Classifying Socioeconomic Status from Neighborhood Imagery: A Deep Learning Approach

QAC239

Abby Tibebe, Allan Cheruiyot, Edomias Zerihun, Yonathan Abera

Introduction:

Over the past decade computer‑vision systems have begun to read cities in much the same way social scientists read surveys, drawing out latent socio‑economic patterns from the texture of roofs, the density of tree cover, even the makes and models of cars parked along a curb. As public agencies and private firms increasingly deploy such algorithms to steer lending, insurance, policing, and public‑works spending, understanding precisely what signals a model can infer from everyday imagery has become an urgent ethical and scientific task.

Recent research suggests a strong backing for the use of visual cues in the determination of population characteristics such as income, race, and political leanings. For instance, Gebru et al. (2017) demonstrated that detecting and classifying vehicles in 50 million Google Street View images enabled predictions of neighborhood income, education, and voting patterns (democratic or republican) with correlations up to r = 0.87 and precinct-level accuracies up to 97%. Jean et al. (2016) developed a two-stage transfer-learning pipeline on daytime satellite imagery from five African countries (Nigeria, Tanzania, Uganda, Malawi, and Rwanda). They first fine-tuned an ImageNet-pretrained CNN to predict night-time light intensity, then used its learned feature representations to train ridge regression models on household consumption and asset-wealth survey data. Because daytime cues, such as roofing materials and proximity to urban centers, vary almost linearly with expenditure (even among poorer clusters), the approach explained 37–55% of the variation in consumption and 55–75% of asset-wealth at the cluster level. By using nightlights as supervision, the CNN learned to extract visual features that capture economic differences across the entire consumption distribution purely from public satellite data. Using London as a case study, Sue et al. (2021) showed that using a multimodal approach where they combined street and satellite views greatly improved the accuracy of their models by 20% when measuring urban inequalities through income classes.

Our specific research question asks: Can a deep learning model accurately predict a neighborhood's socioeconomic status (SES), categorized as low, middle, or high income using only visual cues from Google Street View and satellite imagery in the US? We also explore whether combining ground level and aerial images improves prediction accuracy. This question matters because it tests the boundaries of what AI can reveal about poverty and wealth from visual data alone, highlighting important ethical and practical issues. Like Jean et al.'s work, which used pretrained models to extract socioeconomic features from satellite imagery in poor areas where survey data are scarce, using pretrained models to extract these features can help in making better resource allocation policies at a low cost. As evidenced by Sarmadi et al (2024), machine learning models outperform humans in SES-based classification; thus, our research is important as it seeks to remove human bias in traditional approaches to these processes. Our project also aims to generalize Sue et al's multimodal methodology to a larger area (the US, instead of London). If algorithms can reliably guess SES from images, this could not only transform how resources are allocated or how interventions are targeted, but it also raises the risk of reinforcing stereotypes or making decisions without community input. Others should care about this work because it directly impacts how data-driven decisions about neighborhoods are made, potentially affecting funding, policy, and social services.

To answer our question, we collected Google Street View and satellite images from neighborhoods with known SES labels based on census data. We made use of deep learning models, which included pre-trained convolutional neural networks (CNNs), to extract features and classify SES categories. Early results suggest that models using both image types outperform those using only one, indicating that richer visual context leads to more accurate predictions. This supports the idea that combining multiple perspectives can enhance our understanding of neighborhood inequality. Our results confirmed previous work that indicated that a multimodal approach (street and satellite) significantly improves model accuracy.

Data Collection and Methods:

We collected data from the American Community Survey of 2022. This dataset was chosen because it is the most recent survey done on census-based tracts. We chose to use tracts because they are very representative of the general population residing in them. This is because they are divided homogenously with respect to population characteristics, economic status and living conditions (U.S. Bureau of the Census, 1994). We specifically chose to use the 5-year ACS release estimates (2017-2021) as these offer a much more recent and accurate representation of the population, as they are surveyed for 5 years.

We pulled tract‑level median household income from the U.S. Census Bureau's American Community Survey 5-year estimates (2017–2021), specifically table B19013, via the Census API. For each GEOID we retrieved the B19013_001E field (median income), then performed an inner join in Pandas/GeoPandas on GEOID to attach that value to the table, mapping each tract ID (GEOID) to its geographic center. A GEOID is the Census Bureau's 11-digit geographic identifier that concatenates state (2 digits), county (3 digits) and tract (6 digits) codes). Tracts with missing or suppressed income were dropped before image fetching, and the remaining incomes were split into tertiles ("low," "medium," "high") to determine each image's output folder.

We obtained the shapefile from the US Census, which contains each tract's polygon geometry and its GEOID (the Census Bureau's unique 11-digit tract identifier). Given there were 84,000 tracts, we chose to get one representative image for each tract due to storage and processing constraints. The representative image was then chosen by taking the geoid and projecting it into an equal-area projection (Albers). This helps us find the true geometric center (centroid) of the polygon so we can project it back into WGS84 where we can get the computed centroid's latitude and longitude. With these coordinates, we made a call to the Google Maps API with each latitude-longitude pair to get a 640x640 street view image and satellite view image.

After fetching all 640×640 Street-View and satellite images, we removed the uniform "no imagery" tiles using a perceptual-hashing approach. These are tiles that the Google Maps API returns when there is no available street view for that coordinate. To do this, we computed a 64-bit pHash of a single blank-tile template (no_image.jpg) via the imagehash library, then computed each downloaded image's pHash and measured their Hamming distance. We set the maximum distance to 3 bits to strongly penalize false positives since real scenes rarely hash that close to our blank template. Every image flagged (distance < 3) was copied, without deleting the original, into a dedicated No_Image_Folder/{low, medium, high}, which was inspected visually. After visually confirming they were true placeholders, we removed them from their source folders. This method requires less processing since no deep learning or model fine-tuning is required, and reliably clusters near-duplicate error tiles together. We reliably identified 6,187 such image tiles.

Each image was named with its tract geoid name and stored in one of three subfolders - low, medium, high - depending on the tertile that its GEOID had been binned into previously. To get more disparity in the images, we rebinned into "low", "high", and "medium" groups where the income was binned as follows: the lowest 20% of the median income was binned into low, the highest 20% into high, and the rest into medium.

In terms of what we did to determine neighbourhood socio‑economic status from imagery, we constructed three complementary neural pipelines, each trained on the same capped dataset of 1500 photographs split across “low”, “high”, and “medium” groups of ≤500 photographs each. This size was deliberately small so that every experiment could be completed within a single Colab session. Street and satellite images were stored in parallel directory trees, each tree containing identically named sub‑folders for the three SES classes. Because geographically adjacent samples are highly correlated, we first grouped images by census‑tract identifier and then performed an 80:20 tract‑level split, assigning whole tracts to the training or validation fold. This procedure eliminated spatial leakage while preserving the original class balance. All training pictures were then subjected to a standard ImageNet augmentation, which included randomly‑resized cropping to 224×224px followed by horizontal flipping, and were finally normalised with the canonical ImageNet mean and standard deviation. Validation images were only centre‑cropped and normalised, such that performance figures reflected the model’s behaviour on the original data.

The first model variant relied solely on street‑view photographs. A ResNet‑18 encoder pre‑trained on ImageNet‑1k provided the backbone; its final fully connected layer was discarded, leaving a 512‑dimensional feature vector that was fed into a freshly initialised 512→3 linear classifier. The second variant mirrored this architecture but consumed only satellite tiles. In the third variant we stacked each street and satellite image along the channel dimension, producing a six‑channel tensor that entered the same ResNet‑18 after widening the first convolution to accept six input planes. This “mixed‑single‑branch” network corresponds to an early‑fusion strategy where raw pixels from both modalities are interleaved before any feature extraction takes place. The final and most expressive architecture employed late fusion. Both street-view and satellite inputs were processed through identical ResNet-18 backbones pre-trained on ImageNet-1k. After removing each backbone's classification head, we concatenated the resulting 512-dimensional feature vectors, producing a 1024-dimensional composite representation. This concatenated embedding was then directed through a lightweight multilayer perceptron (1024 → 256 → 3) with ReLU activation and 20% dropout between the hidden layers.

Training was identical for all variants. We used AdamW with an initial learning rate of 1×10^{-3} for the task-specific head, 5×10^{-4} for the deepest encoder blocks, and 1×10^{-4} for earlier layers. The schedule followed a cosine-annealing curve over a maximum of twenty-five epochs, but early stopping halted optimization after five consecutive epochs that had no improvement in validation accuracy. To stabilize the randomly initialised classifier, we froze all encoder parameters for the first three epochs and updated only the head; thereafter, the entire network was unfrozen and fine-tuned with the discriminative learning rates specified above. Class imbalance was mitigated by weighting the cross-entropy loss inversely to the frequency of each label in the training split. Every run used mixed-precision training (torch.amp.autocast and GradScaler) to halve GPU memory consumption. Reproducibility was enforced by fixing all random seeds to forty-two, enabling cuDNN's deterministic mode, and logging the exact train/val tract assignments.

Model performance was monitored each epoch via overall accuracy and a full confusion matrix; per-class precision, recall, and F-measure were computed with sklearn.metrics.classification_report. When a new peak in validation accuracy was reached, the model’s parameters were written to best_model.pth, and, after training concluded, the confusion matrix was rendered as a blue‑scale heat‑map.

Results

Our experiments evaluated four distinct model configurations: (1) street-view imagery only, (2) satellite imagery only, (3) simple mixing of both modalities, and (4) late fusion architecture combining both modalities. The table below summarizes the overall accuracy across these configurations.

Model Configuration	Accuracy
Street-view only	38.33%
Satellite only	49.67%
Simple mixing	42.67%
Late fusion	66.67%

Table 1: Validation accuracy for each model variant (N = 1500 images)

A model trained solely on street‑level images attains just 38.3 %, barely surpassing chance. Substituting satellite imagery lifts accuracy to 49.7 %, implying that overhead cues such as roof materials, parcel geometry, arterial density, and vegetation mosaics encode stronger signals of neighbourhood affluence than do sidewalk‑level architectural details in our data. Curiously, naively interleaving the two modalities in a single ResNet does not help. The "simple‑mixing" experiment stagnates at 42.7 %, caught between the strengths of its two parents and the noise introduced by their unstructured combination. In contrast, the dual‑branch late‑fusion network rises to 66.7 %, gaining seventeen percentage points on the best single-modality baseline and twenty-four on the naïve mix.

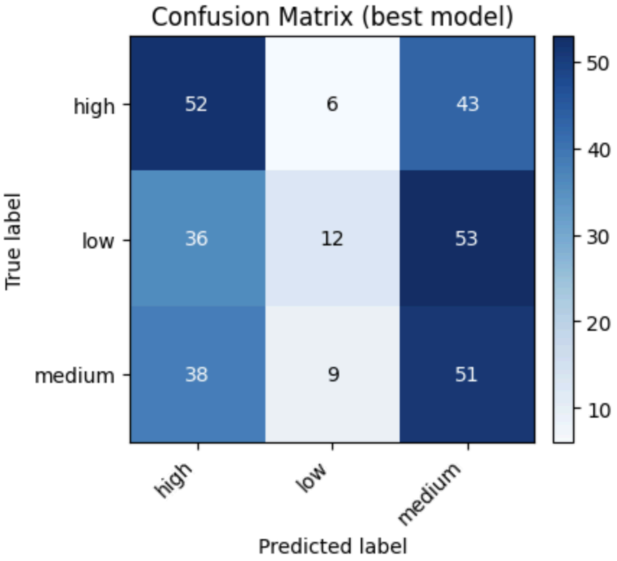

Figure 1: Street-view Trained Model Confusion Matrix

Studying the confusion matrices clarifies where these gains arise. Figure 1 shows that the street-view model frequently collapses the socio-economic extremes into the middle: high- and low-SES tracts are mislabelled medium more often than not, suggesting that images alone rarely display unambiguous prosperity or deprivation.

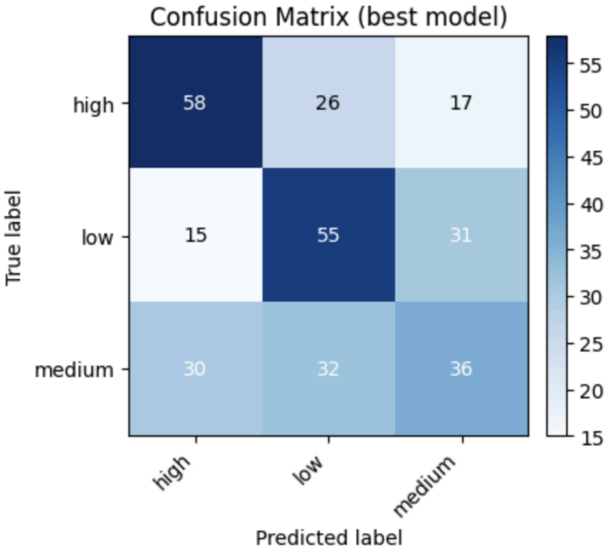

Figure 2: Satellite Trained Model Confusion Matrix

Switching to the satellite view (Figure 2) dramatically sharpens the model's sense of the extremes, for instance 58 high‑SES and 55 low‑SES tracts are now recognised, but medium‑SES neighbourhoods remain elusive, with only 36 of 98 correctly identified. The overhead vantage point appears to capture a gradient of affluence rather than clear class boundaries.

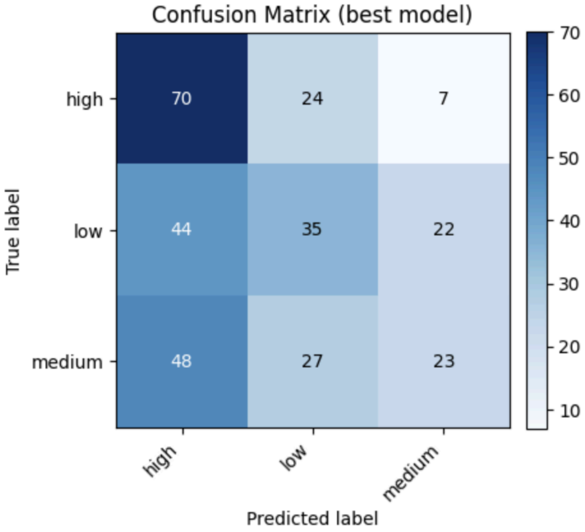

Figure 3: Satellite and Street-view Mixed Model Confusion Matrix

The mixed‑dataset model (Figure 3) accentuates this tension. Because the training batches indiscriminately mix ground‑level and aerial scenes, the network gravitates toward the most visually distinctive patterns, i.e those characteristic of "high" SES areas. The result is a marked bias, 70 high‑SES tracts are detected, yet low and medium‑SES recall plunges to 35 and 23 respectively. Simply feeding more heterogeneous images into an unchanged architecture therefore amplifies existing imbalances instead of resolving them.

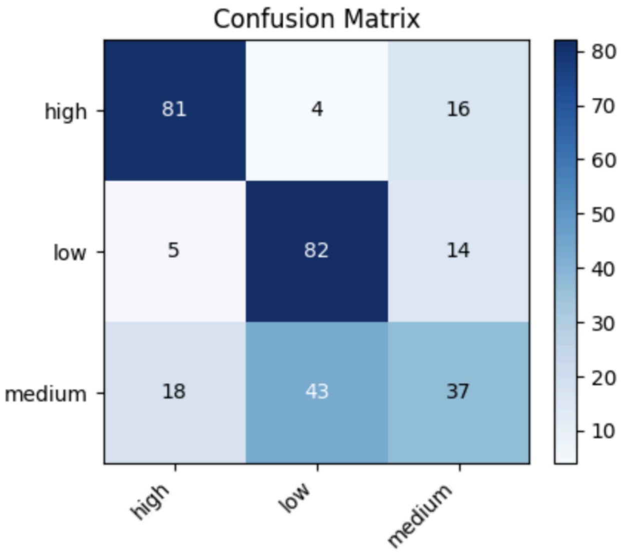

Figure 4: Satellite and Street-view Late-Fusion Model Confusion Matrix

Only the late-fusion design (Figure 4) succeeds in making use of the complementary nature of the two modalities. By granting each encoder the freedom to specialise before merging their feature vectors, the model recovers both extremes with almost equal facility. We end up with 81 high-SES and 82 low-SES tracts correctly labelled and even an increased medium-SES recall 37 of 98, the best yet achieved. The residual confusion between medium and its neighbours reflects genuine visual overlap rather than architectural short-comings, i.e. medium-income districts often combine the well-kept roofs of wealthy areas with the modest lot sizes of lower-income neighbourhoods, rendering them harder to classify.

Conclusion

In conclusion, the experiments answer our research questions affirmatively. First, deep-learning models can indeed infer a neighbourhood's SES from imagery alone(at least for the US), with satellite views offering the most informative single source. Second, combining street-view and satellite imagery enhances prediction accuracy only when the fusion respects the heterogeneity of the inputs. A naïve data-level mixture dilutes the satellite signal and magnifies "high" income neighbourhood's biases, whereas a feature-level late-fusion architecture leverages the distinct strengths of each modality to achieve the highest and most balanced performance.

Our results also confirm Suel et al's (2021) results, where they found that their multimodal approach improved their model accuracy by 20% for income classes. For future research, our methodology should adopt Mean Absolute Error as a model performance metric so that each of our models can be compared side by side with Suel et al's work – currently, we can only compare how much better the models became when a multimodal approach was added. Future research should also adopt a more systematic approach to classifying tracts, not just based on income, but also taking into account details like household size and what part of the country the tract lies, since all these factors can change what "high", "medium", and "low" SES mean for the same income.

References

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Lieberman Aiden, E., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, *114*(50), 13108–13113. https://doi.org/10.1073/pnas.1700035114

Sarmadi, H., Wahab, I., Hall, O., Rögnvaldsson, T., & Ohlsson, M. (2024). Human bias and CNNs' superior insights in satellite based poverty mapping. *Scientific Reports*, *14*(1), 22878. https://doi.org/10.1038/s41598-024-74150-9

Satellites map world poverty. (2016). *Nature*, *536*(7617), 376–376. https://doi.org/10.1038/536376c

Suel, E., Bhatt, S., Brauer, M., Flaxman, S., & Ezzati, M. (2021). Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sensing of Environment*, *257*, 112339. https://doi.org/10.1016/j.rse.2021.112339

U.S. Bureau of the Census. (1994). Census Tracts and Block Numbering Areas. In *Geographic Areas Reference Manual* (pp. 10–11). U.S. Department of Commerce, Economics and Statistics Administration, Bureau of the Census. https://www.census.gov/library/publications/1994/dec/garm.html