From spectral indices to actionable insights: Sensitivity analysis of a multispectral U-Net for spatially-optimized urban heat island mitigation

Oğuz Deniz*[a], Hüseyin Pekkan[a], and A. Murat Özbayoğlu[a]

[a]TOBB University of Economics and Technology, Söğütözü St. No:43, 06560, Ankara, Türkiye

ABSTRACT

Urban heat island (UHI)[1] effects present significant challenges for sustainable urban development. Within the UDENE[2] project, a deep-learning framework was established to model land surface temperature (LST) from remotely sensed urban indices, assess the effectiveness of green and blue infrastructure interventions such as linked park systems, and identify thermally analogous regions across cities. A compact U-Net[3] architecture was trained on six spectral indices obtained from Google Earth Engine[4] (NDVI,[5] NDBI,[6] NDWI,[7] SAVI,[8] IBI,[9] EVI[10]), with corresponding LST values serving as the prediction target. The primary case study was conducted on 50×50-pixel windows at 30 m resolution from İstanbul, with further validation in San Francisco and Abu Dhabi to ensure robustness across different climatic contexts. To enhance interpretability and minimize spatial autocorrelation, a spatial block partitioning strategy was employed, and a multi-method sensitivity analysis was implemented. This included channel ablation, input perturbation, spatial occlusion, and gradient-based explainability techniques. Results consistently highlighted the dominant cooling influence of vegetation indices (EVI, SAVI, NDVI), with preservation of existing green areas demonstrating greater impact than equivalent new additions. Context-dependent effects were also observed, particularly for NDWI, reflecting complex interactions between urban water features and the surrounding built environment. Overall, the proposed framework provides robust and interpretable evidence to support urban planners in the design and strategic placement of green and blue infrastructure for effective UHI mitigation across diverse urban landscapes.

Keywords: Urban Heat Island Mitigation, Sensitivity Analysis, Similarity Analysis, Environmental Remote and Proximal Sensing

1. INTRODUCTION

The urban heat island (UHI) effect represents one of the most pressing environmental challenges facing contemporary cities, with urban areas experiencing temperatures 1-3°C higher than surrounding rural regions, and in extreme cases, differences exceeding 10°C. This phenomenon, first comprehensively analyzed by Oke[11] through its energetic basis, results from the complex interplay of modified surface materials, altered urban geometry, reduced vegetation cover, and anthropogenic heat release. As global urbanization accelerates, with over 68% of the world's population expected to reside in cities by 2050, understanding and mitigating UHI effects has become critical for sustainable urban development, public health, and climate adaptation strategies. Two decades of urban climate research have evolved from simple observational studies to sophisticated modeling approaches,[12] with remote sensing emerging as a fundamental tool for monitoring urban thermal dynamics. Weng[13] demonstrated how thermal infrared remote sensing enables systematic observation of urban surface temperatures across multiple spatial and temporal scales, providing unprecedented insights into UHI patterns. Subsequently, global analyses have revealed the widespread nature of this phenomenon, with Peng et al.[14] documenting UHI effects across 419 global cities and Manoli et al.[15] establishing that UHI magnitude is largely predictable based on climate and population factors, particularly in arid and temperate regions. These foundational studies underscore both the universality of UHI challenges and the potential for systematic, data-driven mitigation strategies.

The role of green and blue infrastructure in mitigating UHI effects has been extensively documented through empirical studies. Urban parks have been shown to function as "cool islands," with temperature reductions

*Corresponding author: Oğuz Deniz, E-mail: oguzdeniz179@gmail.com ranging from 2-8°C compared to surrounding built environments.[16,17] The cooling intensity depends not only on park size but also on spatial configuration, with Lin and Lin[18] demonstrating that the arrangement and distribution of green spaces significantly influence their collective cooling effect. Gunawardena et al.[19] synthesized evidence showing how vegetation's evapotranspiration and water bodies' high heat capacity contribute synergistically to urban cooling, while Norton et al.[20] provided frameworks for prioritizing green infrastructure deployment based on population density, existing vegetation, and temperature patterns. These studies collectively support the concept of linked park systems as a strategic approach to urban heat mitigation. Remote sensing-based assessments have consistently identified vegetation indices as primary indicators of UHI intensity. The Normalized Difference Vegetation Index (NDVI) shows strong negative correlation with land surface temperature (LST), while the Normalized Difference Built-up Index (NDBI) exhibits positive correlation.[21–23] Cetin et al.[21] demonstrated these relationships across multiple districts in Kayseri, Turkey, establishing methodological benchmarks for index-based UHI analysis. Similarly, Kikon et al.[22] confirmed these patterns at kilometer-grid scales, providing empirical validation for using spectral indices as proxies for urban thermal conditions. The consistent identification of vegetation indices—particularly NDVI, Enhanced Vegetation Index (EVI), and Soil-Adjusted Vegetation Index (SAVI)—as dominant cooling factors across diverse geographic contexts underscores their fundamental role in urban thermal regulation.

The advent of deep learning has revolutionized remote sensing applications, offering unprecedented capabilities for analyzing complex spatial patterns and relationships.[24,25] Convolutional Neural Networks (CNNs), particularly U-Net architectures originally developed for biomedical image segmentation, have proven exceptionally effective for pixel-wise prediction tasks in Earth observation.[26] Recent advances in deep learning for Earth system science emphasize not only predictive accuracy but also process understanding and interpretability.[27,28] The application of these techniques to LST estimation has evolved rapidly, with Bouaziz et al.[29] surveying spatio-temporal fusion approaches including U-Net variants with attention mechanisms and temporal modules. Kustura et al.[30] demonstrated that combining multiple data sources—including Sentinel-2 indices, land cover, and meteorological variables—within deep learning frameworks significantly improves LST estimation accuracy, validating multi-channel approaches for capturing urban thermal complexity. However, despite these technological advances, critical gaps remain in translating deep learning predictions into actionable urban planning insights. While models achieve high predictive accuracy, understanding the causal relationships between urban features and temperature patterns remains challenging. Recent work has begun addressing this through explainable AI approaches, with Huang et al.[31] employing XGBoost[32] with SHAP[33] values to quantify per-feature contributions to LST, identifying NDVI as the top cooling driver across urban districts. Syeda et al.[34] demonstrated how machine learning can integrate physical and social determinants to develop zone-specific mitigation strategies, emphasizing the need for interpretable models that support evidence-based policy. These studies highlight the importance of moving beyond black-box predictions toward models that provide mechanistic insights for urban planning applications.

The spatial heterogeneity of urban environments presents additional challenges for model generalization and transferability. Urban thermal patterns vary significantly across climatic zones, urban morphologies, and development intensities.[14,15] Models trained on single cities often fail to generalize to different urban contexts, limiting their practical applicability. Furthermore, the complex interactions between urban features—such as the context-dependent effects of water bodies that can either cool through evaporation or warm through heat storage depending on surrounding conditions—require sophisticated analytical approaches that capture non-linear relationships while maintaining interpretability. Within this context, our research addresses these challenges through a comprehensive deep learning framework that combines predictive accuracy with causal understanding. We present a compact U-Net architecture trained on six carefully selected spectral indices (NDVI, NDBI, NDWI, SAVI, IBI, EVI) to predict LST at 30-meter resolution, enabling fine-grained analysis of urban thermal patterns. Our approach uniquely integrates multiple sensitivity analysis methods—including ablation studies, perturbation analysis, spatial occlusion, and gradient-based explainability techniques—to systematically examine the causal relationships between urban features and temperature patterns. This multi-method framework provides statistically robust, interpretable evidence for the effectiveness of green and blue infrastructure interventions, particularly linked park systems, in mitigating UHI effects.

The primary contributions of this work are fourfold: (1) We develop a spatially-aware deep learning model that maintains high predictive accuracy while preventing spatial autocorrelation through strategic block partitioning;

(2) We implement a comprehensive sensitivity analysis framework that quantifies the individual and interactive effects of urban features on LST, providing mechanistic insights beyond correlation; (3) We validate our approach across climatically diverse cities (San Francisco, İstanbul, Abu Dhabi); and (4) We translate model insights into actionable planning recommendations, quantifying the relative benefits of preserving existing green spaces versus creating new ones, and identifying optimal locations for green infrastructure deployment. By bridging the gap between advanced deep learning capabilities and practical urban planning needs, this research provides a foundation for evidence-based strategies to combat UHI effects through strategic green and blue infrastructure development. Our findings offer urban planners and policymakers quantitative tools to evaluate and optimize heat mitigation interventions, supporting the development of more resilient and sustainable urban environments in the face of ongoing climate change.

2. METHODOLOGY

Our methodology employed a multi-city approach to ensure robust model generalization across diverse climatic and urban contexts. San Francisco served as the primary case study due to its well-documented UHI patterns and Mediterranean climate characteristics. To validate model transferability, we extended our analysis to İstanbul, Turkey (temperate oceanic climate with continental influences) and Abu Dhabi, UAE (hot desert climate), representing distinct climatic zones and urban morphologies that collectively span a broad spectrum of global urban environments. All remotely sensed data were acquired through Google Earth Engine (GEE),[4] leveraging its comprehensive archive of Landsat 8 imagery with 30-meter spatial resolution. The temporal window for data collection spanned 2013-2025, with cloud-free images selected during peak summer months (June-August) to capture maximum UHI intensity. For each study area, we systematically sampled 50×50 pixel windows at the native 30-meter resolution to preserve fine-scale urban thermal patterns while ensuring computational tractability, excluding tiles that were degenerate (i.e., containing entirely black or white channels) or that did not conform to the required 50×50 shape.

2.1 Spectral Index Calculation and Feature Selection

We derived six spectral indices for each sampled window chosen for their demonstrated relationships to urban thermal dynamics and for their complementary ability to represent vegetation, built surfaces, soil effects, and surface water. Vegetation indices capture photosynthetic activity and canopy structure; the Normalized Difference Vegetation Index (NDVI) follows the conventional form:

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}}, \tag{1}$$

while the Enhanced Vegetation Index (EVI) augments NDVI by accounting for atmospheric scattering and canopy background:

$$\mathrm{EVI} = 2.5 \cdot \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + 6 \cdot \mathrm{Red} - 7.5 \cdot \mathrm{Blue} + 1}. \tag{2}$$

To reduce soil-brightness effects in sparsely vegetated pixels we used the Soil-Adjusted Vegetation Index (SAVI) with the commonly used soil correction factor L:

$$\mathrm{SAVI} = \left(\frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red} + L}\right) \times (1 + L), \quad L = 0.5. \tag{3}$$

Indices that emphasize impervious and built surfaces complement the vegetation measures. The Normalized Difference Built-up Index (NDBI) highlights built-up areas by contrasting short-wave infrared and near-infrared reflectances:

$$\mathrm{NDBI} = \frac{\mathrm{SWIR} - \mathrm{NIR}}{\mathrm{SWIR} + \mathrm{NIR}}. \tag{4}$$

We also computed the Index-Based Built-up Index (IBI) to provide a normalized indicator of built cover that combines SWIR, NIR, Red and Green information; written compactly:

$$\mathrm{IBI} = \frac{2 \cdot \frac{\mathrm{SWIR1}}{\mathrm{SWIR1}+\mathrm{NIR}} - \left(\frac{\mathrm{NIR}}{\mathrm{NIR}+\mathrm{Red}} + \frac{\mathrm{Green}}{\mathrm{Green}+\mathrm{SWIR1}}\right)}{2 \cdot \frac{\mathrm{SWIR1}}{\mathrm{SWIR1}+\mathrm{NIR}} + \left(\frac{\mathrm{NIR}}{\mathrm{NIR}+\mathrm{Red}} + \frac{\mathrm{Green}}{\mathrm{Green}+\mathrm{SWIR1}}\right)}. \tag{5}$$

Surface water features were captured with the Normalized Difference Water Index (NDWI):

$$\mathrm{NDWI} = \frac{\mathrm{Green} - \mathrm{NIR}}{\mathrm{Green} + \mathrm{NIR}}, \tag{6}$$

which is sensitive to open water and moisture-related signatures that can modulate local thermal behavior.

All indices were computed from atmospherically corrected surface reflectances and resampled to the common 30 m grid of analysis; band names refer to the usual multispectral channels (Blue, Green, Red, NIR, SWIR/SWIR1). Land surface temperature (LST), used as the model target, was retrieved from Landsat 8 thermal infrared bands via the radiative transfer approach with atmospheric correction parameters obtained from NCEP reanalysis fields. Where available, LST retrievals were validated against in-situ temperature measurements to verify absolute accuracy and to calibrate any systematic offsets prior to model training.

2.2 Spatial Block Partitioning Strategy

To address spatial autocorrelation—a critical challenge in geospatial machine learning—we implemented a spatial block partitioning strategy rather than random sampling. The study regions were systematically divided into **non-overlapping** spatial blocks, with each block containing multiple 50×50 pixel windows. Blocks were then randomly assigned to training (70%), validation (10%), and testing (20%) sets, ensuring that spatially proximate samples were grouped within the same partition. This approach prevents data leakage between training and testing sets, where spatial proximity could artificially inflate model performance metrics. The block size was optimized to balance spatial independence while maintaining sufficient sample diversity within each partition. This methodology ensures that model evaluation reflects true generalization capability to unseen geographic areas rather than interpolation within spatially correlated regions.

2.3 Dataset Composition and Preprocessing

Following the spatial block partitioning and quality control procedures, the final datasets comprised windows distributed across the three study cities as detailed in Table 1. The sampled windows underwent preprocessing to eliminate samples affected by clouds, technical issues, and other distortion artifacts. Additionally, windows corresponding to regions outside the scope of urban development were systematically excluded—such as extensive desert areas in Abu Dhabi—to maintain focus on urban thermal dynamics and ensure model relevance to built environment applications.

Table 1: Per-city dataset sizes and train/validation/test splits (post-filtering).

City	Total windows	Training	Validation	Testing
İstanbul	40,828	28,513	4,098	8,217
San Francisco	3,352	2,329	337	686
Abu Dhabi	13,821	9,675	1,382	2,764
Total	**58,001**	**40,517**	**5,817**	**11,667**

These city-wise test sets were held out for final sensitivity analyses and statistical evaluation to ensure unbiased assessment of causal relationships. The training and validation partitions were used exclusively during model development (training, hyperparameter tuning, and early stopping), while all figures, metrics and statistical summaries reported in the Results section are computed on the held-out test data shown in Table 1. The substantial difference in sample sizes across cities reflects variations in urban extent, data availability, and the effectiveness of quality control procedures in different climatic contexts.

2.4 U-Net Architecture Design

We implemented a compact U-Net tailored for six-channel input and per-pixel Land Surface Temperature (LST) regression, designed to preserve fine spatial detail while remaining computationally efficient and interpretable. The encoder part consists of three resolution stages; each stage applies two successive 3×3 convolutions (with ReLU activations[35] and batch normalization[36]) and then reduces spatial resolution with a 2×2 max-pool. Filter counts increase with depth ($64 \rightarrow 128 \rightarrow 256$) so that the encoder progressively captures higher-level features while retaining locality through the paired convolutions. The bottleneck uses two 3×3 convolutions with 512 filters, each followed by batch normalization and ReLU, to learn complex spatial relationships efficiently.

The decoder component of our U-Net symmetrically reconstructs spatial detail through a series of learned upsampling operations. Each stage of the decoder begins with a 2×2 transposed convolution, which progressively increases the spatial resolution while simultaneously reducing the feature dimensionality (with 256, 128, and 64 filters in successive stages). To enrich the upsampled representation, the output of each transposed convolution is concatenated with the corresponding encoder feature maps via skip connections,[37] thereby restoring spatial context that might otherwise be lost during downsampling. This concatenated tensor is subsequently refined through two successive convolutional blocks, each consisting of a 3×3 convolution, batch normalization, and ReLU activation, which together enhance the representational power of the network. To ensure that skip-connected encoder features align precisely with decoder outputs, we implemented a custom tensor-cropping mechanism that removes minor spatial mismatches introduced by pooling and upsampling operations. This guarantees a strict correspondence between encoder and decoder features across resolution levels. Finally, a 1×1 convolution projects the refined decoder output into a single continuous-valued channel. This channel represents the land surface temperature (LST) in degrees Celsius for each pixel. The model uses a linear activation function, which is suitable for the regression-based prediction of temperature values. Each predicted pixel corresponds to a 30x30 meter region in the real world.

To stabilize and speed up training, every convolution is followed by batch normalization. The design also uses symmetric skip connections and paired convolutions to preserve spatial context for accurate predictions. The network was trained with the Adam optimizer[38] (initial learning rate 1×10^{-4}) for up to 50 epochs, with early stopping triggered by validation loss to avoid overfitting. Mean squared error was used as the primary training objective, while complementary diagnostics (MAE, RMSE) were monitored during training and evaluation. The described architecture is illustrated in Figure 1.

2.5 Sensitivity Analysis Framework

Our sensitivity analysis integrated five complementary methodological approaches, providing a robust and multi-faceted understanding of the causal relationships between urban features and thermal patterns captured by the U-Net model.

2.5.1 Channel ablation analysis

The ablation study systematically evaluated each spectral index's contribution by selectively removing individual channels while maintaining all other inputs constant. For each of the six indices, a modified input was created by setting the target channel values to zero across all spatial locations within the 50×50 window. The resulting LST predictions were compared against baseline predictions using the complete six-channel input.

The Mean Difference (MD) metric quantified the impact of each channel's removal:

$$\mathrm{MD} = \frac{1}{n} \sum_{i=1}^{n} (\mathrm{LST}_{\mathrm{modified},i} - \mathrm{LST}_{\mathrm{baseline},i}) \tag{7}$$

where n is the total number of pixels in the test dataset and i denotes the spectral channel being ablated. Positive MD_i values indicate that removing channel i leads to higher predicted temperatures (warming effect), while negative values correspond to cooling effects.

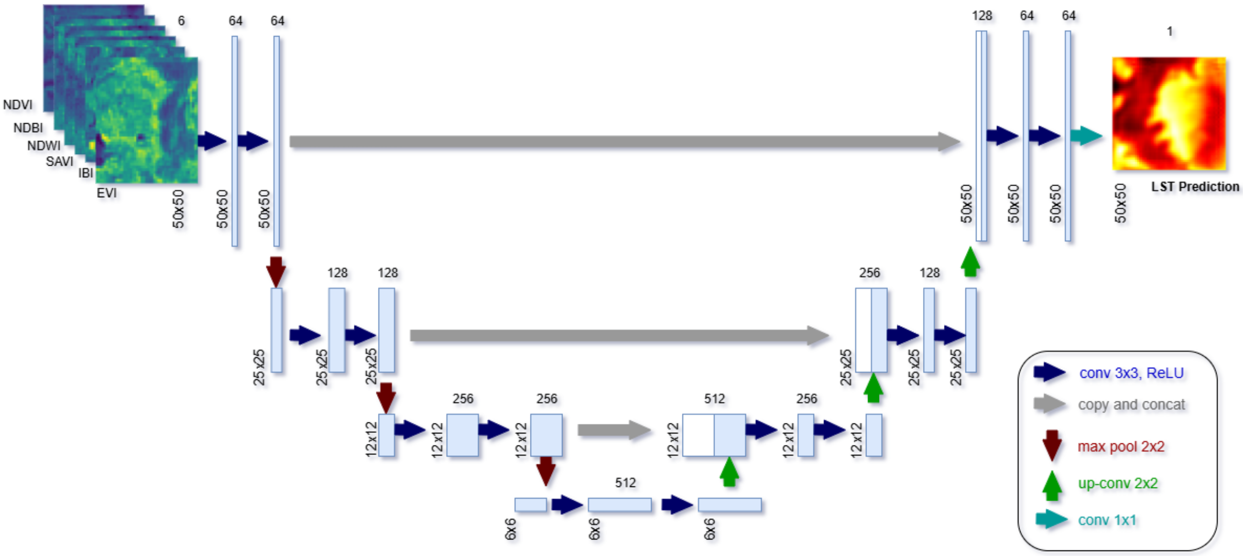

Figure 1: Compact U-Net architecture for per-pixel LST regression: six-channel input → three encoder stages (64→128→256 filters, two 3×3 conv + BN + ReLU per stage, 2×2 max-pooling) → 512 filter bottleneck → symmetric decoder (2×2 transposed convolutions with 256, 128, 64 filters, skip connections) → 1×1 linear output to a single LST channel.

2.5.2 Perturbation-based sensitivity assessment

To evaluate model stability and the magnitude of its responses to subtle urban feature changes, we conducted a controlled perturbation analysis using three values of ε ($\varepsilon \in \{0.01, 0.05, 0.1\}$). For each spectral index and perturbation level, we generated modified inputs by either incrementing or decrementing:

$$c_\varepsilon^+ = c + \varepsilon \cdot c \tag{8}$$

$$c_\varepsilon^- = c - \varepsilon \cdot c \tag{9}$$

where c denotes the original channel value, c_ε^+ represents the perturbed value obtained by incrementing c with a fraction ε, and c_ε^- represents the perturbed value obtained by decrementing c with the same fraction.

This procedure simulates realistic scenarios such as gradual vegetation growth or decline over time, as well as incremental urban development changes. Sensitivity was quantified using the same MD metric, thereby providing insights into both the direction and magnitude of predicted temperature responses to small-scale urban modifications.

2.5.3 Statistical significance framework

All sensitivity analyses were subjected to rigorous statistical validation to support robust inference. We adopted the following hypothesis-testing framework:

$\mathcal{H}_0$**:** No significant change in model output due to channel modification.

$\mathcal{H}_1$**:** A significant change in model output due to channel modification.

The significance level was set to $\alpha = 0.01$. When multiple channels or comparisons were tested, reported p-values were adjusted to control the false discovery rate using the Benjamini–Hochberg procedure,[39] unless otherwise stated.

Parametric tests

- Two-sample (or paired, where appropriate) t-tests comparing the modified and baseline prediction distributions.
- Cohen's d[40] for effect-size quantification:
$$d = \frac{\mu_1 - \mu_2}{\sigma_{\text{pooled}}}, \tag{10}$$
where the pooled standard deviation is
$$\sigma_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}. \tag{11}$$
Here μ_k, σ_k, and n_k denote the sample mean, sample standard deviation and sample size for group $k \in \{1, 2\}$.

Non-parametric validation

- Wilcoxon[41] signed-rank tests for paired comparisons that do not rely on distributional assumptions.

Robust estimation

- Bootstrap confidence intervals (95%) for MD estimates computed with $B = 1{,}000$ bootstrap resamples.
- Bias-corrected and accelerated (BCa) intervals were used to improve interval accuracy in the presence of skewness or bias.

2.5.4 Spatial occlusion analysis

Our spatial occlusion analysis adapts the occlusion-sensitivity approach of Zeiler and Fergus[42] to multi-channel remote-sensing input. Specifically, we slide a 5×5 occlusion patch across each 50×50 input window, zeroing the patch simultaneously across all six channels at every location. For each occlusion position we compute the change in the model's per-pixel LST predictions and then aggregate those changes across the entire test set to produce a population-level spatial sensitivity map. This procedure highlights the local regions whose removal most strongly alters predicted temperatures, allowing us to identify spatially critical features for downstream interpretation and intervention planning. Let

$$X \in \mathbb{R}^{C \times H \times W}$$

denote an input sample with $C = 6$ channels and $H = W = 50$. Let the occlusion patch size be $k = 5$, and let (p, q) denote the top-left location of the patch with $p \in \{1, \ldots, H - k + 1\}$ and $q \in \{1, \ldots, W - k + 1\}$. Define the binary occlusion mask $M_{p,q} \in \{0, 1\}^{C \times H \times W}$ by

$$M_{p,q}[c, h, w] = \begin{cases} 1, & \text{if } h \in \{p, \ldots, p + k - 1\},\ w \in \{q, \ldots, q + k - 1\},\ \forall c, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The occluded input is obtained by zeroing all channels inside the patch:

$$X^{(p,q)} = X \odot (1 - M_{p,q}), \tag{13}$$

where $\odot$ denotes element-wise multiplication.

For each occlusion location (p, q) we compute a sensitivity value $\Delta_{p,q}$ as the mean change in model prediction over the test set. Let the test set contain N samples, and let the model produce a per-pixel output

$\mathrm{LST}(\cdot) \in \mathbb{R}^{H\times W}$. Denoting the baseline prediction for test sample t by $\mathrm{LST}(X_t)$ and the occluded prediction by $\mathrm{LST}(X_t^{(p,q)})$, we define

$$\Delta_{p,q} \;=\; \frac{1}{n}\sum_{t=1}^{N}\sum_{u=1}^{H}\sum_{v=1}^{W}\Big(\mathrm{LST}\big(X_t^{(p,q)}\big)_{u,v} - \mathrm{LST}\big(X_t\big)_{u,v}\Big), \tag{14}$$

where $n = N \cdot H \cdot W$ is the total number of output pixels across the test set (here $H\!=\!W\!=\!50$). The set of $\Delta_{p,q}$ values forms a spatial sensitivity map of size $(H-k+1)\times(W-k+1)$ (here 46×46). For visualization, this sensitivity map is optionally upsampled or centered to the original 50×50 grid to create heatmaps.

The resulting heatmaps visualize the spatial distribution of model sensitivity to local occlusions. Positive $\Delta_{p,q}$ values indicate locations where occluding local features increases predicted temperature (warming effect), while negative values indicate locations where occlusion reduces predicted temperature (cooling effect). This analysis helps identify local urban elements that most strongly influence modeled thermal patterns and informs placement of targeted green infrastructure interventions.

2.5.5 Gradient-based explainability methods

We used two complementary gradient-based techniques to probe model behaviour: Gradient-weighted Class Activation Mapping (Grad-CAM)[43] and Integrated Gradients.[44] Let $F(X)$ denote the scalar model output (per-pixel LST may be aggregated e.g. by spatial mean to produce a scalar for Grad-CAM) and let $A_k \in \mathbb{R}^{H\times W}$ be the k-th feature map from the final convolutional layer with spatial entries $A_k^{i,j}$.

Grad-CAM: The importance weight for feature map k is computed as the spatially averaged gradient of the output with respect to that feature map:

$$\alpha_k \;=\; \frac{1}{Z}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\partial F(X)}{\partial A_k^{i,j}}, \qquad Z \;=\; H\times W. \tag{15}$$

The Grad-CAM heatmap is then obtained by a weighted linear combination of feature maps followed by a ReLU to retain positive influences:

$$L_{\text{Grad-CAM}} \;=\; \mathrm{ReLU}\left(\sum_k \alpha_k\, A_k\right). \tag{16}$$

The resulting map is upsampled to the input resolution for visualization and, if required, normalized across the dataset for population-level comparisons.

Integrated Gradients: Integrated Gradients attribute each input feature x_i by integrating gradients along a straight-line path from a baseline input x' (typically the all-zero image) to the input x:

$$\mathrm{IG}_i(x) \;=\; (x_i - x_i')\int_0^1 \frac{\partial F\big(x' + \alpha(x-x')\big)}{\partial x_i}\, d\alpha. \tag{17}$$

In practice we approximate the integral using m Riemann steps:

$$\mathrm{IG}_i(x) \;\approx\; (x_i - x_i')\,\frac{1}{m}\sum_{k=1}^{m}\frac{\partial F\big(x' + \frac{k}{m}(x-x')\big)}{\partial x_i}. \tag{18}$$

Typical choices use $m \in \{50, 100\}$ to balance accuracy and compute cost.

Aggregation: Both methods were applied per-sample across the full test set and aggregated (e.g. by per-pixel mean or median and by normalizing maps) to identify consistent spatial and channel-wise patterns of importance across diverse urban contexts.

2.6 Model Training and Evaluation Protocol

The final dataset comprised 50 × 50 windows drawn from three study cities. The held-out test set contained city-wise samples as follows: İstanbul ($N_{\text{test}} = 7{,}549$), Abu Dhabi ($N_{\text{test}} = 12{,}377$), and San Francisco ($N_{\text{test}} = 15{,}000$), for a total of $N_{\text{test}} = 34{,}926$ test windows. Spatial partitioning was performed column-wise to maintain geographic separation between training, validation and test sets. Data ingestion was implemented with a custom PyTorch Dataset class; numerical anomalies were handled at load-time via `torch.nan_to_num` (NaN, $+\infty$, $-\infty \mapsto 0$) to preserve tensor consistency during training.

Training configuration: Models were trained with a batch size of 8 using the Adam optimizer with initial learning rate $\eta_0 = 1 \times 10^{-4}$. We ran up to 50 epochs with early stopping monitored on validation loss; the checkpoint with the lowest validation RMSE was selected for downstream sensitivity analyses. When necessary, bilinear interpolation was used to align prediction and ground-truth resolutions. Training and inference were conducted on GPU when CUDA was available.

Loss function: To accommodate missing target values, we used a masked mean-squared-error loss computed over valid pixels. Let $\mathcal{M}$ denote the set of valid (non-NaN) pixel indices and $|\mathcal{M}|$ its cardinality. Then the masked MSE is

$$\text{MSE}_{\text{mask}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(\hat{y}_i - y_i\right)^2, \tag{19}$$

where $\hat{y}_i$ and y_i are the predicted and ground-truth temperature values for pixel i, respectively. This formulation ensures loss contributions only from reliable ground-truth measurements.

Evaluation metrics: Model performance was assessed using multiple complementary regression metrics computed over the held-out test set:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}, \tag{20}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{21}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{22}$$

where n is the number of evaluated pixels (or pixel-aggregates) and $\bar{y}$ is the mean of the ground-truth values. In addition to global scores, we report metric breakdowns across temperature bands and across key urban-feature strata to verify consistent performance across the thermal range and heterogeneous urban contexts.

Model selection and numerical stability: Early stopping, checkpointing, and a cosine annealing learning-rate scheduler[45] were used to avoid overfitting and to stabilize training. NaN detection and masking were enforced throughout forward and backward passes to prevent propagation of invalid values. Where predictions and labels required spatial alignment, bilinear resampling was applied prior to metric computation.

3. RESULTS

3.1 Overall model performance

Model performance was assessed on the held-out test sets from Abu Dhabi, İstanbul, and San Francisco. Table 2 reports standard regression metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

In Abu Dhabi, the model achieved an R^2 of 0.7375, indicating that it successfully captured a substantial portion of the variance in LST. Performance in İstanbul was moderate, with an R^2 of 0.4352. By contrast, results in San Francisco showed a negative R^2 value of -0.1483, implying that the model's predictions were less accurate than a naïve baseline that predicts the mean LST. This divergence underscores the limitations of direct model transferability across urban regions with distinct climatic, geographic, and morphological characteristics, emphasizing the need for context-specific adaptation.

Table 2: Performance of the U-Net model on held-out test sets for each study city.

City	**Valid Pixels**	**RMSE (°C)**	**MAE (°C)**	R^2
Abu Dhabi	6,188,500	2.78	2.10	0.738
İstanbul	20,542,500	2.50	2.04	0.435
San Francisco	1,715,000	2.54	2.23	-0.148

Qualitative comparisons, presented in Figure 2, illustrate the spatial agreement between ground-truth and predicted LST maps for representative samples from each city. Predictions for Abu Dhabi and Istanbul closely matched observed thermal patterns, accurately mapping hotspots over dense built-up areas and cooler zones associated with parks and water bodies, which was consistent with the quantitative evaluation.

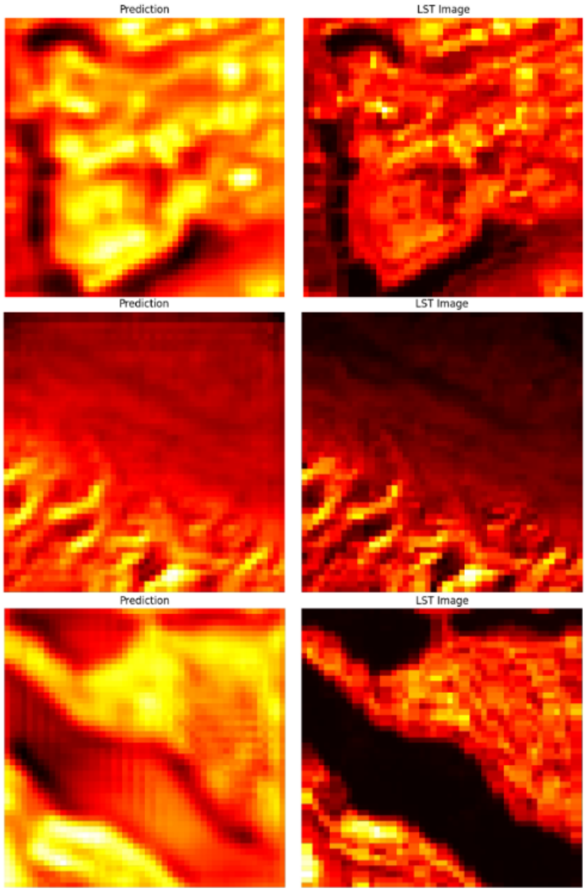

Figure 2: Qualitative comparison of predicted and ground-truth LST maps for representative samples from three cities (top: İstanbul, middle: Abu Dhabi, bottom: San Francisco). In each row, the U-Net prediction is shown on the left and the ground-truth map on the right, illustrating the model's ability to reproduce key urban heat patterns.

3.2 Sensitivity Analysis Results

To better understand how the model leverages spectral indices in predicting land surface temperature (LST), we applied our five-method sensitivity analysis framework to the held-out test sets from İstanbul, Abu Dhabi, and San Francisco. The subsequent sections present results from channel ablation, controlled input perturbations, statistical significance testing, spatial occlusion, and gradient-based explainability techniques.

3.2.1 Analysis for İstanbul

We evaluated the İstanbul test set ($N_{\text{test}} = 8{,}217$ windows) using five complementary sensitivity analysis methods: channel ablation, controlled perturbation, statistical testing, spatial occlusion, and gradient-based explainability. Results quantify the influence of individual spectral indices on per-pixel LST predictions. Reported p-values are compared against a significance threshold of $\alpha = 0.01$, with effect sizes given as Cohen's d. For MD estimates, bootstrap 95% BCa confidence intervals are also provided.

Channel Ablation: The channel ablation analysis for İstanbul revealed distinct contributions of each spectral index to LST predictions. Vegetation-related indices had strong cooling effects when removed: EVI showed the largest impact with a mean difference (MD) of −2.63°C, followed by SAVI (−1.14°C) and NDVI (−0.75°C). The negative MD values indicate that removing these indices led to higher predicted temperatures, confirming their cooling influence on the urban thermal environment.

In contrast, built-up indices contributed to warming. IBI had the strongest warming effect with an MD of +2.87°C, while NDBI showed a moderate cooling effect with an MD of −0.33°C. The differing responses of IBI and NDBI suggest sensitivity to distinct aspects of urban development patterns. Notably, NDWI displayed a substantial warming effect when removed (MD +2.75°C), highlighting the significant cooling influence of water features in İstanbul's urban landscape.

Perturbation analysis: We applied multiplicative perturbations $c \mapsto c \pm \varepsilon c$ with $\varepsilon \in \{0.01, 0.05, 0.1\}$ and measured MD for both + and − perturbations. Table 3 summarizes the average MDs across the test set; detailed statistical test outputs for each perturbation (t/Wilcoxon, Cohen's d, bootstrap CI) are provided afterward.

Table 3: İstanbul — Perturbation MDs (average over test set). Positive MD = predicted temperature increases after the perturbation.

Channel	$\text{MD}_{+0.01}$	$\text{MD}_{-0.01}$	$\text{MD}_{+0.05}$	$\text{MD}_{-0.05}$	$\text{MD}_{+0.1}$	$\text{MD}_{-0.1}$
NDVI	0.2922	−0.2678	1.5318	−0.9096	2.9827	−1.3519
NDBI	0.1239	−0.1062	0.7332	−0.3295	1.6151	−0.3331
NDWI	−0.3239	0.3614	−1.2372	1.8442	−2.4648	3.9700
SAVI	0.2140	−0.2048	1.0853	−0.8330	2.3270	−1.2088
IBI	−0.0314	0.0373	−0.0950	0.2457	−0.0502	0.6385
EVI	0.2837	−0.2699	1.5227	−1.1139	3.4400	−1.7627

Across $\varepsilon \in \{0.01, 0.05, 0.1\}$ the model responses scale with perturbation magnitude. Vegetation indices (NDVI, SAVI, EVI) produce cooling when increased and warming when decreased, with EVI the most sensitive (up to +3.44°C at +10%, down to −1.76°C at −10%) and NDVI showing an approximately linear response (about −0.27°C at −1% to +2.98°C at +10%). Built-up indices are more complex: NDBI behaves roughly linearly (e.g., ~ 1.62°C at +10%, ~ -0.33°C at −10%), while IBI is nearly insensitive at $\varepsilon = 0.01$ ($\sim \pm 0.03$°C) but becomes asymmetric at larger perturbations. NDWI shows the strongest inverse effect (cooling up to −2.46°C for +10%, warming up to +3.97°C for −10%).

Statistical Significance: Perturbations and ablations were evaluated using paired t-tests and Wilcoxon signed-rank tests. All sensitivity analyses were significant at $\alpha = 0.01$, with the strongest effects yielding p-values near machine precision ($p \ll 10^{-160}$). Cohen's d values indicate large effects for vegetation and select non-vegetation channels. In particular, EVI ablation exhibited the largest vegetation effect ($d \approx -1.75$, very large), followed by SAVI ($d \approx -1.68$) and NDVI (approximately $d \approx -1.77$). Among non-vegetation indices, IBI (built-index) reached $d \approx +2.87$ and NDWI (water-related) reached $d \approx +2.21$. Bootstrap resampling produced narrow 95% BCa confidence intervals, supporting the stability of these estimates. For example, EVI ablation: 95% CI = $[-2.77, -2.59]$ K, and IBI ablation: 95% CI = $[+2.78, +2.91]$ K. Parametric (paired t) and non-parametric (Wilcoxon signed-rank) tests were concordant across channels, reinforcing the robustness of the inferences.

Perturbation experiments showed monotonic scaling of effect size with perturbation magnitude ε ($c \mapsto c \pm \varepsilon c$). Effect sizes increased rapidly for the strongest channels; for instance, NDVI at $\varepsilon = 0.1$ yielded Cohen's $d \approx 6.82$, highlighting the model's high sensitivity to large vegetation changes. Combined with narrow bootstrap intervals and agreement between parametric and non-parametric tests, these large effect sizes provide strong statistical support for the causal interpretations from our multi-method sensitivity framework.

Spatial Occlusion: The spatial occlusion analysis revealed that the model's sensitivity is not uniformly distributed across the input windows. Sensitivity values ranged from −4.46°C to +4.74°C (mean 0.11°C, std 0.44°C), indicating substantial spatial variability in feature importance. The slightly positive mean suggests that occluding most spatial locations tends to increase predicted temperatures modestly, consistent with the predominance of cooling features such as vegetation and water in typical urban scenes.

The aggregated sensitivity map in Figure 3 highlights hotspots where the model is most sensitive, typically at land cover transitions such as green space–residential interfaces or water edges, identifying critical locations for targeted urban heat island mitigation.

Gradient Based Methods: Grad-CAM analysis produced attention maps with a mean activation of 0.17 (SD = 0.23), indicating focused sensitivity on specific spatial regions. Figures 4 and 5 highlight locations most influencing the model's temperature predictions. Figure 4 shows the average Grad-CAM map, while Figure 5 presents average Integrated Gradients maps for the six channels, providing interpretable insight into the model's decisions.

Integrated Gradients maps across six channels showed a mean attribution of 0.0003 (SD = 0.0032). Vegetation indices generally indicate cooling, and built-up indices indicate warming, with low magnitude and high spatial variability reflecting context-dependent urban thermal dynamics. Grad-CAM highlights regions of high occlusion sensitivity, while Integrated Gradients align with channel-wise perturbation effects.

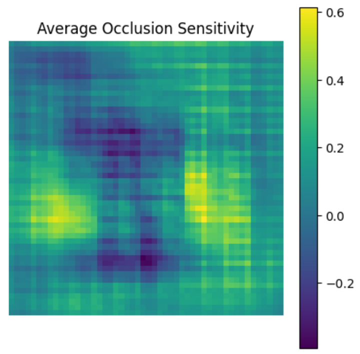

Figure 3: Aggregated spatial sensitivity map for İstanbul from the 5 × 5 patch occlusion analysis.

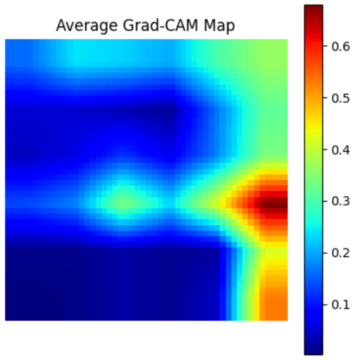

Figure 4: Aggregated spatial Grad-CAM sensitivity map for İstanbul.

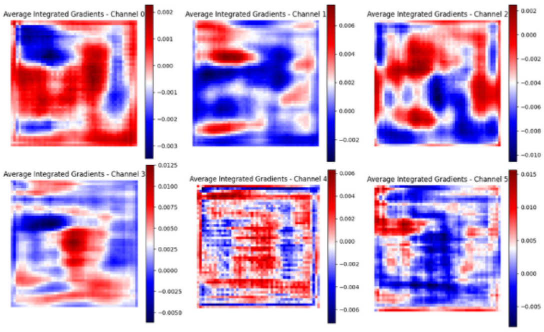

Figure 5: Integrated Gradients maps for İstanbul in a 2 × 3 grid: NDVI, NDBI, NDWI (top) and SAVI, IBI, EVI (bottom).

3.2.2 Analysis for Abu Dhabi

We evaluated the Abu Dhabi test set ($N_{\text{test}} = 2{,}764$ windows) using the same five-method sensitivity analysis framework applied to İstanbul and San Francisco. The arid climate and distinct urban morphology of Abu Dhabi yielded markedly different and often counter-intuitive sensitivity patterns, with consistently larger effect magnitudes across all spectral indices compared to the temperate İstanbul context.

Channel Ablation: The channel ablation analysis in Abu Dhabi revealed a striking pattern: the removal of any spectral index led to a significant increase in predicted LST, indicating that all six channels were interpreted as exerting a cooling influence. This suggests that the model may have learned highly context-specific relationships within the desert environment.

The strongest cooling effects were unexpectedly linked to built-up indices. IBI showed the largest impact by far (MD = +35.97°C), followed by NDBI (MD = +22.59°C). This counter-intuitive result may reflect the model's association of these indices with urban features such as building height and self-shading, which can provide localized cooling in intensely hot conditions. Vegetation indices also exhibited substantial cooling effects, with SAVI (MD = +16.88°C) and EVI (MD = +16.81°C) showing nearly identical magnitudes. NDWI (MD = +8.89°C) and NDVI (MD = +4.44°C) contributed as well, though to a lesser extent.

Perturbation analysis: Table 4 lists the average MDs under multiplicative plus/minus perturbations $c \mapsto c \pm \varepsilon c$ for $\varepsilon \in \{0.01, 0.05, 0.1\}$.

Table 4: Abu Dhabi — Perturbation MDs (average over test set). Positive MD = predicted temperature increases after the perturbation

Channel	$\text{MD}_{+0.01}$	$\text{MD}_{-0.01}$	$\text{MD}_{+0.05}$	$\text{MD}_{-0.05}$	$\text{MD}_{+0.1}$	$\text{MD}_{-0.1}$
NDVI	2.7427	1.3299	7.6961	3.1211	20.4321	4.2618
NDBI	1.1318	1.7688	4.2724	9.8335	5.3494	20.0772
NDWI	0.7294	0.4552	2.8427	2.6689	4.5662	4.2598
SAVI	2.5223	1.6771	3.5135	4.3488	4.5442	5.2015
IBI	0.8453	0.5893	3.6801	5.1859	4.9035	11.9629
EVI	0.9158	0.7657	3.3013	4.7757	3.7533	7.2919

Perturbation analysis for Abu Dhabi revealed asymmetric responses with larger effects than in İstanbul, reflecting the desert's extreme thermal gradients.

For built-up indices, negative perturbations produced particularly strong cooling: NDBI reached MD = +20.01°C versus +5.47°C for positive perturbations at $\varepsilon = 0.1$, while IBI showed +11.86°C versus +4.90°C. Vegetation indices were more symmetric but still substantial; NDVI scaled strongly, with MD = +20.24°C for positive versus +4.21°C for negative perturbations, suggesting vegetation enhancement provides disproportionate cooling. Water-related indices (NDWI) showed moderate, relatively symmetric effects, ranging from +0.45°C to +4.65°C.

Statistical Significance: All sensitivity analyses achieved statistical significance well below the $\alpha = 0.01$ threshold, with effect sizes substantially exceeding those observed in İstanbul. The most pronounced effect was observed for IBI ablation, yielding Cohen's $d = 8.58$ (extremely large), followed by NDBI ($d = 4.76$) and EVI ($d = 4.11$), among the largest effects typically reported in environmental remote sensing studies.

Bootstrap confidence intervals confirmed the stability of these estimates despite the extreme effect magnitudes. IBI ablation showed a 95% CI of $[35.69, 36.22]$°C, while NDBI ablation had $[22.28, 22.90]$°C. The narrow intervals relative to effect sizes indicate robust estimation across resampling iterations.

Perturbation experiments exhibited systematic scaling, with Cohen's d reaching exceptional levels for the most sensitive channels. For example, NDVI at $\varepsilon = 0.1$ produced $d = 5.04$ for positive perturbations, while NDBI reached $d = 4.46$ for negative perturbations at the same level. Both parametric and non-parametric tests yielded p-values near machine precision ($p < 10^{-160}$) for all major effects, providing unequivocal statistical support. The agreement across statistical tests, combined with extremely large effect sizes and narrow confidence intervals, provides strong evidence for the causal influence of urban features on thermal patterns in desert environments.

Spatial Occlusion: Spatial occlusion analysis for Abu Dhabi revealed considerable variability in model sensitivity, ranging from −9.84°C to +8.63°C. The overall mean sensitivity was positive and higher than in İstanbul (mean 0.36°C, SD 1.90°C), consistent with the ablation results showing that most landscape features exert a net cooling effect. Occluding a random patch is thus more likely to remove a cooling feature, increasing predicted LST. The aggregated sensitivity map (Figure 6) highlights key areas of sensitivity, likely at interfaces between irrigated green spaces, dense building clusters, and the surrounding desert.

Gradient Based Methods: Gradient-based methods offered visual insight into the model's behavior in Abu Dhabi. Grad-CAM analysis produced maps with a mean activation of 0.23, highlighting the model's focus on key features within the input windows. The aggregated Grad-CAM and Integrated Gradients (IG) maps (Figures 7 and 8) illustrate these areas of focus. The IG maps, with a mean attribution of 0.0013, visually confirm the quantitative results, showing negative attributions (cooling) across all six channels, including the built-up IBI and NDBI indices.

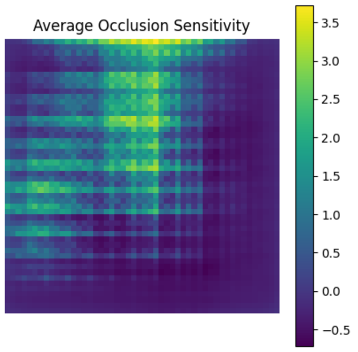

Figure 6: Aggregated spatial sensitivity map for Abu Dhabi from the 5 × 5 patch occlusion analysis.

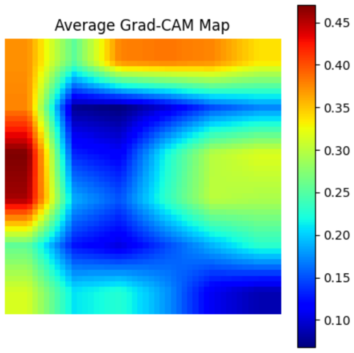

Figure 7: Aggregated spatial Grad-CAM sensitivity map for Abu Dhabi.

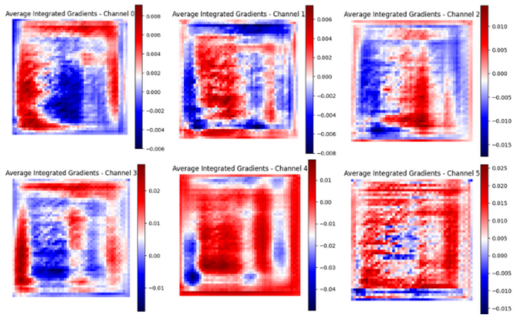

Figure 8: Integrated Gradients maps for Abu Dhabi in a 2 × 3 grid: NDVI, NDBI, NDWI (top) and SAVI, IBI, EVI (bottom).

3.2.3 Analysis for San Francisco

We evaluated the San Francisco test set ($N_{\text{test}} = 686$ windows) using the same five-method sensitivity analysis framework. The temperate, maritime-influenced climate and heterogeneous urban morphology of San Francisco produced sensitivity patterns that differ from both İstanbul and Abu Dhabi, with several indices showing strong, and in some cases counter-intuitive, directional effects.

Channel Ablation: The channel ablation analysis for San Francisco revealed a mixed pattern of cooling and warming effects when spectral indices were removed, more similar to İstanbul than Abu Dhabi but with distinct characteristics. Built-up indices demonstrated the strongest cooling effects when removed: NDBI showed the largest impact with $MD = -8.44\,°C$, indicating substantial warming influence of built surfaces in San Francisco's temperate climate. EVI exhibited a similarly strong cooling effect when removed ($MD = -7.01\,°C$), confirming the importance of enhanced vegetation for thermal regulation.

NDVI showed a modest cooling effect when removed ($MD = -0.37\,°C$), suggesting limited sensitivity to basic vegetation measures in this Mediterranean climate context. In contrast, water-related and soil-adjusted vegetation indices showed warming effects when removed: NDWI contributed $MD = +1.42\,°C$ and SAVI contributed $MD = +1.24\,°C$, indicating their cooling influence. IBI demonstrated a moderate warming effect when removed ($MD = +0.87\,°C$), suggesting built-up areas identified by this index provide some cooling relative to other urban features.

The pattern in San Francisco differs notably from both İstanbul and Abu Dhabi, with built-up indices (NDBI) showing strong warming effects and vegetation indices (EVI) showing strong cooling effects, but with SAVI unexpectedly contributing to warming when removed.

Perturbation analysis: We applied multiplicative perturbations $c \mapsto c \pm \varepsilon c$ ($\varepsilon \in \{0.01, 0.05, 0.1\}$) and recorded average MDs for plus/minus perturbations (Table 5).

Table 5: San Francisco — Perturbation MDs (average over test set). Positive MD = predicted temperature increases after the perturbation.

Channel	$\mathrm{MD}_{+0.01}$	$\mathrm{MD}_{-0.01}$	$\mathrm{MD}_{+0.05}$	$\mathrm{MD}_{-0.05}$	$\mathrm{MD}_{+0.1}$	$\mathrm{MD}_{-0.1}$
NDVI	1.3008	−1.2808	2.4467	−5.9379	2.1948	−1.7583
NDBI	1.8412	−7.4159	2.6524	−6.5390	2.5676	−5.2586
NDWI	0.0088	0.0587	1.1259	2.0436	2.7548	2.6315
SAVI	−6.3470	1.7172	−6.3392	2.9699	−3.9819	3.2897
IBI	−0.8426	0.8442	−1.9292	1.7803	−4.2471	2.2541
EVI	1.6625	−5.9562	3.0437	−6.9877	3.5853	−5.1267

Perturbation analysis revealed highly asymmetric responses across different spectral indices, with some channels showing extreme sensitivity to directional changes. Built-up indices (NDBI) exhibited strong asymmetry: negative perturbations (reduced built-up area) consistently produced large cooling effects ranging from $MD = -7.42\,°\mathrm{C}$ at $\epsilon = 0.01$ to $MD = -5.26\,°\mathrm{C}$ at $\epsilon = 0.1$, while positive perturbations showed modest warming effects ($MD = +1.84\,°\mathrm{C}$ to $+2.57\,°\mathrm{C}$). This asymmetry suggests that reducing built-up intensity provides disproportionately large cooling benefits in San Francisco's climate.

SAVI demonstrated counterintuitive behavior with positive perturbations (vegetation enhancement) producing cooling effects ranging from $MD = -6.35\,°\mathrm{C}$ to $MD = -3.98\,°\mathrm{C}$, while negative perturbations (vegetation reduction) showed warming effects ($MD = +1.72\,°\mathrm{C}$ to $+3.29\,°\mathrm{C}$). This pattern, opposite to conventional expectations, may reflect complex soil-vegetation interactions in San Francisco's Mediterranean environment.

EVI showed expected vegetation patterns with positive perturbations yielding warming effects ($MD = +1.66\,°\mathrm{C}$ to $+3.59\,°\mathrm{C}$) and negative perturbations producing cooling effects ($MD = -5.96\,°\mathrm{C}$ to $-5.13\,°\mathrm{C}$). The asymmetry suggests that vegetation loss has disproportionately large warming impacts compared to vegetation gains. Water indices (NDWI) showed minimal sensitivity at low perturbation levels but increasing symmetric responses at higher levels, reaching MD values around $+2.75\,°\mathrm{C}$ for both positive and negative perturbations at $\epsilon = 0.1$.

Statistical Significance: All major sensitivity effects were significant at $\alpha = 0.01$, though effect sizes were smaller than in Abu Dhabi: NDBI ablation produced Cohen's $d = -3.07$ (very large) and EVI ablation $d = -1.97$ (large). Bootstrap CIs were stable despite the modest sample size ($N = 686$): NDBI 95% CI $[-8.63, , -8.25],°$ C and EVI $[-7.28, , -6.76],°$ C, supporting robust estimation. Perturbation tests similarly returned highly significant results for most channel ε combinations. Notable effect sizes include NDBI (large negative effects for $-$ perturbations, e.g., $d \approx -4.12$ at $\varepsilon = 0.1$ minus) and EVI (large negative effect sizes for some minus perturbations, e.g., $d \approx -3.59$ at $\varepsilon = 0.1$ minus). NDWI at the smallest perturbation ($\varepsilon = 0.01$) was not significant (Wilcoxon $p \approx 0.286$) indicating negligible sensitivity to very small NDWI changes in some contexts.

Spatial Occlusion: Spatial occlusion analysis in San Francisco revealed the most constrained sensitivity range among all three cities, spanning from −1.88 °C to +0.42 °C (mean = −0.10 °C, std = 0.15 °C). The negative mean ($\approx -0.10, \mathrm{K}$) indicates that occluding a random patch is, on average, slightly more likely to reduce predicted temperature — i.e., many patches contain features the model treats as warming in San Francisco, in contrast to İstanbul and Abu Dhabi where occlusion means were positive. The relatively narrow occlusion-sensitivity range (min $\approx -1.9, \mathrm{K}$ to max $\approx +0.42, \mathrm{K}$) points to less extreme local sensitivity than Abu Dhabi, reflecting San Francisco's more heterogeneous but less extreme urban/thermal contrasts. At the population level, the occlusion map (Figure 9) pinpoints the spatial locations—dense built-up corridors, large impervious surfaces, and microclimates near water and parks—where occlusion produces the largest local impacts.

Gradient Based Methods: Grad-CAM activations are the strongest among the three cities (population mean ≈ 0.305), indicating that the model focuses on well-defined spatial structures in San Francisco when predicting LST. Integrated Gradients show small raw magnitudes but, when aggregated channel-wise and spatially, reveal that built-up (NDBI, IBI) and vegetation (EVI, NDVI) channels contribute substantially to the model's attributions, consistent with the ablation and perturbation results. Areas highlighted by Grad-CAM generally correspond to occlusion hotspots and to locations where perturbations produce large MDs, providing convergent spatial evidence of the urban elements driving the model's predictions.

The per-channel attribution patterns (Figures 10 and 11) further reveal complex spatial relationships that align with the mixed positive and negative effects observed in the ablation analysis.

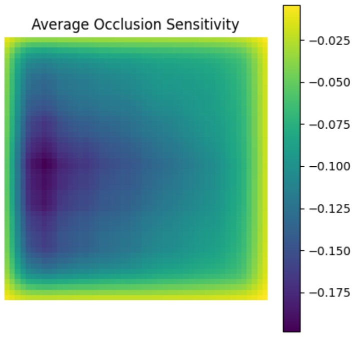

Figure 9: Aggregated spatial sensitivity map for San Francisco from the 5 × 5 patch occlusion analysis.

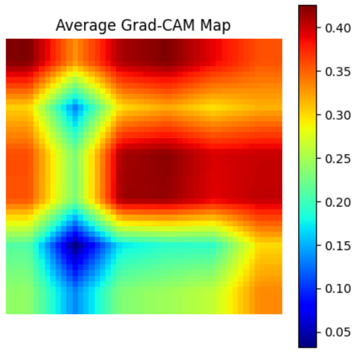

Figure 10: Aggregated spatial Grad-CAM sensitivity map for San Francisco.

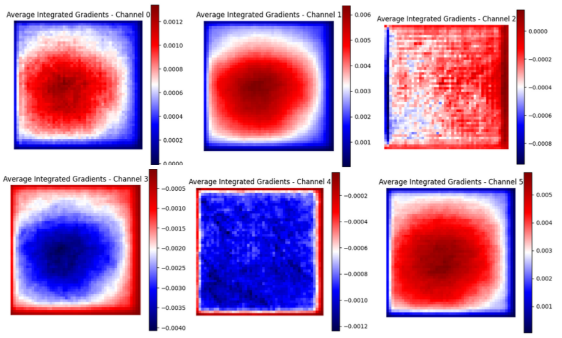

Figure 11: Integrated Gradients maps for San Francisco in a 2 × 3 grid: NDVI, NDBI, NDWI (top) and SAVI, IBI, EVI (bottom).

3.3 Cross-City Synthesis and Implications for UHI Mitigation

Collectively, the three-city analysis demonstrates both consistent patterns and strong context dependence. Across all methods vegetation signals were generally associated with cooling (e.g., İstanbul: EVI MD ≈ −2.63°C, SAVI MD ≈ −1.14°C; San Francisco: EVI MD ≈ −7.01°C), and these effects are supported by large effect sizes and narrow bootstrap intervals. However, the magnitude and even the spectral drivers differ markedly by climate and morphology: Abu Dhabi yielded substantially larger sensitivities (e.g., IBI ablation MD ≈ +35.97°C, NDBI MD ≈ +22.59°C; Cohen's d up to ~ 8.6), while İstanbul and San Francisco show moderate-to-large but far smaller effects (İstanbul EVI $d \approx -1.75$, San Francisco NDBI $d \approx -3.07$). Spatial diagnostics further emphasize these contrasts — mean occlusion sensitivities are positive in İstanbul (≈ +0.11°C) and Abu Dhabi (≈ +0.36°C), but slightly negative in San Francisco (≈ −0.10°C), reflecting different prevalences of cooling versus warming features and the model's localized attention (Grad-CAM mean activations: İstanbul ≈ 0.17, Abu Dhabi ≈ 0.23, San Francisco ≈ 0.31). Taken together, the results show that the model is highly sensitive to vegetation and water features, highlight strong and sometimes counter-intuitive built-environment effects in arid contexts, and emphasize the need to tailor interventions to local climate, morphology, and spatial heterogeneity.

4. CONCLUSION

Our research developed and evaluated a deep learning framework for modeling urban heat island effects across diverse climatic contexts, using İstanbul, San Francisco, and Abu Dhabi as case studies. The compact U-Net architecture achieved variable performance across cities (Abu Dhabi $R^2 = 0.74$, İstanbul $R^2 = 0.44$, San Francisco $R^2 = -0.15$), highlighting both the potential and limitations of cross-regional thermal modeling. The comprehensive five-method sensitivity analysis framework revealed consistent cooling influences from vegetation indices across all contexts, with effect magnitudes varying dramatically by climate—EVI removal increased temperatures by 2.63 °C in İstanbul compared to 16.81 °C in Abu Dhabi, illustrating how desert environments amplify the importance of limited vegetation. Built-up indices showed complex, context-dependent relationships, with counterintuitive cooling effects in Abu Dhabi (IBI removal: +35.97 °C) possibly reflecting building shading or thermal mass effects in extreme heat conditions. The spatial occlusion analysis revealed heterogeneous sensitivity patterns that provide actionable guidance for targeted urban planning interventions, while gradient-based explainability methods offered convergent evidence supporting the causal interpretations. Despite limitations in cross-regional transferability, the robust statistical validation (including bootstrap confidence intervals and effect size quantification) establishes the framework's value for evidence-based UHI mitigation strategies. The findings underscore the critical importance of context-specific approaches to urban climate adaptation, demonstrating that while universal solutions remain elusive, careful analysis of local thermal dynamics can provide actionable insights for resource-efficient mitigation strategies tailored to the unique characteristics of different urban environments and climatic zones.

ACKNOWLEDGMENTS

This work was developed within the framework of the UDENE project and funded by its initiative. The authors would like to express their sincere gratitude to Ali Türker, PMP, UDENE Project Coordinator, for his guidance and support throughout the project.

REFERENCES

[1] Myrup, L. O., "A numerical model of the urban heat island," *Journal of Applied Meteorology and Climatology* **8**, 908–918 (1969).

[2] UDENE, "Urban development explorations using natural experiments." https://udene.eu/. Accessed: 24 August 2025.

[3] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).

[4] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R., "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote sensing of Environment* **202**, 18–27 (2017).

[5] Kriegler, F. J., "Preprocessing transformations and their effects on multspectral recognition," in [*Proceedings of the sixth international symposium on remote sesning of environment*], 97–131 (1969).

[6] Zha, Y., Gao, J., and Ni, S., "Use of normalized difference built-up index in automatically mapping urban areas from tm imagery," *International journal of remote sensing* **24**(3), 583–594 (2003).

[7] Gao, B.-C., "Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space," *Remote sensing of environment* **58**(3), 257–266 (1996).

[8] Huete, A. R., "A soil-adjusted vegetation index (savi)," *Remote sensing of environment* **25**(3), 295–309 (1988).

[9] Xu, H., "A new index for delineating built-up land features in satellite imagery," *International journal of remote sensing* **29**(14), 4269–4276 (2008).

[10] Jiang, Z., Huete, A. R., Didan, K., and Miura, T., "Development of a two-band enhanced vegetation index without a blue band," *Remote sensing of Environment* **112**(10), 3833–3845 (2008).

[11] Oke, T. R., "The energetic basis of the urban heat island," *Quarterly journal of the royal meteorological society* **108**(455), 1–24 (1982).

[12] Arnfield, A. J., "Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island," *International Journal of Climatology: a Journal of the Royal Meteorological Society* **23**(1), 1–26 (2003).

[13] Weng, Q., "Thermal infrared remote sensing for urban climate and environmental studies: Methods, applications, and trends," *ISPRS Journal of photogrammetry and remote sensing* **64**(4), 335–344 (2009).

[14] Peng, S., Piao, S., Ciais, P., Friedlingstein, P., Ottle, C., Bréon, F.-M., Nan, H., Zhou, L., and Myneni, R. B., "Surface urban heat island across 419 global big cities," *Environmental science & technology* **46**(2), 696–703 (2012).

[15] Manoli, G., Fatichi, S., Schläpfer, M., Yu, K., Crowther, T. W., Meili, N., Burlando, P., Katul, G. G., and Bou-Zeid, E., "Magnitude of urban heat islands largely explained by climate and population," *Nature* **573**(7772), 55–60 (2019).

[16] Jha, P., Joy, M. S., Yadav, P. K., Begam, S., and Bansal, T., "Detecting the role of urban green parks in thermal comfort and public health for sustainable urban planning in delhi," *Discover Public Health* **21**(1), 236 (2024).

[17] Chang, C.-R., Li, M.-H., and Chang, S.-D., "A preliminary study on the local cool-island intensity of taipei city parks," *Landscape and urban planning* **80**(4), 386–395 (2007).

[18] Lin, B.-S. and Lin, C.-T., "Preliminary study of the influence of the spatial arrangement of urban parks on local temperature reduction," *Urban Forestry & Urban Greening* **20**, 348–357 (2016).

[19] Gunawardena, K. R., Wells, M. J., and Kershaw, T., "Utilising green and bluespace to mitigate urban heat island intensity," *Science of the total environment* **584**, 1040–1055 (2017).

[20] Norton, B. A., Coutts, A. M., Livesley, S. J., Harris, R. J., Hunter, A. M., and Williams, N. S., "Planning for cooler cities: A framework to prioritise green infrastructure to mitigate high temperatures in urban landscapes," *Landscape and urban planning* **134**, 127–138 (2015).

[21] Cetin, M., Ozenen Kavlak, M., Senyel Kurkcuoglu, M. A., Bilge Ozturk, G., Cabuk, S. N., and Cabuk, A., "Determination of land surface temperature and urban heat island effects with remote sensing capabilities: the case of kayseri, türkiye," *Natural Hazards* **120**(6), 5509–5536 (2024).

[22] Kikon, N., Kumar, D., and Ahmed, S. A., "Quantitative assessment of land surface temperature and vegetation indices on a kilometer grid scale," *Environmental Science and Pollution Research* **30**(49), 107236–107258 (2023).

[23] Grover, A. and Singh, R., "Monitoring spatial patterns of land surface temperature and urban heat island for sustainable megacity: A case study of mumbai, india, using landsat tm data," *Environment and Urbanization ASIA* **7**(1), 38–54 (2016).

[24] Zhang, L., Zhang, L., and Du, B., "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and remote sensing magazine* **4**(2), 22–40 (2016).

[25] Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A., "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing* **152**, 166–177 (2019).

[26] Kattenborn, T., Leitloff, J., Schiefer, F., and Hinz, S., "Review on convolutional neural networks (cnn) in vegetation remote sensing," *ISPRS journal of photogrammetry and remote sensing* **173**, 24–49 (2021).

[27] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, F., "Deep learning and process understanding for data-driven earth system science," *Nature* **566**(7743), 195–204 (2019).

[28] Camps-Valls, G., Tuia, D., Zhu, X. X., and Reichstein, M., [*Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*], John Wiley & Sons (2021).

[29] Bouaziz, S., Hafiane, A., Canals, R., and Nedjai, R., "Deep learning for spatio-temporal fusion in land surface temperature estimation: A comprehensive survey, experimental analysis, and future trends," *arXiv preprint arXiv:2412.16631* (2024).

[30] Kustura, K., Conti, D., Sammer, M., and Riffler, M., "Harnessing multi-source data and deep learning for high-resolution land surface temperature gap-filling supporting climate change adaptation activities," *Remote Sensing* **17**(2), 318 (2025).

[31] Huang, C., Liu, K., Ma, T., Xue, H., Wang, P., and Li, L., "Analysis of the impact mechanisms and driving factors of urban spatial morphology on urban heat islands," *Scientific Reports* **15**(1), 18589 (2025).

[32] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system," in [*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*], 785–794 (2016).

[33] Lundberg, S. M. and Lee, S.-I., "A unified approach to interpreting model predictions," *Advances in neural information processing systems* **30** (2017).

[34] Syeda, A. Q., Castillo-Villar, K. K., and Alaeddini, A., "Sustainable urban heat island mitigation through machine learning: Integrating physical and social determinants for evidence-based urban policy," *Sustainability* **17**(15), 7040 (2025).

[35] Nair, V. and Hinton, G. E., "Rectified linear units improve restricted boltzmann machines," in [*Proceedings of the 27th international conference on machine learning (ICML-10)*], 807–814 (2010).

[36] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*International conference on machine learning*], 448–456, pmlr (2015).

[37] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[38] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[39] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300 (1995).

[40] Cohen, J., [*Statistical Power Analysis for the Behavioral Sciences*], Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd ed. (1988).

[41] Wilcoxon, F., "Individual comparisons by ranking methods," in [*Breakthroughs in statistics: Methodology and distribution*], 196–202, Springer (1992).

[42] Zeiler, M. D. and Fergus, R., "Visualizing and understanding convolutional networks," in [*European conference on computer vision*], 818–833, Springer (2014).

[43] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*Proceedings of the IEEE international conference on computer vision*], 618–626 (2017).

[44] Sundararajan, M., Taly, A., and Yan, Q., "Axiomatic attribution for deep networks," in [*International conference on machine learning*], 3319–3328, PMLR (2017).

[45] Loshchilov, I. and Hutter, F., "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983* (2016).