Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hidden and fragmented patterns of morphological informality across African cities revealed by open geospatial data

WENBIN Zhang

wb.zhang@soton.ac.uk

University of Southampton https://orcid.org/0000-0002-9295-1019

Weixuan Fang

University of Southampton

Alessandro Sorichetta

Università degli Studi di Milano https://orcid.org/0000-0002-3576-5826

Borys Nosatiuk

University of Southampton

Rhorom Priyatikanto

University of Southampton

Tom McKeen

University of Southampton

Longwu Liang

University College London

Maksym Bondarenko

University of Southampton https://orcid.org/0000-0003-4958-6551

Shengjie Lai

University of Southampton https://orcid.org/0000-0001-9781-8148

Jessica Espey

University of southampton

Peter Atkinson

Lancaster University

Andrew Tatem

University of Southampton

Natalia Tejedor-Garavito

University of Southampton

Article

Keywords:

Posted Date: June 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9877361/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Informal settlements are among the fastest-growing forms of urbanization, yet their spatial distributions within cities remain under-represented and inconsistently documented in official statistics and global urban datasets. Conventional assessments often identify large, contiguous informal neighborhoods but miss smaller and fragmented forms embedded within otherwise formal urban areas. This spatial invisibility limits the ability of cities to identify where informality intersects with infrastructure deficits, environmental risk and service provision. Here, we develop a scalable framework to map morphological informality using open satellite imagery and global building datasets. City-specific machine learning models are combined using a similarity-weighted ensemble that enables cross-city transfer while retaining local sensitivity. Applying the approach across 84 African cities, we generate 10 m resolution probability surfaces describing the likelihood of morphologically informal settlements. The resulting maps reveal substantial intra-urban heterogeneity in the spatial expression of morphological informality. In addition to identifying extensive informal settlements, the framework captures fragmented clusters embedded within more formal neighborhoods, which are patterns frequently overlooked by existing global mapping products and coarse statistical indicators. By revealing these hidden spatial structures, the approach provides a harmonised and scalable workflow with global applicability, enabling rapid updates to monitor change and supporting the integration of contextual, place-based knowledge into assessments of informal settlement status and impacts.

Introduction

Rapid urbanization is transforming cities, in many regions, faster than planning and governance systems can respond, leaving large and growing populations living in informally developed urban areas with limited access to secure housing, infrastructure, and services[1,2]. Yet the spatial distribution of these informal settlements remains systematically underrepresented in official data, rendering them partially invisible to urban planning, policy, and crisis response[3]. Conventional data sources, including censuses, household surveys, and administrative audits, are costly, infrequent, and poorly suited to capturing the fine-scale spatial organization of urban deprivation[4–6]. As a result, many existing assessments identify large, contiguous informal neighborhoods but overlook smaller and fragmented forms embedded within otherwise formal urban areas.

This invisibility is further compounded by conceptual ambiguity. Terms such as *slums*, *informal settlements*, *favelas*, *shanty towns*, and *unplanned* or *unserviced* neighborhoods are often used interchangeably, despite referring to diverse urban conditions[7]. Informality therefore resists neat categorization and reflects overlapping conditions of tenure insecurity, infrastructure deficits, and irregular development. While *slums* are defined as household-level deprivations designed for survey-based monitoring (e.g. SDG 11.1.1), *informal settlements* are place-based conditions characterized by irregular planning, insecure tenure, and uneven service provision[8,9]. Although informal settlements are not uniformly poor, they frequently exhibit distinctive morphological characteristics, such as dense

clusters of small buildings and irregular layouts, that reflect exclusion from formal urban systems[10]. Mapping morphological informality at fine spatial resolution therefore provides a complementary perspective on urban deprivation, enabling identification of where these spatial patterns are embedded within the urban fabric.

Recent advances in Earth observation provide a promising path forward to identify street-level morphological informality and to update its spatial footprint rapidly as new imagery becomes available. Freely available, fine-resolution satellite sensor imagery (e.g., Sentinel-2) and open building datasets, such as Google Open Buildings and those from OpenStreetMap, now enable scalable and repeatable mapping of informal urban settlements across large regions. Several studies have demonstrated the feasibility of detecting morphologically informal settlements using satellite imagery paired with machine learning models, offering a practical alternative to costly field-survey data[11]. However, most existing efforts are confined to a limited number of cities[12–15], while for large-scale studies, many models are trained using pooled label data across multiple sites[16–18], implicitly assuming that informal settlements share common morphological characteristics across regions. In practice, the morphology of informal settlements varies widely across countries and over time, reflecting self-organizing growth and local socio-economic conditions[19], while labelled reference data remain scarce. Despite recurring morphological signatures, this pooling approach limits the ability of conventional models to generalize across diverse urban contexts. Community-driven mapping efforts such as the Atlas of Informality[20] and Know Your City[21] have begun to address this gap, and recent initiatives like IDEAtlas and its IDEABench dataset now provide carefully curated, multi-city benchmarks of deprived urban areas based on Sentinel-1 and Sentinel-2 patches to advance model development and comparison[22]. However, their coverage remains geographically limited and, in some cases, outdated with the documented settlements now upgraded or redeveloped[20,23]. Together, these limitations point to the need for both scalable and timely mapping methods and a flexible, generalizable framework that accounts for the spatial and contextual diversity of morphological informality across cities and diverse urban contexts.

Here, we address this gap by testing whether open satellite imagery and global building datasets can be used to map urban morphological informality across diverse cities using transferable machine learning models. Specifically, we evaluate whether multispectral satellite imagery and open building-footprint data can jointly capture physical characteristics commonly associated with morphologically informal development, including building density, vertical structure, vegetation context, settlement layout, and urban texture, at 10 m spatial resolution. Using available informal settlement boundaries from the Know Your City initiative for eight cities, we construct city-specific training datasets and fit separate Random Forest classifiers. To enable cross-city transfer while retaining sensitivity to local urban form, we introduce a similarity-weighted ensemble framework in which predictions for a target city are generated by combining models trained in other cities, with weights determined by similarity across indicators of urban form, growth, and service access. This design explicitly relaxes the assumption of a single, globally consistent morphology of informality. Model outputs are represented as continuous probability

surfaces describing the likelihood of morphological informality, allowing both extensive and fragmented expressions of informality to be retained without imposing arbitrary class boundaries.

The framework is then applied across a large and heterogeneous set of African cities to evaluate its scalability and transferability across diverse urban contexts. The resulting probability surfaces provide spatially coherent representations of morphological informality that capture both extensive settlements and fragmented clusters embedded within more formal neighborhoods. Rather than serving as definitive maps of informal settlement extent, these outputs are intended as decision-support and hypothesis-generating tools that can guide validation, support community-based mapping and surveys, prioritize upgrading and infrastructure investments, and contribute to more equitable urban governance.

Results

A scalable framework for mapping morphological informality

We developed a scalable workflow in Fig. 1 to map morphological informality across cities using open satellite imagery and global building datasets. Training data were compiled through manual delineation of morphologically informal and formal settlement polygons in eight cities. Built-environment and environmental covariates were extracted within urban extents defined by the Degree of Urbanisation framework[24] and used to train city-specific Random Forest models. To enable prediction in cities without local training data, we implemented a similarity-weighted ensemble that transfers model predictions across cities based on their similarity in urban form and development context. The resulting probability surfaces were calibrated using national statistics on urban informal settlement populations and validated using independent survey datasets and AI-assisted image classification. This workflow enables the generation of fine-scale probability maps of morphological informality across diverse urban contexts.

Continental patterns of urban morphological informality

To examine spatial patterns of morphological informality, Fig. 2 illustrates our generated 10 m resolution probability surfaces for 84 African cities with projected populations exceeding one million in 2025. These probability surfaces should be interpreted as indicators of morphological informality rather than definitive classifications of settlement status, highlighting spatial patterns in the built environment that may be associated with informal development. Across cities, the resulting maps reveal substantial heterogeneity in the spatial expression of morphological informality, both within and between urban areas. In some cities, such as Freetown, Lagos, and Nairobi, large and spatially coherent zones of high probability correspond to extensive areas of dense and irregular built form. In contrast, cities such as Cairo and Durban exhibit smaller and more fragmented clusters of elevated probability embedded within predominantly formal urban fabric. These patterns demonstrate that morphological informality can

occur both as large contiguous settlements and as dispersed clusters within otherwise formal neighborhoods, highlighting the diversity of informal urban development across cities.

At the intra-urban scale, the probability surfaces further demonstrate the sensitivity of the approach to distinct morphological expressions. In Lagos, for example, the model assigns consistently high probabilities to the Makoko area (Fig. 3a–b), identifying a dense and spatially coherent settlement characterized by compact roof patterns, limited formal road structure, and a complex land–water interface visible in satellite imagery. By contrast, in the Mabushi area of Abuja (Fig. 3c–d), the model identifies a smaller and more isolated cluster of elevated probability, surrounded by regularly planned neighborhoods with larger building footprints and ordered street layouts. This contrast illustrates the ability of the framework to detect both extensive settlements and highly localized clusters of morphologically informal development across different urban contexts.

Transferability of morphological informality predictions across cities

To assess the transferability of the similarity-weighted ensemble across cities, we conducted leave-one-out cross-validation using the eight cities with labelled training data (Freetown, Lagos, Nairobi, Kampala, Accra, Dar es Salaam, Cape Town and Port Harcourt). Across all transfer experiments, the ensemble achieved consistently strong predictive performance. Area under the receiver operating characteristic curve (AUC) values ranged from 0.91 to 0.99, indicating high separability between morphologically informal and formal settlement areas under cross-city transfer. Binary classifications derived from probability thresholds optimized for F1-score also showed strong agreement with the held-out reference data. F1-scores ranged from 0.74 to 0.95, with lower values primarily observed in cities exhibiting more fragmented or atypical settlement morphology. These results indicate that the ensemble framework can generalize across diverse urban contexts while retaining sensitivity to local morphological variation. Optimized evaluation thresholds varied across cities, reflecting differences in how morphological signals translate into informal settlement likelihood in different urban environments. Full performance metrics, including evaluation thresholds, F1-scores and AUC values, are reported in Table 1.

Table 1
Performance metrics from leave-one-out cross-validation of the similarity-weighted ensemble across eight training cities.

City	Binary Comparison		Probability level comparison
	Threshold	F1	AUC
Accra	0.50	0.93	0.99
Cape Town	0.71	0.82	0.94
Dar es-Salaam	0.48	0.95	0.99
Freetown	0.40	0.90	0.96
Kampala	0.22	0.94	0.98
Lagos	0.50	0.84	0.96
Nairobi	0.41	0.92	0.98
Port Harcourt	0.83	0.74	0.91

Context-dependent signals of morphological informality and threshold calibration

Feature-importance patterns in Fig. 4a reveal substantial variation in the built-environment characteristics associated with signals of morphological informality across cities. In some contexts, such as Kampala, these signals arise from a combination of spectral indicators (NDBI and NDVI variability), intermediate building presence and vertical irregularity. In others, such as Freetown, high probabilities are primarily associated with exceptionally low building presence, reflecting fragmented and sparsely built settlement morphology. In several large cities, including Lagos and Dar es Salaam, vertical irregularity emerges as the dominant signal, whereas in Nairobi and Accra exceedingly high building presence provides the strongest indicator. In Cape Town, predictions are driven almost entirely by completeness of the built footprint, with minimal contribution from height-related metrics. These patterns indicate that the ensemble model does not rely on a single morphological template of informality but instead responds to context-specific expressions of informal urban form.

This heterogeneity implies that a single global probability threshold would perform inconsistently across cities. Instead, thresholds must account for differences in how morphological signals translate into the likelihood of informality across urban contexts. To address this, we developed a calibration strategy that adjusts probability thresholds using national statistics on the share of urban populations living in informal settlements. Figure 4b compares thresholds derived from leave-one-out cross-validation in cities with training data with population-adjusted thresholds estimated using a nonlinear regression model. The close correspondence between these estimates indicates that national statistics can be translated into operational probability thresholds for interpreting the probability surfaces.

Further comparison shows that thresholds derived solely from population alignment can systematically under- or over-estimate morphological informality depending on how population is distributed within built-up areas (see Fig. 4c). Model-based recalibration reduces this bias, producing thresholds that better reconcile probability surfaces with both aggregate statistics and model behavior. The spatial distribution of threshold adjustments in Fig. 4d shows that larger corrections occur in countries where morphologically informal areas are dense or vertically complex and therefore underrepresented in gridded population datasets, whereas smaller adjustments occur where informal morphology aligns more closely with population allocation patterns.

Validation

To evaluate model performance, we validated predicted probabilities against independent reference points across four countries including Ghana, Kenya, Malawi, and Nigeria (Fig. 5a). The distributions for the predicted probability of morphologically informal settlement show clear separation between groups: informal settlement points cluster at high predicted probabilities (mean = 0.81), while formal settlement points are concentrated near zero (mean = 0.15). This difference is highly significant (Welch's two-sample t-test, $t = -69.7$, $p < 2 \times 10^{-16}$), confirming the validity of the model. When applying country-specific updated thresholds, overall classification accuracy increased from 85.1% to 91.0% compared to the use of baseline thresholds derived from national informal-settlement population shares. This improvement is particularly pronounced for formal settlement points, where the misclassification rate was reduced from 30.9% under the original thresholds to 9.7% with the updated thresholds (see Fig. 5b). Accuracy for informal settlement points remained high for the updated thresholds, with a recall value of 91.5%. The updated thresholds also resulted in higher precision (91.5% vs 78.5%) and F1-scores (0.92 vs 0.88) compared to the baseline thresholds, reflecting improved balance between correctly identified informal settlements and reduced false positives.

Independent validation using AI-assisted image classifications from OpenAI further corroborated these patterns. Model predicted probabilities increased systematically across settlement categories, with mean probabilities rising from 0.17 for *Formal*, to 0.28 for *Probable formal*, 0.35 for *Mixed*, 0.57 for *Probable informal*, and 0.72 for *Informal* (see Fig. 5c). For each pair of adjacent categories, mean probabilities differed significantly (all $p < 0.01$, Welch's t-test; see Supplementary Table 2), confirming the model's ability to discriminate among nuanced settlement types. These results demonstrate both the strong discriminative capacity of the model and the value of recalibrated thresholds in reducing false positives.

Discussion

Cities, and the global institutions that shape urban policy, financing, and monitoring, cannot plan for what they cannot see. Informal settlements are frequently underrepresented in official statistics, not because they are uncommon, but because they are spatially fragmented, embedded within formal neighborhoods,

and poorly captured by conventional data sources[4–6]. As a result, existing datasets and mapping approaches tend to identify large, contiguous informal settlements while overlooking smaller clusters dispersed across the urban fabric. Recent advances in open geospatial datasets and machine-learning methods now make it possible to quantify the morphological dimensions of informality at fine spatial scales, creating new opportunities not only for municipal planning but also for global decision-making, advocacy, resource allocation, and urban research.

Here, we have tested whether a harmonised combination of open satellite imagery, building-structure datasets, and transferable machine-learning models can reliably capture urban morphological informality across diverse contexts. Specifically, we evaluated whether Sentinel-2 spectral information, Google Open Buildings 2.5D temporal building metrics, and a similarity-weighted ensemble of city-specific classifiers can identify spatial patterns of morphological informality under cross-city transfer conditions. Rather than aiming to produce definitive informal-settlement boundaries, the framework was designed to generate fine-resolution probability surfaces that preserve uncertainty and heterogeneity in urban form. Crucially, the framework is designed to complement, not replace, survey-based measures of deprivation. Household indicators such as SDG 11.1.1 provide indispensable information on service access and housing conditions[1], but they are spatially coarse and sensitive to the population surfaces used for disaggregation. By localizing morphological informality at street scale, the probability surfaces provide a missing spatial link between aggregate monitoring indicators and the place-based decisions that determine where investments, upgrades, and services are delivered[25].

Across a large and heterogeneous testbed of 84 African cities, our results show that such surfaces can recover both large contiguous informal areas and small, isolated clusters embedded within otherwise formal neighborhoods, which are routinely missed by coarser global mapping products[6,26]. Feature-importance analyses further show that the morphological signals associated with high informality likelihood vary substantially across cities, with different combinations of building density, vertical structure, and spectral characteristics dominating in different contexts[10,27]. This diversity highlights the risks of assuming a universal informal morphology and underscores the value of representations that preserve spatial heterogeneity.

Beyond mapping informality directly, the probability surfaces also highlight a pathway for improving gridded population modelling. Existing products such as WorldPop and GHSL typically disaggregate census counts using proxies that perform poorly in informal settlements, such as night-time lights, road networks, and formal infrastructure, leading to systematic underestimation of population density in precisely those neighborhoods most exposed to service deficits and environmental risk[28,29]. As a result, dense informal neighborhoods are often assigned too few people. The street-scale probability surfaces provide a place-based prior that can be integrated into population disaggregation workflows to re-weight density toward under-lit and under-serviced blocks, with implications for exposure estimates, service catchments, fiscal targeting, and sub-city benchmarking. In this sense, morphological informality mapping offers a spatial bridge between high-level monitoring statistics and the localized decisions that

determine where investments are made. This linkage is directly relevant to SDG 11.1.1, as morphological signals help identify where planning irregularity and service deficits are most likely to intersect, enabling sub-city benchmarking over time.

Our study also highlights the analytical value of emerging open building datasets. The Google Open Buildings 2.5D Temporal dataset provides large-area, temporally consistent information on building presence, fractional coverage, and height across much of the Global South[30]. These attributes are directly informative for identifying compact, low-rise, and irregular built fabrics that cannot be reliably inferred from two-dimensional footprints alone[31,32], and they enable tracking changes in settlement morphology through time[33]. More broadly, the results underscore the importance of building layers that combine spatial completeness, vertical structure, and temporal continuity; extending these characteristics to widely used datasets such as OSM, Microsoft, and Ecopia would substantially improve their utility for monitoring urban informality, particularly when combined with community-generated data and local expertise.

Several limitations must be acknowledged. First, the probability surfaces capture built-environment morphology, not social, legal, or economic conditions. High probability reflects morphological informality rather than tenure insecurity, service deprivation, or household poverty, and both false positives and false negatives are inevitable where morphology and lived conditions diverge[34,35]. Second, the optional binarisation step aligns city-scale morphology with national SDG 11.1.1 estimates, introducing ecological inference risk where individual cities differ substantially from national averages. Resulting binary maps should therefore be interpreted as indicative rather than definitive representations of informal settlement extent. Third, street-level imagery used for plausibility checks remains incomplete and uneven across Africa[36], limiting its representativeness for continent-wide validation.

Taken together, these results do not constitute a definitive census of informal settlements. Rather, they demonstrate how much of their spatial invisibility can be reduced using new open datasets and transferable computational methods. The probability surfaces are explicitly framed as exploratory, updateable layers that invite local validation, co-development with decision-makers, and integration with surveys and community-based mapping. The feasibility of this transition is significantly enhanced by established frameworks like the IDEAMAPS Data Ecosystem, which provide a ready-made platform for hosting, validating, and sharing these results. To operationalize this, would take the integration of these models into existing pipelines, alongside targeted local capacity building to ensure stakeholders can interpret and refresh the data independently. As geospatial data coverage improves and collaborative validation expands, such approaches offer a scalable way to support more inclusive urban governance, research, advocacy, and resource allocation, while recognizing that informality is multidimensional and ultimately context-specific.

Methods

Data collection and processing

Informal settlement polygons

Training data were assembled from the Know Your City program[21], which provides delineations of informal settlement polygons for eight African cities: Dar es-Salaam (Tanzania), Port Harcourt (Nigeria), Lagos (Nigeria), Nairobi (Kenya), Accra (Ghana), Freetown (Sierra Leone), Kampala (Uganda), and Cape Town (South Africa). Because some mapped areas pre-date recent redevelopment, we reviewed and updated these boundaries using the most recent cloud-free high-resolution Google Earth imagery available for 2023, allowing recent upgrading and infill growth to be incorporated. We then manually delineated formal settlement areas in each city to provide a balanced training dataset. These were identified from imagery as neighborhoods with more regular street layouts, durable building materials and visible infrastructure, and were selected outside the mapped informal areas. All polygons were digitized in QGIS and are shared for reproducibility. To improve consistency, each polygon was independently reviewed by at least one co-author other than the original digitizer, and disagreements were resolved through discussion. Per-city maps of the final training data are provided in Supplementary Information Section A.

Urban boundaries

Urban boundaries were defined following the UN-endorsed Degree of Urbanisation (DEGURBA) method[24], which provides a harmonised delineation of urban areas worldwide. The data were collected from the Global Human Settlement - Urban Centre Database (GHS-UCDB 2024A) released by the European Commission's Joint Research Centre[37]. From these boundaries we selected all African cities with projected populations exceeding 1 million in 2025, resulting in a sample of 84 cities across the continent. For each city, we assembled a set of descriptor variables characterizing its demographic, morphological, and development context. These descriptors, including total population, urban extent, building height, built-up area, population growth rate, building-stock volume, and indices of development and service readiness, were used to quantify inter-city similarity in our ensemble modelling approach (see Supplementary Information Section B). Full variable definitions and data sources are provided in Supplementary Table 1 for reproducibility.

Covariates

Built-environment characteristics were derived from the Google Open Buildings 2.5D Temporal Dataset[30] for the year 2023, which provides building footprints and estimated building heights based on Sentinel-2 imagery. These data were combined with satellite-based environmental indicators, including the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Built-up Index (NDBI), calculated from Sentinel-2 surface reflectance imagery. NDVI and NDBI are widely used proxies for vegetation cover and built-up surfaces in urban remote sensing. To characterize the physical structure of settlements, we summarized these indicators within neighborhood-scale areas of approximately 130 m around each location. This scale captures the block-level spatial structure of urban neighborhoods while

remaining compatible with the spatial resolution used in global population datasets such as GHSL and WorldPop. Within each neighborhood we derived a set of indicators:

- **Building presence**: Building coverage within each neighborhood was grouped into 10% intervals (0–10%, 10–20%, …, 90–100%). The number of pixels within each interval was then counted to describe the local distribution of built-up structures, capturing differences between sparsely built areas and dense building clusters.
- **Building height**: We summarized building height within each neighborhood by calculating the average height and the degree of variation in height between buildings. We also derived the coefficient of variation (CV) to capture the relative heterogeneity of building heights within the neighborhood.
- **NDVI and NDBI**: For both indices, we calculated the average value within each neighborhood together with measures of spatial variability. The coefficient of variation was again used to describe how vegetation cover and built-up surfaces vary across the neighborhood.

In summary, we quantified the distribution of building presence within neighborhoods, the mean and variability of building height, and the average and spatial variability of NDVI and NDBI. These measures capture characteristics commonly associated with morphologically informal development, including dense clusters of small buildings, irregular building height patterns and limited vegetation. In total, 16 covariates were derived and used as inputs to the classification models.

Morphological informal settlement mapping

City-specific random forest models

For each of the eight cities with labelled training data, we trained a separate Random Forest (RF) model using the neighborhood-scale indicators described above. Random Forest was chosen because it can capture complex and non-linear relationships among Earth observation variables while remaining robust to correlated predictors. Each model operates at the pixel level, using the set of 16 covariates derived from building and environmental characteristics. Training labels were created by converting the settlement polygons into a 10 m grid aligned with the predictor layers. Pixels intersecting the labelled informal or formal polygons were assigned the corresponding class. During prediction, pixels with no detected building presence were excluded, and the model produced a probability surface describing the likelihood of morphological informality at 10 m resolution. The feature calculation and model development were conducted in R 4.4.1.

Similarity-weighted ensemble predictions

For cities without training labels, we predicted pixel-level informal-settlement probability with an ensemble of the eight city-specific RF regressor models. We first used city-level descriptor vectors to measure how similar each unlabeled city is to each of the eight labelled cities and converted those similarities into weights that sum to one. More details can be found in Supplementary Information

Section B. We then formed the final map by taking a weighted average of the eight per-pixel probability surfaces at 10 m, which were produced by the city-specific RF regressor models with the target city predictor stack as input, so that predictions for each city draw more heavily from models trained on source cities with similar socio-economic and demographic contexts. For the eight labelled cities, we used the same ensemble scheme to increase robustness: for each city, we excluded its own model and combined the other seven models using the similarity-derived weights. This mitigates against overfitting to local labels and yields probability maps that are methodologically consistent across all 84 cities. While informal settlement forms vary across Africa, a single pooled model implicitly assumes one morphology of informal settlement and, thus, can transfer poorly. The similarity-weighted ensemble addresses inter-city heterogeneity by prioritizing models from cities with similar urban characteristics, which reduces misclassification errors that often arise when models are applied to very different contexts. Additionally, this method is data-efficient since it requires no local labels in the target city. It is also transparent where descriptors are interpretable, weights are explicit, and adding a new labelled city simply adds another expert to the ensemble without full retraining.

Cross-validation of similarity-weighted ensemble modelling

To evaluate the robustness of the similarity-weighted ensemble approach, we implemented a leave-one-city-out cross-validation across the eight cities with training data. In each iteration, one city was held out as the target, while the remaining seven city-specific random forest models were used to generate predictions. Ensemble weights were computed from the inter-city similarity matrix, excluding the target city itself, and applied to combine the seven probability surfaces into a single ensemble prediction for the held-out city.

Predicted probability maps were compared against our manually labelled formal settlement and informal settlement polygons for the held-out city. Performance metrics included the area under the receiver operating characteristic (ROC) curve (AUC) and the F1-score derived from threshold-optimized classifications. This procedure ensured that evaluation was based entirely on transfer learning from other cities, reflecting the conditions under which the ensemble would be deployed in cities without local training data. Results confirmed that the similarity-weighted ensemble consistently produced high AUC and F1-scores across diverse settlement morphologies, demonstrating the robustness of the methodology.

Estimating threshold for binary classification

A binary map of informal settlement allows for direct comparison with official figures and enables clear, actionable delineation of informal areas for planning and intervention. To convert continuous informal probability surfaces into binary classifications, we developed a country-level threshold estimation framework that aligns gridded probability maps with reported national statistics on the share of urban populations living in informal settlements.

For each country, we extracted a constrained gridded population surface at 100 m resolution from the WorldPop dataset (2023 release) and aggregated the 10 m informal-settlement probability maps to the same grid to ensure spatial alignment with population counts. We then evaluated candidate probability thresholds ranging from 0 to 1 in increments of 0.01. For each threshold, pixels with probabilities equal to or exceeding the threshold were classified as informal, and the predicted informal settlement population share was computed as the proportion of the national urban population residing in those pixels. The optimal threshold was selected as the value that minimized the absolute difference between this predicted share and the officially reported national urban informal settlement population share, as reported by UN-Habitat.

To refine these baseline estimates, we incorporated a model-based calibration step using cities with independent reference data. For the eight cities with training data, city-specific calibrated thresholds (Supplementary Table 2) were derived by selecting the probability cut-off that maximized agreement between the resulting binary classification and manually delineated informal settlement polygons. This step provides an empirical benchmark for how probability thresholds relate to observed settlement morphology, independent of gridded population redistribution.

We then fitted a nonlinear logistic regression model that relates these empirically calibrated city-level thresholds to their corresponding baseline country-level thresholds, the reported national urban informal settlement population shares, the population shares estimated from gridded data, and the ratio between reported and estimated shares. This modelling step accounts explicitly for the systematic under-allocation of population to informal areas in gridded datasets identified above. Finally, we applied the fitted model to predict updated thresholds for all countries in the dataset. These adjusted thresholds were used in subsequent analyses to generate binary settlement classifications, ensuring consistency with both local morphological patterns and national-scale population statistics.

Validation

We validated the probabilistic maps of informal settlement presence using two independent reference datasets.

Informal and formal settlement reference points

To assess classification accuracy, we assembled an independent validation dataset of georeferenced informal and formal settlement points across four countries: Ghana, Nigeria, Kenya, and Malawi. These points were obtained from the UN-Habitat's Informal Settlements Vulnerability Mapping Program and supplemented by local expert annotation and visual interpretation of fine-resolution satellite imagery and Google Street View data[38]. Although these reference points provide high-quality ground reference data, they are available for only a limited number of countries and are spatially sparse, making them unsuitable for training city-scale models intended to generalize across diverse urban contexts. We, therefore, reserved them exclusively for validation to ensure an unbiased assessment of model performance. The final validation dataset comprised 1,910 points within the study area in total, including 1016 labelled as informal and 894 as formal settlement. Each point was intersected with the 10 m

informal-settlement probability surface corresponding to its city, and the extracted probabilities were compared with the ground-reference labels. Model performance was assessed by examining the degree of probability separation between formal and informal classes using Welch's two-sample t-test; and the binary classification performance was assessed using standard metrics, including overall accuracy, precision, recall, F1-score, and class-specific misclassification rates, under both original and updated thresholds.

OpenAI-based settlement classification

To further evaluate external validity, we conducted a second validation exercise using automated settlement classification powered by the OpenAI GPT-4o multimodal model[39]. For each of the 84 cities in the study, we randomly sampled 20 points, stratified by predicted probability: 10 points from the [0, 0.5) range and 10 points from the [0.5, 1] range. This produced a total of 1,680 locations across the continent. For each point, Street View imagery (where available) or fine-resolution satellite images collected from Google Maps in 2025 were analyzed, and the model assigned a categorical settlement status (more details can be found in Supplementary Section E). To ensure consistency, we excluded unknown and no buildings outputs and reclassified the remaining categories into five groups: *formal*, *probable formal*, *mixed*, *probable informal*, and *informal*. We then examined the distribution of informal settlement probabilities for each pair of adjacent categories using Welch's t-test.

Declarations

Ethics approval

This research utilized only publicly available, secondary geospatial datasets and remote sensing imagery. No human participants, personally identifiable information, or sensitive personal data were involved in any part of the study. All procedures complied with institutional and international ethical guidelines for research involving secondary data. Therefore, ethics approval was not required.

Acknowledgements

This work was supported by funding from the Gates Foundation (INV-088965), GAVI (MEL11779722) and Wellcome Trust (308679/Z/23/Z).

Data availability

Urban boundaries and city descriptor variables were obtained from the Global Human Settlement - Urban Centre Database (GHS-UCDB 2024A), available at https://human-settlement.emergency.copernicus.eu/download.php?ds=ucdb. Built-environment covariates were derived from the Google Open Buildings 2.5D Temporal Dataset (https://sites.research.google/gr/open-buildings/temporal/). Environmental indices (NDVI and NDBI) were calculated from Sentinel-2 Level 2A

surface reflectance data, which can be accessed via Google Earth Engine at https://developers.google.com/earth-engine/datasets/catalog/sentinel-2/. WorldPop Global 2 population data (2025 release) were used, available at https://www.worldpop.org. National reference statistics on the percentage of urban populations living in slums were compiled from UN-Habitat at https://data.unhabitat.org/pages/housing-slums-and-informal-settlements. The manual labelling informal and formal settlement polygons are available via GitHub https://github.com/WB-zhang94/Morphological-informality-in-Africa. Google Maps and Google Street View imagery cannot be redistributed due to copyright restrictions, but were available to access using the Google Maps API.

Code availability

All codes are publicly available via GitHub at https://github.com/WB-zhang94/Morphological-informality-in-Africa.

References

1. Housing S, Settlements I https://data.unhabitat.org/pages/housing-slums-and-informal-settlements
2. Lilford RJ et al (2017) Improving the health and welfare of people who live in slums. Lancet 389:559–570
3. Lilford R et al (2019) Because space matters: conceptual framework to help distinguish slum from non-slum urban areas. BMJ Glob Health 4:e001267
4. Mitlin D, Satterthwaite D (2013) Urban Poverty in the Global South: Scale and Nature. Routledge
5. Carr-Hill R (2013) Missing Millions and Measuring Development Progress. World Dev 46:30–44
6. Thomson DR et al (2020) Need for an Integrated Deprived Area Slum Mapping System (IDEAMAPS) in Low- and Middle-Income Countries (LMICs). Soc Sci 9:80
7. Dovey K, Shafique T, Van Oostrum M, Chatterjee I (2021) Informal settlement is not a euphemism for 'slum': what's at stake beyond the language? Int Dev Plan Rev 43:139–150
8. SDG Indicators — SDG Indicators https://unstats.un.org/sdgs/metadata
9. UN-Habitat. SDG Indicator 11.1.1 Training Module: Adequate Housing and Slum Upgrading (2018)
10. Gulyani S, Talukdar D (2008) Slum Real Estate: The Low-Quality High-Price Puzzle in Nairobi's Slum Rental Market and its Implications for Theory and Practice. World Dev 36:1916–1937
11. Fan J, Heiskanen J, Vuolteenaho J, Sa H (2026) Game-changers in global slum detection? A review of remote sensing for assessing complexity of informal settlements. Int J Remote Sens 0:1–23
12. Yang H et al (2026) A transferable and interpretable approach to slum mapping using building morphometrics and optical imagery. GIScience Remote Sens 63:2649308
13. Chamunorwa B, Shoko M, Magidi J (2025) Low-cost and scalable detection of sparse informal settlements using machine learning in Gcuwa, Eastern Cape, South Africa. Afr Geogr Rev 44:287–303

14. Wurm M, Taubenböck H, Weigand M, Schmitt A (2017) Slum mapping in polarimetric SAR data using spatial features. Remote Sens Environ 194:190–204
15. Mahabir R, Croitoru A, Crooks AT, Agouris P, Stefanidis AA (2018) Critical Review of High and Very High-Resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities. Urban Sci 2:8
16. Verma D, Jana A, Ramamritham K (2019) Transfer learning approach to map urban slums using high and medium resolution satellite imagery. Habitat Int 88:101981
17. Li C et al (2025) Mapping urban slums and their inequality in sub-Saharan Africa. Nat Cities 2:1037–1048
18. Büttner N, Stalder S, Volpi M, Suel E, Harttgen K (2025) Large-scale slum mapping in sub-Saharan Africa's major cities: Remote sensing and deep learning reveal strong slum growth in the urban periphery between 2016 and 2022. Habitat Int 161:103403
19. Friesen J (2023) Towards a Link between Quantitative and Qualitative Sciences to Understand Social Systems Using the Example of Informal Settlements. Entropy 25:262
20. Samper J, Shelby JA, Behary D (2020) The Paradox of Informal Settlements Revealed in an ATLAS of Informality: Findings from Mapping Growth in the Most Common Yet Unmapped Forms of Urbanization. Sustainability 12:9510
21. Beukes A Know Your City: community profiling of informal settlements. https://www.iied.org/17244iied
22. Tereke BW et al (2025) IDEABench: Benchmark Dataset for Mapping Deprived Urban Areas. DANS Data Station Physical and Technical Sciences https://doi.org/10.17026/PT/X4NJII
23. Boanada-Fuchs A, Kuffer M, Samper J (2024) A Global Estimate of the Size and Location of Informal Settlements. Urban Sci 8:18
24. Dijkstra L et al (2021) Applying the Degree of Urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation. J Urban Econ 125:103312
25. Understanding (2025) improving informal settlements in African cities. Nat Cities 2:1001–1002
26. Kuffer M et al (2020) The Role of Earth Observation in an Integrated Deprived Area Mapping System for Low-to-Middle Income Countries. Remote Sens 12:982
27. Bettencourt LMA, Marchio N (2025) Infrastructure deficits and informal settlements in sub-Saharan Africa. Nature 645:399–406
28. Breuer JHP, Friesen J, Taubenböck H, Wurm M, Pelz PF (2024) The unseen population: Do we underestimate slum dwellers in cities of the Global South? Habitat Int 148:103056
29. Thomson DR et al (2021) Evaluating the Accuracy of Gridded Population Estimates in Slums: A Case Study in Nigeria and Kenya. Urban Sci 5:48
30. Sirko W et al (2023) High-Resolution Building and Road Detection from Sentinel-2. *arXiv.org* https://arxiv.org/abs/2310.11622v3

31. Gevaert CM, Persello C, Sliuzas R, Vosselman G (2017) Informal settlement classification using point-cloud and image-based features from UAV data. ISPRS J Photogramm Remote Sens 125:225–236
32. Hafner S et al (2025) Towards an Indicator-Based Morphological Informality Model for Sub-Saharan Africa Using Open Building Footprint and Road Data (Version 1). Int Arch Photogramm Remote Sens Spat Inf Sci XLVIII–M–7–2025:221–228
33. Veeravalli SG, Haas J, Friesen J, Georganos S (2025) Understanding Informal Settlement Transformation through Google's 2.5D Dataset andStreet View based Validation. in vol. XLVIII-M-7–2025. Copernicus, pp 245–251
34. Gulyani S, Talukdar D, Jack D, Poverty (2010) Living Conditions, and Infrastructure Access: A Comparison of Slums in Dakar, Johannesburg, and Nairobi. SSRN Scholarly Paper at https://papers.ssrn.com/abstract=1650479
35. Friesen J et al (2025) Comparative Analysis of Manual Slum Identification Using Satellite Data: A Case Study of Medellin, Colombia. in *2025 Joint Urban Remote Sensing Event (JURSE)* vols CFP25RSD-ART 1–4
36. Warren JL, Perez-Heydrich C, Burgert CR, Emch ME (2016) Influence of Demographic and Health Survey Point Displacements on Distance-Based Analyses. Spat Demogr 4:155–173
37. Kemper T, GHS-UCDB R, A - GHS Urban Centre Database (2024) 2025. European Commission, Joint Research Centre (JRC) https://doi.org/10.2905/1A338BE6-7EAF-480C-9664-3A8ADE88CBCD (2024)
38. Li C et al (2023) Slum and urban deprivation in compacted and peri-urban neighborhoods in sub-Saharan Africa. Sustain Cities Soc 99:104863
39. Hello GPT-4o *OpenAI* https://openai.com/index/hello-gpt-4o/

Figures

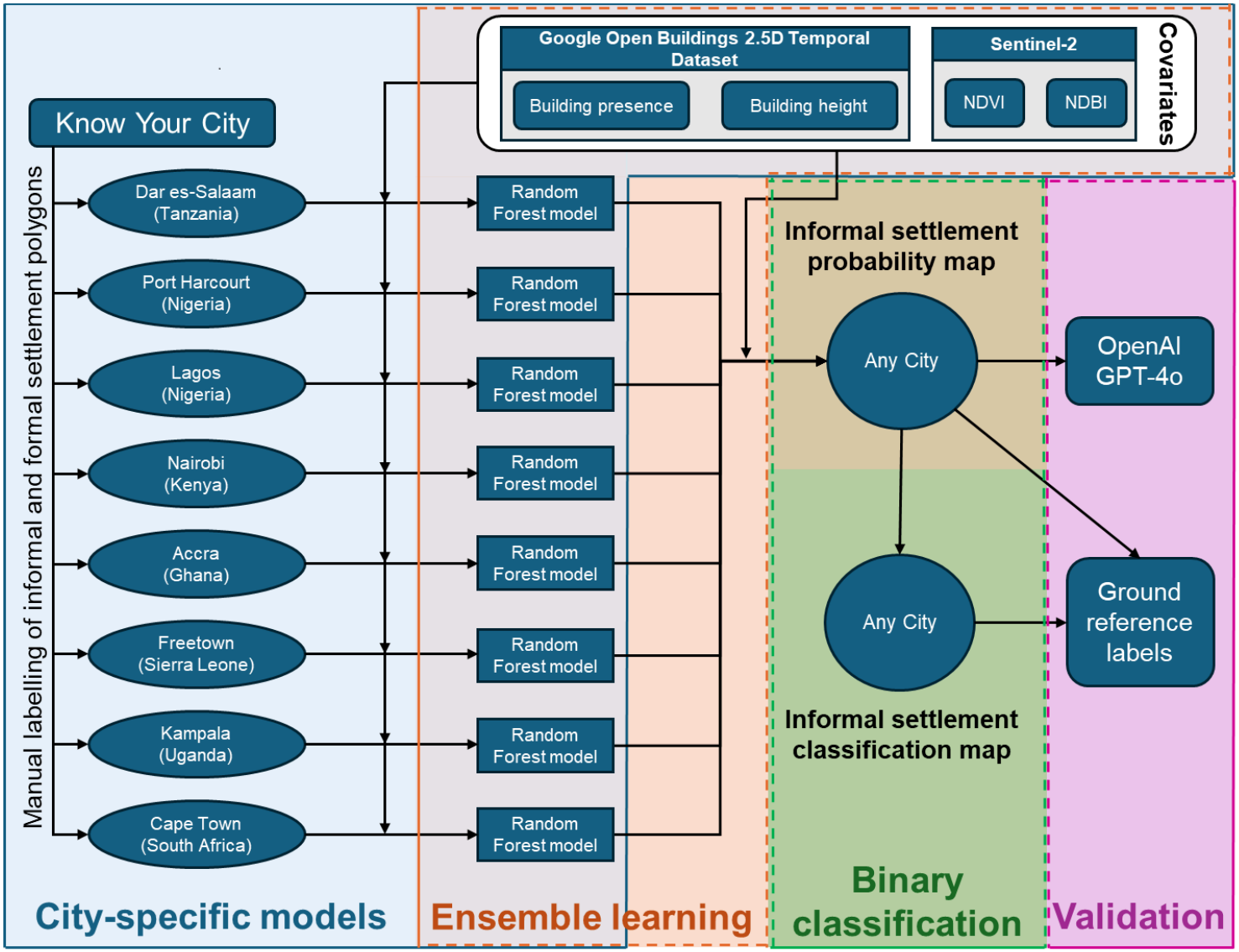


Figure 1

Overview of the multi-step workflow for mapping morphologically informal settlements in African cities. Training data are compiled and refined from manually labelled informal and formal settlement polygons. Built environment and environmental covariates are extracted within urban boundaries defined by the GHSL Urban Centre Database. City-specific random forest models are trained where data are available, and a similarity-weighted ensemble approach transfers model predictions to unlabeled cities, generating continent-scale city-level probabilistic informality maps. Classification thresholds are then calibrated to align predicted informal population shares with national statistics, producing interpretable binary maps. Final outputs are validated against both independent survey datasets and automated AI-based settlement classification.

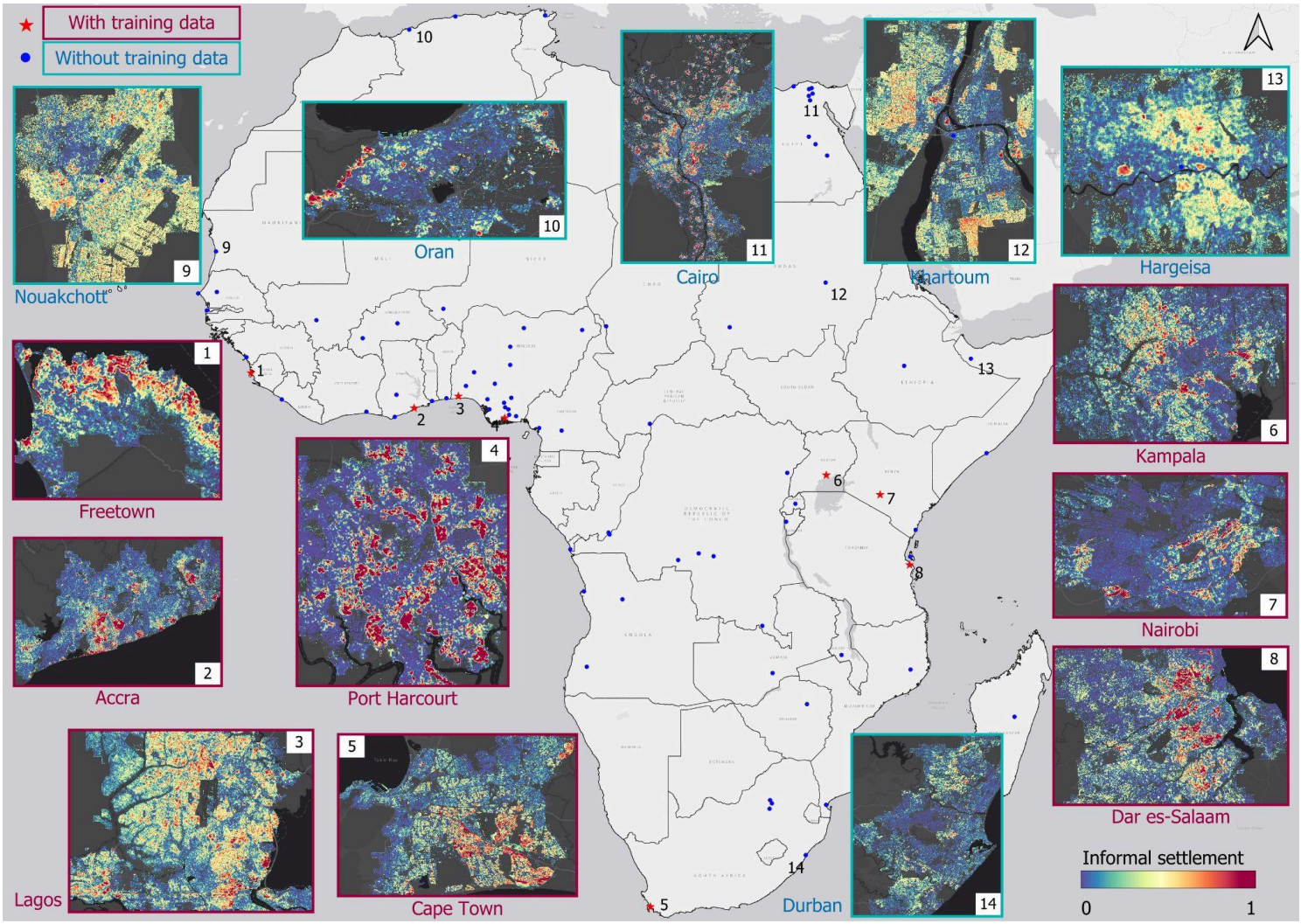


Figure 2

Spatial distribution of morphological informality probabilities across African cities in 2023. The morphologically informal settlement probabilities are shown at 10 m resolution for 84 African cities which have populations over one million in 2025. Red stars mark cities with training data and blue points show other study cities without training data. Insects illustrate examples from different regions, including all eight cities with training data and six cities without training data, highlighting spatial heterogeneity in settlement morphology and probability structure.

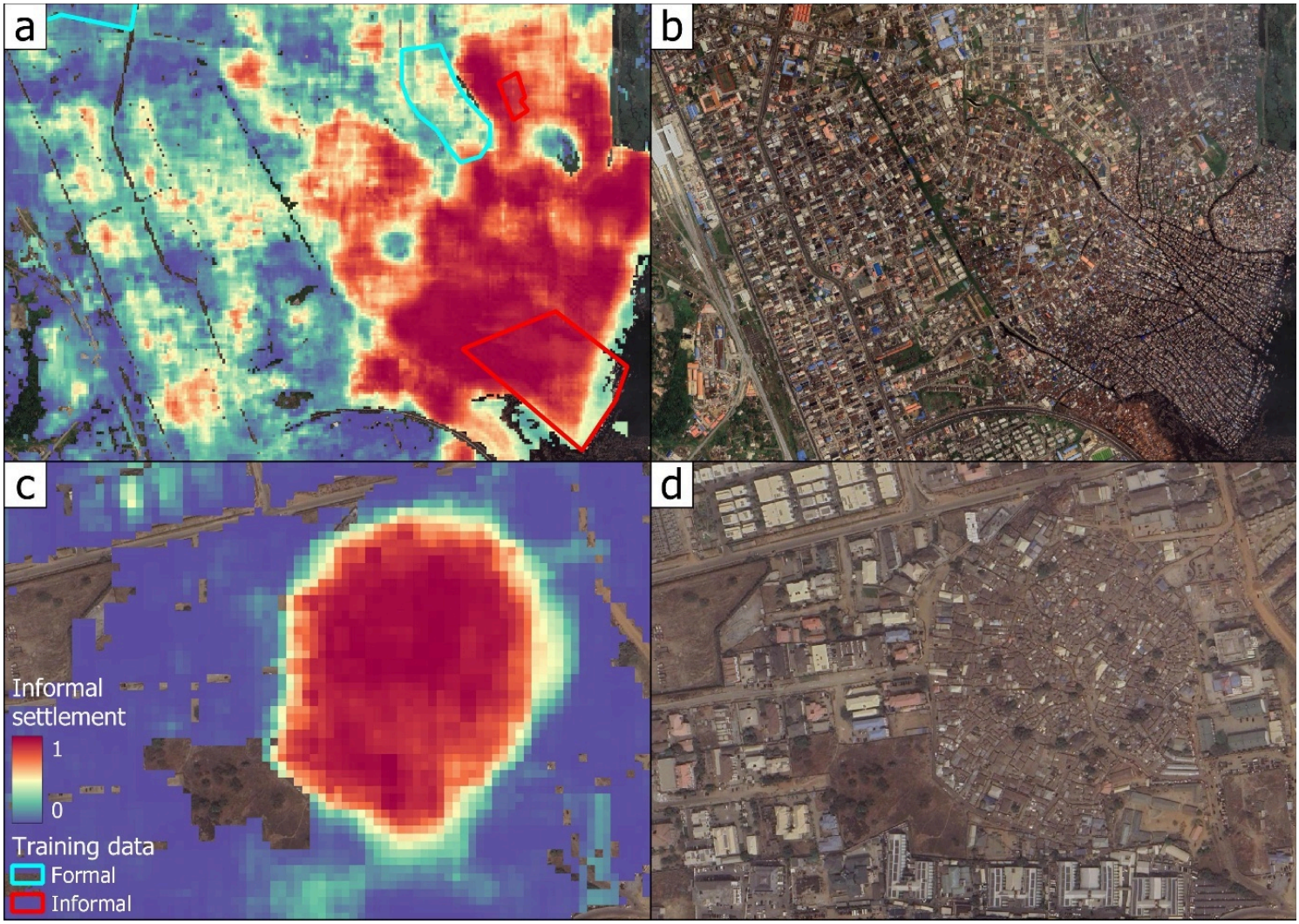


Figure 3

Comparison of morphologically informal settlement patterns in Lagos and Abuja, Nigeria. a. Probability map of morphologically informal settlements in Makoko, which is an informal settlement across the 3rd Mainland Bridge located on the coast of mainland Lagos, Nigeria. b. Corresponding satellite imagery for 2025. c. Probability map of a smaller cluster of morphologically informal development at the north side of the Federal Ministry of Housing and Urban Development, Mabushi, Abuja, Federal Capital Territory, Nigeria. d. Corresponding satellite imagery for 2025. Training data were available for Lagos but not for Abuja, illustrating model behavior under both trained and transferred conditions.

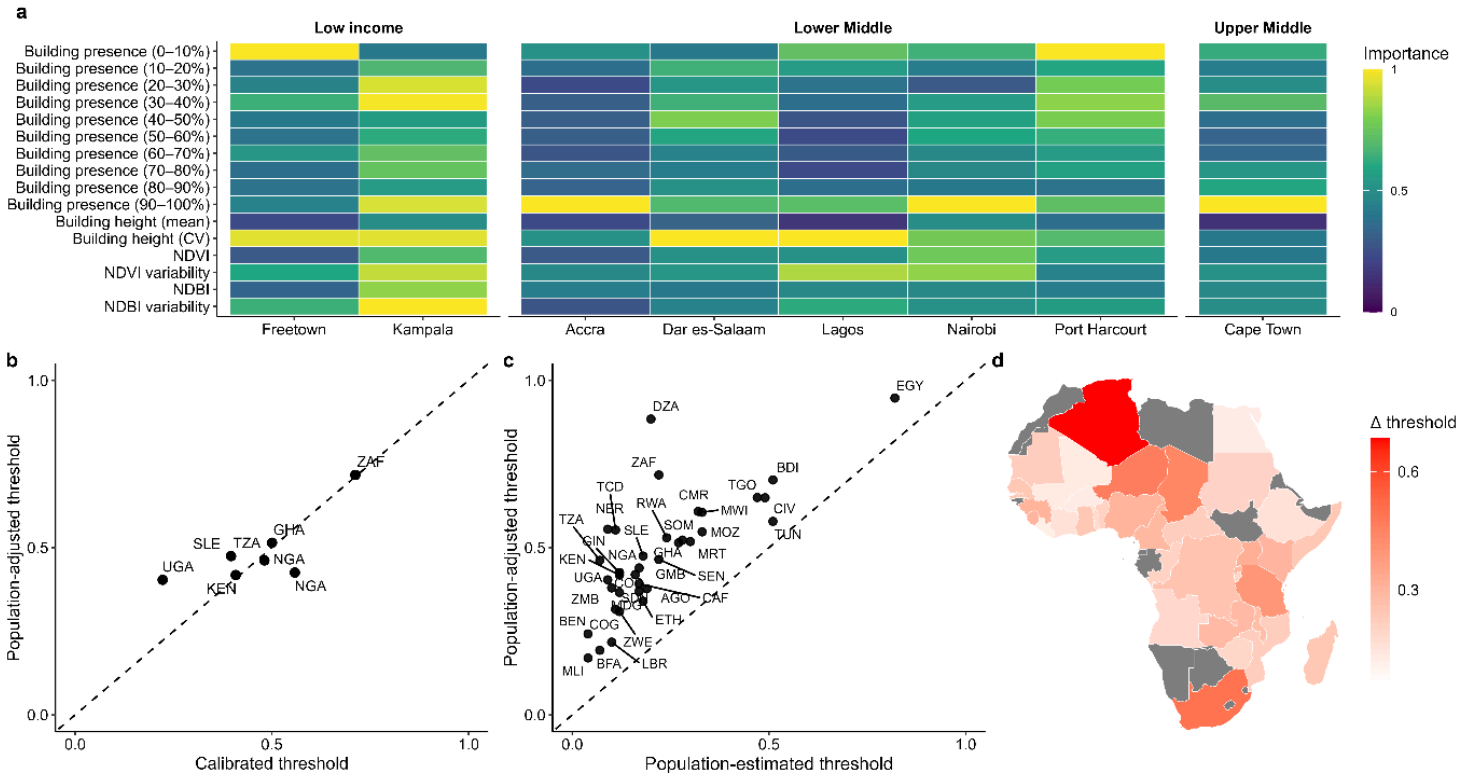


Figure 4

Spatial heterogeneity in signals of morphological informality and calibration of probability thresholds. a, Normalized variable importance from city-specific random forest models, grouped by national income category, showing how different built-environment features contribute to identifying morphological informality across urban contexts. Importance values are scaled within each city. b, Relationship between calibrated thresholds derived from leave-one-out cross-validation and population-adjusted thresholds, illustrating consistency between model-based calibration and population-informed adjustment across countries with training data. c, Comparison between population-estimated thresholds (derived by aligning high-probability pixels with reported national urban informal population shares) and population-adjusted thresholds after model-based recalibration, highlighting substantial cross-country variation and the absence of a single universal cutoff. d, Spatial distribution of threshold adjustments (Δ threshold = population-adjusted − population-estimated) at the national level, indicating where stronger recalibration is required to reconcile probability surfaces with aggregate population statistics.

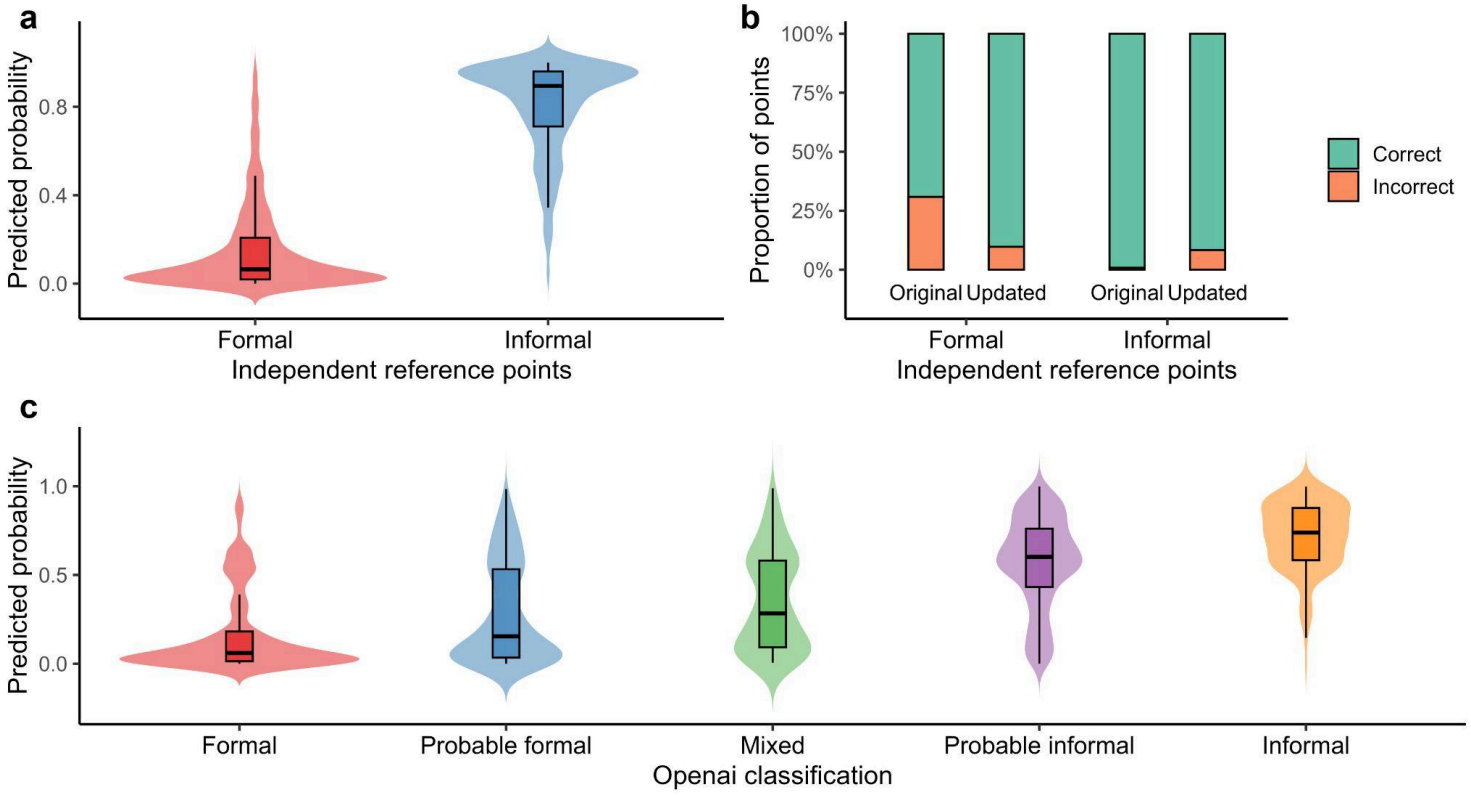


Figure 5

Validation of probability predictions of morphological informality. a. Distribution of predicted probabilities for formal and informal validation points, showing clear separation. b. Classification performance under the original and updated thresholds. Updated thresholds substantially reduce false positives in formal settlements, while maintaining high accuracy in informal settlements. c. Validation against AI-assisted street view or remote sensing image classifications from OpenAI. Predicted probabilities align with the continuum of categories from formal to informal, with intermediate probabilities for probable formal, mixed, and probable informal.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.docx