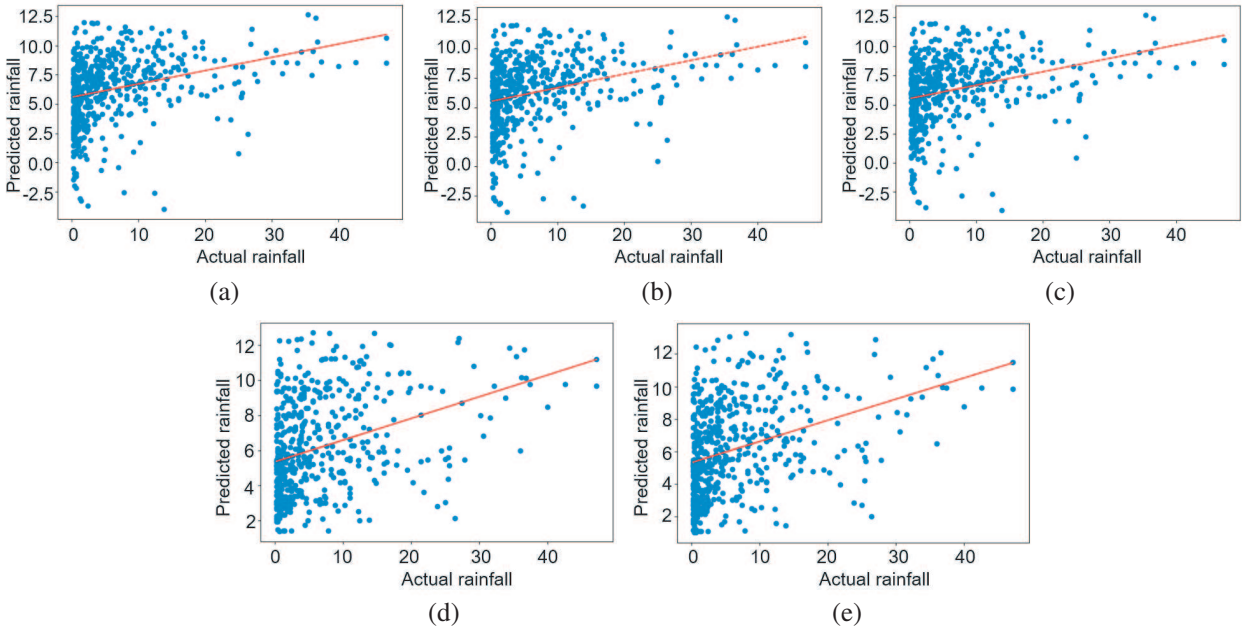


Figure 5: Fitting curve using: (a) MLR, (b) MDS+MLR, (c) PCA+MLR, (d) UMAP+MLR, and (e) t-SNE+MLR.

5. CONCLUSION

In conclusion, precise rainfall forecasting is essential for many disciplines and industries, including transportation, construction, tourism, healthcare, and the protection of wildlife. This study proposed a new method to enhance the accuracy of rainfall forecasting by utilizing weather datasets ranging the past 5 to 10 years to capture recent weather data patterns. In addition, we used dimensionality reduction techniques to ensure that only significant features were used for rainfall prediction, which is essential for overcoming the curse of dimensionality, which can negatively affect model performance and lead to overfitting.

Our experimental findings indicate that the proposed method, specifically the UMAP+MLR and t-SNE+MLR methods, outperformed other methods in terms of reduced MSEs and higher r^2 scores. These results indicate that our proposed method has the potential to maximize resource utilization and mitigate the effects of rainfall in a wide range of disciplines and industries.

Although our study has limitations, such as the use of a specific dataset and geographical region and the focus on MLR models, our findings provide important insights into how dimensionality reduction techniques can improve the accuracy of rainfall prediction. Future research should explore the proposed approach in other contexts and with other machine learning algorithms to further validate its effectiveness.

ACKNOWLEDGMENT

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. The ADAPT Centre for Digital Content Technology is partially supported by the SFI Research Centres Programme (Grant 13/RC/2106_P2) and is co-funded under the European Regional Development Fund.

REFERENCES

1. Abbass, K., M. Z. Qasim, H. Song, M. Murshed, H. Mahmood, and I. Younis, "A review of the global climate change impacts, adaptation, and sustainable mitigation measures," *Environmental Science and Pollution Research*, Vol. 29, No. 28, 42539–42559, 2022.
2. Lu, X., F. K. S. Chan, W.-Q. Chen, H. K. Chan, and X. Gu, "An overview of flood-induced transport disruptions on urban streets and roads in chinese megacities: Lessons and future agendas," *Journal of Environmental Management*, Vol. 321, 115991, 2022.
3. Schuldt, S. J., M. R. Nicholson, Y. A. Adams, and J. D. Delorit, "Weatherrelated construction

delays in a changing climate: A systematic state-of-the-art review," *Sustainability*, Vol. 13, No. 5, 2861, 2021.
4. Atzori, R., A. Fyall, and G. Miller, "Tourist responses to climate change: Potential impacts and adaptation in florida's coastal destinations," *Tourism Management*, Vol. 69, 12–22, 2018.
5. Le, K. and M. Nguyen, "In-utero exposure to rainfall variability and early childhood health," *World Development*, Vol. 144, 105485, 2021.
6. Pandey, V., M. R. Ranjan, and A. Tripathi, "Climate change and its impact on the outbreak of vector-borne diseases," *Recent Technologies for Disaster Management and Risk Reduction: Sustainable Community Resilience & Responses*, 203–228, Springer, 2021.
7. Mills, F. L. and E. A. Iniama, "Rainfall prediction for agriculture and water resource management in the United States Virgin Islands," *Environment and Labor in the Caribbean*, 19–49, Routledge, 2021.
8. Pathan, M. S., M. Jain, Y. H. Lee, T. A. Skaif, and S. Dev, "Efficient forecasting of precipitation using LSTM," *PhotonIcs and Electromagnetics Research Symposium*, 2312–2316, IEEE, 2021.
9. Manandhar, S., S. Dev, Y. H. Lee, Y. S. Meng, and S. Winkler, "A data-driven approach for accurate rainfall prediction," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 11, 9323–9331, 2019.
10. Manandhar, S., S. Dev, Y. H. Lee, S. Winkler, and Y. S. Meng, "Systematic study of weather variables for rainfall detection," *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 3027–3030, IEEE, 2018.
11. Van der Maaten, L., E. Postma, and H. Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research — JMLR*, Vol. 10, Jan. 2007.
12. Abhishek, K., A. Kumar, R. Ranjan, and S. Kumar, "A rainfall prediction model using artificial neural network," *2012 IEEE Control and System Graduate Research Colloquium*, 82–87, IEEE, 2012.
13. Qiu, M., P. Zhao, K. Zhang, J. Huang, X. Shi, X. Wang, and W. Chu, "A short-term rainfall prediction model using multi-task convolutional neural networks," *2017 IEEE international Conference on Data Mining (ICDM)*, 395–404, IEEE, 2017.
14. Aswin, S., P. Geetha, and R. Vinayakumar, "Deep learning models for the prediction of rainfall," *2018 International Conference on Communication and Signal Processing (ICCSP)*, 0657–0661, IEEE, 2018.
15. Basha, C. Z., N. Bhavana, P. Bhavya, and V. Sowmya, "Rainfall prediction using machine learning & deep learning techniques," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 92–97, IEEE, 2020.
16. Cramer, S., M. Kampouridis, A. A. Freitas, and A. Alexandridis, "Predicting rainfall in the context of rainfall derivatives using genetic programming," *2015 IEEE Symposium Series on Computational Intelligence*, 711–718, IEEE, 2015.
17. Darji, M. P., V. K. Dabhi, and H. B. Prajapati, "Rainfall forecasting using neural network: A survey," *2015 International Conference on Advances in Computer Engineering and Applications*, 706–713, IEEE, 2015.
18. Mohapatra, S. K., A. Upadhyay, and C. Gola, "Rainfall prediction based on 100 years of meteorological data," *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 162–166, IEEE, 2017.
19. Chatterjee, S., B. Datta, S. Sen, N. Dey, and N. C. Debnath, "Rainfall prediction using hybrid neural network approach," *2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)*, 67–72, IEEE, 2018.