Sustainable Cities and Society 149 (2026) 107734

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

Semantic enrichment of 3D city models via roof material classification for urban greening and heat island mitigation

Elmehdi Kanna *, Jannik Matijevic, Lukas Arzoumanidis, Huynh Duc An Son Nguyen, Youness Dehbi

HafenCity University Hamburg, Computational Methods Lab, Henning-Voscherau-Platz 1, Hamburg, 20457, Germany

ARTICLE INFO

Dataset link: Github

Keywords:
Roof material classification
Semantic enrichment
CityGML
Urban heat island
Land surface temperature

ABSTRACT

Semantic enrichment extends 3D city models, which already encode object-level semantics such as building parts and surfaces, with thematic attributes that make them suitable for a broad range of urban analyses, and building energy and microclimate studies. This article addresses a frequently missing but vital attribute in such models, namely roof material, by combining high resolution RGB image classification with footprint guided background suppression. Building footprints from OpenStreetMap (OSM) are rasterised into masks that remove non-roof pixels, allowing the classifier to focus on material cues rather than surrounding urban context. An image classification network is then trained to distinguish five roof material classes. Predicted materials are written back into a semantically and geometrically rich City Geography Markup Language (CityGML) dataset as per-building attributes and used to drive a streamlined urban heat island screening workflow. This enables estimation of baseline roof temperatures as well as scenario-based changes under cool-roof and green-roof substitutions. A citywide greening simulation of suitable roof candidates across Hamburg (Germany), Paris (France), and Madrid (Spain) indicates an overall cooling on roofs of approximately 0.83 K across Hamburg, 0.16 K across Paris, and 0.6 K across Madrid in the most favorable scenarios. The proposed pipeline is data-efficient, reproducible, and deployable at city scale. To the best of the authors' knowledge, this is the first pipeline to explicitly combine footprint-guided roof material classification with city-scale urban heat island screening on CityGML data in a single end-to-end workflow. The full pipeline code is available at: Github.

1. Introduction

Semantic 3D city models support a wide range of urban analysis tasks, owing in large part to the rich semantic information they carry beyond purely geometric representations. Yet many attributes essential to such analyses are often absent from publicly available datasets. Urban heat island (UHI) screening is a prominent example: without knowledge of building roof materials, heat-absorbing and reflective elements cannot be distinguished, causing fundamental uncertainty in the resulting simulations that additional geometric detail alone cannot resolve. This type of information is far from academic; it has direct implications for urban climate adaptation, which ranks among the most pressing climate challenges cities face today.

Global surface temperatures from 2011 to 2020 averaged 1.1°C above the 1850–1900 baseline, and heat extremes have grown more frequent and more severe (IPCC, 2023). This causes more heat related illnesses, added strain on power grids, and lost work on the hottest days. Therefore, cool roofs and urban greening became standard requirements in city planning (Dodman et al., 2022). It is against this backdrop that UHI screenings earn their importance by giving decision-makers the quantitative grounding needed to target interventions, weigh trade-offs, and evaluate the cooling potential of greening strategies at meaningful scales.

Delivering that quantitative grounding, however, requires a data substrate capable of carrying both geometry and the thematic attributes that physical models demand and this is precisely the role that semantic 3D city models, and CityGML in particular, are designed to play.

CityGML is a widely adopted Open Geospatial Consortium (OGC) standard for semantically rich, interoperable 3D city models (Gröger et al., 2012). Yet the practical value of any public CityGML dataset depends on what individual providers choose to release, and content and quality vary considerably between countries and federal states. Prior audits have documented recurring geometric and topological issues, along with inconsistencies in semantic labeling that can degrade downstream simulation (Biljecki et al., 2016; Ledoux, 2018), but these

* Corresponding author.
E-mail addresses: elmehdi.kanna@hcu-hamburg.de (E. Kanna), jannik.matijevic@hcu-hamburg.de (J. Matijevic), lukas.arzumanidis@hcu-hamburg.de (L. Arzoumanidis), son.nguyen@hcu-hamburg.de (H.D.A.S. Nguyen), youness.dehbi@hcu-hamburg.de (Y. Dehbi).

https://doi.org/10.1016/j.scs.2026.107734
Received 21 January 2026; Received in revised form 22 July 2026; Accepted 23 July 2026
Available online 4 August 2026

Nomenclature

Abbreviations

3DCityDB	3D City Database
ADE	Application Domain Extension
CityGML	City Geography Markup Language
CRS	Coordinate Reference System
DOP20	Digitales Orthophoto, 20 cm ground sampling distance
ECE	Expected Calibration Error
F1	Harmonic mean of precision and recall
GMLID	Geography Markup Language identifier
IPCC	Intergovernmental Panel on Climate Change
LoD	Level of Detail
LST	Land Surface Temperature
NDBI	Normalized Difference Built-up Index
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
OGC	Open Geospatial Consortium
OSM	OpenStreetMap
RDBMS	Relational Database Management System
SQL	Structured Query Language
SWIR	Shortwave Infrared bands
TIRS	Thermal Infrared Sensor
UHI	Urban Heat Island
UML	Unified Modeling Language
WMS	Web Map Service
XML	Extensible Markup Language
YOLO-CLS	You Only Look Once classification variant

studies are concerned with modeling correctness rather than with the domain attributes that applications actually need. The gap we address is of a different kind: the absence of roof material information, a semantic prerequisite for parameterizing the thermo-optical behavior that UHI screening depends on. When such attributes are missing, thermal outcomes degrade accordingly (López-Cabeza et al., 2022). CityGML provides two technical paths to carry such semantics: defining an Application Domain Extension (ADE), or attaching generic attributes to existing objects. For a lightweight, interoperable deployment across existing 3DCityDB installations; which is an open-source relational database and toolset that maps CityGML into normalized tables, we adopt the latter (Yao et al., 2018).

This choice in fact matters in practice. Over the past decade, the geoinformatics community has shown that 3D city models become effective decision-making tools only when semantically enriched (Biljecki & Ito, 2021; Biljecki, Stoter et al., 2015). Roof material in particular is a very important semantic information for urban heat and energy analyses. However, it is often absent from the CityGML datasets released by mapping agencies, not because the information does not exist somewhere, but because it is typically not curated or distributed within CityGML, For urban climate analyses, this attribute therefore needs to be carried within the CityGML dataset itself, attached to RoofSurface features as a generic attribute, which extends the data content of an instance document without modifying the CityGML conceptual model.

Recent work has begun to close this gap by inferring roof materials directly from imagery and injecting the results into semantic city models (Arzoumanidis, Nguyen et al., 2025). Yet two factors still limit classification accuracy, and by extension, the fidelity of UHI screening that depends on those labels. First, aerial tiles or standard orthophoto image patches used for training and inference are visually noisy and they frequently contain non-roof content such as roads, trees, parked cars, and façades, which dilutes the discriminative signal for roofing materials. Second, roof material distributions are highly imbalanced, with *roof_tiles* generally dominating and bituminous classes rare, which biases learned classifiers toward more represented classes.

The methodology developed in this paper is designed to address both failure modes at the source. First, background leakage is handled at ingestion by utilizing building footprints to suppress any non-roof pixels before the classifier ever sees them, so that training and inference operate on the same clean support region and the network learns material cues rather than urban context. Additionally, class imbalance is handled at the data level, through targeted oversampling of under-represented materials, which lifts minority recall without touching the loss or the architecture. The result is a pipeline that keeps the learning task strictly at the image level while remaining aligned between training and deployment.

Taken together, these choices yield an end-to-end pipeline for large-scale roof material enrichment of authoritative 3D city models. 3D city models enriched with material information support a wide range of downstream analyses that footprints, even when tagged with material like OSM, cannot address. This is mainly because these physical analyses are inherently three-dimensional and depend on the explicit geometry of the 3D city models such as:

- Ray-tracing-based radio propagation modeling for cellular and emerging 6G networks, where reflection, scattering, and diffraction coefficients depend on both the surface geometry and the electromagnetic properties of the material (Testolina et al., 2024).
- Electromagnetic reflection and line-of-sight analysis for unmanned aerial vehicles, in particular for above-rooftop flight, swarm coordination, and landing manoeuvres, where the reflective behavior of roof materials directly affects radar and communication signal predictions (Zhekov et al., 2020).
- Building energy and solar-gain modeling, where the angular geometry of pitched roofs interacts with material thermo-optical properties to determine incident solar load (Biljecki, Heuvelink et al., 2015).

The UHI screening demonstrated in this work is one such application of the enriched 3D city models, but not the only one. Enriching CityGML therefore situates the contribution within a broader ecosystem of urban applications that are heavily dependent on semantically rich and accurate 3D city models.

Our main contributions are then:

- A masking-aware roof material classifier that suppresses non-roof context using footprint-derived masks to improve material discrimination.
- A deterministic multi-material inference module that aggregates image-level predictions into per-building material assignments.
- Direct integration with 3DCityDB, enabling semantic enrichment of existing 3D city models at scale.
- A UHI screening workflow that converts roof material semantics into scene-specific land surface temperature (LST) estimates for both before and after greening assessment.
- A reproducible, city-scale demonstration, including evaluation on Hamburg, Germany, as well as other cities beyond Germany such as Madrid and Paris to quantify the potential of greening and its cooling benefits.

The remainder of this paper situates our work within related literature (Section 2), details our masking-aware classification procedure (Section 3), quantifies performance and demonstrates the integration into a 3D city repository with a UHI screening case study (Section 4), and finally discusses type–material correlation and its effect on UHI screening (Section 5).

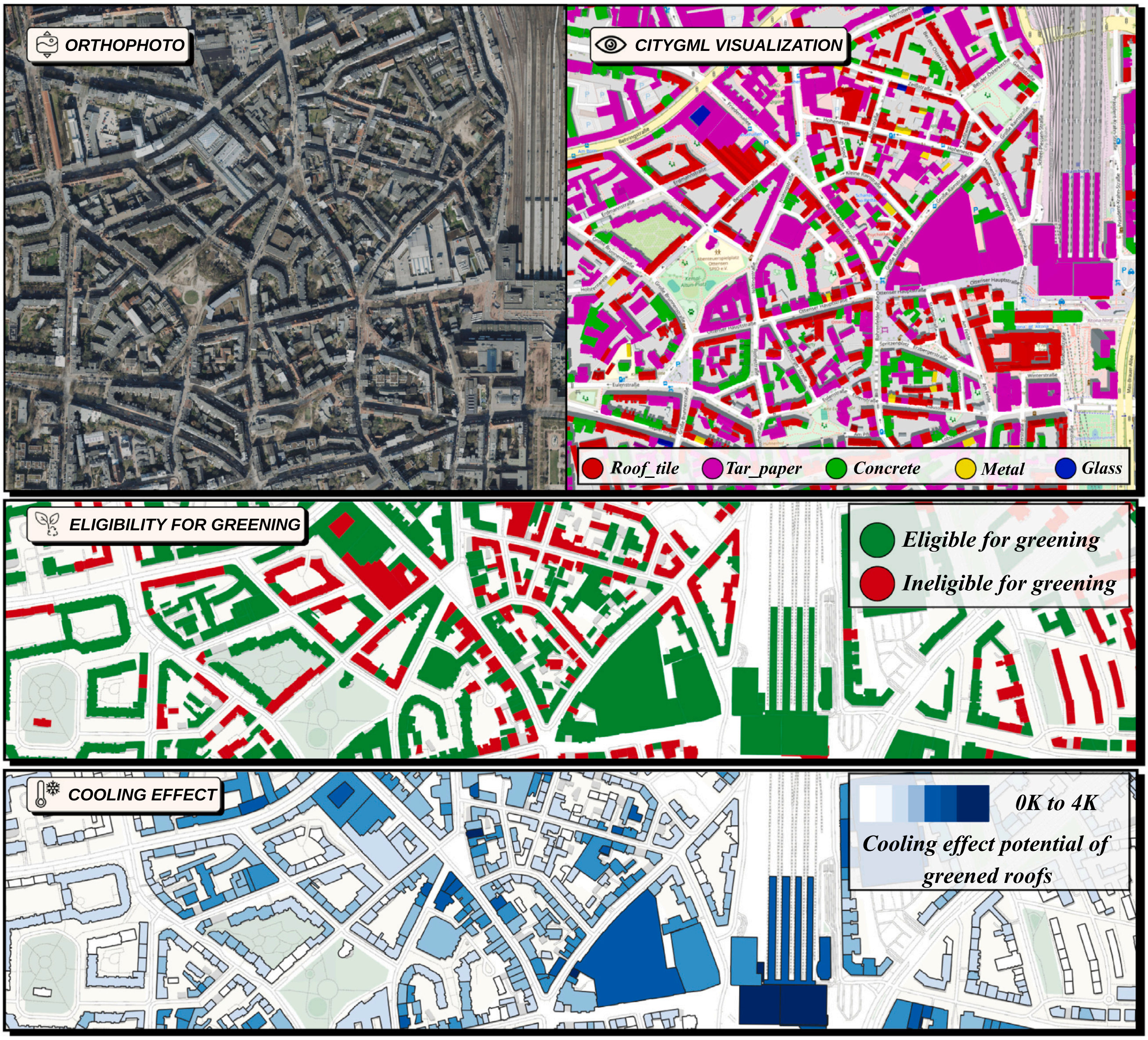


Fig. 1. Visualization of the test area around Altona train station, Hamburg, Germany. (Top left): Representative areal image of the Altona train station area. (Top right): Buildings with semantically enriched roof materials, color-coded as follows: *concrete* (green), *metal* (yellow), *glass* (blue), *roof tiles* (red), *tar paper* (purple), and *no data* (gray). The visualization was created using CesiumJS and the 3DCityDB Web Map Client, with OpenStreetMap as the basemap. (Middle): Suitability assessment for roof greening, where green indicates a suitable roof and red indicates an unsuitable roof. (Bottom): Simulated cooling effect intensity of potential roof greening, represented by a blue hue gradient, from white (no cooling) to dark blue (4 K).

2. Related work

In the context of roof material classification and enrichment, Arzoumanidis, Nguyen et al. (2025) demonstrated that city scale roof material mapping does not require extensive surveys. Starting from two openly available resources, OSM roof attributes and high-resolution RGB orthophotos over Germany, they assembled a training dataset by querying OSM, gridding space, and pairing candidate buildings with imagery, while discarding weak samples through geometric checks. With this approach, they trained a YOLOv11 detector to localize and classify roof segments, then integrated the detections into a 3D city model. The result was an end-to-end, repeatable pipeline: images in, materials out, and attributes returned to a database that planners can query. Two key steps made the approach operational. First, detections were integrated into 3DCityDB (v5.0) by spatially matching predicted bounding boxes to CityGML building footprints using a modified Jaccard index, transforming per-tile predictions into per-building attributes with lineage. Second, a Bremen case study demonstrated coverage at scale roof materials attached to 83% of 20,764 buildings.

Despite these methodological advances, the approach inherits and highlights several persistent challenges in the field. Tiles and detections suffer from background leakage, where roads, trees, and façades dilute the roof signal and depress recall, often manifesting as "background" classifications rather than incorrect material assignments. The label source amplifies class imbalance: volunteered attributes skew toward common materials, leaving rare classes statistically underrepresented even with large sample counts. Finally, the write-back process relies on box–footprint overlap. However, loose bounding boxes, as well as small and complex roofs can lead to incorrect spatial matching, and training on object detections while processing tiles with mixed context can subtly miss-align training and inference.

Where RGB cues blur distinctions between *metal*, *roof_tiles*, and bituminous surfaces, multi-spectral and thermal channels enable sharper discrimination. Ilehag et al. (2018) segmented coarse material families from high-resolution multi-spectral and thermal data and injected attributes into a 3D city model, illustrating how additional spectral bands regularize cases that defeat RGB classification. At the other extreme, purely RGB pipelines report local segmentation and classification over small areas (Kim et al., 2021; Solovyev, 2020), while color-infrared data have been used to identify asbestos–cement roofs (Krówczyńska et al., 2020). The consistent lesson is that richer spectral bands help, but their availability and coverage are uneven at city scale. Consequently, methods relying solely on ubiquitous RGB imagery remain a critical, if constrained, pathway to scalability.

A parallel research direction recovers roof structure to guide semantic interpretation. Dehbi et al. (2021) applied an active sampling to classify and reconstruct complex roofs from LiDAR point clouds. Dehbi et al. (2019) identified roof superstructures like dormers using statistical descriptors of inclination and residuals, demonstrating strong performance on both synthetic and real point clouds. More recent work couples transformer-style segmentation with mesh prediction to regularize roof geometry from high-resolution imagery (Yu et al., 2025). While these methods do not label materials directly, they yield clean partitions and type priors that many enrichment systems exploit qualitatively. Our work, by contrast, focuses more on the material classification task itself, arguing that improving the core image classification signal is a prerequisite for effective integration with such structural priors.

When areal imagery cannot capture areas under trees or along façades, ground-level views offer complementary evidence. Controlled comparisons suggest that decision-level fusion of aerial and street-view models can outperform early feature fusion, partly because cross-view misalignment corrupts learned features (Hoffmann et al., 2019). In parallel, community datasets and benchmarks have begun to standardize façade semantics at scale, most notably the OpenFacades initiative by Liang et al. (2025), which curates street-view façades with reproducible annotations and evaluation protocols, enabling training and comparison for material and element recognition. These gains, however, come with practical costs, uneven street-view coverage, licensing constraints, and persistent alignment issues, which limit direct, city wide deployment even as they expand the semantic signal available for modeling (Sun et al., 2025). Closely related, Tan et al. (2025) developed AdvUNet3+, a deep-learning segmentation model that recovers heterogeneous façade materials from street-view imagery and projects the recovered material information onto 3D building models, explicitly motivating the work with light-pollution, urban heat island, and solar-potential analyses.

When it comes to the semantic enrichment of 3D city models, the scientific scene has seen multiple efforts to automate the integration of attributes to CityGML. Early works heavily relied on geometric matching of attributes and buildings (Ogawa et al., 2024), formal concept analysis (Ahmadian & Pahlavani, 2022), or knowledge graphs (Ding et al., 2024), to align external data, such as OSM tags, with the CityGML schema. Some other studies deviated from these approaches and proposed to leverage Large Language Models (LLMs) to automate the extraction of unstructured data sources like PDF files and the semantic mapping into CityGML database (Kanna & Kolbe, 2025). This LLM-based framework is a shift towards more flexible and accessible enrichment pipelines. However, there have been also a rise of multiple works that complements the vision of operational, end-to-end systems, as seen in roof material mapping (Arzoumanidis, Nguyen et al., 2025) and cross-view fusion generalizable tool for populating a wider range of semantic attributes directly into the city model's database (Hoffmann et al., 2019). These reviews position their efforts within the broader landscape of urban digital twins. Weil et al. (2023), for example, systematically reviewed the challenges of operationalising urban digital twins for sustainable smart cities and semantic richness, and the connection between models and real world decision making as recurring bottlenecks.

Beyond detection benchmarks, recent work by Arzoumanidis, Li et al. (2025) tackles the automatic detection of tiny rooftop infrastructure, drainage outlets and ventilation caps, on flat roofs in high-resolution aerial imagery. Their contribution is explicitly enrichment-oriented, turning image evidence into maintainable, queryable assets using utilities that are embedded into GIS/BIM models. From a sustainability standpoint, such drainage attributes are directly relevant to greening feasibility and UHI mitigation, as adequate flow governing load, and a watering capacity estimator for green roofs. Our pipeline's material labels and their drainage utilities are therefore complementary, materials control thermo-optical parameters used in screening, while drainage controls hydrological safety and operability of greening measures.

A separate research thread extends semantics directly to sustainability indicators. Roofpedia, for instance, is an open roof registry mapping the spatial distribution and area of solar and green roofs across multiple cities (Wu & Biljecki, 2021). Park et al. (2024) on the other hand, proposed an application driven study that estimates cool-roof potential via rooftop detection and color-based albedo proxies, explicitly connecting appearance and material attributes to energy savings and UHI-oriented retrofits (Park et al., 2024). These systems underscore the value of queryable write-back and provenance, while also exposing variation in how reliably detections are attached to buildings and propagated into city model schemas.

Bulatov et al. (2020) took a closely related direction to ours. They fused multi-sensor aerial data with semantic 3D city models to forward-simulate thermal and infrared responses and then localize urban heat islands with high spatial fidelity. Their study shows that UHI identification improves when radiative and material properties are tied to explicit city objects rather than to imagery alone, showing the value of semantic enrichment for downstream climate analyses.

When approaching UHI mitigation works, one major study comes to light (McConnell et al., 2022), a quasi-experimental study in Chicago found larger, intensive green roofs with diverse plantings cooled more than small, extensive systems, offering a low-cost template based on Landsat and open-source analysis.

Recent contributions have also refined this picture by quantifying how cool roofs and urban vegetation translate into measurable thermal outcomes under different climates and urban morphologies. Lu et al. (2023) evaluated the effectiveness of cool roofs and urban vegetation during extreme heat events in three Canadian regions (Ottawa-Montreal, Calgary, and Vancouver) and reported substantial albedo and vegetation cooling effectiveness that varied strongly between cities. Al Kafy et al. (2024) study is the closest one to this work. They estimated the cooling potential of green roofs in downtown Austin by combining LiDAR derived 2D and 3D urban morphological parameters with an artificial neural network, and reported a daytime surface UHI reduction for greened building. However, their workflow does not propagate the result back into a semantic 3D city model. A key contribution of this work bridges that gap by enriching CityGML with roof material attributes that determine which retrofits are physically plausible, then feed the retained candidates into the LST scenario.

Methodologically, Remote Sensing and Machine Learning (ML) enable scalable targeting and evaluation. City and district level studies use LST as a metric to predict cooling benefits. In that sense, we see major studies report different LST. Yan et al. (2025) for example reported 0.47 K city level and 2.5 K local air temperature drops. Calhoun et al. (2024) estimates neighborhood scale evening temperature drops of 0.7–0.9 K from greening and cool-roof interventions. Similarly (Sánchez-Cordero et al., 2025) assessed green roof benefit in Granada to reach 1.45 K LST reduction. This highlights in evidence that the size of the cooling effect depends both on the local climate and on where in the city the green roofs are installed. Additionally, Lin et al. (2021) formalize the behavior of diminishing-returns for planning scenarios, and explicitly quantify how additional green or cool roof retrofits produce progressively smaller extra cooling benefits, providing planners with curves to evaluate different retrofit scenarios.

Furthermore, it is important to take into consideration design trade-offs, courtyard simulations caution that very high albedo can raise heat stress despite lowering surface temperature, motivating balanced reflectance plus vegetation rather than reflectance alone (López-Cabeza et al., 2022). Together, these findings align with operational enrichment goals, to use machine learning in service of UHI screening (Seeberg et al., 2022).

Beyond material attributes, Li and Zeng (2024) reviewed multidisciplinary parameters used to characterize 3D urban morphology and found that core indicators such as mean building height and percentage of landscape dominate, while vegetation and material related indicators remain comparatively underrepresented. Building on this line of work, Cai et al. (2025) proposed a bi-directional mapping between morphology metrics and 3D city blocks to support performance driven sustainable urban design. At the city scale, Mashhoodi and Unceta (2024) studied LST inequality across 683 European Functional Urban Areas and found that nearly 70% of intra-city LST inequality can be explained by urban form variables, including impervious surfaces, density, compactness, vegetation and albedo. At the block scale, Wan et al. (2025) used spatial regression to identify the most pivotal block level environmental characteristics shaping LST in a high density city, and Wang et al. (2024) showed that block morphology effects on the urban thermal environment are spatially heterogeneous and need to be modeled accordingly. Taken together, these studies frame why semantic enrichment matters operationally. Defensible thermal screening requires building's thermo-optical attributes such as roof material to translate morphology into viable climate studies.

Across these research directions, The proposed pipeline is positioned to address the gap of missing roof material information, and the issues raised from previous works notably background leakage and class imbalance to concretely make roof material semantics both accurate, queryable into a standard 3DCityDB, and operational at city scale for UHI mitigation.

3. Methodology

3.1. Overview

Classifying roof materials from aerial imagery is on the surface a standard image recognition task. In practice however, it is not. It culminate a wide combination of difficulties that standard classification pipelines are not designed to absorb. Since roofs occupy only a fraction of any aerial tile, sharing the frame with roads, vegetation, vehicles, and city furniture, all of which carry textures that a classifier can easily confuse with roofing materials. The materials themselves are distributed unevenly across cities. *Roof_tile* and bituminous surfaces (i.e.*tar_paper* and *concrete*) dominate, while metal, and glass roofs appear far less frequently. Since, aerial imagery offers only a 3 channel RGB view, visually similar materials under different lighting or weathering conditions can be nearly indistinguishable even to the naked eye.

These difficulties concentrate into two dominant failure modes that recur across prior work. The first, context confusion, where models trained on mixed tiles learn to recognize 'urban fabric' rather than 'roof material'. The second is class imbalance where rare materials are statistically overwhelmed by common ones, and their visual diversity makes them hard to learn. Both failures degrade not only classification accuracy but also every downstream analysis that depends on it, most acutely UHI screening, where misclassifying a reflective metal roof as a heat absorbing tile roof directly distorts the predicted surface temperature.

This proposed pipeline is organized around addressing these two failures at their source, rather than compensating for them further downstream. Context confusion is eliminated at ingestion. OpenStreetMap (OSM) building footprints are used to remove irrelevant pixels or non building pixels before the classifier ever sees the image. Class imbalance is addressed at the data level, through aggressive oversampling of under represented material classes, so that minority classes gain enough per epoch exposure to be learned without the need for loss reweighting or architectural changes.

The workflow illustrated in Fig. 2 shows how to leverage aerial images that are partitioned into tiles from state provided dop20 orthophotos, combined with footprints from each building from OSM, and rasterized into one single roof mask per building to create a dataset ready for training. This process helps to mask any noise at the source (non-building pixels set to black), improving the receptive field of the learning model and producing the support region used both in training and inference. Afterwards, this cleaned input is processed by a YOLO-CLS image classifier trained on a five way material label. The aggressive oversampling employed to balance minority classes is set to 12,000 samples each. This choice deliberately prioritizes recall on critical but rare classes like *glass* and *metal*, and accepting a slight performance draw back on moderately represented classes like *tar_paper*. Our ablation studies in Section 4 validate this engineering decision where accuracy gains is noticed on minority classes. However, this aggressive oversampling comes at a cost. The confusion between texturally similar classes increases. Nevertheless, this confusion have a very low impact UHI screenings results. This is discussed further in Section 5.

Classification performance is evaluated at the image level using Top-1 accuracy, balanced accuracy, macro F1, per-class precision, recall, and F1, complemented by a normalized confusion matrix and an Expected Calibration Error analysis. Predictions are then written back into the 3D city repository as per-building attributes, enabling large-area semantic enrichment. These attributes are then used in the UHI screening, by identifying which roofs are eligible for greening and estimating the LST changes alongside the cooling effect of greening interventions

The guiding logic throughout is to correct the dominant error at its source rather than addressing it with more model capacity or expensive data collection. A decently sized dataset of 87,605 buildings, combined with masking and aggressive oversampling, produces substantially cleaner predictions than a larger but contextually noisy training regime would, and as Section 4 shows, those cleaner predictions propagate directly into more reliable baseline LST estimates and more defensible greening deltas.

3.2. Data collection

Each training and inference patch in our pipeline contains exactly one building. High resolution orthophotos are fetched via Web Map Service (WMS) across more than ten German cities at a ground sampling distance of 20 cm, extracted as 512 by 512 px patches sized to the YOLO classifier input. Building footprints are retrieved from OSM as described in Section 3.3. For every candidate building, we extract a roof only support region by masking non building pixels to black, so that the resulting patch carries a single roof as its only visual signal. The learning task is therefore strictly image level, with one label per masked tile.

The resulting dataset, summarized in Table 1 spans as mentioned 87,605 masked buildings. The class distribution is heavily skewed, but the signal is cleaner because each image carries a single roof, not a mixture of roofs, roads, trees, and façades.

3.3. Background suppression

Detector-first pipelines often attribute misses to "background". Tiles dominated by roads, trees, or façades nudge predictions toward background or dark bituminous classes. In our context, footprints serve as a geometric prior to remove non-roof context before learning. For each georeferenced image in the training data, the footprints of all buildings

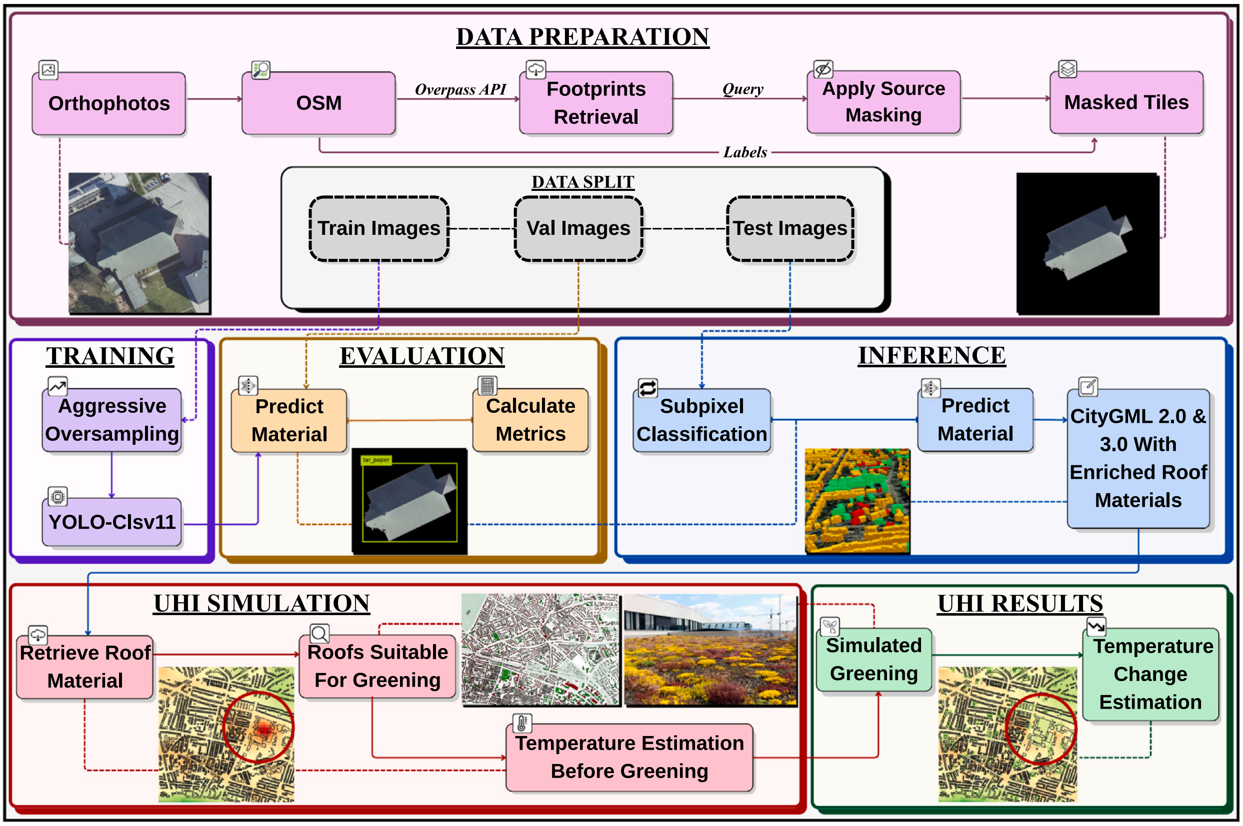


Fig. 2. Overview of the pipeline methodology used for classification, integration into CityGML and urban heat island analysis.

Table 1
Targeted classes and their respective distribution in the dataset.

Class	Train	Val	Test	Total
concrete	959	119	121	1199
glass	472	59	59	590
metal	1448	181	181	1810
tar_paper	4089	511	512	5112
roof_tiles	63,115	7889	7890	78,894
Total	70,083	8759	8763	87,605

within the image are retrieved and reprojected to the image Coordinate Reference System (CRS). Each building is then processed individually to create one building roof per image by setting all non-building element's pixels to black (RGB = 0), that is,

$$\tilde{I} = M \odot I \tag{1}$$

where:

- $I \in \mathbb{R}^{H\times W\times 3}$ is the original RGB image.
- $M \in \{0,1\}^{H\times W}$ is the binary roof mask.
- $\odot$ denotes element-wise multiplication between the roof mask and the aerial image.

Equivalently, the output image can be written element-wise as:

$$\tilde{I}_{H,W,C} = \begin{cases} I_{H,W,C}, & \text{if } M_{H,W} = 1 \\ 0, & \text{if } M_{H,W} = 0 \end{cases} \tag{2}$$

This masking step removes non-building context, which has been shown to improve class separability and classification accuracy without requiring substantial changes to the model's architecture (Ilehag et al., 2018; Kim et al., 2021; Krówczyńska et al., 2020).

However, the most noticeable concern regarding this masking strategy is its dependency on footprint quality. Positional errors in OSM footprints can shift the mask relative to the actual roof boundary, letting non-roof pixels into the support region or clipping roof edges, which propagates into the training data as a form of geometric label noise. This concern is well-documented for the specific data combination used by Fan et al. (2014) to systematically compare OSM building footprints in Germany against the authoritative ALKIS cadastre, and quantified the positional and completeness discrepancies between the two.

On the other side, Cordeiro and Carneiro (2020) provided a comprehensive survey of training under noisy labels, and showing that well-designed pipelines can be made robust to moderate noise levels even when individual annotations are imperfect. In our pipeline, the choice of OSM as the training footprint source is a deliberate trade-off informed by these findings: OSM introduces some positional noise, but its global coverage enables assembly and annotation of a substantially larger and more diverse training dataset than would be possible from authoritative cadastres alone. At inference time, the use of CityGML as a footprint provider however limits the practical impact of OSM positional error on the predicted labels. This choice is discussed in detail in both Sections 3.5 and 5

3.4. Classifier choice and imbalance handling

The pipeline leverages the YOLO11l-cls image classifier on a five-way label space (*concrete, glass, metal, tar_paper, roof_tiles*), keeping training and deployment strictly on image-level. In each sample, a single building roof is visible and all non-building pixels are set to black, so the network learns material cues from the roof surface rather than surrounding context.

Optimization follows the classifier's default schedule with a standard cross-entropy objective. The training proceeds for 100 epochs with mini-batches of 32 image, and oversampling under-represented classes reaches 12,000 images. The target of 12,000 samples per class was chosen to match the size of the largest minority class multiplied by a factor of three. Lower values degraded evaluation accuracy, while higher values biased the model toward the oversampled classes and led to overfitting

Oversampling aggressively in this way is a trade off between minority class representation and intra-class accuracy. However, it can be very effective because it operates at the data level, it increases the prevalence of minority classes per epoch without modifying the loss. The added exposure improved minority recall. Oversampling is applied to the training split only, validation and test are kept at natural frequencies, and light augmentation is used to reduce overfitting to duplicated samples. Since oversampling changes the effective class prior, it helps significantly with per class accuracy.

Alternative strategies for handling class imbalance, such as class-weighted cross-entropy, were tested and compared in this study. Overall, the oversampling approach was chosen for its simplicity and because it operates at the data level without modifying the loss function or requiring architectural changes. Moreover, it yields significantly better results as discussed in detail in Section 4.

The design described so far has one inherent limitation, where a roof is composed of multiple material patches. Since our model is not a object detector, it can only classify the material of the roof as one instance per image. This is due to 2 main reasons: First, because training exposed the model to only one roof per image with a single material label as ground truth, and second, because YOLO-CLS predicts a single class per image by design. This limitation motivated the solution presented in Section

3.5. Inference and multi-material classification

The classifier described above assigns exactly one material label per image, which means that roofs composed of multiple materials are reduced to whichever class dominates the prediction. To address this, a deterministic, class agnostic inference pipeline was developed that produces multiple material labels per building from masked roof chips. Without changing the pipeline, each input chip is a georeferenced image where non roof pixels are blacked out, and no additional training or external segmentation is required.

To reliably identify non dominant materials, the default roof class prediction is first obtained, and then refined by sweeping a sliding grid to classify local patches, creating per pixel class and confidence maps where the highest confidence is kept. The sliding window operates at a tile size of 192 px with a stride of 96 px, padded by 32 px at the borders to provide context. A patch contributes to the per-pixel material map only when at least 55% of its area falls within the building footprint, which prevents windows straddling the roof boundary from injecting non-roof signal into the prediction. Non-default materials (i.e. non dominant materials which occupies a smaller areas of the roof) are retained when their computed confidence exceeds 0.6. The choice of this value is established in detail in Section 4. This latter illustrate that predictions below the desired accuracy, while predictions at or above 0.6 enter the regime where the model's stated confidence becomes a reliable proxy for accuracy. Fig. 3 shows a prime example where the original classification predicted *metal* as the dominant material, but after the multi-material step it was able to detect 2 different segment of materials, the first being metal with a confidence of 0.98 and *roof_tiles* with a confidence of 0.6. This design's strength relies on its retrain free nature, and its capability to produces concise building level labels.

While OSM is used to label the training data and guide the masking process, inference is driven entirely by CityGML footprints because they are geometrically more accurate than their OSM counterparts and they carry stable, persistent object identifiers (GMLIDs) that are maintained across dataset versions by the official mapping agencies. This persistence is important for enrichment workflows, as predictions that are tied to a GMLID remain associated with the same physical building, even when the underlying CityGML dataset has been updated, which is not guaranteed for OSM identifiers that change with each contributor edit. As a result, CityGML enables straightforward enrichment without the need for a footprint matching process such as the one used by Arzoumanidis, Nguyen et al. (2025), eliminating the spatial uncertainty that would arise from matching OSM-derived bounding boxes to CityGML buildings.

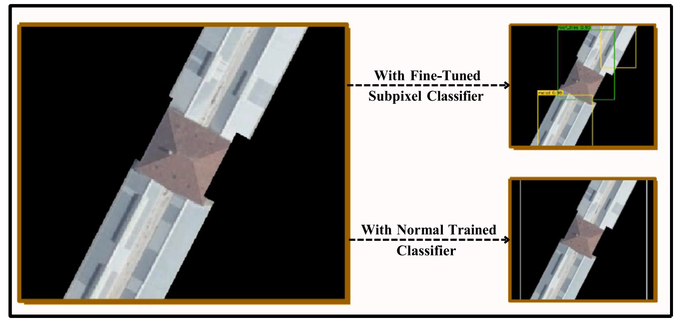


Fig. 3. Highlight of the difference between normal classifier and the tweaked sub-pixel classifier.

The footprint guided masking approach proposed in this paper only suppresses non-building context (e.g., roads, vegetation, vehicles), and any occlusions on roof level such as chimneys, rooftop equipment, and solar panels remain visible within the masked region and are processed by the classifier as if they were part of the roof surface. In most cases, this does not cause any misclassifications, since most of these objects occupy a small fraction of the roof area and are absorbed into the dominant material prediction without measurably affecting accuracy. However, these occlusions may trigger a secondary material prediction, particularly on roofs with large equipment and installations. For example, a roof made of *roof_tile* with a chimney made of *concrete* would either be classified completely as *roof_tile* or a mix of *roof_tile* and *concrete* if the chimney has a coverage area big enough to be flagged as a secondary material.

For inference, Fig. 4 shows a fully CityGML driven, masked first pipeline at city scale. The starting point is building footprints exported from the CityGML stack as a GeoJSON in the CRS EPSG:25832. For each orthophoto tile, the overlapping footprints are selected, each footprint is rasterized to a binary roof mask aligned with the image grid, and all non roof pixels inside a tight bounding window around the mask are zeroed out as described in Section 3.3. Finally, for every building footprint, the predicted material is written in the GeoJSON properties while keeping the original CityGML ground surface polygons unchanged.

3.6. Enriching semantic 3D city models

CityGML is an OGC-standardized data model and exchange format developed to support the representation and transfer of the majority of 3D urban and landscape features. Unlike other 3D city modeling paradigms that prioritize visual realism, CityGML integrates geometric, topological, and appearance information with rich semantic information of city objects. This combination facilitates the development of robust, general-purpose semantic 3D city models, which have been broadly adopted across numerous application areas (Kolbe et al., 2008; Schwab et al., 2020; Willenborg et al., 2017). The most recent version of the standard, CityGML 3.0, was released by the OGC between 2021 and 2023 (Kolbe et al., 2021; Kutzner et al., 2023). In this work, we apply both CityGML 2.0 and 3.0.

In real-world scenarios, applications rarely rely on general-purpose CityGML datasets in their native text-based form. Rather, these datasets

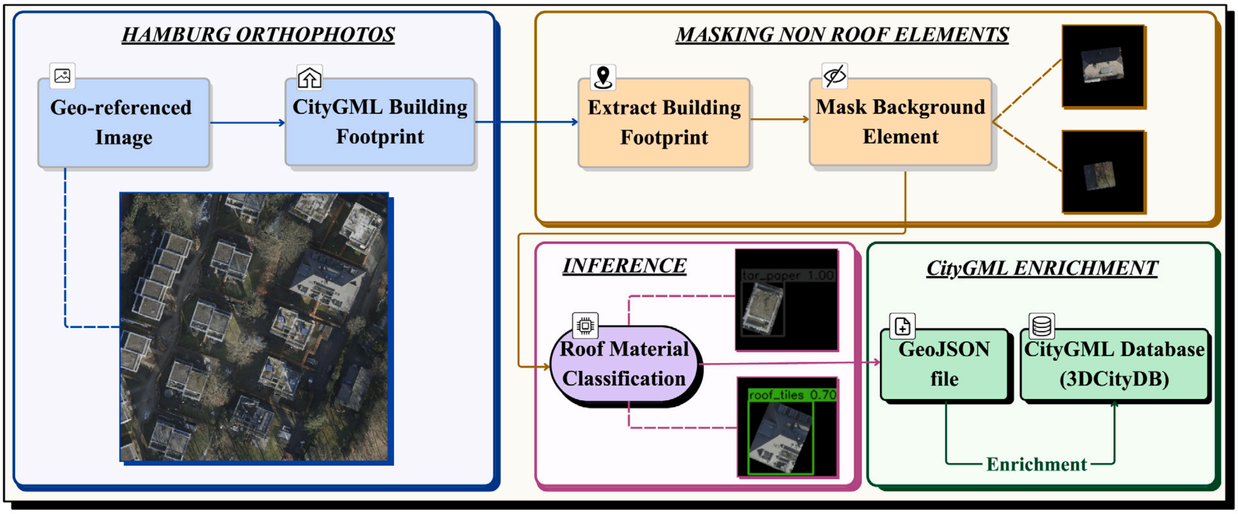


Fig. 4. Masked-first inference pipeline 3D city model enrichment.

are typically further parsed and converted into application-specific representations that support more efficient storage, querying, and analysis. Such derived models often adopt either graph-based structures (Nguyen, 2024; Nguyen & Kolbe, 2024) or relational database schemas (Yao et al., 2018).

In Germany, LoD2 CityGML building models are produced by state mapping agencies from cadastral building footprints and reconstructed using airborne laser scanning point clouds. Each Building object typically has one GroundSurface, multiple WallSurface objects, and one or more RoofSurface objects. In CityGML, a simple gabled roof may be represented by two RoofSurface polygons, while complex roofs with dormers or mixed pitches may be composed of considerably more. This subdivision is geometric and does not always align with material boundaries, and matching the material with the RoofSurface is proven difficult since our machine learning based model is only a classifier is not a detector. This is why the material classification in this work is performed at the Building level rather than per RoofSurface. Assigning material to each RoofSurface requires additional logic to associate the material pixels and map them with individual RoofSurface geometries, which is not currently supported and is identified as a direction for future work.

Since the advent of Structured Query Language (SQL), Relational Database Management Systems (RDBMSs) have been widely adopted across numerous domains. In a typical relational database, data is organized into tables consisting of rows and columns, and semantic relationships among entities, such as those between city objects, are maintained through primary–foreign key associations that link tables together as illustrated in Fig. 5. Many widely used RDBMSs, including Oracle and PostgreSQL, provide native support for XML and GML data and can therefore store CityGML objects given a predefined schema. Spatial extensions, most notably PostGIS for PostgreSQL, further enable the handling of the rich spatial information embedded in the CityGML data model. One of the most popular solutions in this domain is the 3D City Database (3DCityDB) (Yao et al., 2018), an open source, high-performance, spatially enhanced database system designed specifically for managing, visualizing, and exporting large scale CityGML datasets.

As of May 2025, the 3DCityDB has been released in its latest major version, 5.0, which provides full support for CityGML 3.0 and introduces a thoroughly redesigned and simplified database schema. Given the maturity of the 3DCityDB software suite[1] and the improved functionalities offered in this new release, we adopt version 5.0 and its

feature

id	objectid	envelope	...
1707	DEHHALKAP0000p6g	01030000_	...
2234	DEHHALKA10005Y41	01030000_	...
5696	DEHHALKAP0000q9s	01030000_	...

Primary key

Foreign key

property

id	feature_id	name	val_string	...
16420735	1707	PredictedRoofMaterial	metal	...
16420743	2234	PredictedRoofMaterial	concrete	...
16420800	5696	PredictedRoofMaterial	glass	...

Fig. 5. An excerpt of the enriched tables *feature* and *property* and their relationship for the Hamburg dataset.

associated utilities *citydb-tool*[2] to store, enrich, and export our CityGML test datasets.

The updated 3DCityDB schema *citydb* adopts a simplified design: all feature types (for example, buildings) are stored in a single table named *feature*, while most thematic attributes are captured via a central table named *property* (see Fig. 5). Because the CityGML data model does not include a dedicated attribute for roof materials, we store the predicted roof materials for each building using the concept of generic string attributes permitted by the CityGML data model. Fig. 6 shows an excerpt of the CityGML 3.0 UML class diagram illustrating the relationship between the class *Building* and its generic string attributes. The following example illustrates how such attributes can be integrated into CityGML 3.0 for each building:

```
<genericAttribute>
    <gen:StringAttribute>
        <gen:name>PredictedRoofMaterial</gen:name>
        <gen:value>roof_tiles</gen:value>
    </gen:StringAttribute>
</genericAttribute>
```

The predicted building roof materials are integrated into the 3DCityDB v5.0 instance of the test CityGML dataset of Hamburg as follows:

1. For each building GMLID in the CityGML dataset, retrieve the corresponding predicted roof material from the GeoJSON file generated in the previous step.

[1] https://github.com/3dcitydb.

[2] https://docs.3dcitydb.org/latest/citydb-tool.

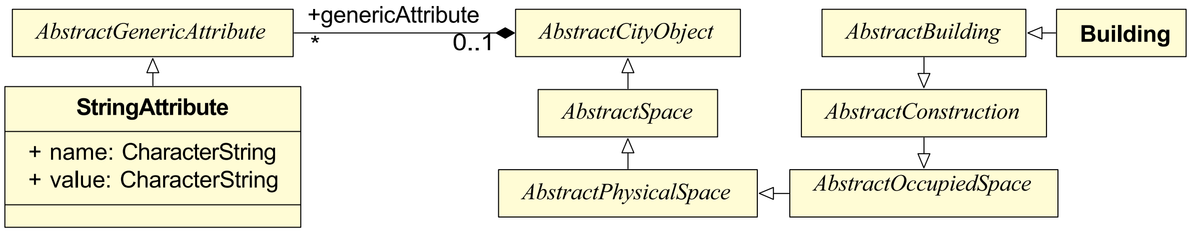


Fig. 6. Excerpt UML class diagram in CityGML 3.0 showing generic string attributes as properties of *Building* objects, inherited from the abstract class *AbstractCityObject*. For simplicity, class stereotypes and package namespaces are omitted.
Source: Adapted from Kolbe et al. (2021).

2. Using the building GMLID, join the tables *feature* and *property* and assign a generic string attribute *PredictedRoofMaterial* to the building.

Aligning per-pixel predictions with individual RoofSurface objects remains a challenging problem at this stage, since the building footprint often encompasses multiple roof patches whose boundaries cannot be reliably matched to image-derived material distributions. The multi-material inference itself also carries residual noise, particularly along boundary regions and on roofs with substantial occlusions. The dominant material per building is therefore used as the target for enrichment, and each CityGML building is assigned at most one predicted roof material. The following SQL statement is used to insert a generic string attribute *PredictedRoofMaterial* for a specific building into the database:

```sql
INSERT INTO citydb.property(feature_id,
    datatype_id, namespace_id, name, val_string)
-- val_string can be 'concrete', 'metal',
    'glass', 'roof_tiles', or 'tar_paper'
VALUES (id, 5, 3, 'PredictedRoofMaterial',
    'roof_tiles');
```

The table *property* is defined within the *citydb* schema. Each property is linked to its corresponding building through *feature_id*, which refers to a GMLID stored in the table *feature* within the same schema, as illustrated in Fig. 5. For a generic string attribute in 3DCityDB version 5.0, the values of *datatype_id* and *namespace_id* are set to 5 and 3, respectively. The enriched city model can then be exported to both CityGML 2.0 and 3.0 using the 3DCityDB utilities *citydb-tool*.

3.7. Urban heat island analysis

Urban heat island screening treats surface overheating as a scene specific prediction problem grounded in satellite thermography, where land surface temperature is modeled from physically interpretable predictors that capture vegetation, built form, and surface radiative properties. Landsat-8 Level-2 products are utilized, which provide atmospherically corrected surface reflectance and temperature, to evaluate the potential thermal benefits of green roof retrofitting and estimate the LST change throughout different thermal scenarios. The Normalized Difference Vegetation Index (NDVI), broadband albedo, and the Normalized Difference Built-up Index (NDBI) are computed as dominant determinants of urban LST. NDVI quantifies vegetation greenness by contrasting red light absorption with near infrared reflectance, NDBI highlights built-up structures due to the high reflectance of impervious materials in the shortwave infrared, and albedo represents broadband surface reflectance, a key determinant of radiative heating. The selection of these specific indices follows recent remote sensing studies that demonstrated the predictive strength of these indices for LST estimation in urban settings (Joshi et al., 2023; Mansourmoghaddam et al., 2024; Sánchez-Cordero et al., 2025; Seeberg et al., 2022).

Adapting the counterfactual intervention simulation logic of Calhoun et al. (2024), altering predictor pixel values at a location of interest and comparing model outputs before and after. Rooftop NDVI, NDBI, and albedo are modified to reflect hypothetical greening, and LST is re-predicted with a random forest regressor trained on the baseline scene. In this context, the predicted roof material serves as the semantic filter that determines which roofs are eligible for greening. Roofs classified as *concrete* and *tar_paper* were chosen as the primary host of greening because these two materials are structurally and functionally the most amenable to green roof retrofit. *Concrete* offers the load bearing capacity needed to support greening, while *tar_paper* surfaces already provide a waterproofing membrane that is compatible with standard green roof assemblies. *metal* and *glass*, by contrast, are challenging to greening because of load transfer, thermal expansion, and pollutant leaching, which make them poor candidates for vegetation (Cascone, 2019; Castleton et al., 2010; Forschungsgesellschaft Landschaftsentwicklung Landschaftsbau e.V. (FLL), 2018).

In addition to this material filtering, a geometric eligibility check is also applied. Roof slope is computed directly from the LoD2 RoofSurface geometries of the CityGML dataset. Consequently roofs with a slope greater than 15 degrees are excluded from the greening simulation, in line with the Hamburg Green Roof Guideline (Hamburgische Investitions- und Förderbank (IFB Hamburg), 2024), which identifies this threshold as the limit above which extensive green roof installation requires special structural reinforcement and incurs significantly higher construction and maintenance costs (cf. Fig. 7)

For Madrid and Paris, roof slope is not an actual breaking criteria, it determines whether a green roof can be installed with a standard assembly or whether slope stabilization measures become necessary. At or below 20 percent grade, which corresponds to 11.31 degrees, no slope stabilization provision is required, and this is the limit set by the French professional rules for green roofs (Adivet et al., 2018) for Paris and by the Spanish NTJ 11C standard (Fundació de la Jardineria i el Paisatge, 2012) for Madrid. Between this value and 30 degrees, a green roof remains feasible but requires anti shear protection, and above 30 degrees a separate structural analysis becomes necessary. The screening is therefore reported under two scenarios in Section 4. For both these cities outside of Germany, where city scale CityGML is not distributed, slope is derived from openly available national LiDAR surface models, the IGN LiDAR HD surface model at 0.5 m resolution for Paris and the PNOA LiDAR surface model at 2 m resolution for Madrid.

Following the green eligible filtering, and for each cloud free scene, the Landsat-8 surface reflectance and thermal products are ingested. To characterize the surface properties, three key spectral parameters are computed using the shortwave integration method proposed by Liang (2001). The same integration method has been applied successfully in prior LST estimation work (Mansourmoghaddam et al., 2024) and yielded good results in their case. These indices rely on reflectance from specific Landsat-8 Operational Land Imager (OLI) bands, each capturing energy from a distinct region of the electromagnetic spectrum:

- The Blue band (2, 0.45–0.51 μm) is sensitive to atmospheric scattering and water detection.
- The Red band (4, 0.64–0.67 μm) strongly absorbed by chlorophyll and therefore useful for vegetation discrimination.

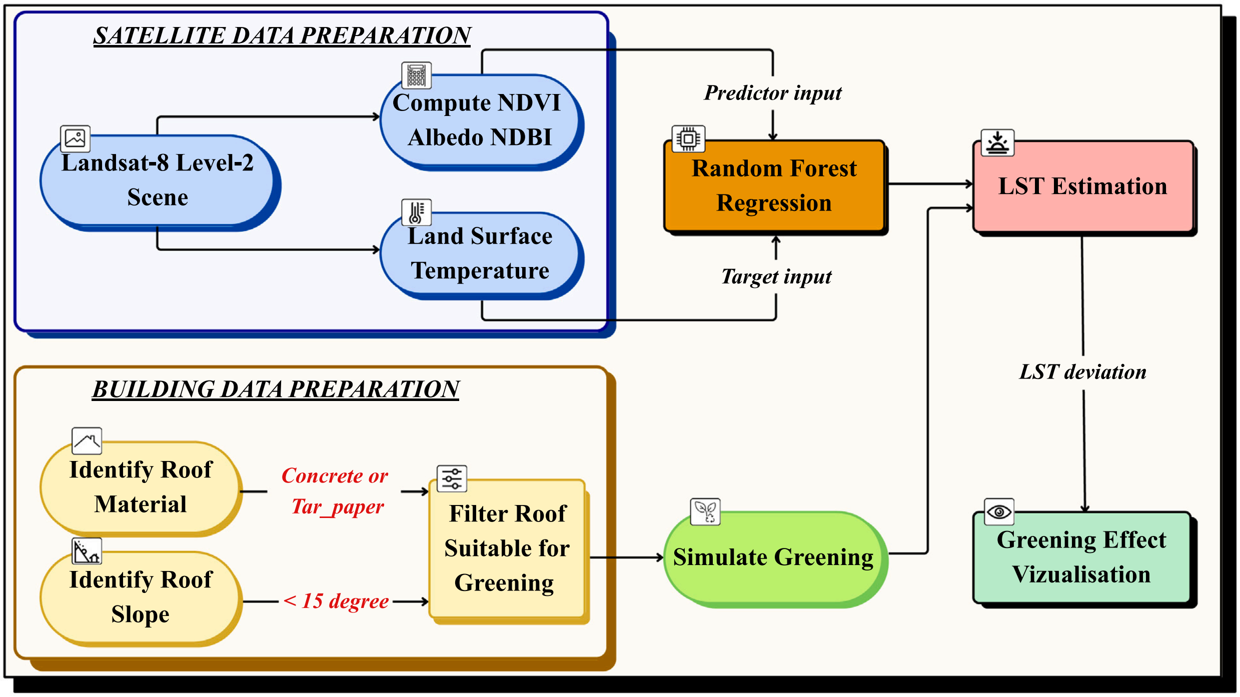


Fig. 7. UHI screening pipeline, from data preparation to greening effect visualization on LST.

- The Near-Infrared (NIR) band (5, 0.85–0.88 μm) highly reflective in healthy vegetation.
- The Shortwave Infrared 1 (SWIR1) band (6, 1.57–1.65 μm) is responsive to soil and vegetation moisture.
- The Shortwave Infrared 2 (SWIR2) band (7, 2.11–2.29 μm) is effective for differentiating built-up and dry surfaces.

Using these bands, the spectral indices are then computed as follows:

$$\mathrm{NDVI} = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}} \quad (3)$$

$$\mathrm{NDBI} = \frac{\rho_{SWIR1} - \rho_{NIR}}{\rho_{SWIR1} + \rho_{NIR}} \quad (4)$$

$$\mathrm{Albedo} = 0.356\rho_{Blue} + 0.130\rho_{Red} + 0.373\rho_{NIR} + 0.085\rho_{SWIR1} + 0.072\rho_{SWIR2} - 0.0018 \quad (5)$$

where ρ represents the surface reflectance of the respective Landsat-8 bands. In the simulated greening workflow illustrated in Fig. 7, modifying a rooftop pixel corresponds to assigning it spectral values (NDVI, albedo, NDBI) that match the radiometric signature of a typical green roof.

Afterwards, a Random Forest regressor is trained to learn the local radiative context, using these parameters as predictors and the observed LST as the response variable. The Random Forest algorithm is selected for its robustness in satellite based UHI prediction and its ability to model non linear spectral thermal relationships without parametric assumptions. This specific workflow has been previously validated for green roof assessment by Joshi et al. (2023) and Sánchez-Cordero et al. (2025).

To ensure the screening results reflects realistic green roof performance, a set of target spectral values are derived from existing green infrastructure, using OSM footprints to extract all green buildings and compute the average spectral parameters. The screening targets are set as follows: an NDVI of 0.5099, an NDBI of −0.1531, and an albedo of 0.1436.

With the baseline model established, counterfactual greening scenarios are generated. A sub pixel mixing approach is employed to account for the spatial resolution of Landsat-8 (approximately 30 m) relative to building footprints. The fraction of each pixel covered by the target roof materials is calculated. The trained Random Forest predicts the scenario LST, and the delta between the scenario prediction and the baseline prediction is calculated. The resulting per pixel temperature deltas, aggregated to both building and city scale, are presented in Section 4.

4. Experimental results

To evaluate the robustness of the roof material classifier, a normalized confusion matrix (Fig. 8(c)), together with Top-1 and Top-5 accuracy curves (Fig. 9(b)) are computed and analyzed jointly to assess not only classification accuracy but also how model confidence evolves over training. Over 100 epochs the network moves steadily into a highly decisive regime: Top-1 accuracy rises to approximately 96.2%, while Top-5 accuracy remains essentially 1.00 throughout training and validation, meaning that the correct label is virtually always among the top five predictions.

Fig. 9(b) demonstrate that training cross-entropy decreases monotonically as the optimizer continues to shape a boundary that separates the classes on the training distribution. Validation cross-entropy, in contrast, follows a pronounced U-shape and drops early to roughly 0.145 and then drifts upward toward 0.20–0.21, even as Top-1 accuracy continues to improve. This apparent paradox is called logit sharpening and it means that the model becomes too confident in its predictions, even when it is wrong. In this dataset, the effect is caused by class imbalance and amplified by the aggressive oversampling applied during training, as well as by genuine borderline texture similarities, most dominantly apparent between *metal* and *roof_tiles* or between *tar_paper* and *concrete*, which explains why Top-5 saturates early while Top-1 keeps inching upward as confidence is reweighted among an already correct ranking.

To broaden the evaluation scope, the model performance has also been evaluated using a broader set of metrics such as balanced accuracy, recall, F1 and per-class precision. The results are summarized in Table 2. As highlighted above, the Top-1 accuracy of 95.1% reflects

Table 2
Per-class classification metrics on the test set (8763 samples).

Class	Precision	Recall	F1	Support
concrete	0.644	0.463	0.538	121
glass	0.594	0.644	0.618	59
metal	0.478	0.470	0.474	181
roof_tiles	0.986	0.983	0.984	7890
tar_paper	0.698	0.768	0.731	512
Macro avg	0.680	0.665	0.669	8763

the dominance of the *roof_tiles* class, which constitutes the majority of the test set and is recognized with both precision and recall above 0.98. Balanced accuracy and macro F1, which weight all classes equally regardless of support, fall to 0.665 and 0.669 respectively. *Tar_paper* achieves the second strongest per-class F1 at 0.731, followed by *glass* at 0.618 and *concrete* at 0.538. However, *Metal* remains the most challenging class with an F1 of 0.474 because of its visual similarity to the most dominant class *roof_tiles*.

Moreover, to better quantify the model's performance, the calibration analysis was issued to measure whether the model's confidence scores reflect actual accuracy. This index is particularly useful because it shows that the reported confidences stay within roughly three percentage points of true accuracy. The reliability diagram (Fig. 9(a)) shows slight overconfidence in the mid-range bins (0.4–0.8), but the highest confidence bin tracks closely with actual accuracy where most test predictions fall. This is relevant to the multi-material inference step in Section 3.5. In that step, the confidence threshold was set to 0.6, and the calibration evidence supports the choice of a relatively strict cutoff rather than treating all predictions as equally trustworthy.

To isolate and assess the contribution of each class imbalance handling strategies, an ablation was conducted on the test split of Table 1 under identical setup as the primary training. Three configurations are compared:

- Masking only baseline, where the classifier is trained on footprint masked tiles at the natural class frequencies of Table 1.
- Masking combined with class-weighted cross-entropy, where the per-class weights are set inversely proportional to the training support.
- Masking with aggressive oversampling to 12,000 training images per minority class.

Fig. 8 reports the normalized confusion matrices for the three configurations. On the baseline masking only approach (Fig. 8(a)), the biggest issue noticed is the absorption of minority classes by the dominant class *roof_tile*. The confusion of *roof_tile* with the minority classes *metal* and *glass* indicate that nearly half of the glass and metal evidence is collapsed onto the majority class. The diagonal yields per-class recall of 0.32 for *concrete*, 0.42 for *glass*, 0.28 for *metal*, 0.99 for *roof_tile*, and 0.73 for *tar_paper*, with a macro recall of 0.548.

Class weighted cross entropy however, changes the optimization signal without altering the data distribution. Illustrated in Fig. 8(b), *glass* benefits the most with a recall of 0.53, and *tar_paper* and *roof_tile* are essentially preserved. *Concrete* and *metal* show no meaningful improvement, and the confusions of minority class and *roof_tile* and *glass* still persist. The macro-recall reaches 0.564, which is only 0.016 above the baseline. Re-weighting the loss therefore proves insufficient on its own to overcome the imbalance induced by the dominance of *roof_tile* in the training distribution.

The oversampling configuration proposed in this paper improves per-class recall on every minority class simultaneously, reaching 0.52 for *concrete*, 0.58 for *glass*, 0.49 for *metal*, and 0.83 for *tar_paper*, while *roof_tile* preserving its accuracy as is expected. The off diagonal mass also redistributes in a structurally meaningful way: leakage into *roof_tile* is reduced for every minority class and the *concrete* and *tar_paper* confusion drops from 0.43 to 0.22. The full confusion matrix is illustrated in Fig. 8(c)

To sum up, the ablation supports the interpretation that the gains reported throughout this paper reflect the joint contribution of masking and oversampling rather than either component in isolation. Masking removes the urban context that would otherwise dilute the discriminative signal for roofing materials, and oversampling ensures that minority class evidence is seen sufficiently often during training to be learned.

Beyond classification metrics, the pipeline produces ready-to-use geospatial outputs. Each building footprint is exported with its predicted roof material attributes attached, eliminating any extra post-processing before CityGML integration. For each roof, per-pixel predictions are summarized into material coverage fractions (percentage of roof pixels per class), all materials above a small coverage threshold are kept, and they are sorted in descending order. The GeoJSON output therefore contains one polygon geometry per roof that matches the CityGML footprint and carries the same GMLID together with a list of predicted roof material IDs and their coverage values. Importantly, each building is represented as a single polygon rather than as separate geometries per material class. The matched CityGML building is then enriched with the dominant material, practically the one with the highest coverage fraction, from the list of material in the GeoJSON file, this material is then written as a per-Building generic attribute, for the data modeling reasons discussed in Section 5.

This design deliberately combines geometry (one roof polygon in EPSG:25832) with semantics (potentially multiple materials). This makes the output natively compatible with CityGML enrichment workflows that expect a single geometry per building but allow multiple categorical attributes (cf. Fig. 3). Taken together, the learning curves, F1 score, recall, confusion matrices, Top-1 and Top-5 accuracy confirm that the model is practical and well calibrated, and operational for downstream UHI screenings.

The semantic enrichment of roof materials is applied to the CityGML 2.0 LoD2 dataset for Hamburg (publicly available[3]), covering 194,799 of the city's 388,267 registered buildings. This subset is due to two main reasons:

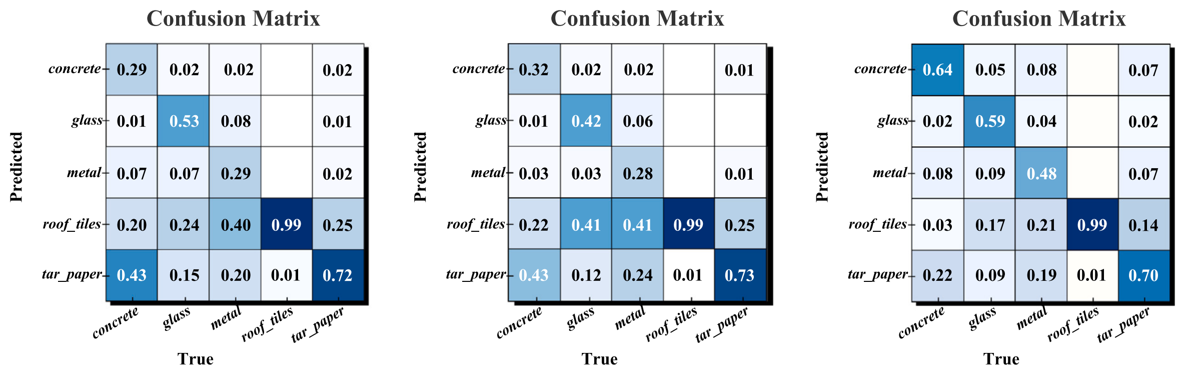
- Operationally, Hamburg's green-roof funding requires a minimum vegetated area of at least $20\,m^2$, so roofs smaller than $15\,m^2$ are excluded from the UHI screening (Hamburgische Investitions- und Förderbank (IFB Hamburg), 2024).
- Technically, the 30 m Landsat pixel means that a roof smaller than $15\,m^2$ adds negligible signal to LST modeling (USGS EROS, 2018).

Together these criteria remove noise without sacrificing planning relevance. The dataset spans roughly 120 km by 60 km, and none of the buildings contained roof material attributes prior to enrichment.

After completing the prediction and enrichment process, a total of 194,799 buildings were assigned a roof material. Among these, 108,682 buildings (56 %) were predicted to have *roof_tile*, 51,564 (26 %) to have *tar_paper* roofs, 23,101 (12 %) to have *concrete* roofs, 8784 (4 %) to have *metal* roofs, and 3264 (2 %) to have *glass* roofs. A color coded 3D visualization of these buildings within a representative district is presented in the top left panel of Fig. 1.

For UHI screening, the methodology detailed in Section 3.7 is applied at roof scale in 3 distinct cities. For Hamburg, the plotted for the before and after greening highlighted in Fig. 10 show consistent cooling around green roof complexes, with the strongest temperature reductions over industrial blocks and dense inner-city districts where impervious fractions are highest. In contrast, predominantly sloped roof areas dominated by *glass* and *metal* show the highest surface temperatures because of their low thermal mass and high reflectance,

[3] https://suche.transparenz.hamburg.de/dataset/3d-gebaeudemodell-lod2-de-hamburg2.

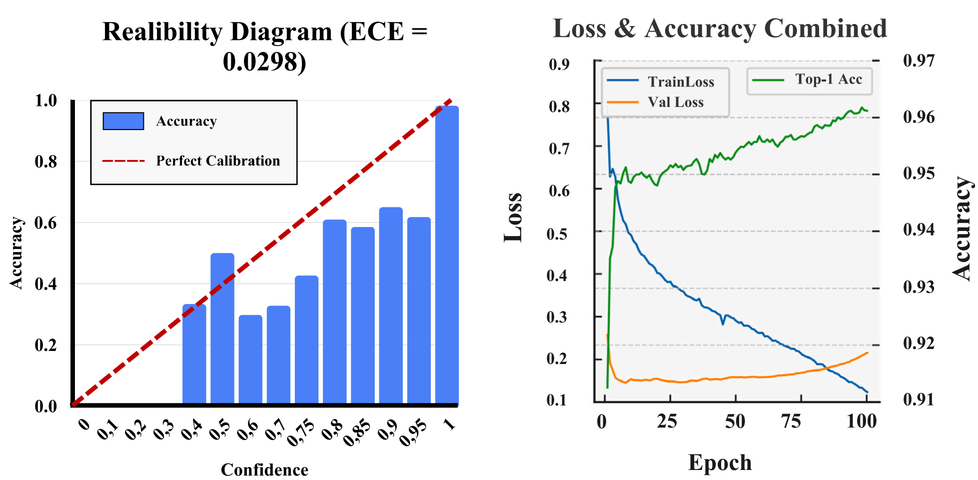


(a) Confusion matrix of the per-class accuracy on the test set of the trained model with no class imbalance handling approach applied.

(b) Confusion matrix of the per-class accuracy on the test set of the trained model with class weighted cross-entropy.

(c) Confusion matrix of the per-class accuracy on the test set of the trained model with oversampling of minority classes.

Fig. 8. Confusion matrix on the test set of 8763 instances on the baseline and two class imbalance handling approaches.



(a) Reliability diagram with an expected calibration error of approximately 0.03.

(b) Training and validation loss alongside Top-1 accuracy across 100 epochs.

Fig. 9. Classification performance indicators of the masked-first classifier with oversampling approach.

which concentrates radiative heat at the surface rather than dissipating it. The spatial pattern is highly encouraging, as cooling effect is visually noticeable around segments that contain green-eligible roofs, this reflects the importance of greening and its wide impact on heat mitigation. Aggregating these per-pixel differences to the city extent reveals that this targeted intervention is climatically beneficial. On the hottest day of the year, the full greening scenario yields an average LST reduction of approximately 0.83 K over all roofs in the entire city of Hamburg, and average cooling for greened buildings of 1.36 K. Interpreted as a mitigation measure, this suggests that greening structurally suitable roofs can contribute meaningfully to moderating projected heat extremes, while simultaneously reducing thermal load on the existing building stock and by extension the surrounding urban districts.

The UHI screening accuracy targeted in this work is directly related to the classification accuracy. However, not all classification errors carry equal weight. The confusion matrix (cf. Fig. 8(c)) shows that the two most frequent misclassifications occur between *concrete* and *tar_paper*, and between *metal* and *roof_tiles*. From a thermo-optical standpoint, *concrete* and *tar_paper* share similar albedo and emissivity characteristics, both being dark, low-reflectance flat roof surfaces (Oke et al., 2017). Confusing one for the other therefore has a negligible effect on the estimated LST, because the spectral parameters assigned to either class are close in value. The confusion between *metal* and *roof_tiles*, however, involves materials that differ substantially in reflectance and thermal mass, but it does not affect the greening simulation because neither class is considered eligible. *Metal* roofs pose structural challenges for green roof installation, and *roof_tiles* appear predominantly on gabled roofs, which are geometrically unsuitable for greening (Cascone, 2019; Forschungsgesellschaft Landschaftsentwicklung Landschaftsbau e.V. (FLL), 2018). As a result, the dominant misclassification pairs in this pipeline have marginal impact on the estimated cooling effect.

This impact of classification uncertainty on the reported cooling is quantified by propagation. Because the random forest regressor operates on scene derived spectral predictors and does not take the predicted material as an input, the material label influences the cooling estimate only through the composition of the eligible roof set. Per building class probability distributions were therefore propagated by Monte Carlo. In each of 5000 realizations a material was drawn for every building from its predicted class probability distribution, the material and slope eligibility filters were re-applied, and the area

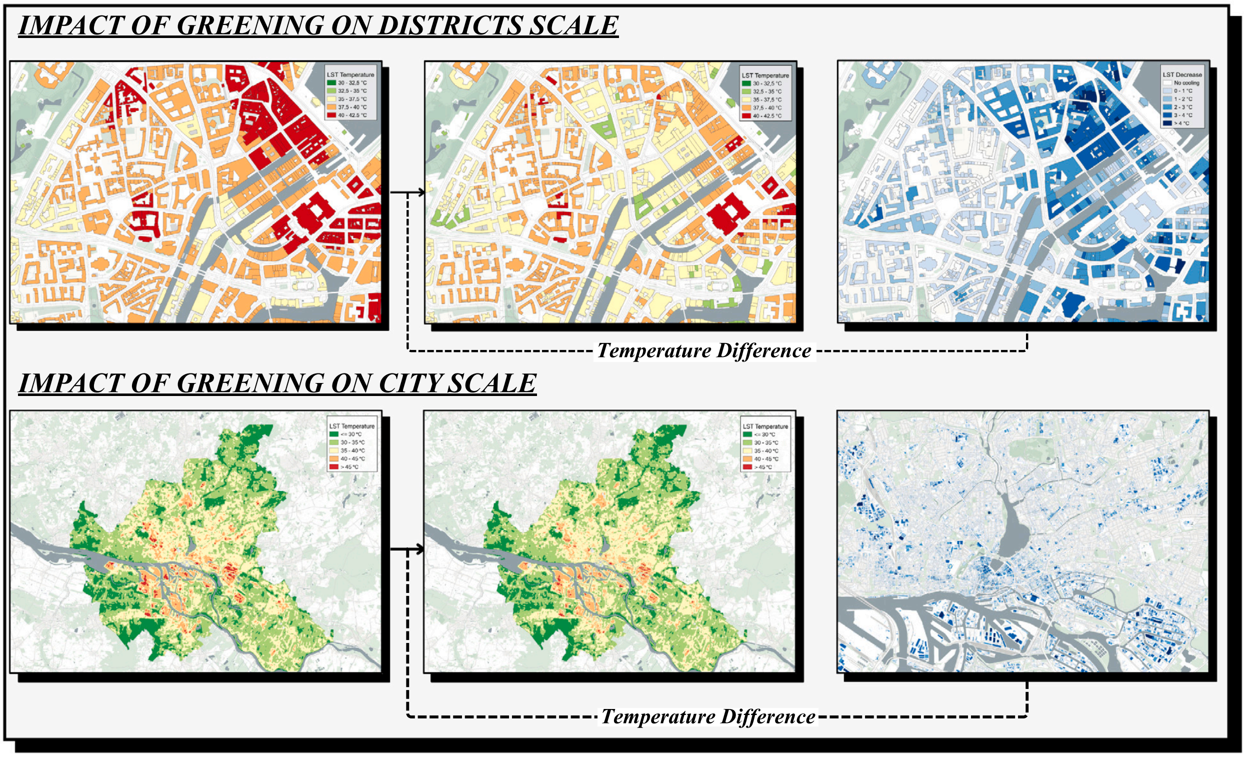


Fig. 10. UHI screening results and the effect of greening roofs on per-district and overall temperature in Hamburg.

Table 3
City-wide cooling impact of green-roof candidate selection across study cities.

City/scenario	Buildings	Candidates	Area (ha)	Cooling (K)
Paris < 11.31°	106,391	8,557	591.8	0.0642
Paris < 30°	106,391	34,997	1,418.1	0.1603
Madrid < 11.31°	121,006	4,895	699.2	0.2283
Madrid < 30°	121,006	25,644	2,030.8	0.6000
Hamburg	194,799	59,949	3,549.0	0.8386

weighted cooling over the resulting greened candidate roofs was recomputed. The distribution sampled is the classifier likelihood re-weighted by the class priors encoding city scale material base rates, whose argmax reproduces the deployed dominant material assignment for 99.8 percent of buildings. For Hamburg the area weighted cooling over greened candidate roofs is 1.3971 K under the deterministic labels, and propagation yields a mean of 1.3570 K with a 90 percent interval of 1.3456 K to 1.3683 K, approximately 1.7 percent of the mean.

The narrowness of this interval follows from the structure of the confusion matrix rather than from classifier confidence. The dominant confusion pair, concrete against tar_paper, leaves eligibility unchanged because both classes are eligible, and the confusion between metal and roof_tiles leaves it unchanged because neither class is eligible. Label uncertainty reaches the cooling estimate only through transitions crossing the eligibility boundary, which are infrequent. The reported interval quantifies the contribution of roof material classification uncertainty alone; the albedo and emissivity values taken from the literature, the random forest regression, and the reference land surface temperature product each carry residual uncertainty that is not propagated jointly.

To evaluate transferability beyond a single city, the pipeline was applied to two additional cities with contrasting urban morphologies and climatic regimes: Madrid and Paris, where greening eligibility was driven by OSM footprints and material classification alone, since openly available LoD2 CityGML datasets are not currently distributed at city scale for these two cities. Reference green-roof spectral parameters were derived per city from existing vegetated roofs detected via OSM tags, producing target NDVI values of 0.4 for Paris, and 0.28 for Madrid, with corresponding target albedo values of 0.14, and 0.17 respectively. These city-specific targets reflect the contrasting baseline vegetation contexts of each scene rather than a single universal greening template, and they enter the counterfactual scenario as area-weighted means.

The screening results across the three cities are summarized in Table 3 and visualized in Fig. 11. As highlighted in Section 3.7, greening for both Madrid and Paris is not as restricted as in Hamburg. A slope of 30° is still viable for greening, even if the cost is slightly higher since it at that slope level the greening installation requires an anti-slip system to keep the greening in place. Under these circumstances, the experiment was done on these 2 scenarios, the upper limit of 30 degrees, which requires anti shear protection, and the standard assembly limit of 11.31 degrees, which requires none. Paris showcased the lowest heat reduction in all 3 experiments with cooling gain of 0.06 K in the most encouraged greening scenario under 11° slope, and 0.16 K on 30° slope. This is mainly because of its cool climate overall and cloudy weather, as well as, its distinct architectural urban signature, as most of the buildings in Paris specially in the city center are gabled roofs and are not eligible for greening. Madrid, however, is a suitable setup for greening since it is warmer overall, and is exposed to the sun a lot more. The LST result of this city showed an overall cooling of 0.22 K on 11° installation, however on the 30° slope regime, the cooling difference becomes more noticeable with an average cooling on roofs of 0.6 K.

This city comparison clarifies that the magnitude of the simulated cooling depends as much on the urban setup and the climatic state as well. Paris, for example, exhibits the smallest area-weighted cooling among the three reported cities despite a candidate-roof area comparable to Madrid. Two factors combine to this lower estimate. First, the central arrondissements are dominated by gabled and mansard roof geometries that are not eligible for greening, which leaves a smaller building area available and eligible for greening compared to Madrid or in northern German cities where flat and *bituminous* roofs are more

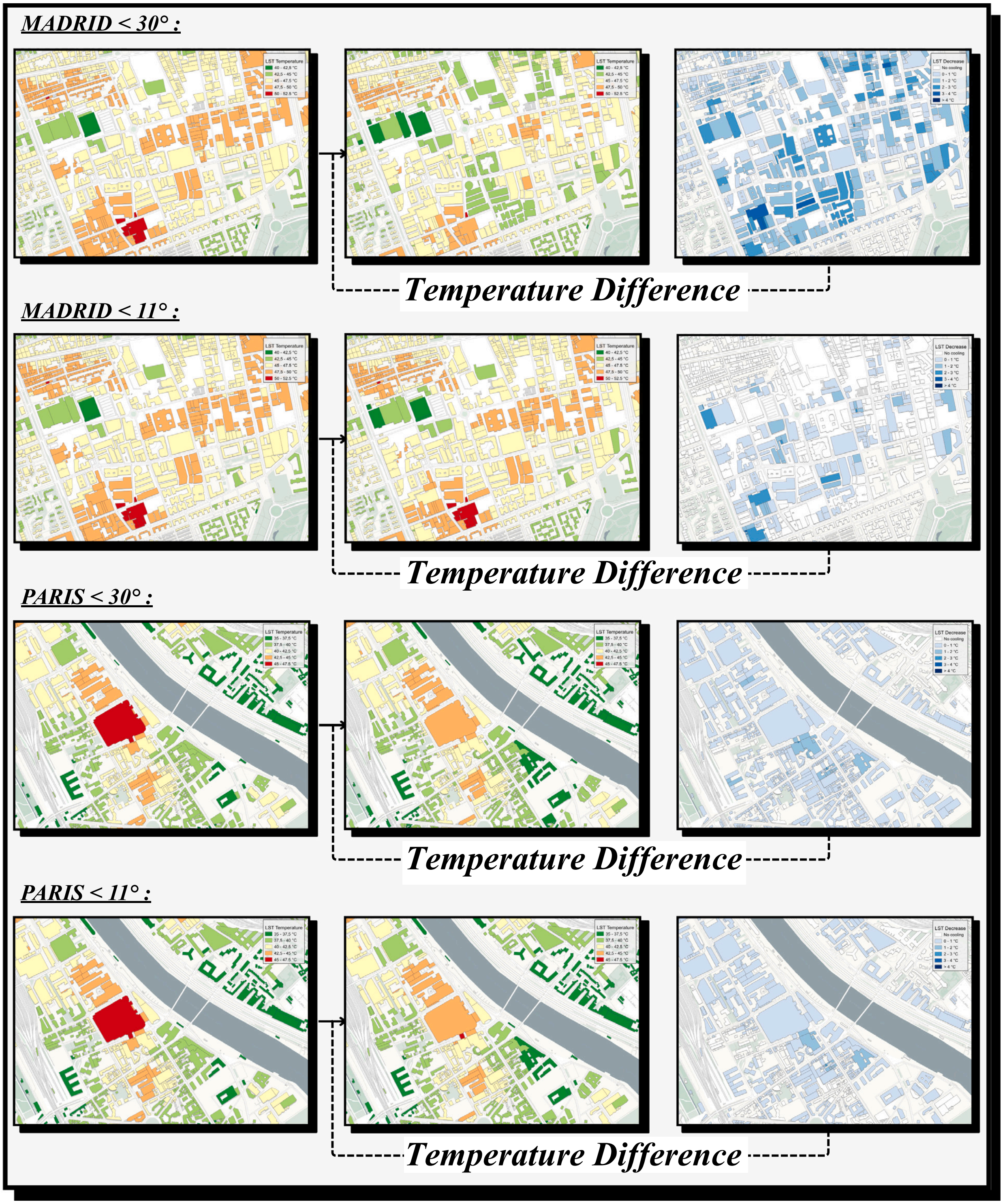


Fig. 11. UHI screening results and LST estimation of the effect of greening roofs on district level for Madrid and Paris.

prevalent on residential and commercial blocks. Second, the Paris scene was acquired at 24.4°C with 55% relative humidity, conditions under which the locally trained Random Forest sees a comparatively narrow thermal range, LST estimates shift to the more lower and negligible values. Madrid, by contrast, was acquired at 33.5°C with 22% relative humidity, which produces substantially more thermal headroom and consequently a stronger roof and city cooling responses. These results should not be read as a universal ranking of cooling potential between cities, since each value reflects a single Landsat scene per city under a fixed counterfactual template. They do, however, support a consistent operational reading: greening yields the largest per-building thermal benefit where the baseline surface is hot, dry and impervious, and where the structural eligibility of the green eligible roof stock is high.

It is worth noting that the accuracy of the material classification model for both Paris and Madrid is low in comparison with Hamburg because the openly available data from OSM for training for these two cities were very scarce. However, the accuracy could be further improved if the model were to be fine tuned or retrained on more data, specially government data if available. Nevertheless, the aim of this experiment is not to claim high performance of the classifier

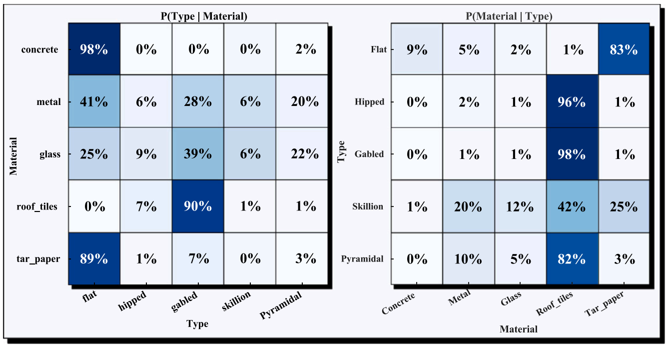


Fig. 12. Distribution correlation between roof type and roof material.

in areas in which data is scarce, but to show the actual benefit of greening a portion of the roofs under different climate settings and the reproducibility of this pipeline in cities where enough data is available.

5. Discussion

The results of this study are encouraging but also expose persistent challenges in extracting roof material semantics from aerial imagery. Although the classifier displays strong performance, and the inference framework succeeds in extracting multi-material partitions across diverse building geometries, the pipeline remains ruled by several strengths and weaknesses of the data. These factors define the operational scope within which this pipeline can be deployed at city scale.

A first strength lies in the generality of the approach. Since the method operates directly on georeferenced orthophotos, without relying on city specific priors, it remains applicable to heterogeneous datasets, including those that vary in illumination, color calibration, or acquisition year. The combination of whole-image priors, grid-based local override allows the classifier to identify secondary materials without retraining and without explicit boundary supervision. This proves especially advantageous in dense urban blocks, where roof complexity often exceeds the expressive power of single-label classifiers. The pipeline also preserves full spatial reference. Every mask, label, and building polygon is propagated through the geotransform and exported into a single GeoJSON file in EPSG:25832, carrying geometries, roof material IDs, and coverage fractions for each building. This enables direct integration into CityGML, and once building footprints are available, inference becomes fully scalable.

These strengths, however, implicate several structural limitations. The accuracy of the masking process is fundamentally tethered to the availability and positional accuracy of openly available footprint data. Missing buildings or mismatched polygons lead to incomplete coverage or misalignment, which renders the enrichments and later on UHI screenings less accurate. Another related limitation arises from textural similarity confusion between *metal* and *roof_tiles*, or between *concrete* and *tar_paper*. These textural ambiguities lead to misclassifications and confusion between classes, even thought the most common confusions are not heavily effecting the UHI screening accuracy, it is still a prominent issue that should be addressed in order to have the most accurate semantic 3D city models.

The transitions between RoofSurface objects in CityGML may also indicate locations where material changes are more likely to be, particularly in large building complexes where height differences separate distinct roof sections. Leveraging this geometric prior could improve multi-material inference by guiding the sliding window override toward structurally meaningful boundaries rather than sweeping a uniform grid. This is especially relevant for datasets from mapping agencies that produce subdivided roof geometries. Even thought this

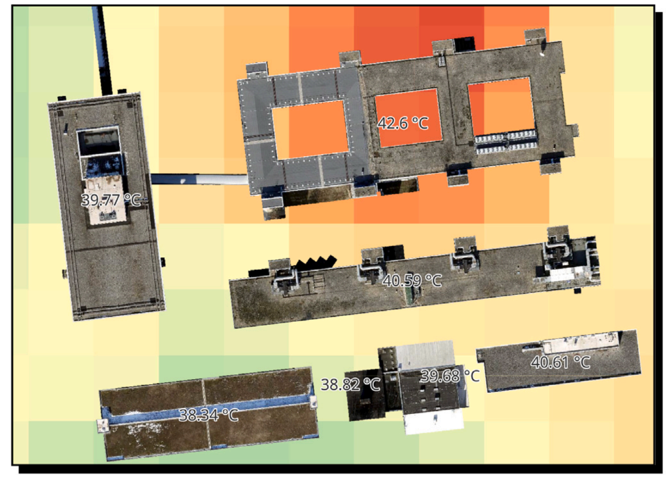


Fig. 13. Landsat land surface temperature evaluation, over identical atmospheric conditions, of a building block in Hamburg, Germany to assess the cooling benefits of green roofs.

integration of RoofSurface level classification into the pipeline is out of the scope of this paper, it is still a very promising direction for future work. However, this design choice will impact the scalability of the pipeline across different region where CityGML data is not available.

Beyond these considerations, the conversion of detected roof segments into geometries ensures that each material class can be stored as a spatially explicit semantic part within 3DCityDB, enabling downstream processes that require thermophysical information. UHI screenings, for instance, rely heavily on albedo, and heat capacity values, all of which vary substantially across the five material classes studied. The ability to assign these parameters at sub-roof scale allows microclimate models to move beyond coarse assumptions, and toward spatially resolved energy balance calculations.

Storing one dominant roof material per Building, rather than a list of material coverage tuples, is a deliberate simplified design. Technically, CityGML supports multi-material representation through nested GenericAttributeSet structures, but the schema has no native tuple type, no ordered list, and no constraint to keep coverage values summing to one. Getting proper ontology level consistency would require a dedicated ADE, which breaks compatibility with standard 3DCityDB installs. The pipeline also cannot reliably assign surface level material distributions when a building footprint spans multiple roof patches. The choice for one dominant material per building was the most reproducible option that still works with a standard setup.

An additional dimension of the discussion emerges when the material predictions are considered alongside the roof type information. Empirically, flat roofs are overwhelmingly dominated by tar-paper or bitumen materials, a pattern that the confusion matrices and probability distributions reflect in Fig. 12. This correlation can be exploited in UHI screening workflows. Wherever flat segments are detected, assigning *tar_paper* as a prior material class can accelerate large scale thermal parameterization when full inference is unavailable.

Yet this shortcut remains probabilistic rather than deterministic. Not all flat roofs are eligible for greening. Structural constraints, load-bearing capacity, and drainage conditions vary significantly between concrete slabs, reinforced bitumen, and metal structures. Relying solely on flatness indicator for greening risks overstating the cooling potential and understating the heterogeneity present within flat surfaces. In the case of flat roofs mixed between *metal* and *tar_paper* it becomes difficult to apply greening, and these limitation instances are not taken into account when deciding based on roof type alone.

The dataset contains a substantial number of flat roofs that, if incorrectly deemed eligible for greening, would noticeably distort the

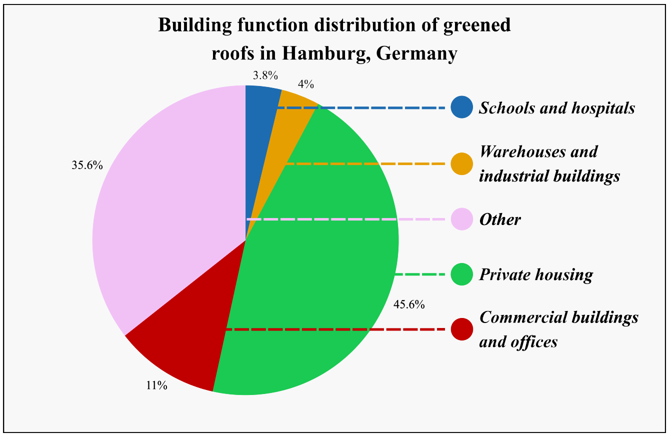


Fig. 14. Distribution of greening-eligible buildings by function in hamburg.

estimated cooling effect. If only type information is chosen as the definitive condition for greening, the potential for cooling is overestimated. However, leveraging correlation as a prediction helper tends to bias the classifier toward the dominant class, potentially degrading accuracy for minority flat-roof materials. Thus, while roof type and material relationships are informative, they should be incorporated as soft priors rather than hard rules.

The roof type information, however, proves most valuable as an auditing tool. Using the correlation matrix between roof type and material, it is possible to test whether the predicted roof is flat. Whenever the classifier predicts *tar_paper* or *concrete*, the roof type can be checked against the expected flat geometry, and if it does not match, the case is flagged as inconsistent with the roof type and material correlation matrix and not taken into consideration as a roof eligible for greening. This consistency check stabilizes the material fractions that feed the LST model, yielding more reliable and accurate UHI screenings.

The screening results carry a known bias arising from incomplete information on roofs that already host vegetation. Reliable records of existing green roofs are not widely available, and in Hamburg many roofs already carry vegetation. Without an attribute flagging these roofs, they cannot be excluded from the counterfactual scenario or down weighted within it. Because the scenario drives rooftop NDVI toward a reference derived from existing vegetated roofs, a roof already greener and cooler than that reference is perturbed toward a warmer state and returns an apparent warming. Such an outcome represents the inverse of the intervention, since the modeled change would remove vegetation rather than add it. These buildings fall outside the intervention being simulated, and their contribution is therefore set to zero rather than entering the aggregate as negative cooling. This keeps the aggregate consistent with the intervention it represents, but the remaining greening potential of those roofs is unresolved, so the reported cooling is conservative with respect to this group. Within these bounds the direction of the result is robust: systematic deployment of comparable greening measures at city scale would lower surface temperatures across the eligible roof stock, with associated reductions in cooling energy demand and, indirectly, greenhouse gas emissions, although the magnitude is conditional on the scenario assumptions and the single Landsat scene.

In that direction, a small qualitative study was conducted to assess the actual benefit of greening in a real life setting. Fig. 13 shows the per-roof Landsat LST values, extracted from the same Landsat scene used in the Hamburg screening, for multiple buildings. The vegetated roofs at the bottom of the panel, have the lowest LST values of 38.34°C, 38.82°C and 39.68°C, and the bituminous and concrete roofs above, have LST values ranging from 39.77°C to 42.6°C. What is straightforward and obvious is that the greened roof has the lowest LST in all of the buildings, with a mean cool offset of approximately 1.8 K. However, what is also interesting is how this green roof affects neighboring building temperatures. This is represented by the hue of green around the building. This shows the practical effect of greening and its potential on reducing the overall temperature on larger scales than only the roof in question.

However, the screening has not been validated against field measured roof level thermal observations. Repeated approaches to the relevant data providers did not yield roof level microclimate or thermal reference data. The absolute magnitudes reported here are therefore conditional scenario outputs rather than validated predictions, and they should be interpreted comparatively. Independent validation against field measured roof level data remains a prerequisite for operational use of the absolute values and a direction for future work.

Drifting the discussion from the real value and advantages of greening, another important thread that can support planning decisions further is a breakdown of greening-eligible buildings by function (residential, commercial, industrial, schools and hospitals). This vital information gathered from our enriched dataset can inform budgetary allocation, subsidy prioritization, and tax incentive design at the municipal level. Fig. 14 illustrate this distribution and can directly inflict real value for decision makers for where and what to green.

Replicating this pipeline in another city requires three inputs: high-resolution orthophotos available via WMS, building footprints from OSM or an equivalent open source, and cloud-free Landsat-8 scene over the target area. The classifier, trained on data from over ten German cities, generalizes across varying urban morphologies and illumination conditions without city specific fine-tuning. For cities outside Germany, retraining or fine-tuning on locally sourced labels may be necessary if the material palette differs substantially from the five classes used here (e.g., in regions where clay, slate, or thatch are common). All code used in this study is publicly available to support reproducibility.

6. Conclusion

This paper presents an end-to-end pipeline that turns open aerial imagery and building footprints into actionable roof material semantics linked directly to 3D city models. By combining a masked-first classification strategy with a deterministic, multi-material inference step, the pipeline achieves high performances, enriches 3D city buildings models with roof material attributes, and estimates the cooling effect of greening eligible roofs across cities. These outputs are directly actionable for urban planners evaluating greening potential and its contribution to climate adaptation.

The pipeline delivers auditable, geo-referenced labels that are immediately usable for CityGML enrichment and urban-heat-island screening. The pipeline demonstrates that city scale material enrichment can be achieved with reproducible, data-efficient methods while preserving full traceability from image evidence to database attribute, bridging the gap between remote sensing classification and urban climate analysis. As demonstrated, this pipeline is not city or region specific. It relies entirely on openly available inputs (orthophotos and OSM or CityGML footprints), making it transferable to any city where these resources exist. This positions the approach as a reusable template for city planners and greening assessments, advancing semantic city modeling towards an operational practice.

Looking ahead, improving discrimination and confusion between classes may require augmenting the model with maps that differentiate between materials at texture level or incorporating illumination invariant features. Higher resolution data would also alleviate some ambiguities. Additionally, roof tilt and orientation, which are already encoded in LoD2 CityGML geometry, could serve as inputs for assessing rooftop photovoltaic integration potential alongside greening suitability, enabling joint optimization of renewable energy generation and thermal comfort.

From an environmental modeling standpoint, future work can build on the present material maps to refine UHI calculations by integrating seasonal albedo variation and heat storage dynamics. Extending the pipeline to additional semantic layers, particularly facade materials, would enable building level energy balance calculations that are currently out of reach for city scale workflows, moving the system closer to a unified material model capable of supporting both operational planning and long term climate adaptation scenarios.

CRediT authorship contribution statement

Elmehdi Kanna: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Jannik Matijevic:** Methodology, Conceptualization. **Lukas Arzoumanidis:** Validation, Methodology, Conceptualization. **Huynh Duc An Son Nguyen:** Writing – review & editing, Visualization, Validation. **Youness Dehbi:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the project 'Next Generation City Networking' (Grant No. 19DZ24004) at the Hanseatic Wireless Innovation Competence Center (HAWICC). The project is funded by the Federal Ministry of Transport via the German Center for Future Mobility (DZM).

Data availability

All data and code used for this study is available at: Github.

References

Adivet, Csfe, & Enveloppe Métallique du Bâtiment (2018). Règles professionnelles pour la conception et la réalisation des terrasses et toitures végétalisées. 3 ed. ISBN: 978-2-9535882-1-7. URL: https://www.hqegbc.org/wp-content/uploads/2018/10/18-05-RP-TTV-ed3.pdf..

Ahmadian, S., & Pahlavani, P. (2022). Semantic integration of OpenStreetMap and CityGML with formal concept analysis. *Transactions in GIS*, *26*(8), 3349–3373. http://dx.doi.org/10.1111/tgis.13012.

Al Kafy, A., Crews, K. A., & Thompson, A. E. (2024). Exploring the cooling potential of green roofs for mitigating diurnal heat island intensity by utilizing LiDAR and Artificial Neural Network. *Sustainable Cities and Society*, *116*, Article 105893. http://dx.doi.org/10.1016/j.scs.2024.105893.

Arzoumanidis, L., Li, W., Knechtel, J., Kosmayadi, Y., & Dehbi, Y. (2025). Automatic detection of tiny drainage outlets and ventilations on flat rooftops from aerial imagery. *vol. X-G-2025*, In *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* (pp. 125–132). http://dx.doi.org/10.5194/isprs-annals-X-G-2025-125-2025.

Arzoumanidis, L., Nguyen, S. H., Johannsen, L., Rothaut, F., Li, W., & Dehbi, Y. (2025). Object detection for the enrichment of semantic 3D city models with roofing materials. *vol. X-4/W6-2025*, In *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* (pp. 9–16). http://dx.doi.org/10.5194/isprs-annals-X-4-W6-2025-9-2025.

Biljecki, F., Heuvelink, G. B., Ledoux, H., & Stoter, J. (2015). Propagation of positional error in 3D GIS: Estimation of the solar irradiation of building roofs. *International Journal of Geographical Information Science*, *29*(12), 2269–2294. http://dx.doi.org/10.1080/13658816.2015.1073292.

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, *215*, Article 104217. http://dx.doi.org/10.1016/j.landurbplan.2021.104217.

Biljecki, F., Ledoux, H., Du, X., Stoter, J., Soon, K. H., & Khoo, V. H. S. (2016). The most common geometric and semantic errors in CityGML datasets. *vol. IV-2/W1*, In *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* (pp. 13–22). http://dx.doi.org/10.5194/isprs-annals-IV-2-W1-13-2016.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., & Çöltekin, A. (2015). Applications of 3D city models: State of the art review. *ISPRS International Journal of Geo-Information*, *4*(4), 2842–2889. http://dx.doi.org/10.3390/ijgi4042842.

Bulatov, D., Burkard, E., Ilehag, R., Kottler, B., & Helmholz, P. (2020). From multi-sensor aerial data to thermal and infrared simulation of semantic 3D models: Towards identification of urban heat Islands. *Infrared Physics & Technology*, *105*, Article 103233. http://dx.doi.org/10.1016/j.infrared.2020.103233.

Cai, C., Li, B., Zhang, Q., Wang, X., Biljecki, F., & Herthogs, P. (2025). Bi-directional mapping of morphology metrics and 3D city blocks for enhanced characterisation and generation of urban form. *Sustainable Cities and Society*, Article 106441. http://dx.doi.org/10.1016/j.scs.2025.106441.

Calhoun, Z. D., Willard, F., Ge, C., Rodriguez, C., Bergin, M., & Carlson, D. (2024). Estimating the effects of vegetation and increased albedo on the urban heat island effect with spatial causal inference. *Scientific Reports*, *14*(1), 540. http://dx.doi.org/10.1038/s41598-023-50981-w.

Cascone, S. (2019). Green roof design: State of the art on technology and materials. *Sustainability*, *11*(11), 3020. http://dx.doi.org/10.3390/su11113020.

Castleton, H. F., Stovin, V., Beck, S. B. M., & Davison, J. B. (2010). Green roofs; Building energy savings and the potential for retrofit. *Energy and Buildings*, *42*(10), 1582–1591. http://dx.doi.org/10.1016/j.enbuild.2010.05.004.

Cordeiro, F. R., & Carneiro, G. (2020). A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? http://dx.doi.org/10.48550/arXiv.2012.03061, arXiv:2012.03061.

Dehbi, Y., Henn, A., Gröger, G., Stroh, V., & Plümer, L. (2021). Robust and fast reconstruction of complex roofs with active sampling from 3D point clouds. *Transactions in GIS*, *25*(1), 112–133. http://dx.doi.org/10.1111/tgis.12659.

Dehbi, Y., Koppers, S., & Plümer, L. (2019). Probability density based classification and reconstruction of roof structures from 3D point clouds. *Vol. XLII-4/W16*, In *The international archives of the photogrammetry, remote sensing and spatial information sciences* (pp. 177–184). http://dx.doi.org/10.5194/isprs-archives-XLII-4-W16-177-2019.

Ding, L., Xiao, G., Pano, A., Fumagalli, M., Chen, D., Feng, Y., Calvanese, D., Fan, H., & Meng, L. (2024). Integrating 3D city data through knowledge graphs. *Geo-Spatial Information Science*, 1–20. http://dx.doi.org/10.1080/10095020.2024.2337360.

Dodman, D., Hayward, B., Pelling, M., Castan Broto, V., Chow, W., Chu, E., Dawson, R., Khirfan, L., McPhearson, T., Prakash, A., Zheng, Y., & Ziervogel, G. (2022). Cities, settlements and key infrastructure. In H. O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegrí a, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, & B. Rama (Eds.), *Climate change 2022: Impacts, adaptation and vulnerability. contribution of working group II to the sixth assessment report of the intergovernmental panel on climate change* (pp. 907–1040). Cambridge, UK and New York, NY, USA: Cambridge University Press, http://dx.doi.org/10.1017/9781009325844.008.

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*(4), 700–719. http://dx.doi.org/10.1080/13658816.2013.867495.

Forschungsgesellschaft Landschaftsentwicklung Landschaftsbau e. V. (FLL) (2018). *guidelines for the planning, construction and maintenance of green roofs*. Bonn, Germany: Forschungsgesellschaft Landschaftsentwicklung Landschaftsbau e.V., URL: https://shop.fll.de/en/fll-english-publications/green-roof-guidelines-2018-download.html.

Fundació de la Jardineria i el Paisatge (2012). NTJ 11C: Cubiertas verdes. Barcelona, Spain. ISBN: 978-84-95372-35-2. URL: https://www.ntjdejardineria.org/ntj/ntj-11c-cubiertas-verdes/.

Gröger, G., Kolbe, T. H., Nagel, C., & Häfele, K.-H. (2012). *OGC city geography markup language (CityGML) encoding standard: OpenGIS® encoding standard 12–019*, Open Geospatial Consortium, Version 2.0.0..

Hamburgische Investitions- und Förderbank (IFB Hamburg) (2024). Förderrichtlinie hamburger gründachförderung. URL: https://www.ifbhh.de/resource/blob/1936/9df6f2f3a972bb33efca6f861f03105e/foerderrichtlinie-hamburger-gruendachfoerderung-data.pdf. Mindestgröße der Nettovegetationsfläche: 20 m^2.

Hoffmann, E. J., Wang, Y., Werner, M., Kang, J., & Zhu, X. X. (2019). Model fusion for building type classification from aerial and street view images. *Remote Sensing*, *11*(11), 1259. http://dx.doi.org/10.3390/rs11111259.

Ilehag, R., Bulatov, D., Helmholz, P., & Belton, D. (2018). Classification and representation of commonly used roofing material using multisensorial aerial data. *vol. XLII-1*, In *The international archives of the photogrammetry, remote sensing and spatial information sciences* (pp. 217–224). http://dx.doi.org/10.5194/isprs-archives-XLII-1-217-2018.

IPCC (2023). Summary for policymakers. In *climate change 2023: Synthesis report. contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change*. http://dx.doi.org/10.59327/IPCC/AR6-9789291691647.001.

Joshi, M. Y., Aliaga, D. G., & Teller, J. (2023). Predicting urban heat island mitigation with random forest regression in belgian cities. In R. Goodspeed, & et al. (Eds.), *The urban book series*, *Intelligence for future cities* (pp. 305–323). Springer Nature Switzerland, http://dx.doi.org/10.1007/978-3-031-31746-0_16.

Kanna, K., & Kolbe, T. H. (2025). Automatic enrichment of semantic 3D city models using large language models. *vol. X-4/W6-2025*, In *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* (pp. 105–112). http://dx.doi.org/10.5194/isprs-annals-X-4-W6-2025-105-2025.

Kim, J., Bae, H., Kang, H., & Lee, S. G. (2021). CNN algorithm for roof detection and material classification in satellite images. *Electronics*, *10*(13), 1592. http://dx.doi.org/10.3390/electronics10131592.

Kolbe, T. H., Gröger, G., & Plümer, L. (2008). CityGML – 3D city models and their potential for emergency response. In *Geospatial information technology for emergency response* (pp. 257–274). Taylor & Francis, http://dx.doi.org/10.4324/9780203928813-25.

Kolbe, T. H., Kutzner, T., Smyth, C. S., Nagel, C., Roensdorf, C., & Heazel, C. (2021). *OGC city geography markup language (CityGML) version 3.0 part 1: conceptual model standard* (p. 463). Open Geospatial Consortium, URL: https://docs.ogc.org/is/20-010/20-010.html. International Standard.

Krówczyńska, M., Raczko, E., Staniszewska, N., & Wilk, E. (2020). Asbestos–cement roofing identification using remote sensing and convolutional neural networks (CNNs). *Remote Sensing*, *12*(3), 408. http://dx.doi.org/10.3390/rs12030408.

Kutzner, T., Smyth, C., Nagel, C., Coors, V., Vinasco-Alvarez, D., Ishimaru, N., Yao, Z., Heazel, C., & Kolbe, T. H. (2023). *OGC city geography markup language (CityGML) version 3.0 part 2: GML encoding standard*. Open Geospatial Consortium, URL: https://docs.ogc.org/is/21-006r2/21-006r2.html. International Standard.

Ledoux, H. (2018). Val3dity: Validation of 3D GIS primitives according to the international standards. *Open Geospatial Data, Software and Standards*, *3*(1), 1–12. http://dx.doi.org/10.1186/s40965-018-0043-x.

Li, K., & Zeng, H. (2024). Multidisciplinary parameters for characterizing the 3D urban morphology: An overview based on the relational perspective. *Sustainable Cities and Society*, *106*, Article 105328. http://dx.doi.org/10.1016/j.scs.2024.105328.

Liang, S. (2001). Narrowband to broadband conversions of land surface albedo I: algorithms. *Remote Sensing of Environment*, *76*(2), 213–238. http://dx.doi.org/10.1016/S0034-4257(00)00205-4.

Liang, X., Xie, J., Zhao, T., Stouffs, R., & Biljecki, F. (2025). OpenFACADES: An open framework for architectural caption and attribute data enrichment via street view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *230*, 918–942. http://dx.doi.org/10.1016/j.isprsjprs.2025.10.014.

Lin, M., Dong, J., Jones, L., Liu, J., Lin, T., Zuo, J., Ye, H., Zhang, G., & Zhou, T. (2021). Modeling green roofs' cooling effect in high-density urban areas based on law of diminishing marginal utility of the cooling efficiency: A case study of Xiamen Island, China. *Journal of Cleaner Production*, *316*, Article 128277. http://dx.doi.org/10.1016/j.jclepro.2021.128277.

López-Cabeza, V. P., Alzate-Gaviria, S., Diz-Mellado, E., Rivera-Gómez, C., & Galán-Marín, C. (2022). Albedo influence on the microclimate and thermal comfort of courtyards under Mediterranean hot summer climate conditions. *Sustainable Cities and Society*, *81*, Article 103872. http://dx.doi.org/10.1016/j.scs.2022.103872.

Lu, H., et al. (2023). Thermal effects of cool roofs and urban vegetation during extreme heat events in three Canadian regions. *Sustainable Cities and Society*, *99*, Article 104925. http://dx.doi.org/10.1016/j.scs.2023.104925.

Mansourmoghaddam, M., Rousta, I., Ghafarian Malamiri, H., Sadeghnejad, M., Krzyszczak, J., & Ferreira, C. S. S. (2024). Modeling and estimating the land surface temperature (LST) using remote sensing and machine learning: Case study Yazd, Iran. *Remote Sensing*, *16*(3), 454. http://dx.doi.org/10.3390/rs16030454.

Mashhoodi, B., & Unceta, P. M. (2024). Urban form and surface temperature inequality in 683 European cities. *Sustainable Cities and Society*, *113*, Article 105690. http://dx.doi.org/10.1016/j.scs.2024.105690.

McConnell, K., Braneon, C. V., Glenn, E., Stamler, N., Mallen, E., Johnson, D. P., Pandya, R., Abramowitz, J., Fernandez, G., & Rosenzweig, C. (2022). A quasi-experimental approach for evaluating the heat mitigation effects of green roofs in Chicago, Illinois. *Sustainable Cities and Society*, *76*, Article 103376. http://dx.doi.org/10.1016/j.scs.2021.103376.

Nguyen, S. H. (2024). *Automatic detection and interpretation of changes in massive semantic 3D city models* (Ph.D. thesis), (p. 380). Technical University of munich, URL: https://mediatum.ub.tum.de/1748695.

Nguyen, S. H., & Kolbe, T. H. (2024). Identification and interpretation of change patterns in semantic 3D city models. In T. H. Kolbe, A. Donaubauer, & C. Beil (Eds.), *Lecture notes in geoinformation and cartography*, *Recent advances in 3D geoinformation science* (pp. 479–496). Springer Nature Switzerland, http://dx.doi.org/10.1007/978-3-031-43699-4_30.

Ogawa, Y., Sato, G., & Sekimoto, Y. (2024). Geometric-based approach for linking various building measurement data to a 3D city model. *PLoS One*, *19*(1), Article e0296445. http://dx.doi.org/10.1371/journal.pone.0296445.

Oke, T. R., Mills, G., Christen, A., & Voogt, J. A. (2017). *Urban climates*. Cambridge, UK: Cambridge University Press, http://dx.doi.org/10.1017/9781139016476.

Park, J., Park, S., & Kang, J. (2024). Detecting and classifying rooftops with a CNN-based remote-sensing method for urban area cool roof application. *Energy Reports*, *11*, 2516–2525. http://dx.doi.org/10.1016/j.egyr.2024.02.001.

Sánchez-Cordero, F., Nanía, L., Hidalgo-García, D., & López-Chacón, S. R. (2025). Assessing the spatial benefits of green roofs to mitigate urban heat island effects in a semi-arid city: A case study in Granada, Spain. *Remote Sensing*, *17*(12), 2073. http://dx.doi.org/10.3390/rs17122073.

Schwab, B., Beil, C., & Kolbe, T. H. (2020). Spatio-semantic road space modeling for vehicle-pedestrian simulation to test automated driving systems. *Sustainability*, *12*, 3799. http://dx.doi.org/10.3390/su12093799.

Seeberg, G., Hostlowsky, A., Huber, J., Kamm, J., Lincke, L., & Schwingshackl, C. (2022). Evaluating the potential of landsat satellite data to monitor the effectiveness of measures to mitigate urban heat Islands: A case study for Stuttgart (Germany). *Urban Science*, *6*(4), 82. http://dx.doi.org/10.3390/urbansci6040082.

Solovyev, R. A. (2020). Roof material classification from aerial imagery. *Optical Memory and Neural Networks*, *29*(3), 198–208. http://dx.doi.org/10.3103/S1060992X20030133.

Sun, K., Li, Q., Liu, Q., Song, J., Dai, M., Qian, X., Gummidi, S. R. B., Yu, B., Creutzig, F., & Liu, G. (2025). Urban fabric decoded: High-precision building material identification via deep learning and remote sensing. *Environmental Science and Ecotechnology*, *24*, Article 100538. http://dx.doi.org/10.1016/j.ese.2025.100538.

Tan, J. K. D., Zhu, R., Song, J., Qin, Z., Xu, Y., & Chen, Y. (2025). AdvUNet3+: Segmentation of city-scale heterogeneous building façade materials from street view images. *Sustainable Cities and Society*, *126*, Article 106407. http://dx.doi.org/10.1016/j.scs.2025.106407.

Testolina, P., Polese, M., Johari, P., & Melodia, T. (2024). Boston twin: The Boston digital twin for ray-tracing in 6G networks. (pp. 441–447). http://dx.doi.org/10.1145/3625468.3652190.

USGS EROS (2018). Landsat 8 OLI (operational land imager) and TIRS (thermal infrared sensor). URL: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-landsat-archives-landsat-8-oli-operational-land-imager-and. OLI multispectral: 30 m; TIRS thermal: 100 m (resampled to 30 m).

Wan, Y., Du, H., Yuan, L., Xu, X., Tang, H., & Zhang, J. (2025). Exploring the influence of block environmental characteristics on land surface temperature and its spatial heterogeneity for a high-density city. *Sustainable Cities and Society*, *118*, Article 105973. http://dx.doi.org/10.1016/j.scs.2024.105973.

Wang, C., Li, Z., Su, Y., Zhao, Q., He, X., Wu, Z., Gao, W., & Wu, Z. (2024). Impact of block morphology on urban thermal environment with the consideration of spatial heterogeneity. *Sustainable Cities and Society*, Article 105622. http://dx.doi.org/10.1016/j.scs.2024.105622.

Weil, C., Bibri, S. E., Longchamp, R., Golay, F., & Alahi, A. (2023). Urban digital twin challenges: A systematic review and perspectives for sustainable smart cities. *Sustainable Cities and Society*, *99*, Article 104862. http://dx.doi.org/10.1016/j.scs.2023.104862.

Willenborg, B., Sindram, M., & Kolbe, T. H. (2017). Applications of 3D city models for a better understanding of the built environment. In *Geotechnologies and the environment*: *vol. 19*, *Trends in spatial analysis and modelling* (pp. 167–191). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-52522-8_9.

Wu, A. N., & Biljecki, F. (2021). Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. *Landscape and Urban Planning*, *214*, Article 104167. http://dx.doi.org/10.1016/j.landurbplan.2021.104167.

Yan, X., Liu, J., Zhang, W., Wang, Z., Jia, R., Wang, L., Liu, N., & Zhu, Y. (2025). Examination of green roofs' cooling effects in a big city based on multi-source remote sensing data: An example from central Xi'an, China. *Sustainable Cities and Society*, *130*, Article 106545. http://dx.doi.org/10.1016/j.scs.2025.106545.

Yao, Z., Nagel, C., Kunde, F., Hudra, G., Willkomm, P., Donaubauer, A., Adolphi, T., & Kolbe, T. H. (2018). 3DCityDB: A 3D geodatabase solution for the management, analysis, and visualization of semantic 3D city models based on CityGML. *Open Geospatial Data, Software and Standards*, *3*(5), 1–26. http://dx.doi.org/10.1186/s40965-018-0046-7.

Yu, D., Ye, F., Yue, P., Chen, M., & Biljecki, F. (2025). RooFormer: Reconstructing detailed 3D roof models from high-resolution remote sensing imagery using transformer. *ISPRS Journal of Photogrammetry and Remote Sensing*, *227*, 745–758. http://dx.doi.org/10.1016/j.isprsjprs.2025.06.010.

Zhekov, S. S., Franek, O., & Pedersen, G. F. (2020). Dielectric properties of common building materials for ultrawideband propagation studies [measurements corner]. *IEEE Antennas and Propagation Magazine*, *62*(1), 72–81. http://dx.doi.org/10.1109/MAP.2019.2955680.