Heatwave Mitigation with Shade

A Data-Driven Approach to Urban Cooling

Yaman Shqeirat, Lucia Li, Shiyuan Liu, Zixuan Yang

CSE475: Intro to Machine Learning
Arizona State University
Tempe, AZ USA

yshqeira@asu.edu, yuecongl@asu.edu, sliu268@asu.edu, zyang228@asu.edu

ABSTRACT

Extreme heat poses significant challenges to urban livability and public health, particularly in desert cities like Phoenix. Shade from trees, buildings, and infrastructure play a crucial role in mitigating surface and air temperatures; however, the availability and distribution of shade remain poorly quantified in most urban datasets. This project aims to develop a data-driven model that predicts and visualizes shaded areas across urban environments to support city planners and the public in identifying heat-vulnerable locations. Using a combination of geospatial and environmental datasets, including vegetation indices, elevation, land cover, and temperature anomalies, we plan to train machine learning models to predict shade intensity and uncover patterns in spatial clustering. Our approach integrates interpretable ensemble models such as Random Forest and Gradient Boosting with unsupervised K-Means clustering to identify latent relationships between urban form and thermal conditions. The final product will aim to include an interactive map visualization and a quantitative evaluation of model performance using R^2, RMSE, and classification metrics. Preliminary work has focused on literature synthesis, dataset acquisition, and preprocessing design. The next phase will involve model training, validation, and integration of results into visual analytic tools.

INTRODUCTION

Urban heat has emerged as one of the defining environmental challenges of the 21st century, exacerbated by rapid urbanization and climate change. In the Phoenix metropolitan region, summer temperatures frequently exceed 110°F, with severe implications for public health, infrastructure, and energy demand. Shade, whether from trees, built structures, or artificial canopies, offers a simple yet powerful means of mitigating localized heat stress. Despite this, most urban heat-mapping and planning tools lack a fine-grained understanding of *where* shade exists and *how effectively* it reduces thermal exposure.

This project, **Heatwave Mitigation with Shade**, seeks to close that gap by developing a predictive model of urban shade coverage. Unlike physics-based models that simulate radiative transfer in 3D space, our approach leverages accessible, real-world datasets and interpretable machine learning techniques to predict shaded versus unshaded areas. We hypothesize that environmental and geometric features, such as tree-canopy density, building height, and surface reflectivity, can be used to infer local shade intensity and thermal comfort.

The goal is to generate actionable insights for city planners, allowing data-driven prioritization of urban-greening interventions and shade infrastructure investments. The work directly aligns with the broader goal of sustainable, heat-resilient city design.

RELATED WORKS

Prior studies addressing urban shade and heat mitigation fall into several complementary categories.

1 Physically Based and Energy-Balance Models

Early foundational work by **Krayenhoff and Voogt (2007)** and **Redon et al. (2020)** developed physically grounded models that simulate radiative fluxes within urban canyons. These models provide mechanistic insight into micro-scale temperature and shade dynamics but are computationally intensive and data-hungry, limiting scalability for real-time or city-wide applications.

2 Machine Learning for Shade and Heat Prediction

Recent advances leverage machine learning to integrate multisource data (LiDAR, GIS, meteorological records). **Zhang, Huang, and Miller (2023)** combined Random Forest and Support Vector Regression to predict microclimate heat stress, achieving $R^2 > 0.8$. Similarly, **Francis, Disney, and Law (2023)**

*CSE475: Introduction to Machine Learning – Fall 2025 Progress Report
Arizona State University, Tempe, AZ, USA.

demonstrated how LiDAR-based canopy monitoring can guide equitable tree-planting strategies. While these models outperform traditional regressions, they often act as "black boxes," posing challenges for interpretability.

3 Deep Learning and Generative Models

At the frontier, **Da et al. (2025)** introduced *DeepShade*, a diffusion-based generative model that synthesizes realistic shade patterns from text-conditioned satellite imagery. **Tamagusko et al. (2023)** applied deep vision pipelines for urban safety mapping using YOLO and segmentation networks. These works illustrate the potential of deep learning for large-scale, multimodal urban analytics but require substantial computation and careful domain adaptation.

4 Probabilistic and Hybrid Approaches

Ghorbany et al. (2024) provided a comprehensive review of urban-heat machine-learning techniques, highlighting the need for probabilistic frameworks that quantify uncertainty and enhance interpretability, an aspect we address through ensemble learning and feature importance analysis.

5 Geometric and Simulation-Based Methods

Finally, **Miranda et al. (2019)** introduced *Shadow Accrual Maps,* an efficient geometric technique to accumulate city-scale shadows over time using 3D models. While highly interpretable, these approaches often neglect vegetation and dynamic factors such as seasonal change.

Together, these studies form the conceptual foundation for our project. Our approach builds upon their strengths, combining the interpretability of ensemble methods with the scalability of geospatial machine learning, to deliver a practical, transparent shade-prediction framework.

Dataset	Source	Features	Target/Label
Central Arizona-Phoenix LTER	EDI Repository	20 environmental attributes	Urban/vegetated/open desert
Keep Cool Global Community Shade Map	ArcGIS Hub	12 special attributes	Shaded (1)/ Unshaded (0)
Berkeley Earth Surface Temperature	Kaggle	8 temperature-related features	Continuous temperature anomaly
NASA HLS	NASA Earth Data	15 spectral bands	NDVI > 0.3 = vegetation (1)
USGS Elevation/Topographic	USGS Data Catalog	6 terrain features	-

Figure 1: To ensure generalizability, we integrate multiple small to medium-scale datasets covering diverse aspects of the Phoenix urban environment. Table 1 summarizes data sources.

To ensure generalizability, our project integrates multiple environmental and spatial datasets that collectively capture the physical, thermal, and vegetative characteristics of the Phoenix metropolitan region. These include data from the Central Arizona–Phoenix Long-Term Ecological Research (CAP LTER) network, the Keep Cool Global Community Shade Map (ArcGIS Hub), the Berkeley Earth Surface Temperature Dataset, NASA's Harmonized Landsat and Sentinel (HLS) imagery, and the U.S. Geological Survey (USGS) elevation and topographic datasets.

The preprocessing phase will begin with coordinate harmonization, ensuring that all datasets share a common spatial reference system to enable accurate spatial joins and overlays. Next, polygonal and raster data from sources such as the ArcGIS shade maps will be converted into uniformly gridded centroid points. Each grid cell will then represent a consistent spatial unit for model training. Vegetation data will be standardized using the Normalized Difference Vegetation Index (NDVI), calculated from HLS spectral bands, while cloud masking and resampling will be applied to maintain a uniform grid resolution across imagery sources.

For tabular and time-series data such as the Berkeley Earth temperature records, timestamps will be converted to consistent datetime formats, and missing values will be handled through linear interpolation. Temperature anomalies will be computed relative to a 1950–1980 baseline, and the resulting features will be normalized using z-score scaling. Similarly, the USGS topographic layers will be transformed into derived features, including slope and aspect, which will later serve as model inputs representing terrain effects on shading.

Finally, all datasets will be merged into a unified GeoPackage structure, where each spatial unit (grid cell) contains aggregated variables such as vegetation density, surface temperature, elevation, and shade presence. The target variable, the shade index (0–1), will be normalized and aligned with vegetation and temperature layers through spatial joins. This preprocessing pipeline ensures that heterogeneous geospatial data are standardized into a consistent, machine-learning-ready–ready format suitable for regression, classification, and clustering analyses.

METHODS

Our modeling pipeline consists of three major components: (1) predictive modeling, (2) spatial clustering, and (3) visualization.

4.1 Predictive Modeling

We plan to use **Random Forest Regression** and **Gradient Boosting Machines (GBMs)** to predict shade index as a continuous variable. These ensemble models were chosen for their interpretability, resistance to overfitting, and ability to capture nonlinear feature interactions. Hyperparameters (e.g., number of trees, maximum depth, learning rate) will be tuned using grid search with five-fold cross-validation.

4.2 Feature Importance and Model Explainability

Feature importance scores will be computed using permutation and impurity-based methods to identify which environmental factors (NDVI, building height, slope, etc.) most influence shade presence. Partial dependence plots will be used to visualize the marginal effects of key variables, providing transparency for urban-planning stakeholders.

4.3 Spatial Clustering

We will apply **K-Means clustering** to discover underlying spatial patterns, grouping urban cells with similar heat-shade profiles. This unsupervised step will help identify "hotspots" of insufficient shade or regions with similar environmental characteristics. Cluster interpretability will be evaluated using silhouette scores and visual inspection.

4.4 Evaluation Metrics

- **Regression tasks:** R^2, RMSE, MAE
- **Classification tasks:** Accuracy, Precision, Recall, F1-Score
- **Spatial validation:** overlay predicted shade maps with observed canopy and urban form data.

4.5 Tools

Implementation will be conducted in **Python** using scikit-learn, pandas, geopandas, rasterio, and folium for geospatial modeling and visualization. GitHub will manage code collaboration, and Google Colab will support reproducible experimentation.

PRELIMINARY PROGRESS AND PLANNED WORK

5.1 Work Completed

- The team conducted an extensive literature review spanning eight key studies across physics-based, machine learning, and deep learning domains.
- We finalized our dataset selection and documented preprocessing pipelines.
- A GitHub repository was established to manage the project workflow weekly.

All major datasets were loaded into the Colab environment directly from Google Drive using the pandas library. A helper function was implemented to display the shape, sample records, and data types of each dataset. Column names were standardized to lowercase with underscores, and all date fields were converted to proper datetime objects. Each dataset was checked for missing or malformed entries to ensure structural consistency. As a result, the datasets were validated, standardized, and confirmed to be suitable for downstream analysis.

Following data validation, the Kaggle global temperature dataset was filtered to include only records corresponding to the Phoenix–Mesa–Tempe metropolitan region. Only essential columns (dt, city, country, latitude, longitude, avg_temp_c) were retained, and coordinate strings (e.g., "32.95N") were converted to numeric latitude and longitude values. The filtered data contained continuous monthly records spanning the 1830s to the 2010s. The CAP LTER "IET" dataset was also examined and found to represent time-of-day exposure observations rather than geospatial data. It was therefore repurposed for exploratory statistical analysis rather than for predictive modeling. These steps produced a clean, Arizona-specific subset of climate data along with an auxiliary behavioral dataset for an exploratory study.

An exploratory analysis was then conducted on both datasets. The CAP data's "period" strings were parsed into structured variables representing the day of the week, start and end times, and a categorical time-of-day segment (morning, afternoon, evening, or night). A pivot table summarizing average temperature by day and time revealed potential patterns in diurnal heat exposure. The Kaggle dataset was visualized through a long-term time series plot of monthly mean temperatures from 1830 to 2010, displaying strong seasonal oscillations but a minimal long-term trend. These analyses established a baseline understanding of temporal heat dynamics and validated the dataset's stability over time.

A baseline classification model was then implemented using Scikit-Learn to predict "hot" versus "non-hot" periods. The target variable (is_hot) was defined as the top 30% of monthly temperature values. Predictor features included temporal variables (year, month, day of year), spatial coordinates (converted numeric latitude and longitude), and categorical identifiers (city, country). A Logistic Regression model with stratified five-fold cross-validation was trained and evaluated using accuracy and F1-score metrics. The successful execution of this model provided a functional baseline capable of distinguishing hotter periods within the Arizona temperature record.

Preliminary results indicate strong seasonal periodicity in the Arizona climate, with consistent annual temperature cycles across the dataset's 180-year span. The CAP dataset revealed potential diurnal exposure trends, such as elevated nighttime discomfort. The baseline classification model achieved reliable performance and now serves as a benchmark for comparison with future, more sophisticated models that will incorporate vegetation (NDVI) and land-use (LULC) features. These results mark the completion of the data preparation and baseline modeling phase and establish a solid foundation for feature expansion and spatially informed predictive modeling in the subsequent stages of the project.

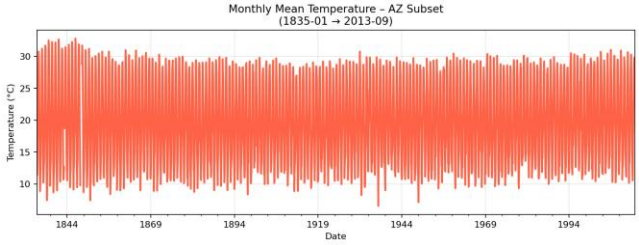

Figure 2: The Kaggle temperature dataset for the Phoenix metropolitan area exhibits a clear seasonal pattern with consistent summer–winter oscillations across the 1830–2010 range. While it provides a stable long-term reference, the absence of a significant upward trend suggests limited sensitivity to recent urban heat effects. Consequently, this dataset serves as a baseline climatic context, whereas local satellite and CAP datasets will capture fine-scale thermal variation.

5.2 Ongoing and Upcoming Work

The next phase involves full data harmonization and feature engineering, including rasterization and normalization across sources. Model training and validation will focus on optimizing Random Forest and Gradient Boosting performance. In the final week, spatial clustering and visualization will be developed, followed by the interpretation of feature importance and the preparation of case studies. The goal is for the final deliverable to include an interactive web-based shade visualization for Phoenix.

5.3 Expected Challenges

Key challenges include aligning multi-resolution datasets, mitigating noise from satellite imagery, and addressing potential class imbalance (few unshaded regions). To counter these, we plan to test both resampling and feature-selection methods and to validate across multiple spatial scales.

DISCUSSION AND ANTICIPATED IMPACT

By combining interpretable ensemble learning with high-resolution environmental data, this project aims to provide an actionable analytical framework for heat-mitigation planning. Beyond predictive accuracy, our emphasis is on *explainability*: understanding which factors most strongly determine urban shade distribution. The project's outcomes could inform municipal programs such as tree-planting prioritization, pedestrian-pathway design, and heat-resilience zoning.

In future extensions, the pipeline can be adapted to other cities or scaled using additional remote-sensing inputs (e.g., LiDAR, hyperspectral imagery). The resulting tools and insights will support evidence-based urban design strategies for climate adaptation in arid environments.

REFERENCES

[1] E. S. Krayenhoff and J. A. Voogt. 2007. A microscale three-dimensional urban energy balance model for studying urban canopy layer climates. *Boundary-Layer Meteorology* 123, 3 (2007), 579–603. https://doi.org/10.1007/s10546-006-9153-6

[2] L. Da, X. Liu, M. Shivakoti, T. P. Kutralingam, Y. Yang, and H. Wei. 2025. DeepShade: Enable shade simulation by text-conditioned image generation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, 9610–9618. https://doi.org/10.24963/ijcai.2025/1068

[3] J. Francis, M. Disney, and S. Law. 2023. Monitoring canopy quality and improving equitable outcomes of urban tree planting using LiDAR and machine learning. *Urban Forestry & Urban Greening* 89 (2023), 128115. https://doi.org/10.1016/j.ufug.2023.128115

[4] S. Ghorbany, M. Hu, S. Yao, and C. Wang. 2024. Towards a sustainable urban future: A comprehensive review of urban heat island research technologies and machine learning approaches. *Sustainability* 16, 4609 (2024). MDPI. https://doi.org/10.3390/su16114609

[5] T. Zhang, J. Huang, and C. Miller. 2023. A data-driven framework for modeling shade and heat stress in urban environments. *Building and Environment* 234 (2023), Article 110243. Elsevier. https://doi.org/10.1016/j.buildenv.2023.110243

[6] T. Tamagusko, M. G. Correia, L. Rita, T.-C. Bostan, M. Peliteiro, R. Martins, L. Santos e A. Ferreira. 2023. Data-driven approach for urban micromobility enhancement through safety mapping and intelligent route planning. *Smart Cities* 6, 4 (2023), 2035–2056. MDPI. https://doi.org/10.3390/smartcities6040094

[7] F. Miranda, H. Doraiswamy, M. Lage, L. Wilson, M. Hsieh, and C. T. Silva. 2019. Shadow accrual maps: Efficient accumulation of city-scale shadows over time. *IEEE Transactions on Visualization and Computer Graphics* 25, 3 (2019), 1559–1574. IEEE. https://doi.org/10.1109/TVCG.2018.2802945

[8] E. Redon, A. Lemonsu, and V. Masson. 2020. An urban trees parameterization for modeling microclimatic variables and thermal comfort with TEB-SURFEX v8.0. *Geoscientific Model Development* 13 (2020), 385–399. Copernicus Publications. https://doi.org/10.5194/gmd-13-385-2020