USC Dornsife Center for Economic and Social Research
USC Schaeffer Leonard D. Schaeffer Center for Health Policy & Economics

Validating machine learning–derived built environment measures from Google Street View for urban aging research in India

Samer Atshan, Min Namgung, Janghyeon Lee, Anushikha Dhankhar, Pranali Khobragade, Aidan Cole, Jennifer Ailshire, Sara Adar, Yao-Yi Chiang, Jinkook Lee, Emma Nichols

Paper No: 2025-009

CESR-SCHAEFFER WORKING PAPER SERIES

Validating machine learning–derived built environment measures from Google Street View for urban aging research in India

Samer Atshan*[1], Min Namgung[2], Janghyeon Lee[2], Anushikha Dhankhar[1], Pranali Khobragade[1], Aidan Cole[1], Jennifer Ailshire[1], Sara Adar[3], Yao-Yi Chiang[2], Jinkook Lee[1], Emma Nichols[1]

[1] Center for Economic and Social Research, University of Southern California, Los Angeles, Los Angeles, California, United States of America

[2] Department of Computer Science & Engineering, University of Minnesota, Minneapolis, Minnesota, United States of America

[3] Leonard Davis School of Gerontology, University of Southern California, Los Angeles, California, United States of America

[4] Epidemiology and Global Public Health, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America

* Corresponding author
E-mail: atshan@usc.edu (SA)

Abstract

Environmental factors play a critical role in healthy aging, yet data sources capturing these exposures remain limited, particularly in low- and middle-income countries (LMICs). To address this gap, we developed a protocol to leverage Google Street View (GSV) images and machine learning (ML) methods to capture features of the physical and built environment (e.g., vehicle congestion, greenspace) in four major Indian cities. We first fine-tuned computer vision models on the Indian Driving Dataset (IDD), where our object detection and segmentation results outperformed state-of-the-art models found in the literature in identifying common street-level features. We then validated ML-derived features from GSV by comparing predictions to human-coded image audits, field-based interviewer assessments, and existing indicators derived from satellites. ML predictions showed strong performance for features such as vehicles, pedestrians, roads, and greenery, and comparatively lower performance for context-specific features such as autorickshaws, sidewalks, and traffic signs. Results underscore limitations of pairing GSV with existing machine learning models given the static and incomplete nature of GSV data and the cultural constraints of applying pre-trained models developed in high-income contexts across cultural settings. Nonetheless, our findings also showcase the reliability of our implemented approach in detecting key features of the urban environment in India.

1. Introduction

The global demographic shift toward older populations is reshaping priorities in public health and aging research [1–3]. Despite advances in medicine, promoting healthy aging increasingly requires attention to non-clinical social, environmental, and behavioral factors [4–6]. One such factor is the physical and built environment in which our societies age. The neighborhoods we live in determine access to greenspace, exposure to traffic and noise, as well as opportunities for physical activity and social interaction, thereby shaping our physical, mental, and cognitive health. Cities concentrate both risks and resources. While access to services and infrastructure may support aging in place, dense traffic, pollution, limited greenery, and inadequate pedestrian environments may accelerate physical and cognitive decline [6]. With the rapidly growing population of older adults in urban areas, there is a pressing need to understand and measure how environments support or impede healthy aging.

The fastest-growing populations of older adults are now found in low- and middle-income countries (LMICs), yet our understanding of how physical and built environment features influence healthy aging in these settings remains limited. A key barrier is the lack of spatial data infrastructure in LMICs. Much of what we know about the relationship between the physical and built environment and healthy aging comes from studies in high-income countries (HICs) whose urban environments differ substantially from those in LMICs. Recent advances in computer vision and the availability of Google Street View (GSV) imagery offer a scalable way to observe features of the physical and built environment. Machine learning models can be trained to identify elements such as vehicles, greenery, and infrastructure within their location-specific contexts using GSV

images. While this approach has been applied in HICs to study various urban exposures such as walkability, greenness [7], and traffic [8], applications in LMICs remain sparse and limited [9–12]. This raises concerns about the performance and validity of these methods in LMICs, where image coverage, visual complexity, and infrastructure design may differ markedly from the settings in which models were originally developed.

To address this gap, this study assessed the validity of using a machine learning model on GSV imagery to capture environmental features in an urban LMIC setting using India as an example. Indian cities provide a compelling case for this validation due to the rapidly urbanizing environment, variable infrastructure quality, and non-standard features that are poorly captured in existing global datasets from HICs. We first fine-tuned existing pretrained machine learning models on an Indian dataset to improve context-specific performance, then utilized a spatial prediction method to predict indicators when images are missing. We then triangulated ML-derived indicators against three complementary sources for validation: manual audits of images by human raters, neighborhood observations from a household aging survey, and secondary geospatial data. Our goal was to evaluate how well the combination of automated image analysis and spatial predictions reflects urban environmental conditions in a data-constrained context, and to highlight practical considerations in applying this method for aging and health research.

2. Literature Review

2.1. Built Environment and Aging

Urban environmental exposures relevant to healthy aging include features such as air quality, green and blue spaces, social dimensions of neighborhoods such as socioeconomic conditions and social cohesion, and human-made “built” infrastructure where we live, work, and age [6,13,14].

Characteristics of the built environment such as street and sidewalk layout, infrastructure quality, and land use can impact walkability, density, exposure to noise and pollution, and access to services and green spaces which may influence aging outcomes through both protective and risk pathways [15,16]. More recently, the concept of *cognability* has emerged to capture the degree to which a neighborhood supports cognitive health among aging residents [16]. This framework emphasizes the combined influence of natural, social, and built features in promoting or hindering healthy cognitive aging. Environments that offer opportunities for physical activity, social engagement, and cognitive stimulation within safe and accessible spaces may support cerebrovascular function and cognitive resilience through neuroprotective mechanisms [8,17].

Despite growing recognition of environmental influences on healthy aging outcomes in general, most studies have been conducted in high-income countries [17–22]. This limits our ability to understand how these exposures impact aging in LMICs and if they may operate differently. Research on the built environment in LMICs, and India in particular, remains sparse due to limited data [23]. Some studies have linked air pollution [24–26] and poor housing quality [27] to declines in physical and cognitive functioning among Indian older adults, highlighting environmental risks that may be relevant to the LMIC context. A few qualitative studies have also noted poor walkability, unsafe pedestrian infrastructure, and limited access to parks and communal spaces as key challenges, emphasizing how neighborhood design shapes opportunities for active and healthy aging [28–30]. Yet, there have been limited large-scale, systematic quantitative investigations of how built environment features in India affect aging outcomes, in part due to the lack of data infrastructure to capture such environmental exposures at scale. This study aimed to address that

gap by developing and validating machine learning-derived measures of the built environment using GSV imagery.

2.2. GSV and Machine Learning in Health and Urban Research:

Google Street View (GSV) has emerged as a scalable resource for assessing built environment characteristics relevant to health, mobility, and social conditions. Early studies in HICs demonstrated the feasibility of using GSV for virtual audits of neighborhood infrastructure, showing high inter-rater reliability for features like sidewalks, crosswalks, land use, and pedestrian safety [31–34]. These findings have since been extended with the use of deep learning and computer vision methods, allowing researchers to automatically extract features such as street greenery, visual complexity, as well as highway and sidewalk presence from GSV imagery [35–37]. GSV-derived indicators of walkability, greenery, and urbanicity have been associated with higher physical activity and lower mortality rates in studies in the U.S. and Hong Kong [7,37,38]. Similar approaches have been used to assess environmental predictors of active play among children in Canada [8], and to model neighborhood socioeconomic conditions across 12 cities in five high-income countries [39]. These applications highlight GSV's capacity to detect features with epidemiological and social relevance across diverse domains.

Although using GSV for environmental sensing has promise, most of the studies demonstrating its validity have been conducted in high-income countries. Two reviews of GSV-based research on urban physical environments found that nearly all applications occurred in cities with robust digital and transportation infrastructure (e.g., street networks that allow geographically dense GSV coverage almost everywhere in an urban environment) [40,41]. While these reviews highlight that GSV is well-suited for capturing static and visible features of the built environment such as

sidewalks, greenspace, and signage, only a few studies have explored similar approaches in LMIC contexts. Research in Brazil, for example, showed strong agreement between GSV-based audits and in-person assessments of neighborhood walkability [42]. In India, emerging work has used satellite imagery, Google Maps APIs, and platforms like Google Earth Engine to estimate traffic volumes, classify urban land use, and analyze infrastructure expansion across major cities [43,44]. These studies underscore the feasibility of using open-source and image-based methods and publicly available datasets to characterize built environments in data-scarce settings. As machine learning applications using GSV continue to expand, validation studies in LMICs are critical, where differences in urban form, infrastructure quality, and GSV coverage may affect the reliability and generalizability of these measures, as well as their validity as exposures in research on links with late-life health outcomes such as dementia.

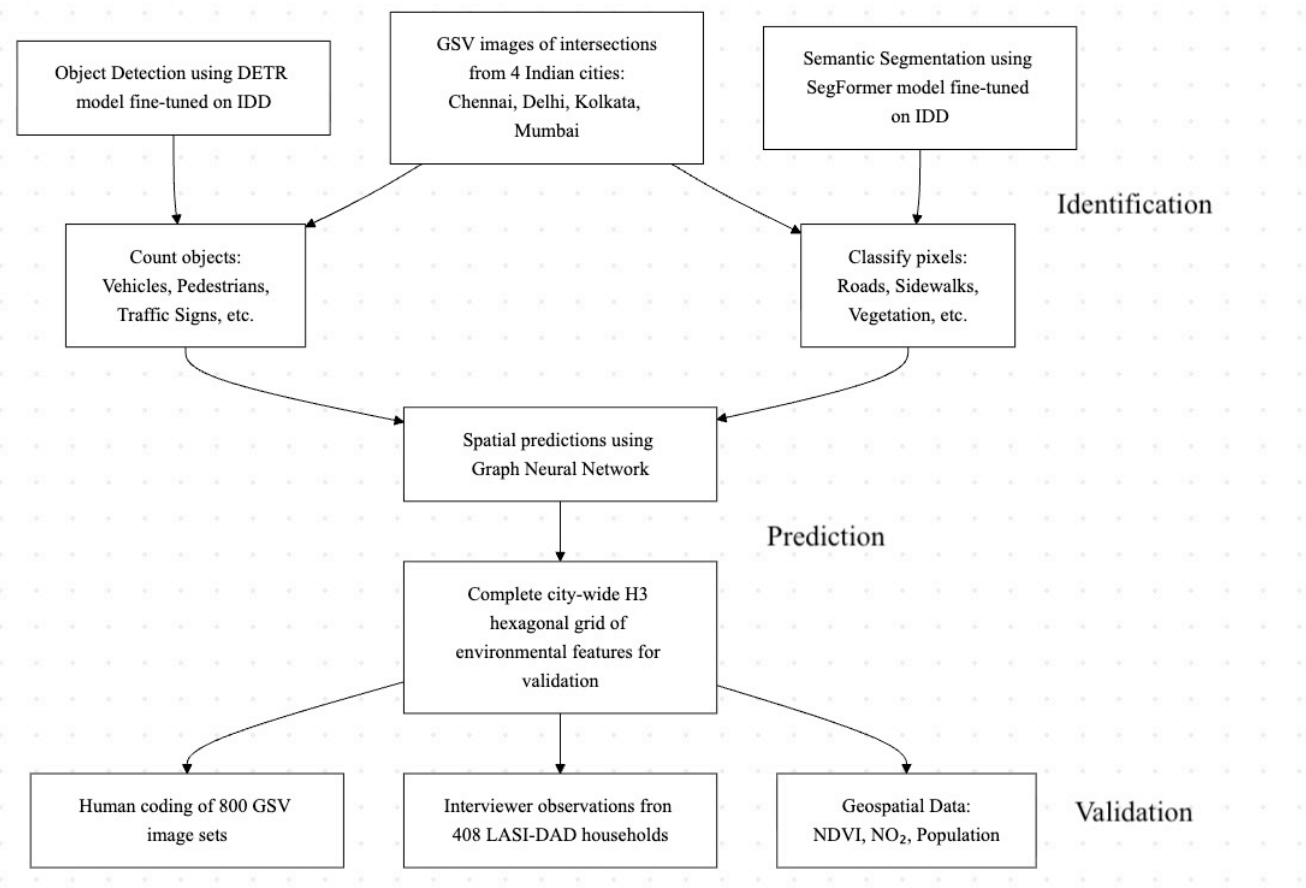
This study sought to fill this gap by assessing the validity of ML-derived indicators of urban environmental features from GSV images of street intersections in four Indian cities. By comparing ML-derived indicators against a range of other data sources, including human audits of the same GSV images, survey-based neighborhood observations, and geospatial data on air quality and greenspace, we seek to gain insights into the strengths and weaknesses of ML-derived GSV-based exposures for use in health research in India.

3. Data and Methods

Our methodological approach consisted of three main phases: identification, prediction, and validation (Figure 1). In the identification phase, we retrieved Google Street View (GSV) images from street intersections across four major Indian cities and applied machine learning models to extract environmental features. The prediction phase involved using spatial modeling techniques

to generate complete city-wide coverage of environmental indicators. Finally, the validation phase compared our machine learning-derived measures against three independent data sources to assess

Figure 1. Methodological workflow for extracting and validating built environment features from GSV images in Indian cities



their accuracy and reliability.

3.1. Machine Learning Feature Extraction

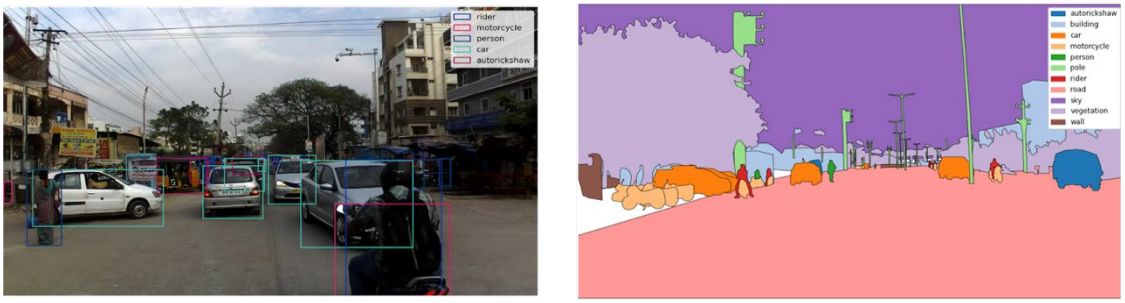
3.1.1. Image Retrieval and Fine-Tuning on Indian Driving Dataset

We retrieved GSV images from four major Indian cities (Chennai, Delhi, Kolkata, and Mumbai) by querying the Google Street View API at known street intersections within each city's boundaries [45]. We included intersections only since they offer broad unobstructed views of the streetscape while reducing the cost of querying additional mid-block images that provide limited incremental information. To ensure a comprehensive view of the built environment, we collected

four directional images at each location, corresponding to 0°, 90°, 180°, and 270° orientations. This approach improves coverage of the surrounding streetscape and increases the likelihood of detecting detailed and relevant physical features [36]. We then extracted environmental features from those images by fine-tuning two deep learning models using the Indian Driving Dataset (IDD). IDD contains annotated road-level images collected under diverse driving conditions in India, and captures unique elements such as vehicle types and various traffic conditions [46].

The first model we fine-tuned using IDD is a Detection Transformer (DETR) model that identifies and counts discrete objects, including persons, riders, cars, trucks, buses, autorickshaws, motorcycles, bicycles, traffic signs, and traffic lights [47]. We started with the detr-resnet-50 weights, which were originally trained on the Common Objects in Context (COCO) dataset. COCO is a benchmark containing over 200,000 images from real-world scenarios and over 80 categories of objects such as people, vehicles, animals, household items, and traffic signs [48]. Figure 2 (a) shows a typical output of object-level predictions on the IDD dataset in the form of bounding boxes. To reduce redundant and noisy predictions, we further conducted non-maximum suppression and small-object filtering (further details in Supplementary Materials S1 Fig. 1) [47,49].

Figure 2. Example annotations from the Indian Driving Dataset



(a) Object Detection Label (b) Semantic Segmentation Label

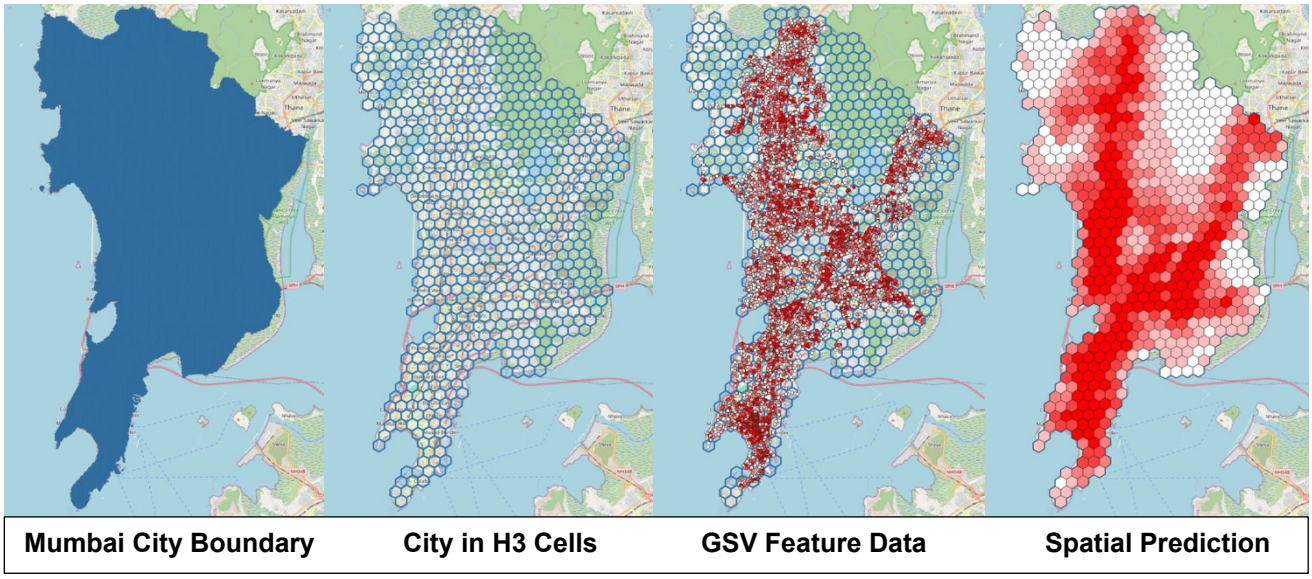
The second model we fine-tuned using IDD is a SegFormer model that performs semantic segmentation by classifying each pixel in the image into predefined categories such as road, sidewalk, vegetation, and pole [50]. We started with the segformer-b1-finetuned-cityscapes-1024-1024 weights, which were originally trained on the Cityscapes dataset. The Cityscapes dataset contains high-resolution street-view images from 50 Western European cities focusing on urban scenes [51]. Figure 2 (b) shows typical output of semantic segmentation on the IDD dataset where each pixel is classified into one category, such as roads, vegetation, and buildings. By fine-tuning both these models on the IDD dataset, we seek to improve comprehensive extraction of built environment features from GSV images from India, capturing both countable elements and broader spatial characteristics unique to the Indian context. The output of this process is feature predictions for object detection and segmentation at each street intersection.

3.1.2. Spatial Predictions

GSV coverage is sparse and geographically unevenly distributed within Indian cities. To address this limitation, we developed a spatial prediction model to predict built environment features in areas lacking images and to smooth model output. As shown in Figure 3, we divided each city into uniform spatial units using Uber's H3 spatial indexing system at resolution level 8, where each hexagonal cell has an average edge length of 500 meters. In cells where GSV images are available, we extracted features using the machine learning models described in Section 3.1.1 and computed mean values per H3 cell. In other cells, we predicted the same set of features using a Graph Neural Network (GNN) [52]. The GNN models each H3 cell as a node and connects it to neighboring cells using bidirectional edges based on geographic adjacency, enabling the model to capture spatial dependencies and local variation. To inform predictions, we used geospatial embeddings generated by Cross-Modal Knowledge Injected Embedding (CooKIE) [53], a region

representation learning model that integrates PLANET satellite imagery [54], Points of Interest (POIs), and Areas of Interest (AOIs) from OpenStreetMap [55] to encode spatial characteristics of each cell, including physical features from satellite imagery and functional attributes from POIs and AOIs. By combining CooKIE embeddings with a GNN, we predicted built environment features in areas without GSV imagery using spatial patterns learned from neighboring cells and created a smoothed output for each environmental feature. The result is a complete, city-wide dataset of built environment features at the H3-cell level, combining observed values from GSV imagery where available and GNN-predicted values elsewhere. Further details about the implementation of the spatial prediction model are presented in Supplementary Materials S1 Section 1.2.

Figure 3. Spatial prediction Workflow



3.2. Human-coding of GSV Images

To validate the performance of the ML tool in extracting and predicting environmental features, we conducted manual coding of 800 GSV image sets, where an image set is the collection of four directional images (0°, 90°, 180°, and 270°) captured at a single location. We sampled 200

image sets per city across the four cities. For each city, we selected intersection locations using a two-step stratified random sampling strategy. First, we randomly selected administrative wards within each city. Then, within each selected ward, we randomly identified a street intersection from which to retrieve a GSV image set. In cities with fewer than 200 wards, some wards were selected more than once, with a different intersection sampled each time to ensure spatial diversity. This approach yielded a balanced and geographically varied set of images for human audit.

Each image set was evaluated by two trained human raters familiar with the urban context in India (details on training process in Supplementary Materials S2 Section 2.1) using an instrument designed to assess objective and subjective features of the built environment relevant to health (full instrument is in Supplementary Materials S2 Section 2.2). The instrument captured information on vehicular and pedestrian traffic, sidewalk and street conditions, crosswalks, greenery, street lighting, neighborhood disorder, and potential hazards to mobility. To ensure balanced geographic coverage among the raters, we distributed the 800 image sets evenly across four raters, such that each image set was independently coded by two different raters and each rater assessed 100 image sets across each city. When two raters had significant disagreement for an image, some items were flagged for review and resolved through consensus discussions led by a trained moderator also familiar with the urban context in India (further details on the consensus process are in Supplementary Materials S2 Section 2.1).

The following items were used in the analysis comparing ML-derived features to human ratings: Raters recorded the total number of vehicles (cars, buses, and trucks), autorickshaws, cycles (bicycles and motorcycles), and pedestrians across the four directional images. They also noted whether street signs and traffic lights were present in any image in the set. Sidewalk quality was

assessed using a five-point ordinal scale reflecting quality and availability, from which we created a binary variable with “no sidewalk” versus “some sidewalk” options for comparison with ML feature extraction since the SegFormer model could only detect sidewalk pixel presence and not quality. Greenspace was assessed with the question: “Are there trees/greenery visible?” with ordinal response options: (1) No, (2) Small amount, (3) Some, (4) More than average, and (5) Very green.

3.3. Interviewer Observations from LASI-DAD Wave 2

As a different form of validation, we used interviewer observations of the physical environment which were also collected as part of Wave 2 of the Harmonized Diagnostic Assessment of Dementia for the Longitudinal Aging Study in India (LASI-DAD) [56]. LASI-DAD is a nationally representative study of adults aged 60 years and above in India that aims to advance understanding of the determinants of late-life cognitive decline and dementia. The second wave of data collection was conducted between December 2022 and May 2024 across 22 Indian states and union territories with a total of 4,635 surveyed older adults. Informed consent was obtained directly from respondents for all study components, including geriatric assessments and environmental monitoring. For respondents with cognitive impairment, consent was obtained from a legally authorized family member, such as a spouse or adult child. Consent forms were available in 12 languages and read aloud when needed. Respondents unable to sign digitally marked the consent form and had a legally authorized representative sign on their behalf. All consent and interviews were conducted in the respondent’s preferred language. For the analysis in this study, we limited, the sample to 408 households that resided within the boundaries of the four cities.

As part of the Wave 2 fieldwork, interviewers completed a structured observation module assessing various aspects of the neighborhood's physical, social, and built environment near each respondent's home. From this broader instrument, we selected four items that aligned conceptually with features extracted from the ML model for comparison (see Table 1). These items assessed vehicular and pedestrian traffic, visible green space, and sidewalk condition. Data was accessed on September 1st, 2024, using a restricted data enclave that included respondents' geocoded residences.

Table 1. Interviewer Observations Responses

Field Interviewer Observation Item	N = 408[1]
Vehicular Traffic: 'What is the volume of car/bus/motor/rickshaw traffic in the area near the home?'	
-No traffic or not much	134 (33%)
-Some	188 (46%)
-More than usual or heavy	85 (21%)
Pedestrian Volume: 'What is the volume of pedestrian traffic in the area near the home?'	
-No traffic or not much	138 (34%)
-Some	181 (44%)
-More than usual or heavy	89 (22%)
Visible Green Space: 'Is a park, garden, or other green space visible from the home?'	
-No	253 (62%)
-Yes	155 (38%)
Sidewalk Condition: 'Which best describes the condition of the sidewalks in the area near the home?'	
-No sidewalks	75 (19%)
-Poor or average condition	292 (74%)
-Above average or very good condition	27 (6.9%)

[1] n (%)

3.4. Geospatial Measures

To further evaluate the validity of ML-derived features, we compared them to three independent geospatial measures: the normalized difference vegetation index (NDVI), nitrogen dioxide (NO_2) concentration, and population density. NDVI is a satellite-based indicator of vegetation density, commonly used as a proxy for greenspace exposure [57,58]. We used data available through the Google Earth Engine derived at 250-meter resolution images captured by the Moderate Resolution Imaging Spectroradiometer aboard NASA's Terra Satellite (MODIS-Terra MOD13Q1) to calculate the mean NDVI in 2020 across each of the four cities [59]. NO_2

concentrations are commonly associated with the transportation sector and traffic in urban areas. For our analysis, we used concentrations estimated at 50 meter resolution from global land use regression models that integrate satellite data from the Ozone Monitoring Instrument (OMI) with information on roads, built environments, and meteorological variables [60]. We used annual estimates of ambient NO_2 for 2020. Population density estimates were from the LandScan 2021 Global Population dataset, which provides gridded estimates of ambient population at a one kilometer resolution [61].

3.5. Statistical Analysis

We first evaluated the fine-tuned object detection and segmentation models and compared their performance with state-of-the-art models on the IDD dataset. Specifically, we compared the object detection model (DETR) with the Faster R-CNN (FRCNN) [62], which was previously identified as the best-performing model on the IDD dataset by Singh et al [63]. We fine-tuned both DETR and FRCNN using 31,000 images from the IDD dataset, and evaluated their performance on 1,151 images. We used the average precision (mAP) per class across all cities to characterize performance. The mAP captures the precision-recall tradeoff, where precision refers to the proportion of correct detections out of all detections made, and recall refers to the proportion of true objects correctly detected.[64] Higher values closer to 1 reflect stronger agreement between model predictions and ground truth labels. We also compared the performance of the semantic segmentation model (SegFormer) with the Dilated Residual Networks (DRN) [65], which was identified as the best-performing segmentation model on the IDD dataset by Singh et al. [63] We fine-tuned both SegFormer and DRN using 7,034 IDD images for training and evaluated models' performance on 1,055 test images using mean Intersection over Union (mIoU) per class. The mIoU quantifies the overlap between the predicted and ground truth segmentations for each class. It is

calculated as the area of overlap divided by the area of union between predicted and true pixel masks, averaged across all evaluated classes. mIoU values closer to 1 indicate greater accuracy in pixel-level classification [64].

Because some areas lacked GSV coverage and instead relied on model-predicted values, we further assessed the performance of the spatial prediction models. We used 5-fold cross validation (80% to 20% training to validation split) by leaving whole intersections out at a time and compared model performance to a simpler Random Forest Regressor also trained on CooKIE embeddings. Evaluation metrics included the coefficient of determination (R^2), mean absolute error (MAE), and Moran's I for spatial autocorrelation in residuals. Metrics were computed across all detected objects and averaged for each city.

Our following analyses focused on evaluating the validity of machine learning derived environment features by comparing the spatial predictions at the H3 cell level to the three complementary external sources: (1) human-coded audits of GSV images, (2) interviewer assessments from a nationally representative aging study, and (3) geospatial measures (NDVI, NO_2, and population). Each comparison targeted overlapping features to assess the extent to which ML-extracted information aligned with human perception, field observation, and external environmental indicators.

Data from the human-coded 800 GSV image sets was first compared to the spatial predictions. For continuous variables captured during the coding of the images (e.g., vehicle and pedestrian counts), we generated scatter plots and calculated linear regression coefficients, mean squared error (MSE), and intra-class correlation coefficients (ICC) to compare ML and human-coded values. For binary indicators (presence of sidewalks, street signs, and streetlights), we computed accuracy, sensitivity

(true positive rate), and specificity (true negative rate) of ML predictions relative to human-coded classifications. For greenspace, we visualized the distribution of ML-derived green view index (GVI) values across ordinal human-coded categories using box plots and reported Spearman's rank correlation and Kruskal–Wallis test results.

To compare ML-derived features with field-based assessments from LASI-DAD, we retrieved respondents' geocoded residential locations and constructed circular buffers at four spatial scales: 0 m (point location), 100m, 250m, and 500m to capture the surrounding environment at varying levels of proximity. The buffers were then intersected with the H3 hexagons containing ML-derived environmental features. For each buffer, we calculated area-weighted averages of ML features based on the proportion of each H3 hexagon that overlapped with the buffer. We then examined the association between buffer-averaged ML-derived measures and interviewer ratings by estimating four linear regression models, each predicting a separate ML feature: vehicular traffic, pedestrian traffic, sidewalk presence, and GVI, based on the corresponding interviewer-assessed item. These models allow us to assess the extent to which interviewer perceptions of a neighborhood aligned with features extracted from street-level imagery in the same neighborhood. We used varying buffer sizes because interviewers were instructed to base their assessments on the immediate surroundings of the household, but the actual spatial scope of these judgments may be uncertain possibly extending beyond the immediate home. Simultaneously, features such as sidewalk quality or green space may vary significantly over short distances. By testing associations across a range of buffers, we account for potential mismatch in the spatial scale of perception versus measurement.

Finally, we compared ML-derived environmental features with the geospatial measures of NDVI, NO_2, and population density aggregated to the same spatial scale. Specifically, the three measures (NO_2, NDVI, and population density) were spatially joined to the H3 hexagonal grids produced for the four cities. For each H3 cell, we identified all intersecting raster pixels from the satellite- or model-based datasets. We then calculated the value for the cell by averaging the values of these pixels, weighted by the proportion of each pixel's area that overlapped with the H3 cell. This process yielded a dataset of environmental exposures and population estimates linked to ML features at the H3 level. To assess correspondence between these data sources and ML-derived spatial predictions, we generated city-specific scatter plots comparing the number of vehicles from images to NO_2 levels, GVI to NDVI levels, and the number of pedestrians to population levels, and calculated Spearman correlations to quantify the strength of association.

4. Results

4.1. Validation against IDD

The number of image sets queried from the GSV API was as follows: 13,481 in Mumbai, 18,591 in Kolkata, 68,409 in Delhi, and 76,888 in Chennai. GSV imagery was available for 11,361 of these intersections in Mumbai (84.2%), 11,618 in Kolkata (62.5%), 52,039 in Delhi (76.1%), and 65,376 in Chennai (85.0%). Over 99% of these images were collected between October 2021 and April 2023.

DETR achieved moderate performance with mAP scores > 0.5 across several classes and consistently outperformed a previously top-performing model (FRCNN) for the Indian Driving Dataset across nearly all object classes. Particularly large gains were observed for detecting vehicles (e.g., +0.39 for trucks) and persons (+0.42). S1 Table 1 in Supplementary Materials shows

mAP for different classes of objects (pedestrians, trucks, cars, etc.) using DETR compared to FRCNN. While both models struggled with traffic lights and signs, DETR still achieved notable gains over FRCNN in those categories. We observed a similar performance trend in semantic segmentation. SegFormer demonstrated excellent performance for road and vegetation segmentation (>0.8), and fair performance for sidewalks (0.59) and poles (0.50). S1 Tabe 2 in Supplementary Materials shows mIoU for each segmentation class. Segformer had superior performance over a previously top-performing model (DRN) across the 4 evaluated classes. Improvements were especially pronounced for features like sidewalks, poles, and vegetation.

4.2. Validation of Spatial Prediction Model

Across all four cities, the GNN consistently outperformed the Random Forest regressor in predicting features in the hold-out set. Average R^2 improved from 0.12 to 0.66 in Mumbai, 0.18 to 0.71 in Chennai, 0.15 to 0.51 in Delhi, and 0.11 to 0.31 in Kolkata. MAE also decreased, from 4.50 to 1.08 in Mumbai, 1.29 to 0.38 in Chennai, 2.26 to 1.22 in Delhi, and 3.91 to 1.23 in Kolkata. MAE reflects the average size of the errors when comparing prediction to ground truth across several classes. The Random Forest regressor shows greater residual variation, suggesting poor spatial generalization and frequent over- or underprediction (Figure 4). In contrast, the GNN's residuals are smaller and more tightly distributed, with fewer cells having large errors. This is further supported by the low spatial autocorrelation in GNN residuals (0.0257 of Moran's I), suggesting reduced spatial bias compared to using the Random Forest model.

Figure 4. Spatial prediction residuals for pedestrian estimates in Kolkata.

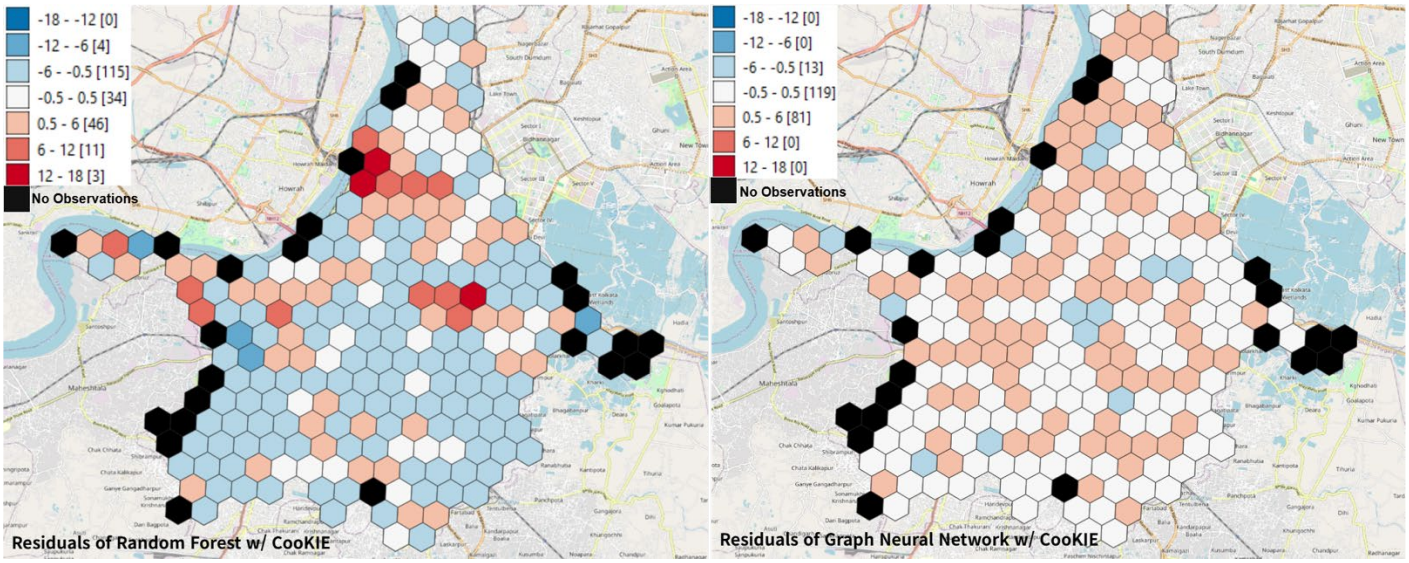


Figure compares Random Forest model (left) to the GNN model (right), both trained on the same input features.

4.3. Validation against Image Audits

Scatter plots illustrated adequate but varying agreement between ML predictions and human-coded counts (Figure 5). We found strong agreement for vehicle and pedestrian counts, with ICCs of 0.79 and 0.94, respectively, and corresponding Spearman correlations (r) of 0.84 and 0.89 ($p < 0.001$ for both). For cycles and autorickshaws, the relationships were more moderate. Cycles, which included both bicycles and motorcycles, yielded an ICC of 0.75 and $r = 0.61$ ($p < 0.001$), while autorickshaws had the weakest alignment (ICC = 0.35, $r = 0.59$, $p < 0.001$). Box plots comparing ML-derived green view index (GVI) scores with human-rated greenness on a 5-point ordinal scale revealed a monotonic trend: both the mean and median GVI scores increased with each level of human-assigned ordinal greenness rating, which ranged from 1 (No greenery) to 5 (Very green) (Figure 6). The Kruskal–Wallis rank sum test confirmed significant differences in GVI across greenness levels ($\chi^2 = 476.99$, $df = 8$, $p < 2.2e{-}16$), and Spearman's rank correlation showed a strong positive association ($\rho = 0.75$, $p < 2.2e{-}16$).

Figure 5. Scatter plots comparing ML features to human coded features.

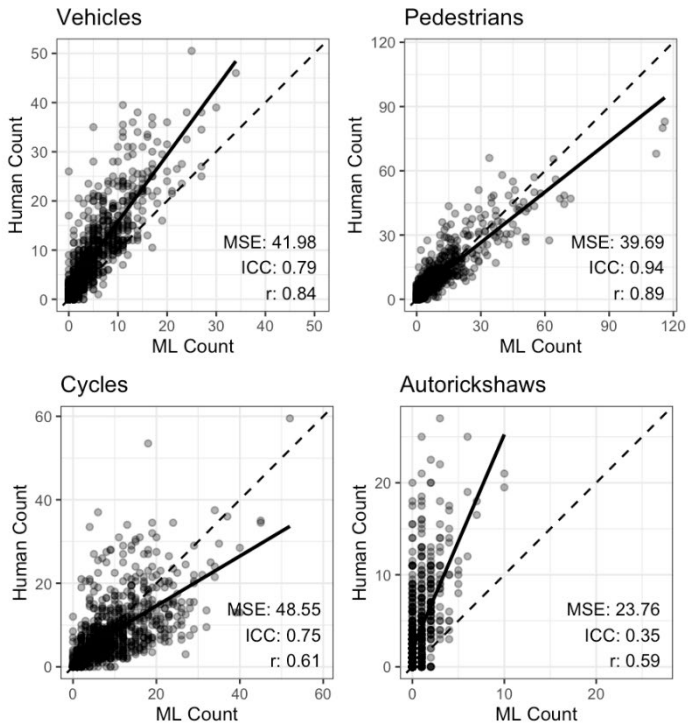
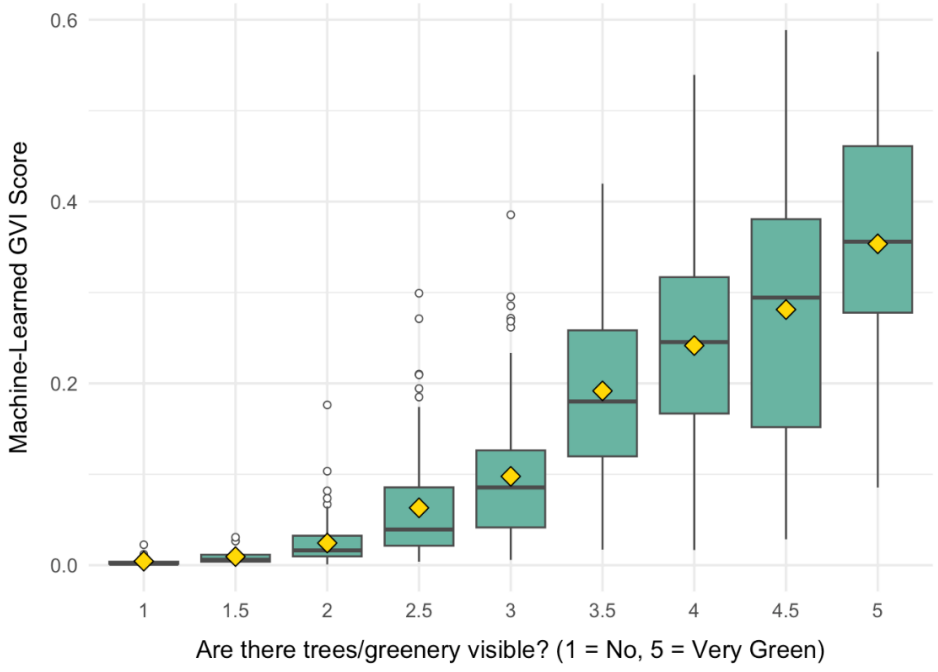


Figure 6. Box plot comparing ML-derived GVI from GSV images to human-rated greenness.



Decimal values reflect average scores between raters when difference ≤ 2

For traffic lights, traffic signs and sidewalks, the two human raters occasionally disagreed on whether images of intersections contained those features. Due to the large number of such cases, we didn't resolve them through consensus but instead excluded them from the analysis to ensure valid comparisons. After these exclusions, ML detection of traffic lights showed high overall accuracy (96%) and sensitivity (99%), but low specificity (11%), indicating a tendency to under-predict presence of traffic lights (Table 2). Here, "accuracy" reflects agreement with this reference label set and should be interpreted as a measure of consistency with human perception, not objective ground truth. Traffic signs showed moderate agreement, with 69% accuracy, 72% sensitivity, and 55% specificity, though 19.1% of images were excluded due to disagreement between raters. Sidewalk detection yielded the lowest accuracy (61%), with sensitivity and specificity of 68% and 59% respectively, and 21.8% of cases excluded for disagreement.

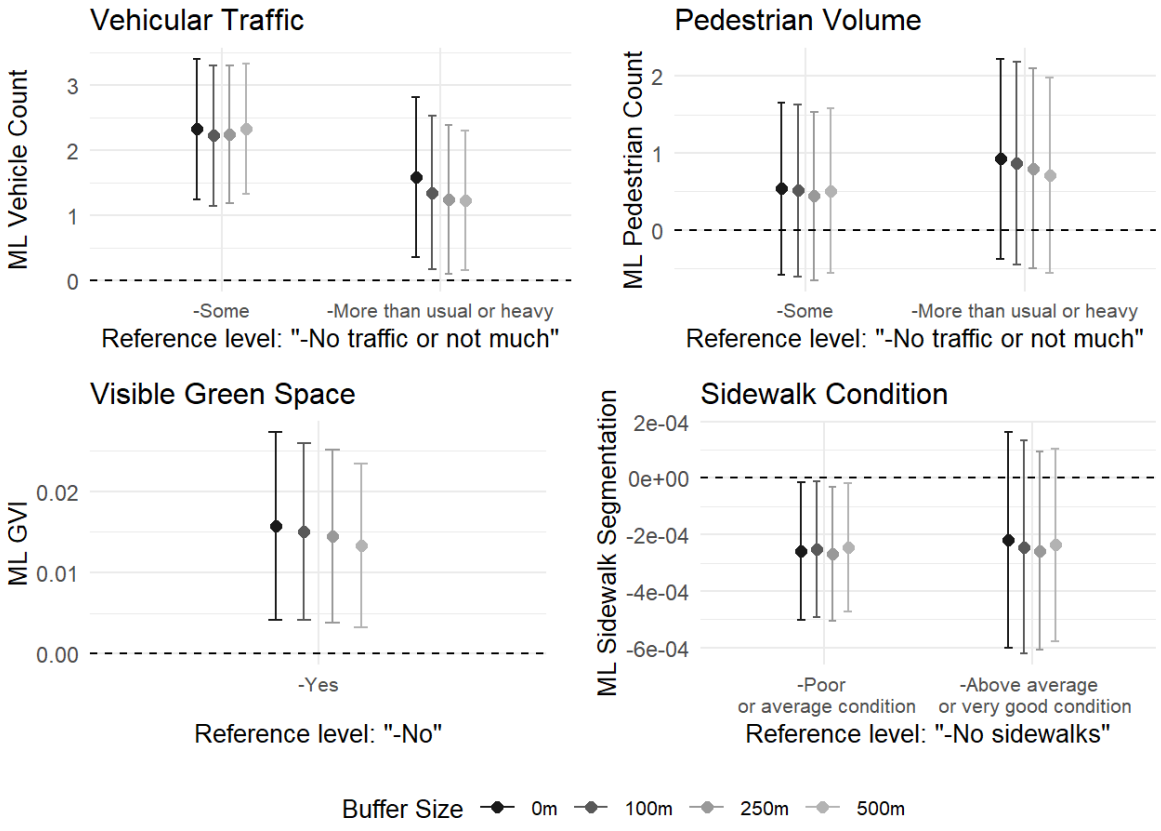
Table 2. Performance metrics for ML binary feature extraction relative to human-coded assessments

Feature	Accuracy	Sensitivity	Specificity	N Rated	Excluded Images (% of 800 image sets with disagreement)	N with Feature Present in Human assessment
Traffic Lights	0.96	0.11	0.99	776	3%	18
Traffic Signs	0.69	0.55	0.72	647	19.1%	142
Sidewalks	0.61	0.59	0.68	625	21.8%	331

4.4. Validation against LASI-DAD Interviewer Assessments

Using buffers of 500 meters around each household, we found that areas rated by interviewers as having "some" vehicular traffic had on average 2.33 more vehicles detected in GSV images compared to those with no or not much traffic (95% CI: [1.32, 3.33], $p < 0.001$) (Figure 7). In areas rated as having more than usual or heavy traffic, the ML-detected vehicle count was 1.23 higher compared to those with no or not much traffic (95% CI: [0.16 to 2.31], $p = 0.025$). While ML-detected pedestrian counts were higher in areas rated as having more foot traffic, these associations were not statistically significant. Visible green space observed by interviewers near participants' homes was associated with higher ML-derived Green View Index (GVI) scores, with a difference of 0.01 in GVI for areas with any greenspace detected by the interviewer (95% CI:

Figure 5. Linear regression estimates comparing ML-derived features and LASI-DAD interviewer assessments

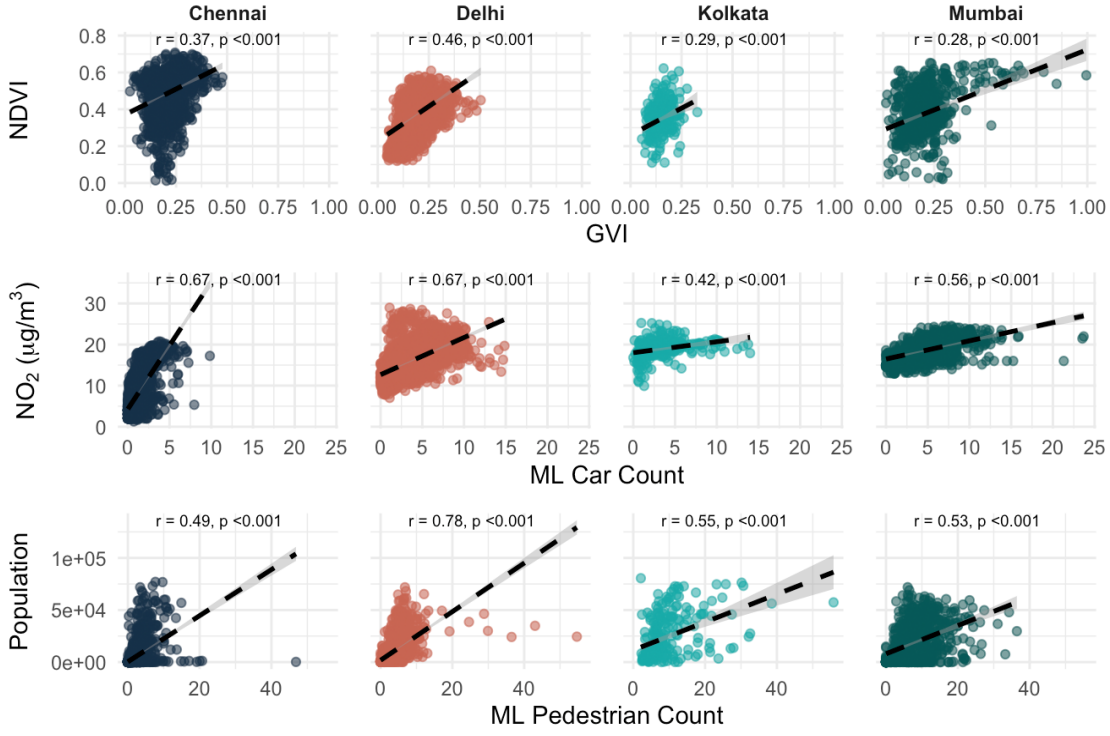


[0.00, 0.02], p = 0.010). Interviewer reports of poor or average sidewalk conditions were linked to significantly lower ML sidewalk pixel segmentation scores in those areas (-2.37e-04, 95% CI: [0.00, -5.10e-04], p = 0.034), while above average or very good sidewalk ratings were not significantly different compared to the reference category of no sidewalks, though the point estimates had similar magnitudes. The results across the four comparisons remained largely consistent for smaller buffer sizes (0m, 100m, 250m).

4.5. Validation against Geospatial Data

Across all four cities, ML-derived features showed meaningful alignment with external geospatial indicators, though the strength and consistency of these associations varied by feature type and location (Figure 8). For NDVI and GVI, we observed positive associations across all cities, though the correlation coefficient (r) was highest in Delhi (r = 0.48) and lowest in Mumbai (β = 0.28), indicating that higher GVI segmentation was moderately associated with more

Figure 6. Validation of ML-derived features against geospatial data by city



vegetation as captured by satellite NDVI. For vehicular traffic, ML-derived car counts were most strongly associated with NO_2 concentrations in Chennai and Delhi ($r = 0.67$), but the relationships were weaker in Kolkata ($r = 0.42$), and Mumbai ($r = 0.56$). Despite differences in magnitude, all associations were statistically significant ($p < 0.001$). Finally, ML-derived pedestrian counts were positively associated with population density in all cities, despite the presence of clear outliers in pedestrian predictions.

5. Discussion

This study assessed the performance of ML models trained on an Indian dataset to detect and quantify environmental features from GSV images using three complementary data sources: human-coded image audits, field-based interviewer observations, and geospatial indicators from satellite and population data. Overall, the ML models produced estimates that are reasonably related to observations from the other datasets, especially for more universally recognizable features such as vehicles, pedestrians, and greenery. In contrast, performance was lower for features that are more ambiguous like sidewalk quality or features specific to LMIC urban environments, such as rickshaws. These results highlight the promises and challenges of deploying ML models for scalable feature detection in complex and heterogeneous urban environments in LMIC.

Before triangulating the models' predictions against human-coded and geospatial data sources, we evaluated model performance using out-of-sample data from the Indian Driving Dataset. Overall, the DETR object detection model outperformed the previously top-performing model (FRCNN) on IDD across all evaluated object classes [62]. Performance gains were particularly notable for commonly occurring street elements such as vehicles and persons, highlighting DETR's ability to generalize to complex Indian streetscapes across these categories. However, gains were smaller

for more ambiguous features like traffic signs and traffic lights. These results demonstrate that while DETR offers improved detection accuracy overall, feature-specific limitations remain. For semantic segmentation, the SegFormer model outperformed the DRN baseline across all evaluated classes, achieving higher accuracy for both broad surfaces like roads and more granular elements such as sidewalks, greenery, and poles.

To extend the geographic reach of GSV-derived features into areas lacking image coverage, we implemented a spatial prediction model using a Graph Neural Network (GNN). Compared to a Random Forest Regressor, the GNN achieved greater predictive accuracy and spatial generalizability in cross-validation, but results differed across cities. It is important to note that the reliability of those predictions for intersections missing images hinges on the assumption that those locations are not systematically different in ways unobservable to the model. If image coverage is missing significantly or non-randomly, this could introduce bias into the predictions.

The comparison with human-coded image audits provides reasonable support for the models' validity in identifying features directly observable by a human eye. ML predictions were highly correlated with rater-derived counts of vehicles and pedestrians, with relatively high ICC and correlation coefficients. Notably, the ML model was more likely to undercount vehicles and cycles while overcounting pedestrians when compared to humans. Refinements to the object detection pipeline, such as the removal of small objects and application of non-maximum suppression, helped reduce noise and improve alignment with the protocol used by human raters. Still, model performance was comparatively lower for features that were less frequent or more visually ambiguous or underrepresented in HIC datasets. For example, autorickshaws were sometimes misclassified as cars, likely due to their similar shapes and visual features, and the model's

extraction of sidewalks, street signs, and poles was inconsistent, reflecting both model limitations and the inherent variability and subjectivity in these features across cityscapes. Other binary features such as traffic lights showed very high specificity but low sensitivity, meaning the model was good at confirming the absence of traffic lights but often failed to detect them when present. This conservative prediction bias is in line with poorer prediction performance on these elements in the IDD dataset re-emphasizing the need for greater representation of rare but relevant built environment elements.

The alignment between interviewer assessments from LASI-DAD and ML-derived features was generally weaker. Measures of vehicular traffic aligned well with interviewer reports whereas pedestrian traffic showed the expected direction of association but was not statistically significant, possibly due to variability in both the model's pedestrian detection and timing mismatches between when GSV images were captured and when interviewers made their observations. Another reason for this mismatch could be that the ML-detection does not differentiate between nearby and far pedestrians, while interviewers might have considered this difference. Surprisingly, despite being a static feature, sidewalk results showed the weakest alignment. ML-derived segmentation scores did not correspond well with interviewer ratings of sidewalk presence. One factor that could explain this is varied definitions of sidewalks as designated by the interviewer compared to the ML. Another factor is the spatial mismatch between the areas being evaluated by the GSV aggregation and the interviewers' observations. While ML features were averaged across larger circular buffers around the household, interviewer assessments typically reflected the immediate area surrounding the home. As a result, interviewer ratings may have a more precise spatial resolution than what is captured by the buffers, leading to discrepancies. While we attempted to ameliorate this by taking smaller buffer sizes, we were ultimately constrained by the resolution of

the H3 cells, which cover 0.74 square kilometers and therefore span several blocks with potentially differing sidewalk conditions. If sidewalk presence varies significantly from block to block, this coarse aggregation could mask meaningful variation. This highlights the need for finer-grained spatial units, non-uniform spatial units (e.g., street blocks), or more targeted sampling within smaller geographic extents when evaluating features like sidewalks that may vary sharply over short distances.

Overall, there was strong agreement between different indicators of greenness, suggesting that ML predictions from GSV provided a reasonable approximation of greenery in this setting. ML-predicted GVI showed strong correlation with human-coded ratings of greenery from the image audits, reinforcing the model's ability to detect eye-level vegetation in street imagery. GVI also aligned well with LASI-DAD interviewer assessments: Respondents living in areas where interviewers reported visible greenery tended to have higher average GVI values in the corresponding buffer. In all four cities, GVI was positively associated with satellite-derived NDVI, although the strength of the association varied by city. Observed variation may reflect the ways in which satellite and street-level imagery capture complementary but distinct dimensions of greenness [66–68]. NDVI reflects vegetation and canopy cover visible from above, while GVI captures human-scale vegetation visible from the street. Differences in urban form, vegetation placement, size, and quality, and building density as well as seasons likely account for variation in how well these measures align across different settings. These results support the use of GVI from GSV as a scalable proxy for neighborhood greenery but also emphasize the benefit of triangulating multiple sources of data to better characterize environmental exposures such as greenery.

Other geospatial comparisons provided further insight into the strengths and limits of ML-derived features. The relatively mixed correlations between car counts and NO_2 concentrations suggest that while vehicular presence is a contributor to air pollution, other sources such as industrial emissions may dilute this relationship. Similarly, associations between pedestrian counts and population density were generally positive, though visual inspection illustrated the presence of outliers. GSV imagery of intersections may disproportionately capture areas of concentrated pedestrian activity, such as intersections, markets, or commercial corridors, which reflect momentary crowding rather than the underlying residential population density. As a result, ML-derived pedestrian counts may overrepresent transient activity in high-traffic nodes rather than providing an even spatial estimate of where people live.

The patterns observed in our results point to the broader challenges in applying ML tools to GSV images [8,34,40,41]. First, incomplete image coverage remains a constraint especially in LMIC. Many intersections within the cities lacked GSV imagery, necessitating the interpolation of feature values across space. While this approach, coupled with ML prediction using multimodal geospatial embeddings with H3-cell aggregation, helped extend geographic coverage, it potentially introduced measurement uncertainty, especially given that GSV image availability is unlikely to be random. Areas with lower infrastructure or informal settlement patterns may be systematically underrepresented, potentially leading to bias in model extrapolation. Second, GSV provides only a static snapshot of the built environment. Features like vehicles and pedestrians are transient and can vary widely by time of day, day of the week, or season. Without incorporating GSV timestamp metadata – which is challenging to collect – temporal misalignment between imagery and other data sources (such as interviewer observations or field assessments) is difficult to account for. Third, the image perspective used by GSV may miss key parts of the streets such as alleys or

spaces behind obstructions limiting the completeness of visual information. In the Indian context, additional challenges arise from the cultural specificity of certain features. Elements like autorickshaws, irregular sidewalk structures, and informal signage are underrepresented in HIC datasets or inconsistently labeled, making them difficult for ML models to detect with accuracy. This study showed that even when models are fine-tuned on locally sourced data, transfer learning alone may not fully suffice. Purpose-built training datasets and annotations tailored to regional visual features will likely be necessary to further improve model performance.

More fundamentally, future research aiming to assess the impact of the built environment on healthy aging in LMICs must grapple with two key challenges. First, while this study addressed limitations in detecting visible built environment features using models trained on data from HICs, accurate detection does not guarantee meaningful interpretation for downstream health outcomes. Features such as congestion, sidewalk availability or greenery may carry different implications for mobility, safety, and cognitive stimulation in settings like India. This points to the need for more contextually grounded definitions of key constructs such as walkability or cognability, that reflect local lived experiences and environmental conditions. Second, uncertainty remains about the spatial scale at which built environment features influence health outcomes related to aging.[69,70] While this study aggregated ML-derived features over standardized buffers broadly capturing a neighborhood, other sources such as interviewer assessments reflect more localized perceptions of the area surrounding a household. The appropriate geographic unit of analysis may vary by feature and context. While theory-driven approaches can help guide decisions about the appropriate spatial scale, future studies should also consider incorporating sensitivity analyses that pairs ML-derived features with complementary data sources at multiple scales.

6. Conclusion

Our findings highlight the validity of ML-derived features in India where we lack data on environmental features. Overall, our results suggest that ML predictions on GSV images, though imperfect, are sufficiently reliable to capture several important urban environmental features. While prediction of ambiguous features and those specific to the Indian urban environment can still be improved, these scalable tools offer valuable potential for characterizing the built environment using ML. Moving forward, ML offers opportunities to identify environmental elements that promote cognitive engagement, reduce stress, or facilitate mobility in LMIC, especially when information on local contexts and theoretical insights are incorporated.

Acknowledgements

This research was supported by the National Institutes of Health/National Institute on Aging under grant numbers R01AG051125, U01AG064948, and R01AG030153. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Langa KM. Cognitive Aging, Dementia, and the Future of an Aging Population. In: Future Directions for the Demography of Aging: Proceedings of a Workshop [Internet]. National Academies Press (US); 2018 [cited 2025 May 20]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513075/

2. Nichols E, Steinmetz JD, Vollset SE, Fukutaki K, Chalek J, Abd-Allah F, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health. 2022 Feb 1;7(2):e105–25.

3. Gianfredi V, Nucci D, Pennisi F, Maggi S, Veronese N, Soysal P. Aging, longevity, and healthy aging: the public health approach. Aging Clin Exp Res. 2025;37(1):125.

4. Menassa M, Stronks K, Khatmi F, Roa Díaz ZM, Espinola OP, Gamba M, et al. Concepts and definitions of healthy ageing: a systematic review and synthesis of theoretical models. eClinicalMedicine. 2023 Jan 12;56:101821.

5. Abud T, Kounidas G, Martin KR, Werth M, Cooper K, Myint PK. Determinants of healthy ageing: a systematic review of contemporary literature. Aging Clin Exp Res. 2022;34(6):1215–23.

6. Song Y, Liu Y, Bai X, Yu H. Effects of neighborhood built environment on cognitive function in older adults: a systematic review. BMC Geriatr. 2024 Feb 27;24(1):194.

7. Keralis JM, Javanmardi M, Khanna S, Dwivedi P, Huang D, Tasdizen T, et al. Health and the built environment in United States cities: measuring associations using Google Street View-derived indicators of the built environment. BMC Public Health. 2020 Feb 12;20(1):215.

8. Boyes R, Pickett W, Janssen I, Swanlund D, Schuurman N, Masse L, et al. Physical environment features that predict outdoor active play can be measured using Google Street View images. Int J Health Geogr. 2023 Sept 28;22:26.

9. Nathvani R, Clark SN, Muller E, Alli AS, Bennett JE, Nimo J, et al. Characterisation of urban environment and activity across space and time using street images and deep learning in Accra. Sci Rep. 2022 Nov 28;12(1):20470.

10. Fry D, Mooney SJ, Rodríguez DA, Caiaffa WT, Lovasi GS. Assessing Google Street View Image Availability in Latin American Cities. J Urban Health. 2020 Aug 1;97(4):552–60.

11. Haddad M, Christman Z, Pearsall H, Sanchez M. Using Google Street View to Examine Urban Context and Green Amenities in the Global South: The Chilean Experience. Front Sustain Cities [Internet]. 2021 July 28 [cited 2025 July 8];3. Available from: https://www.frontiersin.org/journals/sustainable-cities/articles/10.3389/frsc.2021.684231/full

12. Umar F, Amoah J, Asamoah M, Dzodzomenyo M, Igwenagu C, Okotto LG, et al. On the potential of Google Street View for environmental waste quantification in urban Africa: An assessment of bias in spatial coverage. Bloor M, editor. Sustain Environ. 2023 Dec 31;9(1):2251799.

13. Nelson MC, Gordon-Larsen P, Song Y, Popkin BM. Built and Social Environments: Associations with Adolescent Overweight and Activity. Am J Prev Med. 2006 Aug 1;31(2):109–17.

14. Nabaweesi R, Hanna M, Muthuka JK, Samuels AD, Brown V, Schwartz D, et al. The Built Environment as a Social Determinant of Health. Prim Care Clin Off Pract. 2023 Dec 1;50(4):591–9.

15. Trott M, Cleland ,Claire L., Akaraci ,Selin, Valson ,Joanna S, O'Kane ,Niamh, Kee ,Frank, et al. Urban environment exposures and cognitive health: an evidence gap map of systematic reviews. Cities Health. 2025 Jan 2;9(1):129–59.

16. Finlay J, Esposito M, Langa KM, Judd S, Clarke P. *Cognability:* An Ecological Theory of neighborhoods and cognitive aging. Soc Sci Med. 2022 Sept 1;309:115220.

17. Besser LM, McDonald NC, Song Y, Kukull WA, Rodriguez DA. Neighborhood Environment and Cognition in Older Adults: A Systematic Review. Am J Prev Med. 2017 Aug;53(2):241–51.

18. Crous-Bou M, Gascon M, Gispert JD, Cirach M, Sánchez-Benavides G, Falcon C, et al. Impact of urban environmental exposures on cognitive performance and brain structure of healthy individuals at risk for Alzheimer's dementia. Environ Int. 2020 May 1;138:105546.

19. Peters R, Ee N, Peters J, Booth A, Mudway I, Anstey KJ. Air Pollution and Dementia: A Systematic Review. J Alzheimers Dis JAD. 2019;70(s1):S145–63.

20. Astell-Burt T, Navakatikyan MA, Feng X. Urban green space, tree canopy and 11-year risk of dementia in a cohort of 109,688 Australians. Environ Int. 2020 Dec 1;145:106102.

21. Garin N, Olaya B, Miret M, Ayuso-Mateos JL, Power M, Bucciarelli P, et al. Built Environment and Elderly Population Health: A Comprehensive Literature Review. Clin Pract Epidemiol Ment Health CP EMH. 2014 Oct 21;10:103–15.

22. E J, Xia B, Chen Q, Buys L, Susilawati C, Drogemuller R. Impact of the Built Environment on Ageing in Place: A Systematic Overview of Reviews. Buildings. 2024 Aug;14(8):2355.

23. Rodrigues HC, Shekha TAM, Annajigowda HH, Charly D, Jessy A, Suresh S, et al. Current status of research on the modifiable risk factors of dementia in India: A scoping review. Asian J Psychiatry. 2025 Mar 1;105:104390.

24. Shaw S, Kundu S, Chattopadhyay A, Rao S. Indoor air pollution and cognitive function among older adults in India: a multiple mediation approach through depression and sleep disorders. BMC Geriatr. 2024 Jan 22;24(1):81.

25. Adar S, D'Souza J, Shaddick G, Langa KM, Gross AL, Angrisani M, et al. Higher Household Air Pollution Levels Correlate with Poorer Cognitive Function in the Longitudinal Aging Study in India (LASI). Alzheimers Dement. 2022;18(S11):e059940.

26. Jana A, Varghese JS, Naik G. Household air pollution and cognitive health among Indian older adults: Evidence from LASI. Environ Res. 2022 Nov;214(Pt 1):113880.

27. Vaid U. Cognitive Health Costs of Poor Housing for Women: Exploring Executive Function and Housing Stress in Urban Slums in India. Int J Environ Res Public Health. 2024 Dec;21(12):1710.

28. Jahangir S. Perceived Meaning of Urban local Parks and social well-being of Elderly men: A Qualitative study of Delhi and Kolkata. Int J Rev Res Soc Sci. 2018 Sept 30;6(3):243–7.

29. Adlakha D, Krishna M, Woolrych R, Ellis G. Neighbourhood Supports for Active Ageing in Urban India. Psychol Dev Soc. 2020 Sept 1;32(2):254–77.

30. Adlakha D, Chandra M, Krishna M, Smith L, Tully MA. Designing Age-Friendly Communities: Exploring Qualitative Perspectives on Urban Green Spaces and Ageing in Two Indian Megacities. Int J Environ Res Public Health. 2021 Feb;18(4):1491.

31. Rundle AG, Bader MDM, Richards CA, Neckerman KM, Teitler JO. Using Google Street View to Audit Neighborhood Environments. Am J Prev Med. 2011 Jan;40(1):94–100.

32. Clarke P, Ailshire J, Melendez R, Bader M, Morenoff J. Using Google Earth to conduct a neighborhood audit: reliability of a virtual audit instrument. Health Place. 2010 Nov;16(6):1224–9.

33. Kelly CM, Wilson JS, Baker EA, Miller DK, Schootman M. Using Google Street View to Audit the Built Environment: Inter-rater Reliability Results. Ann Behav Med Publ Soc Behav Med. 2013 Feb;45(Suppl 1):108–12.

34. Chiang YC, Sullivan W, Larsen L. Measuring Neighborhood Walkable Environments: A Comparison of Three Approaches. Int J Environ Res Public Health. 2017 June 3;14(6):593.

35. Ki D, Chen Z, Lee S, Lieu S. A novel walkability index using google street view and deep learning. Sustain Cities Soc. 2023 Dec 1;99:104896.

36. Li X, Zhang C, Li W, Ricard R, Meng Q, Zhang W. Assessing street-level urban greenery using Google Street View and a modified green view index. Urban For Urban Green. 2015 Jan 1;14(3):675–85.

37. Nguyen QC, Khanna S, Dwivedi P, Huang D, Huang Y, Tasdizen T, et al. Using Google Street View to examine associations between built environment characteristics and U.S. health outcomes. Prev Med Rep. 2019 June 1;14:100859.

38. Lu Y. The Association of Urban Greenness and Walking Behavior: Using Google Street View and Deep Learning Techniques to Estimate Residents' Exposure to Urban Greenness. Int J Environ Res Public Health. 2018 Aug;15(8):1576.

39. Suel E, Muller E, Bennett JE, Blakely T, Doyle Y, Lynch J, et al. Do poverty and wealth look the same the world over? A comparative study of 12 cities from five high-income countries using street images. EPJ Data Sci. 2023 Dec;12(1):1–14.

40. Kang Y, Zhang ,Fan, Gao ,Song, Lin ,Hui, and Liu Y. A review of urban physical environment sensing using street view imagery in public health studies. Ann GIS. 2020 July 2;26(3):261–75.

41. Rzotkiewicz A, Pearson AL, Dougherty BV, Shortridge A, Wilson N. Systematic review of the use of Google Street View in health research: Major themes, strengths, weaknesses and possibilities for future research. Health Place. 2018 July 1;52:240–6.

42. Silva V, Grande AJ, Rech CR, Peccin MS. Geoprocessing via google maps for assessing obesogenic built environments related to physical activity and chronic noncommunicable diseases: validity and reliability. J Healthc Eng. 2015;6(1):41–54.

43. Agarwal S, and Nagendra H. Classification of Indian cities using Google Earth Engine. J Land Use Sci. 2019 Nov 2;14(4–6):425–39.

44. Bansal C, Singla A, Singh AK, Ahlawat HO, Jain M, Singh P, et al. Characterizing The Evolution Of Indian Cities Using Satellite Imagery And Open Street Maps. In: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies [Internet]. New York, NY, USA: Association for Computing Machinery; 2020 [cited 2025 May 20]. p. 87–96. (COMPASS '20). Available from: https://dl.acm.org/doi/10.1145/3378393.3402258

45. Google. Google Street View API [Internet]. 2024. Available from: https://developers.google.com/maps/documentation/streetview

46. Varma G, Subramanian A, Namboodiri A, Chandraker M, Jawahar CV. IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) [Internet]. 2019 [cited 2025 Aug 6]. p. 1743–51. Available from: https://ieeexplore.ieee.org/document/8659045

47. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers [Internet]. arXiv; 2020 [cited 2025 Aug 6]. Available from: http://arxiv.org/abs/2005.12872

48. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 740–55.

49. Hosang J, Benenson R, Schiele B. Learning non-maximum suppression [Internet]. arXiv; 2017 [cited 2025 Aug 27]. Available from: http://arxiv.org/abs/1705.02950

50. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2021 [cited 2025 Aug 6]. p. 12077–90. Available from: https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

51. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. 2016 [cited 2025 Aug 6]. p. 3213–23. Available from: https://ieeexplore.ieee.org/document/7780719

52. Zhao S, Chen Z, Xiong Z, Shi Y, Saha S, Zhu XX. Beyond Grid Data: Exploring graph neural networks for Earth observation. IEEE Geosci Remote Sens Mag [Internet]. 2024 [cited 2025 Aug 6]; Available from: http://www.scopus.com/inward/record.url?scp=85211639560&partnerID=8YFLogxK

53. Namgung M, Lin Y, Lee J, Chiang YY. Less is More: Multimodal Region Representation via Pairwise Inter-view Learning [Internet]. arXiv; 2025 [cited 2025 Aug 6]. Available from: http://arxiv.org/abs/2505.18178

54. PBC PL. Planet Application Program Interface: In Space for Life on Earth [Internet]. Planet Labs PBC; 2025. Available from: https://api.planet.com

55. Geofabrik GmbH and OpenStreetMap Contributors. GEOFABRIK downloads. [cited 2025 Aug 6]. OpenStreetMap Data Extracts. Available from: https://download.geofabrik.de/

56. Khobragade PY, Petrosyan S, Dey S, Dey AB, Lee J, Team the LDA. Design and methodology of the harmonized diagnostic assessment of dementia for the longitudinal aging study in India: Wave 2. J Am Geriatr Soc. 2025;73(3):685–96.

57. Gandhi GM, Parthiban S, Thummalu N, Christy A. Ndvi: Vegetation Change Detection Using Remote Sensing and Gis – A Case Study of Vellore District. Procedia Comput Sci. 2015 Jan 1;57:1199–210.

58. Rhew IC, Stoep AV, Kearney A, Smith NL, Dunbar MD. Validation of the Normalized Difference Vegetation Index as a measure of neighborhood greenness. Ann Epidemiol. 2011 Dec;21(12):946–52.

59. Didan K, Munoz AB, Huete A. MODIS Vegetation Index User's Guide (MOD13 Series).

60. Larkin A, Anenberg S, Goldberg DL, Mohegh A, Brauer M, Hystad P. A global spatial-temporal land use regression model for nitrogen dioxide air pollution. Front Environ Sci [Internet]. 2023 Apr 18 [cited 2025 June 16];11. Available from: https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2023.1125979/full

61. Lebakula V, Sims K, Reith A, Rose A, McKee J, Coleman P, et al. LandScan Global 30 Arcsecond Annual Global Gridded Population Datasets from 2000 to 2022. Sci Data. 2025 Mar 24;12(1):495.

62. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [Internet]. arXiv; 2016 [cited 2025 Aug 6]. Available from: http://arxiv.org/abs/1506.01497

63. Singh D, Rahane A, Mondal A, Subramanian A, Jawahar CV. Evaluation of Detection and Segmentation Tasks on Driving Datasets. In: Raman B, Murala S, Chowdhury A, Dhall A, Goyal P, editors. Computer Vision and Image Processing. Cham: Springer International Publishing; 2022. p. 512–24.

64. Padilla R, Passos WL, Dias TLB, Netto SL, da Silva EAB. A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. Electronics. 2021 Jan;10(3):279.

65. Yu F, Koltun V, Funkhouser T. Dilated Residual Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. 2017 [cited 2025 Aug 6]. p. 636–44. Available from: https://ieeexplore.ieee.org/document/8099558

66. Larkin A, Hystad P. Evaluating street view exposure measures of visible green space for health research. J Expo Sci Environ Epidemiol. 2019 July;29(4):447–56.

67. Wang R, Helbich M, Yao Y, Zhang J, Liu P, Yuan Y, et al. Urban greenery and mental wellbeing in adults: Cross-sectional mediation analyses on multiple pathways across different greenery measures. Environ Res. 2019 Sept 1;176:108535.

68. Wang R, Feng Z, Pearce J, Zhou S, Zhang L, Liu Y. Dynamic greenspace exposure and residents' mental health in Guangzhou, China: From over-head to eye-level perspective, from quantity to quality. Landsc Urban Plan. 2021 Nov 1;215:104230.

69. Bower M, Kent J, Patulny R, Green O, McGrath L, Teesson L, et al. The impact of the built environment on loneliness: A systematic review and narrative synthesis. Health Place. 2023 Jan 1;79:102962.

70. Kwan MP. The Uncertain Geographic Context Problem. Ann Assoc Am Geogr. 2012 Sept 1;102(5):958–68.

Supporting Information

S1 File. Machine Learning Model Post-processing and Performance Evaluation.

S1 Fig 1. Post-processing steps on object detection predictions from DETR.

S1 Table 1. Object Detection Models Comparison (Mean Average Precision per Class).

S1 Table 2. Semantic segmentation models comparison (mean intersection over union per class).

S2 File. Image Rating Protocol and Instrument.

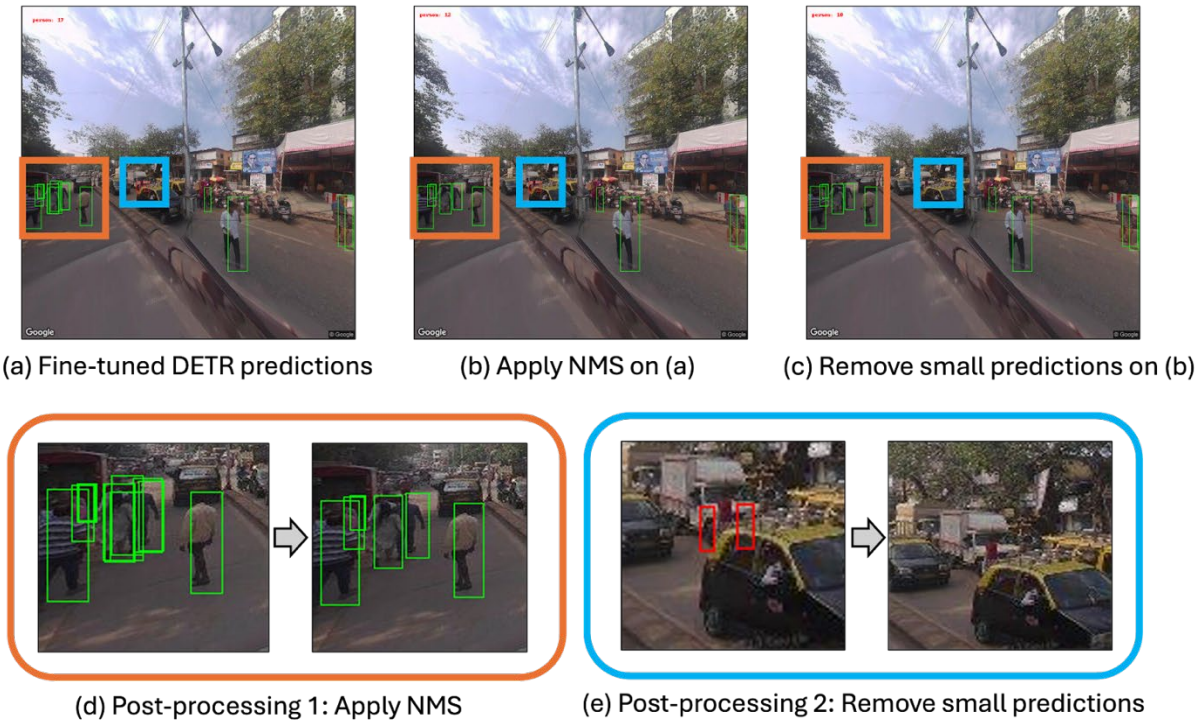
S2 Table 1. Consensus rating thresholds for continuous variables.

S1. Machine Learning Model Post-processing and Performance Evaluation.

1.1 Post-processing on Machine Learning-Based Feature Extraction

Despite the success of automatic feature extraction, we identified instances with redundant and noisy predictions where the DETR object detection model generated overlapping predictions on the same objects. For instance, we show the predictions from the fine-tuned DETR model in Figure 1-(a), highlighting several redundant bounding boxes for overlapping objects.

To address this, we implement post-processing using non-maximum suppression (NMS), which eliminated redundant bounding boxes [47,49]. As shown in Figure 1-(b), we apply NMS to the prediction results from Figure 1-(a) and eliminate these noisy and repetitive predictions. Further refining our results, we implement an additional filtering step to remove small predictions that were challenging for human coders to distinguish the objects in Figure 1-(c).



S1 Fig 1

1.2 Spatial Prediction Model Implementation

The graph neural network used for spatial prediction implements a three-layer GraphSAGE model, with each layer consisting of 128 hidden units, ReLU activations, and a dropout rate of 0.1 applied to the hidden layers. The model is trained from scratch using the Adam optimizer with a learning rate of 0.001 and mean squared error (MSE) as the loss function. Each H3 cell is represented as a node in the graph, and edges are defined based on second-order adjacency, i.e., each node is connected to its immediate and next-nearest neighbors in the H3 grid. To incorporate uncertainty into the model's predictions, we apply Monte Carlo dropout during inference. For each H3 cell, we run the model 50 times with dropout activated and generate a prediction on each pass. We then take the average of these 50 predictions to produce the final output. This process reduces prediction variance and results in more stable and reliable predictions across space.

1.3 Object Detections Results across Classes

Class	**DETR**	**Faster RCNN (FRCNN)**
Person	0.649	0.225
Truck	0.684	0.293
Motorcycle	0.594	0.304
Bicycle	0.482	0.100
Rider	0.550	0.248
Bus	0.523	0.348
Car	0.615	0.401
Traffic light	0.354	0.095
Traffic sign	0.449	0.127

S1 Table 1

From Table 1, DETR significantly outperformed FRCNN across all classes. Specifically, notable performance improvements were observed for trucks with an increase of 0.391 and motorcycles with an increase of 0.290. DETR consistently exhibited superior performance in vehicle detection, showing the performance differences for cars of 0.214, autorickshaw of 0.246, and bus of 0.175. Additionally, DETR substantially outperformed FRCNN in detecting humans, including person with an improvement of 0.424 and rider with an improvement of 0.302. Despite consistently high performance in detecting vehicles and persons, DETR showed comparatively lower mAP values for traffic lights and signs. This was likely due to confusion caused by the similarity of these objects' shapes to poles, frequently present in dense and visually complex Indian urban landscapes. Nevertheless, DETR still significantly improved detection accuracy for traffic signs (difference of 0.322) and traffic lights (difference of 0.259) over FRCNN.These results demonstrate that DETR fine-tuned on IDD more effectively captured physical environment objects compared to FRCNN.

1.4 Semantic Segmentation Results across Classes

Class (IoU)	**Segformer**	**DRN**
road	0.9648	0.9377
sidewalk	0.5871	0.3294
pole	0.4978	0.2170
vegetation	0.8912	0.7972

S1 Table 2

In Table 2, SegFormer significantly outperformed DRN across all four evaluated classes, achieving notable improvements of 0.2577 for sidewalk, 0.2808 for pole, and 0.094 for vegetation. The highest segmentation accuracy was recorded for the road class (mIoU = 0.9648), showcasing SegFormer’s effectiveness in segmenting extensive and clearly defined surfaces compared to DRN

(mIoU = 0.9377). Despite the relatively lower absolute accuracy for poles (mIoU = 0.4978), the considerable gap compared to DRN's performance (mIoU = 0.2170) highlighted SegFormer's superior ability to detect smaller and intricate objects, often complicated by occlusions. Similarly, SegFormer's significant improvement in sidewalk segmentation (from 0.3294 to 0.5871) highlighted its ability in handling complex Indian urban sidewalks characterized by diverse textures, unclear boundaries, or inconsistent construction quality. Vegetation segmentation accuracy also improved notably from 0.7972 to 0.8912, which indicated SegFormer's effectiveness in accurately identifying complicated natural components within urban contexts. Overall, this model's segmentation performance improvement demonstrated SegFormer's robust and consistent advantage over DRN in addressing challenging segmentation tasks in Indian urban environments.

S2. Image Rating Protocol and Instrument.

This appendix provides additional detail on the process used for the human-coding of GSV images. It outlines the training and standardization procedures used to prepare raters, the survey raters undertook, the criteria and process for resolving disagreements through consensus review, and the final scoring rules applied to create a dataset of environmental features in the four cities. These methods were designed to ensure consistency, reliability, and comparability across raters and cities in the study.

2.1 Image Rating Protocol

2.1.1 Rater Training

To ensure consistency across raters, all coders participated in a structured training process prior to the full coding of the complete list of image sets. This began with a collaborative review of a handpicked list of 40 image sets (10 from each city), during which all raters discussed their scoring of the same images as well as their interpretations of respective survey items. All team members participated in this training process including project investigators. This training established shared definitions for subjective or ambiguous categories, such as levels of congestion, sidewalk quality, and perceived greenery which could differ between individuals assessing those features in urban India.

As part of the training, raters were instructed not to exclude duplicate items in their counts of vehicles, pedestrians, cycles, and rickshaws when those features were in multiple images from different angles (0°, 90°, 180°, and 270°) within an image set since it would be too difficult to ascertain unique features across images. Additionally, vehicles parked in areas meant for moving traffic were included in the vehicular count. For counting pedestrians, the scores reflected the

number of people in the image grid, regardless of their activity, such as walking, standing, or sitting. Raters were also asked to avoid zooming in too much on images to discern the exact count of features such as cars and pedestrians. This recommendation was given to avoid capturing what was outside the area interest on the street intersection. Following training, each rater was assigned 200 images (100 per city), ensuring balanced geographic coverage and that each of the 800 image sets was independently rated by two raters.

2.1.2 Consensus Process and Final Scoring

To resolve differences in scoring by the two raters, we defined variable-specific thresholds for triggering consensus review by a third, trained moderator (Table S 3). These thresholds varied by variable type:

Variable	Consensus Threshold
Vehicle Count	If **average ≤ 20**, difference ≥ **5**; if **average > 20**, difference ≥ **10**
Cycles	If **average ≤ 15**, difference ≥ **5**; if **average > 15**, difference ≥ **8**
Rickshaws	If **average ≤ 15**, difference ≥ **5**; if **average > 15**, difference ≥ **8**
Pedestrians	If **average ≤ 20**, difference ≥ **5**; if **average > 20**, difference ≥ **15**

S2 Table 1

- **Continuous count variables** (vehicles, pedestrians, cycles, rickshaws): If the difference between the two raters exceeded a threshold based on the magnitude of their average rating, the image was flagged for review.
- **Ordinal variables** (e.g., greenspace): If raters differed by three or more points on the 1–5 scale, the item was reviewed. Smaller differences were resolved by averaging the two scores (e.g., 2 and 3 became 2.5).

- **Binary variables** (e.g., traffic lights, street signs, sidewalks): No consensus process was used. Cases where raters disagreed on presence/absence were excluded from computation of metrics (accuracy, sensitivity, specificity) comparing performance to ML algorithm.

Consensus scores replaced the original ratings only when disagreements exceeded the above thresholds. For items not triggering review, the final score was determined by averaging the scores across the two rater values. When a consensus review was triggered, the moderator reviewed the original images and both raters' justifications. A final score was then assigned based reference to training guidelines and occasional team discussion. Some items showed consistently large differences between raters. To simplify the process, we monitored which survey items had the highest alignment between individual raters and final moderator-assigned ratings. For these items, we assigned default consensus scores based on the rater who demonstrated the most consistent agreement with the moderator's previous decisions. To ensure the reliability of this approach, we conducted a validation step in which a random 20% of these default consensus ratings were reviewed independently by the moderator. In cases where the default and validation ratings disagreed, the moderator reassessed the item and assigned a final consensus rating. This process helped streamline adjudication while maintaining rigor and internal consistency across the dataset.

2.2 Image Audit Instrument

The GSV image audit was conducted using an instrument developed and deployed through REDCap (Research Electronic Data Capture), a secure web-based platform hosted at the University of Southern California. The survey was designed to systematically capture key features of the built environment relevant to health. Raters completed one survey per image set, which

consisted of four directional GSV images taken from a single street intersection. The instrument included the following questions:

1. Select city where photo was taken
    - Chennai
    - Delhi
    - Mumbai
    - Kolkata
2. What is the density of all types of vehicles (cars, buses, trucks, bicycles, motorbikes & rickshaws)?
    - Very low
    - Low
    - Medium
    - High
    - Very high
3. What is the density of pedestrians?
    - Very low
    - Low
    - Medium
    - High
    - Very high
4. Total number of vehicles (cars, buses, trucks
5. Total number of cycles (bicycles and motorbikes)
6. Total number of rickshaws
7. Total number of pedestrians
8. Are there any streetlights? Yes/No
9. Are there any street signs? Yes/No
10. Are there any traffic lights? Yes/No
11. Amenities/Greenspace: Are there trees/greenery visible?
    - No
    - Small amount
    - Some
    - More than average
    - Very green, park visible

12. Are there pedestrians, motorists, and bikes inter-mingled together on the road, creating potential for conflict?
    - No
    - Low risk
    - Some risk
    - High risk
    - Very high risk

13. Are there walking paths in good condition?

    - No walking paths (sidewalks) available
    - Available but in poor condition with many hazards
    - Available in moderate condition
    - Available in good condition
    - Available with no hazards

14. Are the streets in good condition?

    - Poor condition with many hazards
    - Moderate condition
    - Good condition
    - Excellent condition with no hazards

15. Are there obstructions in the walking paths (beggars, street vendors, motorbikes or cars parked)?

    - No
    - Some
    - Many

16. Are there street crossings (crosswalks) available?

    - Yes
    - No

17. Are the buildings well-maintained?

    - Buildings are in disarray
    - Poorly maintained
    - Somewhat maintained
    - Well maintained
    - Very well maintained
    - N/A (e.g. no visible buildings)

18. Is there visible trash on the streets/sidewalks?

    - None
    - Some
    - A lot

19. Does the area appear to be safe (crime)?

    - Very safe

- Somewhat safe
- Neutral
- Somewhat unsafe
- Very unsafe

20. Rate image confidence:

    - Poor and low confidence in ratings
    - Some issues with ratings due to image quality and moderate confidence
    - Image quality acceptable and moderate to high confidence
    - No issues with image quality and high confidence

21. Issues with image (check all that apply):

    - Blurry portions in image
    - Obstructions blocking view
    - Dark pictures
    - Early morning
    - Not a road intersection
    - No buildings
    - None

22. Any other comments (Text response)