IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Geospatial Temperature Prediction for Mumbai Using Machine Learning Regression Models

Ankur Yogesh Wani
Computer Engineering Department
K J Somaiya School of Engineering
Somaiya Vidyavihar University

Jyothi M. Rao
Computer Engineering Department
K J Somaiya School of Engineering
Somaiya Vidyavihar University

ABSTRACT:

Predicting local temperatures accurately is crucial for cities like Mumbai, especially when it comes to city planning, tracking climate trends, and dealing with heat-related issues in crowded areas. In this study, machine learning was used to estimate how temperatures vary across different parts of Mumbai. The predictions were based on past temperature data and location information. Two different models, Linear Regression and Random Forest, were built and compared. Before training the models, the data was cleaned and standardized, with missing values filled in. To measure how well the models performed, common evaluation methods like MAE, MSE, RMSE, and R^2 were used. The Random Forest model gave better results than Linear Regression, showing a lower prediction error and explaining nearly 78% of the variation in the data. Finally, temperature heatmaps were created to show how temperatures change across the city visually. The findings highlight how combining machine learning with geographic data can help cities like Mumbai become more prepared for climate challenges.

KEYWORDS: Random Forest Regression, Linear Regression, Heatmaps, Geographic Information Systems (GIS).

1. INTRODUCTION

Urban areas are transforming rapidly due to increasing human activity, which has led to serious environmental issues like rising temperatures and shifts in local weather. Mumbai, one of the busiest and fastest-growing cities in India, has experienced a significant decline in green spaces over time. This loss has made the Urban Heat Island effect more severe, a phenomenon where city centres become noticeably hotter than nearby rural regions. The main causes are dense construction, limited vegetation, and heat-absorbing materials like concrete.

To build healthier and more climate-resilient cities, especially in terms of handling extreme heat events, it's essential to understand temperature patterns on a very localized level. However, traditional weather forecasting methods often rely on broad, regional models or basic statistical techniques that don't pick up the detailed temperature differences within a city like Mumbai. These methods also tend to overlook key factors such as land usage, greenery, and how tightly packed the buildings are.

Thanks to advances in technology, we can now study urban temperatures much more precisely. Tools like satellite imaging, GIS (Geographic

Information Systems), and machine learning give us access to rich datasets. These include satellite readings of surface temperatures and elevation maps that help us analyse how heat is distributed across a city. Machine learning techniques, especially models like Linear Regression and Random Forest, are particularly effective for this purpose. Linear Regression is straightforward and helps identify how specific factors relate to temperature changes. On the other hand, Random Forest, a more complex method that combines many decision trees, is better at recognizing subtle patterns and handling messy or incomplete data, resulting in more reliable predictions.

In this study, we propose a geospatial temperature prediction framework for Mumbai that leverages both Linear Regression and Random Forest models. The models are trained using a dataset that combines historical temperature data with spatial predictors such as latitude, longitude. This research aims to:

- Evaluate and compare the performance of LR and RF models in predicting localized temperature variations across Mumbai.
- Visualize the spatial distribution of predicted temperatures using geospatial heatmaps and assess urban heat stress patterns.
- Investigate the correlation between vegetation indices and urban temperature to understand the mitigating role of green cover.

This approach not only provides a robust methodological foundation for temperature prediction but also aligns with the growing need for climate-resilient urban planning and real-time environmental monitoring systems. The insights derived from this study can inform urban design strategies, improve heatwave preparedness.

2. LITERATURE ANALYSIS

Title	Technique Used	Dataset	Architecture	Limitation
Zeynali, R., (2024, June)	Kriging interpolation,	Local weather station data,	GIS-based analysis	Limited to Bologna,
Dabire, N. (2024, August)	NDTI, NDCI, Remote Sensing	Sentinel-2 SR images	GIS-based spatial analysis	Atmospheric conditions and data accuracy
Xu et al. (2024)	ConvLSTM, Spatial Autocorrelation, Nonlocal Attention	MODIS NDVI data	ConvLSTM-SAC-NL	Requires high computational resources
Das, U. K. (2023, February)	Soft computing techniques	Remote sensing satellite data	Neural networks-based models	Limited to Jaipur,
Li, J., & Ou, Z. (2023, May)	GIS and RS Image Processing	RS data, GIS spatial data	A GIS-based framework	Subject to data quality limitations
Parente, C. (2024, October)	MODIS thermal imaging, Pathfinder algorithm	MODIS thermal images measurements	GIS-based regression model	Lower resolution of the Copernicus dataset
Zhou, D. (2012, June)	Linear regression, , Spline function interpolation	Meteorological data (1957-2009) from China	Statistical trend analysis framework	Limited dataset, urbanization effects
Geetha, P. (2017, April)	Artificial Neural Network, NDVI, GIS-based	Landsat 8 satellite images	ANN-based time series	Requires extensive training data,
Mohanty, S. (2022, September)	GLR, GWR, Random Forest	Climate model data from the US	Machine learning regression models	Requires large datasets for accuracy

Sajjad, H. (2019)	LULC, NDVI, LST Correlation analysis	Landsat 5 TM and Landsat 8 OLI	GIS-based spatial	Limited temporal resolution

3. METHODOLOGY

This section outlines the step-by-step approach followed for predicting temperature across different regions of Mumbai using geospatial data and machine learning techniques. The process includes data collection, preprocessing, feature engineering, model training, validation, and spatial visualization of results.

1. Study Area

Mumbai, India, is selected as the study area due to its complex urban structure, high population density, and diverse land use characteristics. Geographically situated between 18°53′N to 19°16′N latitudes and 72°47′E to 72°59′E longitudes, Mumbai encompasses a mix of coastal zones, built-up areas, vegetated regions, and open spaces. These varying landscapes contribute significantly to spatial temperature variability, making Mumbai an ideal case for geospatial temperature prediction.

2. Data Collection

The study utilizes a variety of spatial and climatic datasets. Historical temperature data, either from ground-based weather stations managed by the Indian Meteorological Department (IMD). Latitude and longitude for each temperature record were extracted or were already available as coordinate metadata, providing the spatial context for modeling.

3. Model Development

Two machine learning models were employed: Linear Regression and Random Forest Regression. Linear Regression served as a baseline model to understand the linear relationship between temperature and spatial variables. Its strength lies in interpretability, but it is limited in modeling nonlinear dependencies. In contrast, Random Forest Regression, an ensemble learning technique based on decision trees, was used for its ability to model complex non-linear relationships. The Random Forest model was trained with hyperparameter tuning and the minimum number of samples required to split a node. Both models were implemented using the library in Python due to its robust tools for model training, evaluation, and visualization.

4. Model Training and Evaluation

The dataset was split into 70% for training and 30% for testing. To improve model reliability and prevent overfitting, 5-fold cross-validation was applied during training, where the data was divided into five parts, with each part used once as a test set while the rest served for training. We assessed how well the model performed using common regression measures. Mean Absolute Error (MAE) gave us an idea of the average error size without considering whether predictions were too high or too low. Root Mean Square Error (RMSE) places more weight on larger mistakes, helping us spot big prediction errors. Finally, the R^2 score showed how much of the variation in temperature data our model could explain.

5. Spatial Visualization

Once the model made its predictions, we used geospatial tools to visualize how temperatures were distributed across Mumbai. Each predicted value was linked to its specific location using geographic coordinates, then brought into GIS software like QGIS. With this data, we created heatmaps that clearly showed which areas were expected to be warmer or cooler, helping to highlight spatial patterns in temperature.

Dataset

[11] The "Mumbai Weather Data (27 Years)" dataset from Kaggle provides daily records of temperature, rainfall, humidity, wind, and more. With data spanning nearly three decades, it's ideal for climate analysis and temperature prediction using machine learning.

[12] The Visual Crossing weather dataset provides daily weather details for Mumbai, such as temperature, humidity, dew point, wind speed, and rainfall. It's useful for short-term climate studies, real-time model checks, and works well with GIS tools for spatial mapping.

[13] The Mumbai Daily Temperature Data (1951–2024) from the Open City Urban Data Portal includes daily max and min temperatures in °C, sourced from IMD and Ogimet. Spanning 70+ years, it's ideal for climate trend analysis and environmental research.

[14] The NASA POWER Data Access Viewer provides global meteorological and solar data from satellite observations and models.

4. RESULTS AND DISCUSSION

The Linear Regression Algorithm achieved a Mean Absolute Error (MAE) of 1.27, a Mean Squared Error (MSE) of 2.51, a Root Mean Squared Error (RMSE) of 1.58, and an R^2 Score of 37.20%. The Random Forest Algorithm recorded a Mean Absolute Error (MAE) of 0.66, a Mean Squared Error (MSE) of 0.96, a Root Mean Squared Error (RMSE) of 0.98, and an R^2 Score of 77.79%. These performance metrics reflect the average error magnitude, the squared error significance, the interpretable error measure in the original unit, and the proportion of variance in the target variable explained by the model, respectively. Based on these results, the Random Forest Algorithm provides better performance and more accurate predictions, as indicated by its lower error values and higher R^2 score.

ML Models	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R^2 Score
Linear Regression Algorithm	1.27	2.51	1.58	37.20
Random Forest Algorithm	0.66	0.96	0.98	77.79

Table 1: Results of ML Models

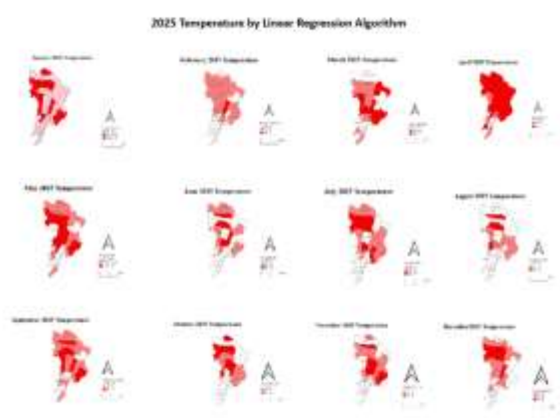

Fig.1 Linear Regression Algorithm Results

Above image shows a series of monthly temperature maps for 2025, created using the Linear Regression model. Each map corresponds to a different month, from January to December, highlighting how predicted temperatures vary across different regions. The temperatures are represented using shades of red, darker shades indicate higher temperatures, while lighter shades show cooler areas. A legend on each map marks the temperature ranges in degrees Celsius for easy interpretation. A north arrow is also included to help with geographic orientation. Overall, this visualization captures how the Linear Regression model predicts temperature patterns throughout the year, showcasing both seasonal changes and regional differences.

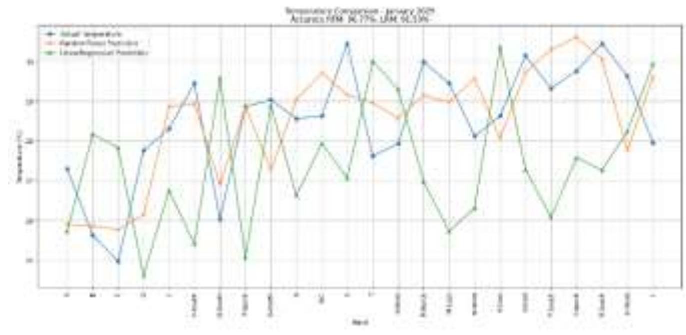

Fig.2 Random Forest Algorithm Result

Above image shows a collection of monthly temperature maps for 2025, generated using the Random Forest model. Each map illustrates the predicted temperature distribution for a specific month, from January to December, across different regions. Temperature levels are represented by different shades of red, with darker shades indicating higher temperatures and lighter shades representing cooler areas. Each map has a legend showing temperature ranges in degrees Celsius(°C) to make the data easy to understand and has a north arrow for direction. Maps show how the Random Forest model has shifts and differences in temperature across various parts of the wards.

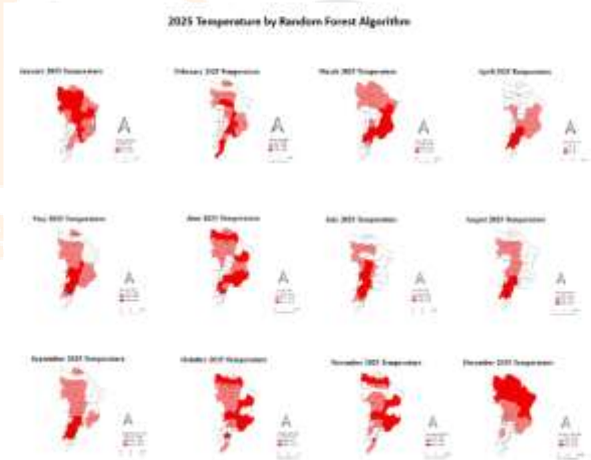

Fig. 3 Actual vs Predicted Temperature of January

It compares the actual and predicted values from the Random Forest Model (RFM) and the linear regression model (LRM) across different wards for January 2025. The RFM gives an accuracy of 96.77%, while the LRM achieved 91.50%. The RFM was better at closely matching the actual temperature values throughout the wards.

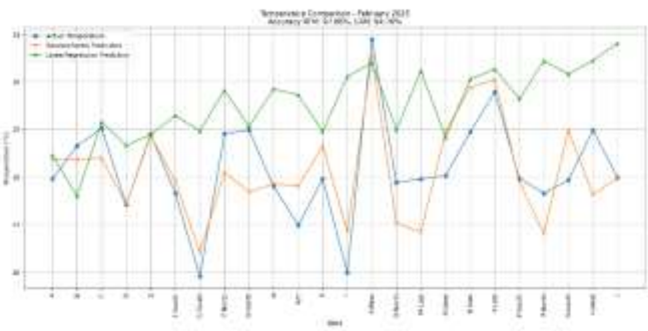

Fig. 4 Actual vs Predicted Temperature of February

It compares the actual and predicted values from the Random Forest Model (RFM) and the linear regression model (LRM) across different wards for February 2025. The RFM gives an accuracy of 97.86%, while the LRM achieved 94.70%. The RFM was better at closely matching the actual temperature values throughout the wards.

5. CONCLUSION

We used machine learning models to predict temperatures across Mumbai wards based on geographic location. We process historical temperature records with latitude and longitude, and use two models, Linear Regression and Random Forest Regression, for training and testing. So, the Result shows that Linear Regression gives less accuracy, and the Random Forest model performs better and has high accuracy for complex patterns too. The predicted temperatures were visualized as heatmaps. It helps to identify urban heat zones to provide useful insights into temperature.

We can explore new techniques using real-time weather analysis, using IoT sensors or time-based satellite feeds for monitoring temperature in cities. Having higher-resolution spatial data and time-based data could make predictions more accurate at the neighborhood level. More spatial features like Humidity, Rainfall, windspeed, etc, could help further improve model accuracy. We can use Advanced techniques such as Convolutional Neural Networks (CNNs) for analyzing previously generated heatmap images. We can also apply this technique to other cities.

REFERENCES

[1] Zeynali, R., Mandanici, E., Sohrabi, A. H., Trevisiol, F., & Bitelli, G. (2024, June). GIS-Based Urban Heat Island Mapping and Analysis: Experiences in the City of Bologna. In 2024 IEEE International Workshop

[2] Dabire, N., Ezin, E. C., & Firmin, A. M. (2024, August). Water Quality Assessment Using Normalized Difference Index by Applying Remote Sensing Techniques: Case of Lake Nokoue. In 2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC) (pp. 1-6). IEEE.

[3] Xu, L., Cai, R., Yu, H., Du, W., Chen, Z., & Chen, N. (2024). Monthly NDVI prediction using spatial autocorrelation and nonlocal attention networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 3425-3437.

[4] Chauhan, S., Jethoo, A. S., & Das, U. K. (2023, February). Duo satellite-based surface temperature comparative study of Jaipur city using soft computing. In 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON) (pp. 1-5). IEEE.

[5] Li, J., & Ou, Z. (2023, May). Construction of urban ecological environment detection system based on gis and rs image processing algorithm. In 2023 International Conference on Networking, Informatics and Computing (ICNETIC) (pp. 684-688). IEEE.

[6] Morale, D., Falchi, U., Mercogliano, P., & Parente, C. (2024, October). GIS Based Analysis and Accuracy Estimation of Sea Surface Temperature from MODIS Thermal Images. In 2024 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea) (pp. 541-545). IEEE.

[7] Meng, D., Gong, H., Li, X., & Zhou, D. (2012, June). Trends in temperature and extreme temperature over the Beijing-Tianjin-Hebei Metropolitan Region During 1957-2009. In 2012 20th International Conference on Geoinformatics (pp. 1-6). IEEE.

[8] Shanmugapriya, E. V., & Geetha, P. (2017, April). A framework for the prediction of land surface temperature using artificial neural network and vegetation index. In 2017 International Conference on Communication and Signal Processing (ICCSP) (pp. 1313-1317). IEEE.

[9] Mohapatra, S., Kundu, M., & Mohanty, S. (2022, September). Climate Downscaling and Prediction Using GIS-Based Machine Learning. In 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.

[10] Sahana, M., Dutta, S., & Sajjad, H. (2019). Assessing land transformation and its relation with land surface temperature in Mumbai city, India, using geospatial techniques. International Journal of Urban Sciences, 23(2), 205-225.

[11] Kaggle, "Mumbai Weather Data (27 Years)," [Online]. Available: https://www.kaggle.com/datasets/kevinnadar22/mumbai-weather-data-27-years

[12] Visual Crossing, "Historical Weather Data for Mumbai," [Online]. Available: https://www.visualcrossing.com/weather-history/Mumbai,%20MH,%20India/us/last15days/

[13] Open City, "Mumbai Daily Temperature Data (1951–2024)," [Online]. Available: https://data.opencity.in/dataset/daily-temperature-70-years-data-for-major-indian-cities/resource/mumbai-daily-temperature-data-1951-to-2024

[14] NASA, "POWER Data Access Viewer," [Online].Available: https://power.larc.nasa.gov/data-access-viewer/