Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

Understanding the Interrelation Between Temperature and Meteorological Factors: A Case Study of Szeged Using Machine Learning Techniques

CHE, Chang [1*] TIAN, Junchi [1]

[1] The George Washington University, USA

**** CHE, Chang is the corresponding author, E-mail: cche57@gwmail.gwu.edu***

Abstract: Temperature serves as a fundamental indicator of thermal conditions, influencing various natural processes and human activities. This study investigates the relationship between temperature and other meteorological factors, including humidity, wind speed, visibility, pressure, and apparent temperature, using historical weather data from Szeged, Hungary (2006-2016). Employing multiple regression models and advanced machine learning algorithms such as XGBoost and Artificial Neural Networks (ANNs), the research aims to elucidate the linear and non-linear dependencies of temperature on these factors. The findings indicate a significant linear correlation, with XGBoost outperforming traditional regression approaches in predicting temperature variations. This study contributes to enhancing temperature forecasting accuracy, which is crucial for improving quality of life and informing climate-related decision-making processes.

Keywords: Machine Learning, Artificial Neural Networks, Regression Model.

Disciplines: Computer Science. **Subjects:** Machine Learning.

DOI: https://doi.org/10.5281/zenodo.13924235 **ARK:** https://n2t.net/ark:/40704/JCTAM.v1n4a06

1 INTRODUCTION

The relationship between temperature and meteorological factors is of significant importance in understanding both short-term weather changes and long-term climate patterns. Accurate temperature forecasting has practical implications for several domains, including agriculture, urban planning, and public health. Meteorological factors such as humidity, wind speed, and pressure not only directly influence thermal comfort but also impact various processes ranging from chemical reactions to transportation efficiency [16]. Moreover, a deeper understanding of how these factors interrelate can inform adaptive measures for mitigating climate change impacts, particularly in urban areas experiencing heat stress. In this study, we aim to leverage advanced data science tools to analyze these relationships, emphasizing the importance of both linear and non-linear modeling approaches in enhancing forecasting accuracy.

The city of Szeged, Hungary, presents an ideal case study due to its varied climatic conditions and historical weather records spanning multiple years. By using a combination of traditional statistical models and machine learning algorithms, this study endeavors to build robust temperature forecasting models that cater to the unique geographical and meteorological characteristics of the region. Ultimately, this research contributes to improved weather predictions and aims to support data-driven decision-making for urban climate adaptation strategies.

This study seeks to address the following research questions:

1. How can temperature be accurately forecasted in a specific geographical area?
2. What is the nature of the relationship between temperature and other meteorological conditions?
3. What insights can this relationship provide to the public and relevant stakeholders?

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on the use of machine learning and statistical models in temperature forecasting and weather prediction. Section 3 details the methodology, including data collection, preprocessing, and model development processes. Section 4 presents the results obtained from both regression-based and advanced machine learning models, along with a discussion of their comparative performance. Section 5 concludes the study by summarizing key findings, discussing limitations, and suggesting future research directions.

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

2 LITERATURE REVIEW

The application of machine learning (ML) in predicting temperature and managing meteorological data has garnered significant attention in recent years. [13] explored various ML techniques, including linear and nonlinear regressions, and clustering methods like K-means and DBSCAN, to forecast indoor temperatures. Their findings highlighted the superior performance of Non-linear Autoregressive Exogenous Multilayer Perceptron in regression tasks, although clustering methods did not enhance predictive accuracy significantly.

[23] introduced a linear stochastic model for forecasting daily soil temperature (DST) through time-series analysis, demonstrating improved performance over multilayer perceptron neural networks. [3] employed the Weather Research and Forecasting (WRF) model to predict urban temperatures in Beirut, achieving optimal outcomes for key urban parameters but acknowledging residual uncertainties.

[22] developed a Building-Resolved Temperature (BRT) model for South Korea, incorporating spatial information to enhance temperature forecasts. Despite its detailed spatial considerations, the model exhibited inconsistent performance due to spatial distribution variability.

Accurate prediction of weather has enormous benefits in building smart cities, improving resilience of infrastructure, alleviating traffic congestions (e.g., [18]-[21]), and transportation planning (e.g., [9]-[11]), post-disaster evaluation [8], delivery (e.g., [12], [25], [24]), improving public transit service [2], and streamlining air traffic (e.g., [6], [7], [14], [4]). These studies underscore the versatility of ML methodologies in diverse applications, highlighting their potential for enhancing predictive models in meteorological contexts as well.

However, existing studies often face limitations related to spatial conditions, dataset quality, and model uncertainties, indicating a need for more robust and comprehensive approaches in temperature forecasting.

3 3 METHODOLOGY

The dataset utilized in this study is sourced from Kaggle and comprises historical weather records of Szeged, Hungary, spanning from 2006 to 2016. The dataset includes hourly and daily measurements of temperature, wind speed, humidity, visibility, pressure, and apparent temperature. To facilitate numerical analysis, irrelevant columns such as time-series data and textual descriptions were removed. Data preprocessing involved eliminating null values and outliers to ensure data integrity and suitability for subsequent analyses. Exploratory data analysis through boxplots revealed normally distributed variables, and correlation analyses indicated strong linear and inverse relationships among the variables, suggesting suitability for regression-based modeling.

This research employs both regression-based and advanced machine learning approaches to model the relationship between temperature and other weather factors.

3.1 Regression Models

3.1.1 Variance Inflation Factor (VIF)

To address multicollinearity among independent variables, VIF analysis was conducted with a threshold of 2. This process led to the exclusion of the 'Apparent Temperature (C)' variable due to high collinearity.

3.1.2 Linear Regression (LR)

A multiple linear regression model was developed to explore the linear dependencies between temperature and the remaining five weather factors. The model was specified as:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Where, Y indicates the dependent variable temperature in Szeged.

3.1.3 Principal Component Regression (PCR)

Recognizing potential limitations of VIF in fully mitigating multicollinearity, PCR was employed as an alternative dimension reduction technique. The process involved standardizing the data, performing PCA to extract principal components, and subsequently conducting linear regression using the selected components. Two principal components were retained based on the eigenvalue distribution, enhancing model efficiency and reducing computational demands.

3.2 Advanced Machine Learning Models

3.2.1 XGBoost

Extreme Gradient Boosting (XGBoost) was selected for its proven accuracy and scalability in predictive modeling [17]. Its ability to handle complex, non-linear relationships makes it a suitable candidate for temperature forecasting.

3.2.2 Artificial Neural Network (ANN)

An ANN, specifically a three-layer Fully Connected Neural Network (FCNN), was constructed to model temperature dependencies. Data normalization using Min-Max scaling was performed to optimize training efficiency. The network architecture included an input layer with six nodes, a hidden layer with ten nodes, and an output layer predicting temperature. The Adam optimizer and Mean Squared Error (MSE) loss function were utilized over twenty training epochs.

4 RESULT

4.1 Regression Models

The linear regression models exhibited varying degrees of multicollinearity, with VIF-adjusted models showing reduced multicollinearity issues. The R-squared value

decreased from 0.991 to 0.548 post-VIF adjustment, indicating diminished explanatory power. Conversely, the PCR model achieved an R-squared of 0.770, outperforming the VIF-adjusted LR model by effectively mitigating multicollinearity without substantial loss in fit. Residual analysis confirmed normal distribution patterns across all regression models, validating their suitability.

4.2 Advanced Machine Learning Models

Both XGBoost and FCNN were evaluated on training (80%) and testing (20%) datasets. XGBoost achieved an R-squared of 0.999585, significantly surpassing FCNN's 0.936483. Prediction versus actual temperature plots for both models demonstrated high accuracy, with minimal Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), underscoring their capability to capture temperature variations effectively.

Among the regression approaches, PCR emerged as the most effective model, balancing complexity and explanatory power with an R-squared of 0.770. Advanced machine learning models, particularly XGBoost, outperformed traditional regression methods, achieving near-perfect predictive accuracy. Despite FCNN's lower R-squared, its lower MSE and RMSE values indicate strong predictive performance, though XGBoost's superior fit and faster training times render it more advantageous for this regression-focused study.

The superior performance of XGBoost suggests that linear relationships heavily influence temperature dynamics, despite the potential presence of non-linear interactions. This aligns with findings from [15] and [1], who demonstrate the efficacy of advanced ML techniques in complex predictive tasks across various domains.

However, the study has limitations, including the absence of temporal and spatial autocorrelation considerations and the focus on a single geographic location. Future research should explore spatio-temporal models and extend analyses to multiple regions to enhance generalizability and robustness.XGBoost's significantly higher R-squared value (0.999585) compared to the Fully Connected Neural Network's (FCNN) R-squared (0.936483) highlights its superior ability to model temperature data, especially in regression tasks where precise predictive accuracy is paramount. This remarkable performance suggests that XGBoost excels at capturing the dominant linear relationships present in the dataset. The extremely high R-squared value indicates that XGBoost can almost perfectly explain the variance in the temperature data, making it a standout choice for regression analysis. Moreover, its faster training times make it even more desirable in practical scenarios where computational efficiency is crucial, particularly in large-scale or real-time applications.

In contrast, although FCNN's R-squared is lower than that of XGBoost, it still demonstrates strong predictive performance. The lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values for FCNN indicate that it captures temperature variations effectively, even though its ability to explain the variance in the data is less comprehensive than XGBoost. FCNN's ability to achieve lower error metrics may be attributed to its flexibility in modeling more complex, non-linear interactions within the dataset. Neural networks, like FCNNs, are designed to handle a wide range of relationships, including non-linear patterns, and often perform well in tasks where complex data interactions are prevalent. However, in this case, the temperature data may have stronger linear components, which is likely why XGBoost, a model that excels in linear relationships, performed better.

This performance difference highlights an important aspect of regression modeling—choosing the right algorithm based on the underlying data structure. XGBoost, as a gradient boosting decision tree model, inherently favors datasets where linear relationships dominate, while FCNNs are generally more suited to capturing complex, non-linear dynamics. In this study, the dominant linear relationships in temperature dynamics provided a perfect opportunity for XGBoost to shine, as reflected in its near-perfect R-squared value and superior predictive accuracy.

In addition to XGBoost and FCNN, the study also evaluated Principal Component Regression (PCR), which, despite being a traditional regression technique, performed reasonably well with an R-squared value of 0.770. PCR balances the complexity and explanatory power, offering a viable option for regression analysis. However, compared to the more advanced machine learning models, PCR's performance lagged behind. The lower R-squared value suggests that PCR is not as adept at capturing the nuances in the temperature data, which could include both linear and non-linear interactions. This result is not unexpected, as PCR simplifies the data by reducing its dimensionality through principal components, potentially leading to a loss of information that more advanced models like XGBoost can retain and utilize.

Despite the strong performance of XGBoost and FCNN, the study acknowledges certain limitations that should be addressed in future research. One key limitation is the lack of consideration for temporal and spatial autocorrelation. Temperature data, particularly in geographic studies, often exhibit temporal dependencies, where past values influence future values, as well as spatial correlations, where nearby geographic locations share similar temperature patterns. These factors can introduce biases if not appropriately modeled. By not accounting for these autocorrelations, the study's findings may be less robust when applied to other contexts where spatio-temporal dynamics are critical. Future studies could benefit from integrating models that specifically handle these types of data dependencies, such as spatio-temporal models, to ensure more generalized and reliable predictions.

Another limitation is the focus on a single geographic

location. While the study's findings are insightful, they may not be generalizable to other regions with different climate patterns or temperature dynamics. Temperature behavior can vary significantly across regions, driven by factors such as altitude, proximity to water bodies, or unique weather patterns. As such, future research should extend the analysis to multiple regions to assess the broader applicability of the findings. Including diverse geographic locations in the analysis would provide a more comprehensive understanding of the models' performance across varied climate conditions and enhance the robustness and generalizability of the conclusions drawn from the study.

5 CONCLUSION

This investigation elucidates the complex relationships between temperature and key meteorological factors in Szeged, Hungary, utilizing both regression and advanced machine learning methodologies. The analysis reveals that temperature has a positive correlation with apparent temperature, wind speed, and visibility, while showing an inverse relationship with humidity and pressure. Among the various models employed, advanced machine learning models, particularly XGBoost, exhibited superior capability in capturing these complex interactions, resulting in highly accurate temperature forecasting.

The findings of this study hold significant implications for urban planning and climate mitigation efforts. Understanding the interplay between meteorological variables allows policymakers to implement more targeted strategies, such as increasing humidity through artificial rainfall or developing vegetation-based solutions to reduce urban temperatures and combat the heat island effect. Moreover, the superior performance of XGBoost highlights the effectiveness of sophisticated, data-driven approaches in managing the increasing complexities of climate change and urbanization.

Despite these promising outcomes, the study is limited by its focus on a single geographical area and the absence of temporal and spatial autocorrelation considerations. Future research should address these limitations by extending the analysis to different regions and incorporating spatio-temporal factors to enhance the robustness and generalizability of the findings. Furthermore, the inclusion of additional variables such as land-use characteristics, solar radiation, and other environmental factors could provide a more holistic understanding of temperature dynamics. Ultimately, this research lays a strong foundation for integrating machine learning into meteorological studies, offering a path forward for developing adaptive and precise climate forecasting systems capable of supporting effective urban and environmental planning initiatives.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

CHE, Chang

The George Washington University, US.

TIAN, Junchi

The George Washington University, US.

REFERENCES

[1] Cheng, X., & Lin, J. (2024). Is electric truck a viable alternative to diesel truck in long-haul operation? Transportation Research Part D: Transport and Environment, 129, 104119.

[2] Cheng, X., Nie, Y. M., & Lin, J. (2024). An autonomous modular public transit service. Transportation Research Part C: Emerging Technologies, 104746.

[3] Ghadban, M., Baayoun, A., Lakkis, I., Najem, S., Saliba, N., & Shihadeh, A. (2020). A novel method to improve temperature forecast in data-scarce urban environments with application to the Urban Heat Island in Beirut. Urban Climate, 33, 100648. https://doi.org/10.1016/j.uclim.2020.100648

[4] Guo, G., Li, X., Zhu, C., Wu, Y., Chen, J., Chen, P., & Cheng, X. (2025). Establishing benchmarks to determine the embodied carbon performance of high-speed rail systems. Renewable and Sustainable Energy Reviews, 207, 114924. https://doi.org/10.1016/j.rser.2021.114924

[5] Kuo, Y. H., Leung, J. M., & Yan, Y. (2023). Public transport for smart cities: Recent innovations and future challenges. European Journal of Operational Research, 306(3), 1001-1026. https://doi.org/10.1016/j.ejor.2023.01.023

[6] Liu, K., Ding, K., Cheng, X., Chen, J., Feng, S., Lin, H., ... & Zhu, C. (2024). Airport Delay Prediction with Temporal Fusion Transformers. arXiv preprint arXiv:2405.08293.

[7] Liu, T., & Meidani, H. (2024). End-to-end heterogeneous graph neural networks for traffic assignment. Transportation Research Part C: Emerging Technologies, 165, 104695.

[8] Liu, T., & Meidani, H. (2024). Graph Neural Network Surrogate for Seismic Reliability Analysis of Highway Bridge Systems. Journal of Infrastructure Systems, 30(4), 05024004.

[9] Liu, T., & Meidani, H. (2024). Heterogeneous Graph Sequence Neural Networks for Dynamic Traffic Assignment. arXiv preprint arXiv:2408.04131.

[10] Liu, T., & Meidani, H. (2024). Neural network surrogate models for aerodynamic analysis in truck platoons: Implications on autonomous freight delivery. International Journal of Transportation Science and Technology (2024).

[11] Mateo, F., Carrasco, J., Sellami, A., Millán-Giraldo, M., Domínguez, M., & Soria-Olivas, E. (2013). Machine learning methods to forecast temperature in buildings. Expert Systems with Applications, 40(4), 1061-1068. https://doi.org/10.1016/j.eswa.2012.09.030

[12] Su, G., Cheng, X., Feng, S., Liu, K., Song, J., Chen, J., ... & Lin, H. (2024). Flight Path Optimization with Optimal Control Method. arXiv preprint arXiv:2405.08306.

[13] Ying, C., Chow, A. H., Yan, Y., Kuo, Y. H., & Wang, S. (2024). Adaptive rescheduling of rail transit services with short-turnings under disruptions via a multi-agent deep reinforcement learning approach. Transportation Research Part B: Methodological, 188, 103067.

[14] Yan, Y., Chow, A. H., Ho, C. P., Kuo, Y. H., Wu, Q., & Ying, C. (2022). Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities. Transportation Research Part E: Logistics and Transportation Review, 162, 102712. https://doi.org/10.1016/j.tre.2022.102712

[15] Yan, Y., Deng, Y., Cui, S., Kuo, Y. H., Chow, A. H., & Ying, C. (2023). A policy gradient approach to solving dynamic assignment problem for on-site service delivery. Transportation Research Part E: Logistics and Transportation Review, 178, 103260. https://doi.org/10.1016/j.tre.2023.103260

[16] Yan, Y., Wen, H., Deng, Y., Chow, A. H., Wu, Q., & Kuo, Y. H. (2024). A mixed-integer programming-based Q-learning approach for electric bus scheduling with multiple termini and service routes. Transportation Research Part C: Emerging Technologies, 162, 104570. https://doi.org/10.1016/j.trc.2023.104570

[17] Yan, Y., Cui, S., Liu, J., Zhao, Y., Zhou, B., & Kuo, Y. H. (2024). Multimodal fusion for large-scale traffic prediction with heterogeneous retentive networks. Information Fusion, 102695. https://doi.org/10.1016/j.inffus.2023.102695

[18] Yi, C., Shin, Y., & Roh, J. (2018). Development of an Urban High-Resolution Air Temperature Forecast System for Local Weather Information Services Based on Statistical Downscaling. Atmosphere, 9(5), 164. https://doi.org/10.3390/atmos9050164

[19] Zeynoddin, M., Bonakdari, H., Ebtehaj, I., Esmaeilbeiki, F., Gharabaghi, B., & Zare Haghi, D. (2019). A reliable linear stochastic daily soil temperature forecast model. Soil and Tillage Research, 189, 73-87. https://doi.org/10.1016/j.still.2019.02.001

[20] Cheng, X., Shen, H., Huang, Y., Cheng, Y. L., & Lin, J. (2024, February). Using Mobile Charging Drones to Mitigate Battery Disruptions of Electric Vehicles on Highways. In 2024 Forum for Innovative Sustainable Transportation Systems (FISTS) (pp. 1-6). IEEE.

[21] Cheng, X., Mamalis, T., Bose, S., & Varshney, L. R. (2024). On Carsharing Platforms With Electric Vehicles as Energy Service Providers. IEEE Transactions on Intelligent Transportation Systems.

[22] Che, C., Li, C., & Huang, Z. (2024). The Integration of Generative Artificial Intelligence and Computer Vision in Industrial Robotic Arms. International Journal of Computer Science and Information Technology, 2(3), 1-

9.

[23] Tianbo, S., Weijun, H., Jiangfeng, C., Weijia, L., Quan, Y., & Kun, H. (2023, January). Bio-inspired swarm intelligence: a flocking project with group object recognition. In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 834-837). IEEE.

[24] Xu, J., Jiang, Y., Yuan, B., Li, S., & Song, T. (2023, November). Automated Scoring of Clinical Patient Notes using Advanced NLP and Pseudo Labeling. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 384-388). IEEE.

[25] Zhang, Q., Cai, G., Cai, M., Qian, J., & Song, T. (2023). Deep Learning Model Aids Breast Cancer Detection. Frontiers in Computing and Intelligent Systems, 6(1), 99-102.

[26] Xu, X., Yuan, B., Song, T., & Li, S. (2023, November). Curriculum recommendations using transformer base model with infonce loss and language switching method. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 389-393). IEEE.

[27] Yuan, B., & Song, T. (2023, November). Structural Resilience and Connectivity of the IPv6 Internet: An AS-level Topology Examination. In Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering (pp. 853-856).

[28] Yuan, B., Song, T., & Yao, J. (2024, January). Identification of important nodes in the information propagation network based on the artificial intelligence method. In 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 11-14). IEEE.

[29] Qian, J., Song, T., Zhang, Q., Cai, G., & Cai, M. (2023). Analysis and Diagnosis of Hemolytic Specimens by AU5800 Biochemical Analyzer Combined With AI Technology. Frontiers in Computing and Intelligent Systems, 6(3), 100-103.

[30] Cai, G., Qian, J., Song, T., Zhang, Q., & Liu, B. (2023). A deep learning-based algorithm for crop Disease identification positioning using computer vision. International Journal of Computer Science and Information Technology, 1(1), 85-92.

[31] Song, T., Zhang, Q., Cai, G., Cai, M., & Qian, J. (2023). Development of Machine Learning and Artificial Intelligence in Toxic Pathology. Frontiers in Computing and Intelligent Systems, 6(3), 137-141.

[32] Liu, B., Cai, G., Qian, J., Song, T., & Zhang, Q. (2023). Machine Learning Model Training and Practice: A Study on Constructing a Novel Drug Detection System. International Journal of Computer Science and Information Technology, 1(1), 139-146.

[33] Che, C., Lin, Q., Zhao, X., Huang, J., & Yu, L. (2023, September). Enhancing Multimodal Understanding with CLIP-Based Image-to-Text Transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies (pp. 414-418).

[34] Che, C., Hu, H., Zhao, X., Li, S., & Lin, Q. (2023). Advancing Cancer Document Classification with R andom Forest. Academic Journal of Science and Technology, 8(1), 278-280.

[35] Huang, Z., Zheng, H., Li, C., & Che, C. (2024). Application of machine learning-based k-means clustering for financial fraud detection. Academic Journal of Science and Technology, 10(1), 33-39.

[36] Huang, Z., Che, C., Zheng, H., & Li, C. (2024). Research on Generative Artificial Intelligence for Virtual Financial Robo-Advisor. Academic Journal of Science and Technology, 10(1), 74-80.

[37] Che, C., Huang, Z., Li, C., Zheng, H., & Tian, X. (2024). Integrating generative AI into financial market prediction for improved decision making. Applied and Computational Engineering, 64, 155-161.

[38] Liu, H., Wang, C., Zhan, X., Zheng, H., & Che, C. (2024). Enhancing 3D Object Detection by Using Neural Network with Self-adaptive Thresholding. arXiv preprint arXiv:2405.07479.