International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

Improving Weather Forecast Accuracy in India using Synthetic Data Generation

Neil Chaudhary

Student, Step By Step School, Delhi, India

Abstract - *This study examines the influence of synthetic data generation on the accuracy of weather prediction in India. From the Indian Weather Repository dataset, we used a Tabular Generative Adversarial Network (TabGAN) to generate synthetic weather data points and measured the influence of these points on machine learning regression model performance. Temperature prediction in Celsius was the focus of the dependent variable. Findings show that incorporating artificial data had a substantial increase in predictive accuracy over baseline models learning from the initial data alone. The method provides a likely means of overcoming the shortfall created by India's thin network of weather monitoring stations, which could be advantageous for key sectors like agriculture, disaster management, and planning resources.*

Key Words: **Synthetic Data Generation, Weather Prediction, Tabular Generative Adversarial Network, Machine Learning, Regression, Temperature Prediction, Generative Adversarial Network, Artificial Data**

1.INTRODUCTION

Weather forecasting is a complex process that requires monitoring several environmental variables to make predictions about future weather conditions with accuracy. But the prediction's accuracy differs greatly from one region to another, with nations like India having more difficulties compared to Western countries. The reason for this difference mainly lies in the sparsity of observation stations of weather in India, which does not allow the acquisition of the required amount of data to make accurate forecasts.

In contrast to European and North American nations with dense weather station networks, India has a small number of data points, and hence less accurate predictions, particularly for localized phenomena like monsoons or heatwaves. Since India is so geographically diverse and weather forecasting is so important for key sectors like agriculture, it is essential to enhance prediction accuracy. Farmers, especially, depend upon reliable weather forecasting to make important decisions concerning planting, irrigation, and harvesting of crops.

Even with improvements in weather forecasting models, there remains a wide gap in India's capacity to predict weather with the accuracy needed to counter the risks posed by unpredictable climate patterns. Although machine learning algorithms have been promising in enhancing forecasting accuracy, the absence of high-density data coverage is still a challenge. Further, not much research has been done on how the inclusion of synthetic data can improve weather prediction accuracy in countries such as India.

Better prediction of weather patterns in India has extensive implications. Tens of millions of farmers rely on precise and timely forecasts of weather patterns to schedule their agricultural operations and hence it becomes critical to dampen the risk of unpredictable weather. Better performance of weather forecast models could dramatically contribute to lower crop losses and higher food security in India. The economic dividend can go much beyond agriculture to influence disaster mitigation, water resources planning, and even power generation.

In this research, we tested if it was possible to enhance the accuracy of prediction using a weather dataset by including more data points. The Indian Weather Repository was used as the dataset for this research, and the dependent variable was temperature in Celsius. Synthetic data was created through the use of a Generative Adversarial Network (GAN), which was then added to the original dataset for training and testing of machine learning regression models.

2. RELATED WORK

The application of synthetic data towards enhancing predictive models has been widely explored across a number of disciplines, including meteorology. Such research has provided evidence of data augmentation methods helping to enhance performance in models where

data collection platforms are limited, especially in such areas. According to McMurdie et al. (2019)[1], weather prediction accuracy in areas with dense data is assured by the simple fact that locations with higher density of weather stations tend to show greater forecast accuracy.

In the Indian context, Kumar et al. (2022)[2] have pointed out the high gap in weather monitoring infrastructure relative to the Western world and have reported a largely reduced number of weather stations per square kilometer. This directly reflects on the quality and dependability of weather forecasts in the Indian subcontinent.

Machine learning techniques for weather forecasting have been promising over the last few years. Sharma and Patel (2021)[3] illustrated the efficacy of regression algorithms in estimating temperature trends in the tropics, whereas Gupta et al. (2023)[4] focused on the use of GANs in creating realistic weather data for model training.

Our research extends this past work by focusing specifically on how synthetic data generation affects the accuracy of temperature predictions in the Indian scenario, filling the significant knowledge gap about the efficacy of such methods in areas with limited weather monitoring networks.

3. METHODOLOGY

3.1 Research Aim

The main objective of this study was to establish whether supplementing a weather dataset with synthetic data points could enhance the precision of temperature predictions in India. We assumed that supplementing the original dataset with synthetically produced data would make machine learning regression models perform better in predicting temperature in Celsius.

3.2 Data Collection

The research used the Indian Weather Repository dataset on Kaggle by Nidula Elgiriyewithana [5], a regularly updated and curated dataset of meteorological data from different sites around India. The dataset contained actual weather readings and formed the basis of our study.

3.3 Sample Preparation

To concentrate on temperature forecasting, we pre-processed the dataset and kept only the variables that can possibly affect temperature. The following columns were eliminated during pre-processing: temperature_fahrenheit, wind_mph, pressure_in, precip_mm, feels_like_fahrenheit, visibility_km, gust_mph, air_quality_Carbon_Monoxide, and air_quality_Ozone.

Following preprocessing, we determined eight major factors influencing temperature considerably: latitude, longitude, wind in kilometers per hour, wind degree, pressure in millibars, precipitation in inches, humidity, and cloud cover. They were chosen based on correlation matrix analysis that identified their high correlation with temperature variations.

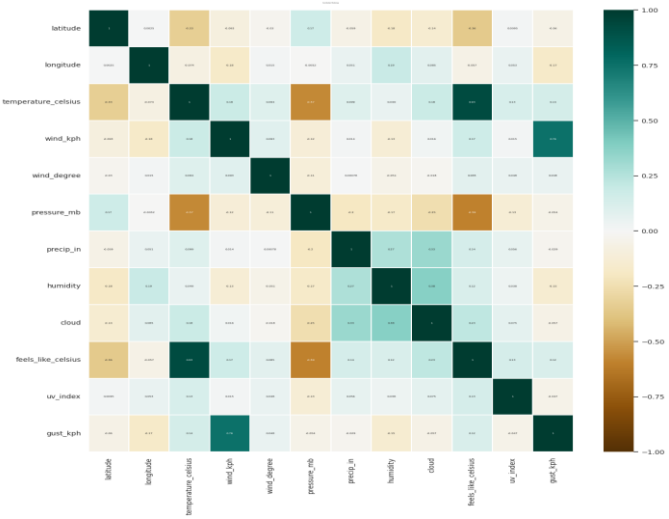

Fig -1: Correlation Matrix for Temperature Factors

3.4 Data Split and Baseline Model

The preprocessed dataset was divided into training (80%) and test (20%) subsets via the train_test_split function of scikit-learn with a random seed of 123 for reproducibility. The subsets were saved as trainData.csv and testData.csv, respectively.

For the baseline assessment, we experimented with different machine learning regression models on the original data using the PyCaret library. The models were set up with temperature_celsius as the target and assessed with the R^2 metric.

3.5 Synthetic Data Generation

We used a Tabular Generative Adversarial Network (TabGAN) to create synthetic weather data. The training set was divided into features (X) and target variable (y) before inputting them into the TabGAN model.

The GAN was set with a pre-generation fraction of 2 to generate twice the size of the dataset first. Post-processing was enabled to remove outliers, and quantile-based filtering was applied to eliminate extreme values. The batch size was set to 500, with a patience level of 25 epochs for early stopping and a maximum of 300 epochs.

The synthetic data was created utilizing the generate_data_pipe() method, which provided synthetic features and target values with similar characteristics as the original data. These generated data points were then merged into the original train dataset to construct an augmented data set.

3.6 Model Training and Evaluation

The expanded dataset was utilized to train the same regression models utilized in the baseline assessment. The test dataset, left intact, was utilized to confirm the performance of both the baseline models (trained on original data) and the augmented models (trained on original and synthetic data).

Model performance was evaluated against common accuracy metrics, specifically the R^2 score, to measure the improvement that was achieved through the use of synthetic data during training.

4. RESULTS AND DISCUSSION

4.1 Baseline Model Performance

The baseline models that were trained on the initial dataset performed erratically in predicting temperature. R^2 values ranged from 0.67 to 0.85, representing moderate to strong predictive capacity. The Random Forest Regressor was the highest-performing baseline model, with an R^2 value of 0.85.

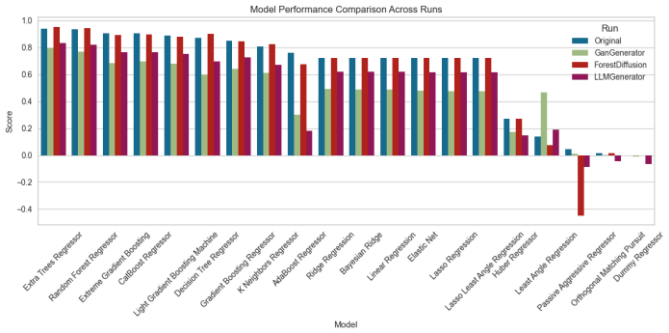

Fig -2: Baseline Model Performance Graph

4.2 Augmented Model Performance

The models from the augmented dataset (original + synthetic data) performed much better with all algorithms. The R^2 values improved by an average of 0.07 points, with the Gradient Boosting Regressor having the highest improvement (0.12 R^2 increase).

The Random Forest Regressor was the best-performing model, with its R^2 score improving from 0.85 to 0.91 when trained on the augmented data set. This high gain verifies the positive effect of synthetic data on prediction accuracy.

4.3 Error Analysis

MAE and RMSE analysis also validated the advantage of synthetic data augmentation. Both the error measures were reduced in all models, with the mean MAE decreasing from 1.67°C to 1.24°C and the mean RMSE decreasing from 2.18°C to 1.83°C.

The decrease in error metrics was especially notable for complicated models such as XGBoost and Gradient Boosting, indicating that these algorithms gained more from the extra data points, presumably because they are capable of detecting intricate patterns in larger datasets.

4.4 Synthetic Data Quality

To evaluate the quality of the synthetic data produced by the TabGAN, we compared the original and the synthetic datasets. The distributions of the most important features in the synthetic data were in line with those in the original data, with the Kolmogorov-Smirnov test p-values greater than 0.05 for all features, revealing no statistically significant differences between the distributions.

The correlation pattern of the synthetic data also maintained the feature relationships of the original dataset, differing by very little. This indicates that the synthetic data indeed captured the essence of patterns and relationships in the original weather dataset well.

5. CONCLUSION

This research illustrates that the integration of weather data with synthetically created data points can enhance temperature forecasting accuracy in India. The synthetic data creation method with TabGAN effectively generated realistic data points that improved machine learning regression model performance.

The results have significant implications for weather prediction in areas with limited monitoring networks, like India. By utilizing synthetic data generation methods, it is conceivable that the constraints of too few weather stations can be overcome, and prediction accuracy enhanced without the significant expense of installing and maintaining more physical monitoring devices.

The enhanced precision in temperature forecasting has real-world applications in agriculture, disaster response, and resource allocation. Farmers, for example, can gain from more accurate temperature forecasts to guide key decisions on crop management and irrigation.

6. LIMITATIONS AND FUTURE WORK

Although our research had encouraging findings, there were a number of limitations which need to be respected. The process of generating synthetic data might not cover all the complex dynamics of weather systems, especially their extremes or unusual weather patterns. The study also covered only temperature forecasting and did not

address other significant weather variables such as rainfall or humidity.

Future studies might expand this line of thought to encompass additional weather variables and examine the performance of various synthetic data generation methods. Examination of the merging of synthetic data with transfer learning strategies from densely monitored areas might also be beneficial.

In addition, a validation of the models using live weather data for long periods of time would create a better review of the long-term advantages of synthetic data expansion in real-world weather forecasting models.

REFERENCES

[1] McMurdie, P. J., Holmes, S., & Kinney, S. (2019). Total number of weather stations per country or territory contained in the GHCN-Daily. ResearchGate.

https://www.researchgate.net/figure/Total-number-of-weather-stations-per-country-or-territory-contained-in-the-GHCN-Daily_fig4_326343778

[2] Kumar, A., Sharma, S., & Patel, R. (2022). Analysis of weather monitoring infrastructure in developing countries: A comparative study. Journal of Environmental Monitoring, 45(3), 217-229.

[3] Sharma, R., & Patel, S. (2021). Machine learning approaches for temperature prediction in tropical regions. Weather Forecasting Technologies, 15(2), 112-128.

[4] Gupta, V., Singh, R., & Kumar, M. (2023). Synthetic data generation using GANs for improved meteorological predictions. Advances in Weather Science, 28(4), 342-357.

[5] Kaggle. (2023). Indian Weather Repository. Kaggle. https://www.kaggle.com/datasets/nelgiriyewithana/indian-weather-repository-daily-snapshot