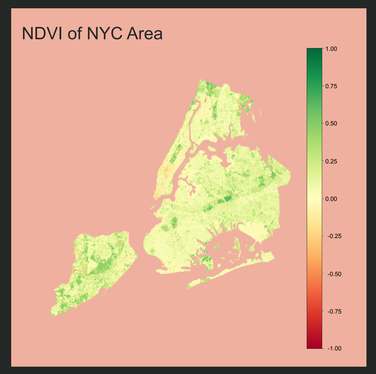
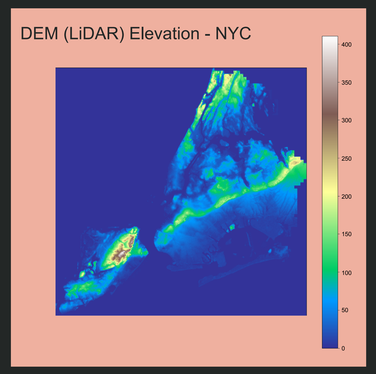
Comparative Analysis of Machine Learning Approaches for Urban Heat Island Prediction: Random Forest, XGBoost, and Convolutional Neural Networks (CNN)

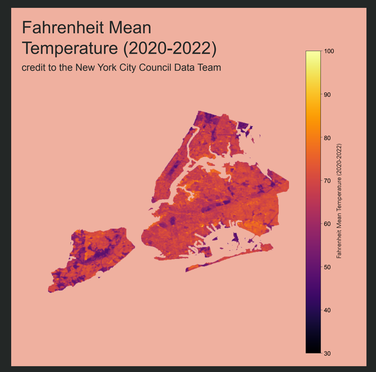
Mina Wei` |Deepesh Dinesh Theruvath Kanhangad* | Xiaoya Pan*
`Department of Urban Planning, Columbia University GSAPP, New York, NY
*Department of Earth and Environmental Engineering, Columbia University, New York, NY

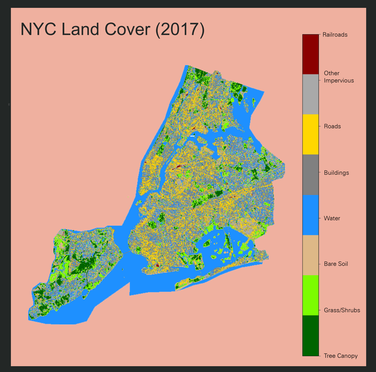
Background

The urban heat island (UHI) effect results in higher temperatures in urban areas compared to surrounding suburban regions, posing challenges to environmental health and residents' quality of life. To better understand and mitigate UHI effects, this study uses a variety of machine learning and deep learning models to predict urban temperature distribution. We focus on three models—Random Forest, XGBoost, and CNN—and evaluate the contributions of different geographical features (such as elevation, vegetation cover, and land use) to UHI.



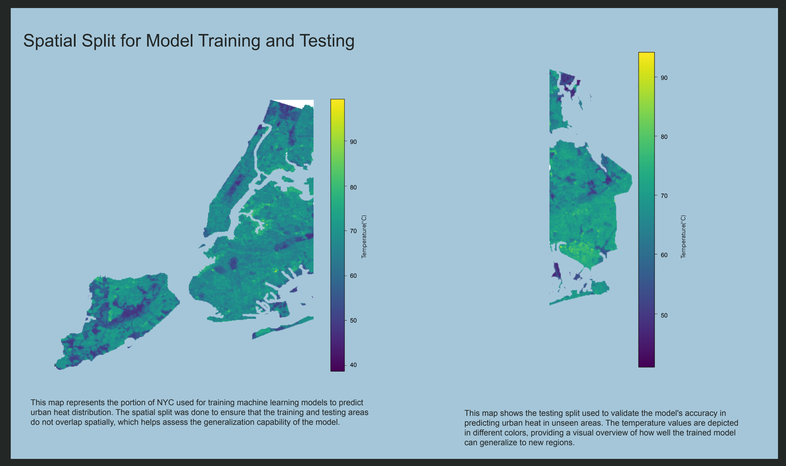






Research Questions

1.Which machine learning model (Random Forest, XGBoost, CNN) performs best in predicting the urban heat island effect?
2.Which geographical features have the greatest impact on urban temperature?
3.How does the deep learning approach (CNN) perform compared to traditional machine learning methods in predicting urban heat?



Conclusion

Model Performance: CNN demonstrated better predictive capability in spatial feature extraction, while Random Forest and XGBoost showed strengths in interpretability.
Feature Impact: Vegetation cover (NDVI) had the most significant impact on the UHI effect, followed by elevation and land cover types.
Urban Planning Recommendations: Increasing vegetation cover and implementing permeable surfaces can effectively mitigate UHI impacts.

Literature Review

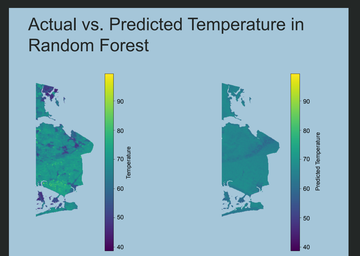
Urban Heat Island Studies: Previous research has shown that the UHI effect is closely related to reduced vegetation, increased building density, and the rise in impervious surfaces (Smith et al., 2021).
Application of Machine Learning in Environmental Prediction: Machine learning and deep learning methods have been widely applied in modeling and predicting environmental data (Johnson & Patel, 2022). Random Forest and XGBoost are favored for their interpretability and ability to handle high-dimensional data, while CNN excels at extracting features from spatial data.
UHI Mitigation Strategies: Vegetation cover and permeable land surfaces are considered effective ways to mitigate UHI effects (Li & Zhang, 2023).

Limitations and Future Work

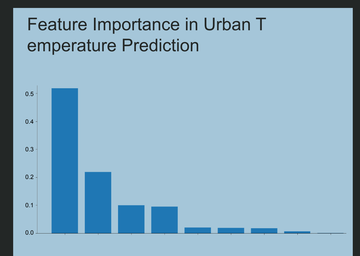
- Limitations:
  - Training data is limited to a specific region, which may not fully represent UHI situations in other cities.
  - Models rely on currently available data, and updates or variations in data quality may impact prediction accuracy.
- Future Work:
  - Explore more environmental features (such as air pollution and building shape) and their impact on UHI.
  - Use larger datasets and additional deep learning architectures to improve model generalizability.

Methodology

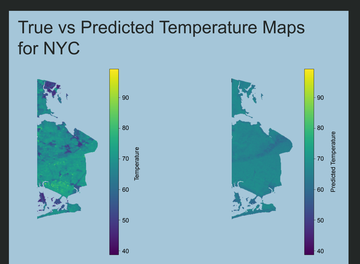
This study assesses the urban heat island (UHI) effect in NYC using Random Forest, XGBoost, and CNN models. Key features include NDVI, elevation, and land cover, all resampled to 1-meter resolution. The study area is split into two halves: one for training and the other for prediction.
Data Analysis : Models are trained to predict urban temperature using NDVI, elevation, and land cover as features. CNN is applied to capture spatial patterns more effectively.
Evaluation : Model performance is compared using R^2 and RMSE metrics, focusing on the impact of each feature on UHI, to determine the most influential factors.



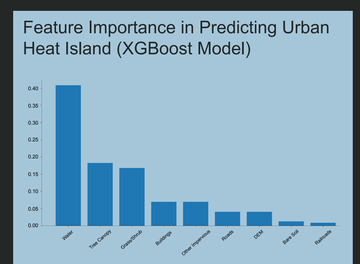
This comparison shows the actual observed temperatures and the predicted temperatures generated by the Random Forest model. The model captures the general spatial temperature distribution, particularly in areas with high variability. However, some local discrepancies exist, resulting in a Mean Absolute Error (MAE) of 4.77, Mean Squared Error (MSE) of 36.38, and an R^2 score of 0.181. These metrics indicate that while the model captures the broader trends, there is still room for improvement in predicting localized temperature variations.



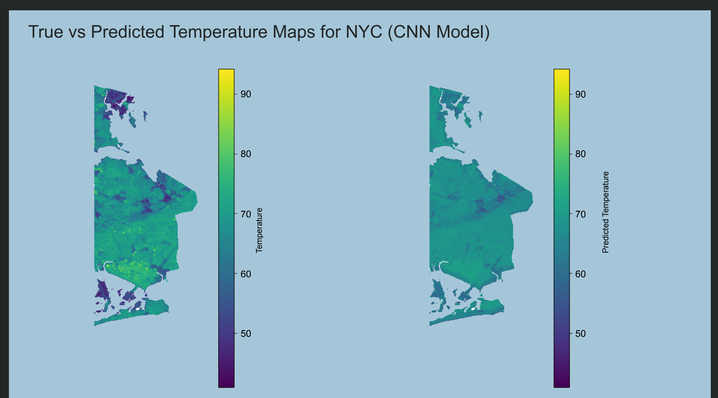
This figure illustrates the importance of various geographic features in predicting urban temperature using a Random Forest model. Elevation (DEM) is found to be the most critical factor, significantly influencing temperature variations across New York City. Water bodies are also crucial, whereas features such as buildings, roads, grass/shrubs, and other impervious surfaces have a comparatively smaller impact on temperature prediction. This insight helps to understand the key contributing factors to the urban heat island effect.



This comparison shows the actual observed temperatures and the predicted temperatures generated by the XGBoost model. The model captures the general spatial temperature distribution, especially in areas with significant variability. However, certain discrepancies remain, particularly in finer details and localized predictions. The performance metrics include a Mean Absolute Error (MAE) of 4.79, Mean Squared Error (MSE) of 36.52, and an R^2 score of 0.18. These metrics suggest that while the model successfully captures overall temperature patterns, further refinement is needed for better accuracy in localized temperature predictions.



This feature importance analysis highlights the influence of different geographic variables on urban temperature prediction. The XGBoost model identifies "Water" as the most critical factor, implying that proximity to water bodies has a significant effect on temperature variation within urban areas. Vegetation features such as "Tree Canopy" and "Grass/Shrubs" also play an important role in mitigating urban heat. This insight underscores the importance of environmental features in urban planning, particularly in efforts to reduce the urban heat island effect and enhance thermal comfort.



This comparison presents the true observed temperatures alongside the predicted temperatures generated by the CNN model. The model effectively captures the broad spatial patterns of temperature variation, particularly in capturing cooler areas and certain hotspots. The Mean Squared Error (MSE) on the test data is 29.59, indicating that the CNN model can reasonably predict urban temperatures, though certain discrepancies remain at a finer scale. The model shows promise for identifying general temperature trends in urban areas, but localized details may require additional refinement or additional features.