iConclusions https://iconclusions.com/ Conclusions in Engineering

Improved Temperature Trend Forecasting in Major Pakistani Cities Using XGBoost and other Ensemble Machine Learning Models

Muhammad Bilal,[1, a)] Mubashir Mumtaz,[2] Memoona Ashraf,[3] Abdullah Nasir,[4] Hanan Nasir,[2] and Musyab Raza[1]

[1)] *Department of Mathematics and Statistics, The University of Lahore, Sargodha Campus, Pakistan*

[2)] *IMBB - Institute of Molecular Biology and Biotechnology. The University of Lahore, Sargodha Campus, Pakistan*

[3)] *Department of Computer Science, The University of Lahore, Sargodha Campus, Pakistan*

[4)] *Department of Computing & IT University of Sargodha, Pakistan*

ABSTRACT: This study employs advanced machine learning algorithms Random Forest, Gradient Boosting Machines (GBM), and XGBoost to predict normalized temperatures in major Pakistani cities, focusing on Lahore. Using a comprehensive dataset divided into training and testing subsets, model performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). Results show that XGBoost excelled with the lowest RMSE (0.0281), MAE (0.0219), and highest R^2(0.9879), demonstrating superior predictive accuracy. Gradient Boosting and Random Forest also performed well, confirming the efficacy of ensemble methods for temperature forecasting. This study highlights the potential of machine learning in environmental analysis, offering a scalable framework for climate predictions and informed urban planning.

Received: 02 December 2024
Accepted: 02 March 2025
DOI: https://doi.org/10.71107/j2djza91

I. INTRODUCTION

This research addresses the critical issue of climate change, focusing on temperature forecasting for Pakistan's major cities—Lahore, Islamabad, Karachi, Multan, Quetta, and Faisalabad. These cities, representing diverse climatic and geographical regions, experience temperature variations influenced by global warming, urbanization, and natural factors[1,2]. Temperature predictions are crucial for addressing the effects of climate change in agriculture, urban development, energy consumption, and disease prevention[3,4]. Traditional statistical methods have struggled to analyze temperature data due to its complexity. To this end, this study employs Random Forest, Gradient Boosting, and XGBoost machine learning algorithms suitable for multivariate analysis, identifying interaction effects, and mitigating overfitting[5,6]. Past daily temperature trends over the last five years are analyzed to identify patterns and make future projections. These models combine results from various base models, yielding improved predictions[7]. The socio-economic significance of temperature forecasting forms the basis of this study's focus. In agriculture, it can assist farmers in selecting appropriate crop varieties, determining optimal sowing or harvesting times, and efficiently utilizing water resources[8]. Urban planners can use this data to develop climate-resilient structures, such as heat-emitting buildings and green spaces[9]. In energy management, temperature predictions help forecast electricity load demand during extreme weather, while healthcare workers can proactively address heat-related ailments, particularly among vulnerable populations[10]. This study aims to enhance the effectiveness of temperature predictions to aid policymakers in developing region-specific climate strategies[11]. Thus, it bridges the gap between traditional approaches and state-of-the-art machine learning, setting a new standard for temperature trend analysis in

[a)] Electronic mail: muhammad.bilal@math.uol.edu.pk

Pakistan[12]. The study also offers insights applicable to other developing countries facing similar climatic challenges and encourages further research and real-world applications of ensemble machine learning across diverse climates[11,13]. Finally, this work enhances the productivity of data analysis in addressing climate change and related issues, promoting sustainable development and further research toward effective solutions to increasing climate variability[6].

II. METHODOLOGY

A. Research Objectives

The primary objectives of this study are:
1. To develop and evaluate predictive models for temperature forecasting in Pakistan's major cities.
2. To compare the effectiveness of Random Forest, Gradient Boosting, and XGBoost models in temperature prediction.
3. To identify key factors contributing to the superior performance of a given model.
4. To assess how the study's results align with or differ from existing literature on temperature forecasting.

B. Justification for Machine Learning Models

Traditional statistical methods struggle to analyze temperature data due to its complexity and non-linearity. Machine learning models such as Random Forest, Gradient Boosting, and XGBoost offer advantages in handling multivariate analysis, identifying interaction effects, and reducing overfitting[6,14]. These models combine the results of multiple base learners to improve predictions[7]. While neural networks and support vector machines (SVM) are also viable options, they often require extensive computational resources and fine-tuning. XGBoost, in particular, is known for its efficiency and high predictive accuracy in structured datasets[9]. This study evaluates whether XGBoost outperforms other models and discusses its advantages compared to deep learning methods.

C. Contribution to Literature

This study aims to bridge the gap between traditional statistical approaches and modern machine learning techniques in temperature forecasting. While previous research has demonstrated the effectiveness of machine learning for climate modeling, limited studies have focused on Pakistan's climatic conditions. Additionally, few studies have systematically compared ensemble models with deep learning methods for temperature prediction in the region. By evaluating and comparing the models, this research establishes a benchmark for future studies and practical applications in climate-related decision-making.

III. RESULTS AND DISCUSSIONS

This section provides detailed results of the temperature data collected from six leading cities of Pakistan. Thus, the concentration is made only on descriptive statistics, normalization for obtaining better comparison, correlation, and visualization to identify the temperature trend and the relation between them.

A. Dataset Overview

The dataset comprises 1,828 observations and seven columns, including the date and daily mean temperature data for six major cities in Pakistan: Lahore, Islamabad, Karachi, Multan, Quetta, Faisalabad. The filter column called the date has been shifted to the proper data type known as Date to improve data filtering and analysis. The temperature variables are the firstdegree Celsius averaged daily mean temperatures for these cities and thus the analysis of these variables captures the changes in temperatures done at more than one location. This set of data can prove useful to investigate temperature fluctuations seen in Table I and their dynamics over periods necessary for climate research in the context of the Pakistani climate.

B. Temperature Trends

Seasonal fluctuations as obtained from diurnal temperature data show different patterns for the cities. The temperature variation that Karachi has is a minimum showing that it has a moderate coastal climate. Whereas, Lahore, Multan and Faisalabad have more variation during the year with high temperature fluctuations showing the symptoms of extreme climate. Out of all the analyzed cities, Quetta has the highest altimetry and the lowest average temperatures.

C. Correlation Analysis

The correlation matrix was computed in Table II to understand inter-city temperature relationships.

From the result it is concluded that Lahore and Multan have the highest coefficient (+0.96 meaning that climate of both cities are correlated)., while it is rather weak when it comes to the inland cities which have different climate conditions than Karachi due to the coastal continental climate.

TABLE I: Temperatures Summary Statistics.

City	Mean (°C)	Standard Deviation (°C)	Min (°C)	Max (°C)
Lahore	24.67	8.37	7.54	40.56
Islamabad	21.58	7.67	5.16	37.98
Karachi	26.15	3.95	14.80	34.55
Multan	27.13	8.76	7.72	41.35
Quetta	15.92	8.48	-5.81	31.25
Faisalabad	25.47	8.68	6.93	40.98

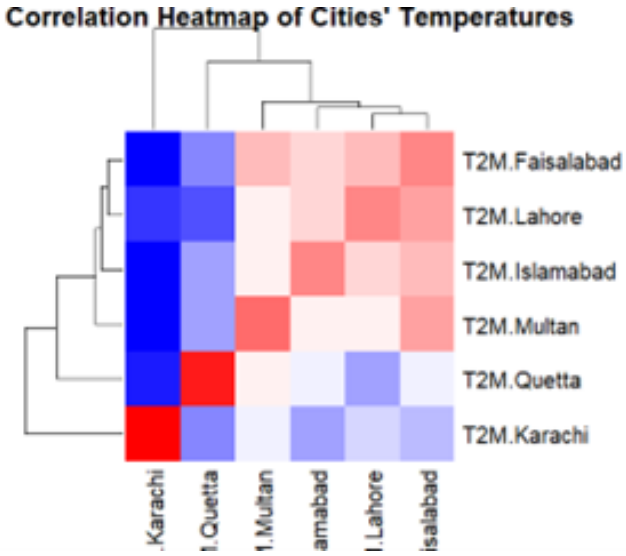

FIG. 1: Correlation Heatmap.

D. Heatmap of Correlations

A heatmap was generated for visualizing the correlation matrix, highlighting the strongest (dark red) and weakest (light blue) correlations in Fig. 1.

The temperature differences of the coastal and inland cities are totally different showing that geography plays an important role in climate (Fig. 2). The tests showing high value indicate the regional climate relationship of Moreno with certain cities like Lahore and Multan. Normalization allows comparing data from different locations and at different times more effectively, and an understanding of temperature fluctuations is visually clear. R's code implementation was used effectively for data preprocessing, analysis and visualization and the developed platform encapsulates a strong foundation to build upon for further study.

E. Evaluating Predictive Models for Lahore's Normalized Temperature Data

For this work, the $T2M_Lahore_norm$ dataset with normalized temperature data of the Pakistani cities was employed with Random Forest (RF), Gradient Boosting (GBM), and XGBoost predictive models. After preprocessing. Hence, the models were trained, tested and a comparative evaluation was made on varying training-test split to the developing 20-70-10 split set of the data. RF made use of the largest amount of decision trees through boot strapping while GBM constructed many weak models in a sequence to reduce prediction mistake. Logistic regression trees were used in XGBoost, which stands for extra Gradient Boosting, as the base of GBM were improved for efficiency. Prediction accuracy indicators such as root mean square error (RMSE, the lower the better precisions), mean absolute error (MAE), and the coefficient of determination (R^2, the closer to 1 , the better) were used.

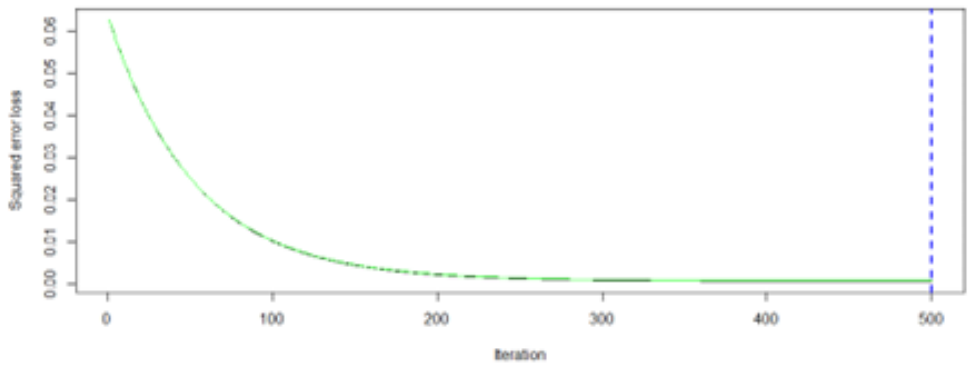

FIG. 2: No. of Iteration(trees) to reduce the error.

TABLE II: Correlation Table.

	Lahore	Islamabad	Karachi	Multan	Quetta	Faisalabad
Lahore	1.00	0.89	0.71	0.96	0.69	0.95
Islamabad	0.89	1.00	0.68	0.86	0.74	0.87
Karachi	0.71	0.68	1.00	0.70	0.55	0.73
Multan	0.96	0.86	0.70	1.00	0.71	0.96
Quetta	0.69	0.74	0.55	0.71	1.00	0.72
Faisalabad	0.95	0.87	0.73	0.96	0.72	1.00

F. Model Evaluation

The evaluation results are summarized in Table III. XGBoost was the most accurate model with lowest RMSE of 0.0281 and MAE 0.0219 and highest R^2 of 0.9879 , therefore the model recommended for temperature prediction. Random forest also had RMSE and MAE substantially greater than XGBoost though slightly better than a constant model, which indicates slightly worse fit in terms of Random Forest model. As expected, Gradient Boosting was similarly efficient though it demonstrated slightly worse accuracy compared to XGBoost and Random Forest in terms of error rate.

TABLE III: Model Evaluation Metrics.

Model	RMSE	MAE	$\mathbf{R^2}$
Random Forest	0.0284	0.0225	0.9877
Gradient Boosting	0.0294	0.0238	0.9870
XGBoost	**0.0281**	**0.0219**	**0.9879**

IV. CONCLUSION

This study highlights the effectiveness of machine learning models in temperature forecasting for Pakistan's major cities, with XGBoost emerging as the most accurate model, surpassing Random Forest and Gradient Boosting in predictive performance. The findings emphasize the significance of ensemble learning techniques in climate modeling and forecasting. XGBoost demonstrated superior accuracy due to its advanced optimization and regularization techniques, while Lahore and Multan exhibited the strongest temperature correlation, reflecting similar climatic patterns. Karachi's coastal influence resulted in lower temperature fluctuations compared to inland cities. The study's outcomes align with previous research, confirming the efficiency of XGBoost for structured climate data. Improved forecasting can benefit various sectors, including agriculture by aiding farmers in planning crop cycles and resource allocation, urban planning by supporting infrastructure resilience strategies, energy management by optimizing energy distribution, and healthcare by mitigating heat-related health risks. Future research should explore hybrid machine learning models integrating deep learning

techniques to enhance temperature prediction accuracy. By setting a benchmark for regional climate forecasting, this study contributes to better decision-making in climate adaptation and policy planning.

DECLARATION OF COMPETING INTEREST

The authors have no conflicts to disclose.

REFERENCES

[1] A. Ali, D. B. Rahut, K. A. Mottaleb, and O. Erenstein, "Impacts of changing weather patterns on smallholder well-being: Evidence from the himalayan region of northern pakistan," International Journal of Climate Change Strategies and Management **9**, 225–240 (2017).

[2] L. Shahzad, A. Tahir, F. Sharif, W. U. D. Khan, M. A. Farooq, A. Abbas, and Z. A. Saqib, "Vulnerability, well-being, and livelihood adaptation under changing environmental conditions: A case from mountainous region of pakistan," Environmental Science and Pollution Research **26**, 26748–26764 (2019).

[3] A. Tariq, F. Mumtaz, X. Zeng, M. Y. J. Baloch, and M. F. U. Moazzam, "Spatio-temporal variation of seasonal heat islands mapping of pakistan during 2000–2019, using day-time and night-time land surface temperatures modis and meteorological stations data," Remote Sensing Applications: Society and Environment **27**, 100779 (2022).

[4] G. Ali, Y. Bao, W. Ullah, S. Ullah, Q. Guan, X. Liu, and J. Ma, "Spatiotemporal trends of aerosols over urban regions in pakistan and their possible links to meteorological parameters," Atmosphere **11**, 306 (2020).

[5] M. Kolambe and S. Arora, "Forecasting the future: A comprehensive review of time series prediction techniques," Journal of Electrical Systems **20**, 575–586 (2024).

[6] D. Li, G. Yan, F. Li, H. Lin, H. Jiao, H. Han, and W. Liu, "Optimized machine learning models for predicting core body temperature in dairy cows: Enhancing accuracy and interpretability for practical livestock management," Animals **14**, 2724 (2024).

[7] S. Babu Nuthalapati and A. Nuthalapati, "Accurate weather forecasting with dominant gradient boosting using machine learning," International Journal of Scientific Research Archive **12**, 408–422 (2024).

[8] N. Khan and S. Shahid, "Urban heat island effect and its drivers in large cities of pakistan," Theoretical and Applied Climatology , 1–20 (2024).

[9] J. Cifuentes, G. Marulanda, A. Bello, and J. Reneses, "Air temperature forecasting using machine learning techniques: A review," Energies **13**, 4215 (2020).

[10] P. Anandan and A. Sundaram, "Unveiling agricultural soil runoff: Remote sensing and ensemble deep learning models to evaluate impact of climate on water quality and human health," Remote Sensing in Earth Systems Sciences , 1–16 (2024).

[11] H. Farman, N. Islam, S. A. Ali, D. Khan, H. A. Khan, M. Ahmed, and A. Farman, "Advancing rainfall prediction in pakistan: A fusion of machine learning and time series forecasting models," International Journal of Emerging Engineering and Technology **3**, 17–24 (2024).

[12] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, and Y. Bengio, "Tackling climate change with machine learning," ACM Computing Surveys (CSUR) **55**, 1–96 (2022).

[13] M. R. Khan, M. Abubakar, A. Tahir, M. W. Dilawar, H. M. A. Hassan, S. R. Ahmad, and M. U. Chand, "Escalating global threat of heatwaves and policy options for adaptation and mitigation," Journal of Asian Development Studies **13**, 980–995 (2024).

[14] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," Engineering Applications of Artificial Intelligence **115**, 105151 (2022).

[15] L. Shahzad, A. Tahir, F. Sharif, W. U. D. Khan, M. A. Farooq, A. Abbas, and Z. A. Saqib, "Vulnerability, well-being, and livelihood adaptation under changing environmental conditions: A case from mountainous region of pakistan," Environmental Science and Pollution Research **26**, 26748–26764 (2019).

[16] Z. J. Khan, "A machine learning analysis of climate change & human health projections in pakistan," (2024).

[17] Y. Lai and D. A. Dzombak, "Use of the autoregressive integrated moving average (arima) model to forecast near-term regional temperature and precipitation," Weather and Forecasting **35**, 959–976 (2020).

[18] A. G. Salman and B. Kanigoro, "Visibility forecasting using autoregressive integrated moving average (arima) models," Procedia Computer Science **179**, 252–259 (2021).

[19] J. Cifuentes, G. Marulanda, A. Bello, and J. Reneses, "Air temperature forecasting using machine learning techniques: a review," Energies **13**, 4215 (2020).

[20] A. Manoharan, K. M. Begam, V. R. Aparow, and D. Sooriamoorthy, "Artificial neural networks, gradient boosting and support vector machines for electric vehicle battery state estimation: A review," Journal of Energy Storage **55**, 105384 (2022).

[21] F. Zennaro, E. Furlan, C. Simeoni, S. Torresan, S. Aslan, A. Critto, and A. Marcomini, "Exploring machine learning potential for climate change risk assessment," Earth-Science Reviews **220**, 103752 (2021).

[22] N. Khan, S. Shahid, T. B. Ismail, and F. Behlil, "Prediction of heat waves over pakistan using support vector machine algorithm in the context of climate change," Stochastic Environmental Research and Risk Assessment **35**, 1335–1353 (2021).

[23] N. Pachauri and C. W. Ahn, "Regression tree ensemble learning-based prediction of the heating and cooling loads of residential buildings," Building Simulation **15**, 2003–2017 (2022).

[24] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: A review," Journal of Hydrology **598**, 126266 (2021).

[25] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, "Ensemble deep learning in bioinformatics," Nature Machine Intelligence **2**, 500–508 (2020).

[26] A. A. Heidari, M. Akhoondzadeh, and H. Chen, "A wavelet pm2.5 prediction system using optimized kernel extreme learning with boruta-xgboost feature selection," Mathematics **10**, 3566 (2022).

[27] Y. Mao, Y. Li, F. Teng, A. K. Sabonchi, M. Azarafza, and M. Zhang, "Utilizing hybrid machine learning and soft computing techniques for landslide susceptibility mapping in a drainage basin," Water **16**, 380 (2024).

[28] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, and G. W. Wei, "Machine learning methods for small data challenges in molecular science," Chemical Reviews **123**, 8736–8780 (2023).

[29] X. Li, Z. Li, W. Huang, and P. Zhou, "Performance of statistical and machine learning ensembles for daily temperature downscaling," Theoretical and Applied Climatology **140**, 571–

588 (2020).
[30]R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning-based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," Engineering Reports **5**, e12599 (2023).
[31]Q. Wang, X. Wang, Y. Zhou, D. Liu, and H. Wang, "The dominant factors and influence of urban characteristics on land surface temperature using random forest algorithm," Sustainable Cities and Society **79**, 103722 (2022).
[32]A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," Machine Learning with Applications **7**, 100204 (2022).
[33]L. Chen, B. Han, X. Wang, J. Zhao, W. Yang, and Z. Yang, "Machine learning methods in weather and climate applications: A survey," Applied Sciences **13**, 12019 (2023).
[34]R. Wang, Z. He, H. Chen, S. Guo, S. Zhang, K. Wang, and S. H. Ho, "Enhancing biomass conversion to bioenergy with machine learning: Gains and problems," Science of The Total Environment , 172310 (2024).
[35]X. Lv, J. Luo, Y. Zhang, H. Guo, M. Yang, M. Li, and R. Jing, "Unveiling diagnostic information for type 2 diabetes through interpretable machine learning," Information Sciences **690**, 121582 (2025).