Analyzing Associations Between Chronic Disease Prevalence and Neighborhood Quality Through Google Street View Images

MEHRAN JAVANMARDI1, DINA HUANG2, PALLAVI DWIVEDI2, SAHIL KHANNA3, KIM BRUNISHOLZ4, ROSS WHITAKER1, QUYNH NGUYEN2, AND TOLGA TASDIZEN1

1Scientific Computing and Imaging Institute, The University of Utah, Salt Lake City, UT 84112, USA
2Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, MD 20742, USA
3Master’s in Telecommunications Program, University of Maryland, College Park, MD 20742, USA
4Healthcare Delivery Institute, Intermountain Healthcare, Salt Lake City, UT 84107, USA

Corresponding author: Mehran Javanmardi (mehran@sci.utah.edu)

This work was supported by the National Library of Medicine of the National Institutes of Health under Award R01LM012849.

ABSTRACT Deep learning and, specifically, convolutional neural networks (CNN) represent a class of powerful models that facilitate the understanding of many problems in computer vision. When combined with a reasonable amount of data, CNNs can outperform traditional models for many tasks, including image classification. In this work, we utilize these powerful tools with imagery data collected through Google Street View images to perform virtual audits of neighborhood characteristics. We further investigate different architectures for chronic disease prevalence regression through networks that are applied to sets of images rather than single images. We show quantitative results and demonstrate that our proposed architectures outperform the traditional regression approaches.

INDEX TERMS Set regression, multi-task learning, permutation invariant network, chronic disease prevalence, Google street view images.

I. INTRODUCTION

Deep convolutional neural networks have been shown to be powerful tools to model sensory data such as speech, images, videos, etc. CNN have been extensively used in the field of computer vision for different tasks, including image classification, image segmentation, and object detection. However, the application of these networks is not limited to a specific field as they have been used in many fields of science to automate or facilitate many tasks that traditionally used to be performed manually and at a significant cost. One instance of the application of CNN is in neighborhood research where they accommodate virtual neighborhood audits. In this work, we address how virtual neighborhood audits can be accomplished through Google Street View images and CNN. Moreover, we leverage the data from the neighborhood virtual audits to examine associations between neighborhood environmental features and chronic disease prevalence.

Neighborhood research has become a fast growing field in the scientific community because we have realized that the environment people live in has a direct influence on their health. Previous research has found associations between neighborhood quality and mortality risk [24], [51], [68], [70], [75], life expectancy [16], mental health [69], self-related health, obesity [7], [33], [52], and diabetes [29], [44] - even after adjusting for individual characteristics of the subjects. Neighborhoods can impact health through multiple pathways. First, disadvantaged neighborhoods may have fewer resources that support physical activity and healthy diets. Poor access to healthy food [15], [50], [73], the presence of fast food chains [9], and the lack of recreational facilities [10], [61] all correlate with higher obesity, diabetes, and blood pressure rates.

Second, neighborhoods may promote poor health through psycho-social pathways. Living in neighborhoods that are unclean, noisy, and violent can be psychologically harmful, through over-activation of the stress response [45]. Negative emotions over time can damage biological systems and lead to obesity, heart disease, diabetes, stroke, and declines in cognitive function [65]. Chronic anxiety and stress can disrupt cardiac function by altering the heart’s electrical stability, promoting atherosclerosis, and increasing inflammation [8].
Indicators of physical disorder (litter, graffiti, unclean parks, streets and sidewalks) have been linked with poorer self-rated health, higher risk of mental health issues (distress [34], depression [26], anxiety [12]), substance abuse, and mortality [25]. Research has also connected physical disorder and physical health outcomes, including obesity [46]. Physical disorder may increase physiology distress which can contribute to poor diet, inadequate sleep, and irregular exercise, thereby leading to worse physical health outcomes [11].

Examining neighborhoods in Seattle, San Diego, and Baltimore, Thorton and colleagues found that neighborhoods with lower socioeconomic status and higher proportions of racial and ethnic groups had poorer aesthetics (e.g., unmaintained buildings, graffiti, broken windows, and litter). Conversely, they found that some neighborhoods with higher socioeconomic status had better pedestrian amenities in terms of sidewalks, crosswalks, and intersection-control features. However, because the study was conducted in only a few cities, the authors caution that generalizability to other locations may be a concern [67]. In other research, Duncan and colleagues demonstrated that neighborhood walkability increased walking among adults in Paris [23], and Rundle and colleagues found that neighborhood walkability was associated with residents’ weekly physical activity and obesity-related health conditions in New York City [63], [64].

As research has corroborated, the quality of the neighborhood where people live is highly associated with different chronic diseases, such as obesity, diabetes, and high blood pressure. In order to have a prospering and healthier society, it is imperative to study what characteristics of the neighborhoods can influence the prevalence of these chronic diseases. The first step in this study is to determine neighborhoods where these chronic diseases are more prevalent than in others. The focus of this work without loss of generality is on obesity as one of the most prevalent and detrimental chronic diseases. We will show how Google Street View (GSV) images can be utilized as an important resource to red-flag neighborhoods with the potential for a high risk of obesity. Over 35% (78 million) of U.S. adults are obese [57]. Obesity is linked with mortality, morbidity, and reduced life expectancy. Comorbidities include type II diabetes, cancers, cardiovascular disease, sleep disorders, and hypertension. Chronic conditions such as obesity are the main drivers of mortality in the U.S. and they endanger the nation’s health, economic strength, and national security (by reducing the number of physically fit people who qualify for military service) [14]. Health care expenditures due to obesity, hypertension, and diabetes have been estimated at $70 billion [18], $110 billion, and $54 billion [22], respectively. These costs are compounded by lost productivity and absenteeism. Additionally, these major chronic conditions are concentrated among poor families and in poor neighborhoods, contributing to health disparities [41]. Numerous investigations have examined individual characteristics and behaviors, but researchers have only begun to establish contextual or structural factors that inhibit or encourage chronic disease health. Genetic variation cannot explain the epidemic rise in obesity and related chronic diseases in the past 20 years [39]. There is a pressing need to investigate societal and cultural processes [11].

II. RELATED WORK

Several studies have used Google Street View images as a resource for virtual audits of neighborhoods. In this section, we review some of the most prominent work in this area and discuss the drawbacks that come with current methods. We will also review related works regarding the invariance of deep networks with respect to group actions.

A. UTILIZING GOOGLE STREET VIEW IMAGES

Google Street View (GSV) images represent a massive data source that can be utilized in characterizing neighborhood built environment. As a virtual neighborhood audit tool, GSV has been validated and proved to be more cost-effective than traditional in-person audits [4], [13], [30], [36], [47], [48], [56], [63], [74]. GSV has also been used in neighborhood research to validate existing neighborhood measurement tools [1], [37], [53]. In addition, GSV cars have been used as a platform for spatial mapping in air pollution studies [2], [3]. Using GSV cars provides possibilities to map street-level traffic-related air pollution within neighborhoods at greater precision [2].

Bader and colleagues developed a Computer Assisted Neighborhood Visual Assessment System (CANVAS) as an innovative platform to perform virtual neighborhood audits [5]. CANVAS is an online application that can interact with Google Application Programming Interface (API) to collect GSV images. Neighborhood auditing toolkits, including the Irvine-Minnesota Inventory, the Pedestrian Environment Data Scan (PEDS), and the Maryland Inventory of Urban Design Qualities (MIUDQ), have been built within CANVAS to help auditors improve the reliability of GSV image labeling. However, CANVAS still relies on human labeling, making it difficult to perform large-scale neighborhood characterization. Mooney et al. utilized CANVAS for a virtual neighborhood audit of 532 intersections within New York City [49]. They found presence of crosswalks, pedestrian signals, nearby billboards, and bus stops were associated with increased pedestrian injuries at street intersections. In addition to CANVAS, the Forty Area Study Street View (FASTVIEW) and SPOTLIGHT-Virtual Audit Tool (S-V AT) are both GSV-based virtual audit tools [6], [28]. A prior study found neighborhood characteristics derived from S-V AT were not associated with total sedentary time for 219 Dutch and 128 Belgian adults who lived in 24 neighborhoods in “Sustainable prevention of obesity through integrated strategies” (SPOTLIGHT) study [19]. We are not aware of any studies have used neighborhood characteristics derived from FASTVIEW to predict health outcomes or health behaviors.

Lu recently published a study assessing the association between urban greenness derived from GSV and walking
behavior in Hong Kong [43]. In this study, deep learning of GSV images was used and greenness was assessed by the Pyramid Scene Parsing Network (PSPNet) [80], which has a high pixel-wise accuracy. Urban greenness was associated with increased walking time in both 400m and 800m buffers. This study takes advantage of publicly available GSV data and the innovative scene segmentation techniques to characterize urban greenness for each participant.

Li and colleagues derived a vegetation index from GSV images through an object-based image analysis [42]. Each image was segmented into homogeneous polygons, and each polygon was then assigned to different feature classes based on the spectral and geometrical properties [76]. The Otsu algorithm was used to optimize the threshold used for differentiating greenness versus non-greenness [58]. Li’s research was more focused on urban landscape planning than on health outcomes research.

Villeneuve suggested GSV measured greenness (GVI) was associated with hours of recreational activity in the summer. Individuals living in neighborhoods with the highest quartile of GVI had, on average, 18.1 hours of recreational activity time every week compared to 12.7 hours for those living in neighborhoods with the lowest quartile of GVI. GVI was not associated with the physical component summary (PCS) score or mental component summary (MCS) score.

Yin and colleagues detected pedestrians using the aggregated channel feature (ACF) algorithm [77]. They collected training data using a camera in a traveling car. The study found good consistency between automatically labeled GSV and human labeled GSV; however, the automatically characterized neighborhood feature was limited to pedestrians and was not linked to health outcomes. Yin et al. also characterized neighborhood walkability using artificial neural networks and support vector machines [78].

B. GAPS IN EXISTING GSV STUDIES

Most studies using Google Street View images have focused on assessing the agreement between virtual audits using GSV images and an in-person field audit. Very few studies have linked neighborhood characteristics captured from GSV to health outcomes. Mooney and colleagues found presence of crosswalk, pedestrian signals, nearby billboards, and bus stops were associated with increased pedestrian injuries at street intersections. However, the study only covered 532 intersection points in New York City and thus yielded limited generalizability. Virtual audits are cost-effective compared to traditional field visits. However, human labeling in virtual audits is still time-consuming when the study requires greater geographical coverage. Existing studies using innovative techniques to construct neighborhood characteristics mainly have focused on neighborhood greenness. Methods to construct neighborhood walkability, neighborhood safety, and neighborhood pleasurability from GSV images are under-developed. In addition, associations between neighborhood characteristics derived from GSV and chronic outcomes are under-studied.

Our research group conducted a pilot study using computer vision models to automatically label neighborhood characteristics including neighborhood greenness, crosswalks, and commercial building in Salt Lake City, Chicago, and Charleston [55]. We sampled images from all road intersections and along street segments at points that were 50m apart. We accessed GSV images from these points using Google’s Street View Image API. Deep convolutional neural networks (CNN) were trained to label unseen images [38]; more details of this approach is described in section III-A. Neighborhood characteristics were aggregated at the zip code level and were then merged to individual-level health data. We found individuals living in neighborhoods (zip codes) with the highest level of greenness, crosswalks, and commercial building had 25%-28% lower prevalence of obesity and 12%-18% lower prevalence of diabetes compared to individuals living in neighborhoods with the lowest levels of these neighborhood characteristics.

We subsequently collected street intersections of all the roads across the United States and retrieved GSV images at these intersections using Google’s Street View API. We randomly sampled two-thirds of the counties and used Google’s Vision API, Out-Of-Box, to perform neighborhood characterization. We obtained 10 neighborhood characteristics using Google’s Vision API and found that greater presence of highways was associated with lower prevalence of chronic conditions and premature mortality at the county level [54]. Individuals living in rural areas (identified from GSV) had a higher prevalence of chronic outcomes, premature mortality, physical distress, physical inactivity, and teen birth rates but a lower prevalence of excessive drinking [54].

C. EQUIVARIANCE AND INVARIANCE IN DEEP NETWORKS

In this paper, we discuss methods we utilized to further process Google Street View images for automated neighborhood characterization and chronic disease prevalence regression. The networks we use to regress the prevalence rates are required to be permutation invariant regarding their inputs. In the literature, we can define group actions on input/output data of a neural layer as a family of transformations on the input/output. Some instances of these group actions are rotation, translation, permutation, and so on. If this family of transformations on the input data does not change the output of the layer, this neural layer is called ‘Invariant’ to this action. However, if this action on the input transforms the output of the layer in a predictable way, this neural layer is called ‘Equivariant’ to this action.

Gens and Domingos [27] address the problem of symmetry groups for object detection by introducing a deep symmetry network where they generalize convnets to form feature maps over arbitrary symmetry groups. Cohen and Welling [17] introduce group equivariant convolutional neural networks (G-CNN). G-CNN increase the expressive capacity of the network by exploiting the symmetries without increasing the number of parameters in the network. Ravanbakhsh et al. [60] address the equivariance of a network through its parameters.
The proposed approach relates the equivariance properties of the neural layer to the symmetries of its parameter matrix. Guttenberg et al. [31] introduce a permutation invariant network to predict trajectories of sets of interacting objects. Vinyals et al. [72] address the problem of equivariance by introducing a “good” ordering of the inputs. In section III-B, we will analyze two popular approaches to make our architecture permutation invariant, which is a requirement for set regression. We will further combine our permutation invariant network with a regular (single image input) network in section III-C.

III. APPROACH

In this section, we will discuss the networks and architectures that make analysis of relations between neighborhood quality and chronic disease prevalence possible. In section III-A, we will discuss how computer vision helps automate the built environmental feature classification process where millions of images are classified according to the characteristics of the neighborhood they visually represent. In section III-B, we will investigate a regression model designed to work on sets of images instead of single input images to predict chronic disease prevalence. In this setting, the network will predict a prevalence index for a given set of images. In section III-C, we will combine models from sections III-A and III-B in a multi-tasking scenario to further increase the precision with which we can predict prevalence rates.

A. BUILT ENVIRONMENTAL FEATURE CLASSIFICATION

Public health scientists are interested in specific indicators in neighborhoods such as whether the neighborhood has plants, trees, and green areas or if the community is comprised of single family houses or more apartment-based housing. These indicators are called built environmental features. Such indicators can be recognized by looking at the Google Street View images of a specific neighborhood, but since these images are in the order of millions, manual annotation of these indicators is not feasible and automation is required. In order to implement this idea, we use powerful CNN-based [40] classifiers that have shown superior performance compared to more traditional classifiers like SVMs [20].

The network used for built environmental feature classification is shown in Figure 1. The built environmental features we are interested in include i) whether more than 30 percent of the images is comprised of green space and street landscaping (a binary label of 0 and 1 is assigned to presence or absence of the features), ii) whether the neighborhood is a single family house community, and finally iii) whether crosswalks are present in the neighborhood.

As can be observed in Figure 1, the network is composed of two main parts, a feature extractor network that extracts visual features from GSV images and a feature classifier network that assigns a binary label of 0 or 1 to each single image for a specific indicator. This predicted label represents whether the corresponding indicator is present in the image. Note that for each indicator mentioned above, we will use a separate classifier network, but the feature extractor part of the network is shared among all the feature classifiers. We notice that sharing the feature extractor among these three results in a slight performance gain of the network as well as a reduction in training time. We use the VGG19 [66] network as our feature extractor network and a single fully connected layer as our feature classifier network. We have also experimented with ResNet [32] as our feature extractor network, however, we did not observe any significant differences between VGG19 and ResNet. For each single image corresponding to a specific neighborhood, we will predict three labels corresponding to the three introduced indicators. For a single given tract, all images corresponding to that tract will be assigned a label for these three indicators. Furthermore these predictions will be aggregated for each indicator and reported as percentile information, i.e., what percentage of the images in a tract contain a specific indicator. This percentile information will be further used in a linear regression model to associate built environmental features with individual chronic disease prevalence rates.

In order to predict the chronic prevalence rates for a new given tract, Google Street View images corresponding to that tract are collected. All the image indicators are predicted and aggregated and the prevalence rate is predicted through a linear regression model. This approach has a significant drawback, and that is learning the built environmental feature classifier. This network is rather large, containing millions of parameters to learn. In order to learn this network, a large amount of manual annotations is needed, which is a costly and time-consuming process. To obviate the need for annotations, we propose a regression network that directly predicts the chronic disease prevalence rates from the GSV images.

B. CHRONIC DISEASE PREVALENCE REGRESSION

In order to be able to directly predict prevalence rates from GSV images at the tract level, we need to modify our current network. The prospective network needs to take as input a set of images rather than single images. Note that this network needs to be permutation invariant regarding its inputs. In other words, changing the ordering of the input set should not change the prediction result for that specific set. To make
the network permutation invariant, we consider two popular approaches in the literature.

1) PERMUTATION INVARIANCE BY ORDERING

Having a set as an input to a network is not a new idea in computer vision. This concept has been the focus of study for many different tasks such as video classification [35] and point cloud segmentation [59]. These works handle the problem of permutation invariance by utilizing a fixed ordering of the inputs based on a measure characterized by the inputs. For instance, for video classification, the visual features are extracted from each frame and concatenated in chronological order before being fed to the classifier. Inspired by this work, we propose to extract visual features from all the images in the set, sort these features based on some fixed measure, concatenate all the feature vectors accordingly, and generate the final tract representative feature vector. Further, this representative feature can be used in the regressor network to predict prevalence rates for a given tract. The network can be seen in Figure 2.

FIGURE 2. The proposed model in section III-B1. Each sample in this setting is a tract, i.e., the input is the set of corresponding images to that tract and the prevalence rate for that specific tract. The feature extractor is a pretrained VGG19 on ImageNet. The resize block is a single fully connected layer to decrease the dimensionality of feature vectors corresponding to single images. The downsized feature vectors are sorted and concatenated based on the intensity of their corresponding image. Finally, the tract representative feature is fed to a regressor network to predict prevalence rate. Final loss is the mean squared error.

This approach comes with two disadvantages. First, in order to sort the input set, we need to map the high-dimensional images into a 1-D real line. This transformation is very unstable to small perturbations since we are reducing the dimension of the input significantly, and therefore no type of sorting exists that is stable when prone to noise and small perturbations. The second problem with this approach is that the concatenation of all the feature vectors corresponding to a set will result in a large and arbitrary size of the representative feature vector for the set. Therefore, the input sets could have inconsistent input sizes. This large and arbitrary size feature vector is not very suitable for training as it can result in very long training times.

2) PERMUTATION INVARIANCE BY AGGREGATION

Deep sets has been introduced by Zaheer et al. [79] in order to manage networks that take a set of inputs rather than individual inputs. Zaheer et al. theorize that a function \( f(X) \) operating on the set \( X \) where \( X \) is countable is a valid permutation invariant set function if and only if it can be decomposed as

\[
f(X) = \rho(\sum_{x \in X} \phi(x))
\]

for suitable transformation functions \( \rho \) and \( \phi \). Motivated by the decomposition of a permutation invariant set function, we design our network accordingly. The first step is to find the mappings of all images in the input set to the feature space, which corresponds to \( \phi \) in the decomposition function. This transformation is handled by the feature extractor in Figure 3. The next step corresponding to \( \sum \) in the decomposition function is to aggregate the mappings in the feature space; this aggregation could be any commutative operand, such as a simple summation. This aggregated feature vector is the representative mapping for the entirety of the input set. Finally, we need to predict the output of the network by feeding the input set representative feature vector into the regressor network. The regressor network in Figure 3 is the counterpart of the \( \rho \) transformation in the decomposition function.

FIGURE 3. Model described in section III-B2. Each sample in this setting is a tract, i.e., the input is the set of corresponding images from that tract and the prevalence rate for that specific tract. The feature extractor is a pretrained VGG19 on ImageNet. The feature vectors of the images in the input set are averaged to generate a tract representative feature, which is fed to the regressor network to predict the prevalence rate. The final loss is the mean squared error.

The choice of the aggregation function for the network is not critical since the whole network is trained end to end, i.e., the feature extractor will adapt accordingly to any choice of commutative operand as the aggregation function.

C. HYBRID MODEL

The network proposed in the former section obviates the need for manual annotations to learn the network, however, at the same time, this elimination also discards the ability of the model to directly associate built environmental features with chronic disease prevalence. Public health scientists are interested in finding associations between built environmental features and chronic disease prevalence in different neighborhoods. In order to introduce the interpretability to the model, in the last section we modified the network in a multi-tasking
paradigm. We combine models in sections III-A and III-B2. We choose the model in section III-B2 over III-B1 because of the technical difficulties associated with it, although it shows slightly better performance. As mentioned before, the model in section III-B2 is massive because of the rather large tract representative feature vector, and therefore, combining it with the model in section III-A will result in a super-large model that we cannot accommodate on our commercial GPUs. In this framework, as can be seen in Figure 4, the feature extractor is influenced by two losses. The first one optimizes the feature extractor to generate features that are more suitable for built environmental feature classification, and the second loss promotes features that are more proportionate to the task of prevalence regression. As we will observe in the experiment section, optimizing these two losses at the same time not only enables more interpretability in the network, but also improves the performance of the network overall. For a more comprehensive study on multi-tasking in deep learning, refer to [62].

![Figure 4. Model described in section III-C. Each sample in this setting is a tract, i.e., the input is the set of corresponding images from that tract, and the prevalence rate for that specific tract. For all images in a tract we have three labels corresponding to the introduced built environmental features. Note that only a small portion of the images are annotated for this purpose, and therefore the classification loss is not backpropagated for unlabeled images. The feature extractor is a pre-trained VGG19 on ImageNet. Each feature classifier is a single fully connected layer, and the losses 1 to 3 are cross-entropy loss. The regression loss is the mean squared error.](image_url)

1) JOINT HYBRID MODEL

The hybrid model can be considered a combination of models in sections III-A and III-B2. In model III-A the regressor (GLM) sees the aggregated indicators for a tract to predict the prevalence rates, however in model III-B2 the aggregated feature from the feature extractor are directly fed to the regressor. Although the input to the regressor part of the models are different, but both seem to be informative. Joining the aggregated indicators from the feature classifiers with the aggregated features from the feature extractor before feeding them to the regressor seems a reasonable next step for this model. However we notice that this approach does not seem to improve accuracy of the network. This can be due to the fact that the features immediately after the feature extractor can be considered low-level features for this task as the aggregated indicators from the feature classifier could be considered high-level. Concatenating these two vectors and feeding them to the regression part of the model is not reasonable. However if we break down the regressor network in model III-C and concatenate the aggregated indicators with the features from the layer immediate to last in the regressor, which contain high-level information about the tract, can increase the accuracy of the model. This model is depicted in Figure 5.

![Figure 5. Model described in section III-C1. This hybrid model benefits from both the aggregated indicators from the built environmental feature classifiers and the aggregated features directly from the feature extractor. The three losses for feature classification are not shown in this graph for simplicity reasons. Note that the yellow blocks shown in the regressor represent feature maps and not fully connected layers.](image_url)

IV. DATA COLLECTION

We obtained roadway data for all road types across the United States using 2017 Census Topologically Integrated Geographic Encoding and Referencing (TIGER). Street centerlines and street intersections were identified using PostGIS plugin built within PostgreSQL (an open-source object-relational database system). PostGIS is a spatial extension that allows location queries to be performed in SQL.

We obtained GSV images at all street intersections across the US using Google Street View’s Application Programming Interface (API). We collected images with a resolution of $640 \times 440$ pixels and collected images from all four directions (the direction the camera is facing: $0 = \text{north}, 90 = \text{east}, 180 = \text{south}$ and $270 = \text{west}$) of each intersection, allowing us to fully capture the neighborhood features at each street intersection point.

We collected over 31 million GSV images from December 15, 2017 to May 14, 2018 using Google API. These 31 million images correspond to 53,921 tracts (neighborhoods) in the United States. On average, more than 500 images are collected for each tract.

V. EXPERIMENTATION

In this section, we discuss quantitative results produced by the different architectures discussed in section III as well as technical details corresponding to each model. Since this problem is naturally a regression problem, we use ‘coefficient of determination’, $R^2$ for quantitative comparison. Considering a dataset with $n$ values $\{(x_i, y_i)\}_{i=1:n}$ with predictions $\{(\hat{x}_i, \hat{y}_i)\}_{i=1:n}$, the coefficient of determination can be
the residual sum of squares is calculated as $SS_{res} = \sum_i^n (y_i - \bar{y})^2$, and the total sum of squares is $SS_{tot} = \sum_i^n (y_i - \bar{y})^2$, where $\bar{y} = \frac{1}{n} \sum_i^n y_i$. $R^2$ values range between 0 and 1, where a higher score demonstrates that the model better represents the variation of the data compared to a simple averaging as prediction. The $R^2$ is interpreted as the proportion of variance in the dependent variable that is predicted by the independent variable.

A. BUILT ENVIRONMENTAL FEATURE CLASSIFICATION

1) MODEL TRAINING

The model introduced in section III-A and demonstrated in Figure 1 is trained in two steps. First, the feature extractor and feature classifiers in Figure 1 need to be trained. Second, a generalized linear model (GLM) is trained to use the aggregated predictions of the feature classifiers and predict the prevalence rates for each tract. Note that in the first step image labels are assigned to all the images, and then the images are categorized according to the tract they belong to and aggregated values for each indicator are calculated. Further, the aggregated indicator values are used to train a linear regression model to predict prevalence rates.

In order to train the feature extractor and feature classifiers, we need annotated data for each indicator introduced in section III-A. Our collaborators have manually annotated images from two cities, Chicago, Illinois and Charleston, South Carolina. About 15,000 images were manually annotated for three indicators to learn network weights in the first step. However, the network used for this purpose is rather large and the amount of annotated data is not enough to learn the network accurately. In order to be able to learn this network, we initialize the feature extractor network from a pre-trained VGG19 network on the ImageNet classification dataset [21] classification dataset. We alter the last layer of the VGG19 network such that it generates a 1-D feature vector of size 4096 to be fed into feature classifiers. Each feature classifier is comprised of one fully connected layer that is learned from scratch.

2) RESULTS

We use the 500 cities dataset [71] released by the Centers for Disease Control and Prevention (CDC). This dataset reports prevalence rates for 27 chronic disease measures, including health outcomes (e.g., asthma, diabetes, cancer), public health prevention metrics (e.g., doctor and dentist visits and cholesterol screenings), and health behaviors (e.g., smoking and physical activity), at the city and census tract level for 500 cities using small-area estimation methods.

The cities selected include the 497 largest cities in the U.S., plus Burlington, VT, Charleston, WV, and Cheyenne, WY, so that all 50 states are represented. The project’s website features reports and interactive maps on the included health measures, documentation on methodology, and downloadable datasets from 2015 to 2018. The data provide estimates on the 27 measures for the 500 cities, as well as for approximately 28,000 census tracts. The sources used to generate the estimates include the Behavioral Risk Factor Surveillance System (BRFSS), Census Bureau 2010 census population data, and American Community Survey (ACS) five-year estimates. In our analysis, we examine obesity rates for 19,562 census tracts that have obesity rates and for which we have Google Street View images. These 19,562 tracts correspond to 7,246,783 images collected from Google Street View. We use 15,000 tracts and their corresponding images for training and the remaining 4,562 tracts for validation.

The model in section III-A results in a $R^2$ measure of 0.06. The bottleneck for this model is that the regression model only sees aggregated indicator values, however if we are able to use more informative features from the feature extractor, we could improve the accuracy of our predictions. The model in section III-B is designed such that it utilizes more informative features for prevalence regression.

B. CHRONIC DISEASE PREVALENCE REGRESSION

The model introduced in section III-B benefits from the fact that it does not need any manual annotation for built indicators. This is important not only because annotating data is time-consuming and costly but also it allows for an automated prevalence regression with minimal human interaction. Technical details and quantitative results for both approaches introduced in section III-B are described below.

1) PERMUTATION INVARIANCE BY ORDERING

The model in section III-B1 utilizes an internal ordering module that makes the network permutation invariant. This internal ordering is based on the average intensity of input images. Basically, a set of images is fed to the feature extractor, with corresponding 4096 dimensional features produced and resized to a smaller 128 dimension through a fully connected layer. These feature vectors are then sorted according to the input image intensities and concatenated to generate the tract representative feature vector. This final feature vector is fed to a regressor to predict prevalence rates. Note that resizing the 4096 dimensional feature vector is critical since concatenating all feature vectors with large dimensions will result in a very large network that cannot be accommodated on commercial GPUs. Although resizing the feature vectors to a smaller dimension helps with training time and alleviates memory burden on GPUs, it is still a training bottleneck for the network. Training time for this network is not optimum because the tract representative feature vector is too large, and the internal ordering of the inputs needs to be performed for each tract, which can be cumbersome. Another issue with this model is its incapability to handle arbitrary sizes of input sets. As can be observed from Figure 2, different input sizes for each tract will result in different tract representative feature vector sizes, which is problematic. To overcome this problem, we fix the size of the input set to 64. If a tract has fewer
images, it will be repeated to construct the input set. For tracts with more than 64 images, we randomly sub-sample the images for each iteration. In order to discard any randomness in our experiments, we train five different models and report the average and standard variation of each model. We achieve an $R^2$ measure of 0.6404 and a standard variation of 0.0024 with this architecture.

2) PERMUTATION INVARIANCE BY AGGREGATION
The model proposed in section III-B2 uses aggregation to take care of the required permutation invariance unlike model III-B1 where it takes advantage of sorting and concatenating the features from each image. This aggregation will result in a smaller tract representative feature which enables faster training time. We choose a simple averaging as aggregation function.

Another technical benefit of this architecture is its ability to handle arbitrary sizes of input sets. The GSV images we have collected ranges from a minimum of 1 image per tract to a maximum of 5,496 images per tract. Not considering the limited memory of commercial GPUs, this model can handle different size of input sets. However using the entirety of images in large tracts is not practically feasible as we cannot fit them into GPU memory. In order to obviate this problem we randomly sub-sample images from tracts with a size larger than 64. This approach results in an $R^2$ measure of 0.6306 and standard variation of 0.0034. As can be observed there is a significant improvement in accuracy compared to the model in section III-A.

C. HYBRID MODEL
The hybrid model introduced in section III-C is designed in a multi-tasking paradigm. In this architecture, the main loss is the mean squared error for the regression loss, but by introducing the classification loss for the built environmental feature classifiers, we are forcing the feature extractor to generate features that are more informative for both tasks. This will result in feature extractor, deriving representations that not only are informative enough for the task of prevalence regression but also are valid for built environmental feature classification. Built environmental features are indicators that are deliberately chosen by public health scientists and are thought to be strongly associated with chronic diseases. Therefore, by applying the classification loss we expect the generated representations to be more explanatory than the representations generated in model III-B2, which will result in a better performance of the hybrid model.

Multi-tasking can increase the performance of the model for several reasons. One can consider the auxiliary task (in this case, the built environmental feature classification) as a regularizer to the network that places an informative prior on the model. Using this auxiliary task can help with faster and better convergence, because it reduces the representation manifold to the intersection of the manifolds generated by two tasks. Intuitively, if a representation is knowledgeable for multiple tasks, we can say it is an informative representation. By adding this auxiliary task, we are able to improve the performance of our model significantly from 0.6306 in model III-B2 to a 0.6906 for the coefficient of determination. The joint model in section III-C1 achieves the highest $R^2$ of 0.7040. A full comparison of the quantitative results for all the proposed architectures is given in Table 1.

| Approach  | Coefficient of Determination ($R^2$) |
|-----------|---------------------------------|
| Section III-A | 0.6900 ± 0.0003 |
| Section III-B1 | 0.6404 ± 0.0024 |
| Section III-B2 | 0.6306 ± 0.0261 |
| Section III-C | 0.6906 ± 0.0034 |
| Section III-C1 | 0.7040 ± 0.0023 |

VI. CONCLUSION
In summary, this research makes significant, relevant contributions to the field of neighborhood research because i) neighborhood environments are increasingly linked to an array of important health outcomes, and ii) this project addresses the limits to research resulting from the lack of neighborhood data by providing new, cost-efficient data resources and methods for characterizing neighborhoods. We significantly contribute to the field by creating national data resources for large-scale examination of neighborhood effects on health. Analyzing the findings of this study may identify community design and public policy as possible levers of change for improving population health. We proposed and evaluated regression models that take as input sets of images as input instead of single entry inputs. Further we combined set regression models with single entry input image classifiers in our hybrid models to increase the overall performance of our regression model.

REFERENCES
[1] J. Adu-Brimpong, N. Coffey, C. Ayers, D. Berrigan, L. R. Yingling, S. Thomas, V. Mitchell, C. Ahuja, J. Rivers, J. Hartz, and T. M. Powell-Wiley, “Optimizing scoring and sampling methods for assessing built neighborhood environment quality in residential areas,” Int. J. Environ. Res. Public Health, vol. 14, no. 3, p. 273, 2017.
[2] S. E. Alexeeff, A. Roy, J. Shan, X. Liu, K. Messier, J. S. Apte, C. Portier, S. Sidney, and S. K. Van Den Eeden, “High-resolution mapping of traffic related air pollution with Google street view cars and incidence of cardiovascular events within neighborhoods in Oakland, CA,” Environ. Health, vol. 17, no. 1, p. 38, 2018.
[3] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. H. Vermeulen, and S. P. Hamburg, “High-resolution air pollution mapping with Google street view cars: Exploiting big data,” Environ. Sci. Technol., vol. 51, no. 12, pp. 6999–7008, 2017.
[4] M. D. Bader, S. J. Mooney, B. Bennett, and A. G. Rundle, “The promise, practicalities, and perils of virtually auditing neighborhoods using Google street view cars,” Ann. Amor. Acad. Political Social Sci., vol. 669, no. 1, pp. 18–40, 2017.
[5] M. D. Bader, S. J. Mooney, Y. J. Lee, D. Sheehan, K. M. Neckerman, A. G. Rundle, and J. O. Teitler, “Development and deployment of the computer assisted neighborhood visual assessment system (CANVAS) to measure health-related neighborhood conditions,” Health Place, vol. 31, pp. 163–172, Jan. 2015.
[49] S. J. Mooney, C. J. D’Maggio, G. S. Lovasi, K. M. Neckerman, M. D. M. Bader, J. O. Teitler, D. M. Sheehan, D. W. Jack, and A. G. Rundle, “Use of Google street view to assess environmental contributions to pedestrian injury,” Amer. J. Public Health, vol. 106, no. 3, pp. 462–469, 2016.

[50] K. Morland, S. Wing, A. D. Roux, and C. Poole, “Neighborhood characteristics associated with the location of food stores and food service places,” Amer. J. Preventive Med., vol. 22, no. 1, pp. 23–29, 2002.

[51] J. Morris, D. Blane, and I. White, “Levels of mortality, education, and social conditions in the 107 local education authority areas of England,” J. Epidemiol. Community Health, vol. 50, no. 1, pp. 15–17, 1996.

[52] M. S. Mujahid, A. V. D. Roux, M. Shen, D. Gowda, B. Sánchez, S. Shea, D. R. Jacobs, Jr., and S. A. Jackson, “Relation between neighborhood environments and obesity in the multi-ethnic study of atherosclerosis,” Amer. J. Epidemiol., vol. 167, no. 11, pp. 1349–1357, 2008.

[53] E. D. Nesoff, A. J. Milam, K. M. Pollack, F. C. Curriero, J. V. Bowie, A. C. Giehen, and D. M. Furr-Holden, “Novel methods for environmental assessment of pedestrian injury: Creation and validation of the inventory for pedestrian safety infrastructure,” J. Urban Health, vol. 95, no. 2, pp. 208–221, 2018.

[54] Q. C. Nguyen, S. Khanna, P. Dwivedi, D. Huang, Y. Huang, T. Tadzisien, K. D. Brunisholz, F. Li, W. Gorman, and T. T. Nguyen, “Using Google street view to examine associations between built environment characteristics and health outcomes,” Preventive Med. Rep., vol. 14, p. 100859, 2019.

[55] Q. C. Nguyen, M. Sajadi, M. McCullough, M. Pham, T. T. Nguyen, W. Yu, H.-W. Meng, M. Wen, F. Li, K. R. Smith, K. Brunisholz, and O. Tadzisien, “Neighbourhood looking glass: 360 automated characterisation of the built environment for neighbourhood effects research,” J. Epidemiol. Community Health, vol. 72, no. 3, pp. 260–266, 2018.

[56] C. L. Ogden, A. Caspi, C. J. Bates, R. J. Sampson, and T. E. Moffitt, “Systematic social observation of children’s neighborhoods using Google street view: A reliable and cost-effective method,” J. Child Psychol. Psychiatry, vol. 53, no. 10, pp. 1009–1017, 2012.

[57] C. L. Ogden, M. D. Carroll, C. D. Fryar, and K. M. Flegal, “Prevalence of obesity among adults and youth: United States, 2011–2014,” Tech. Rep., 2015.

[58] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Trans. Syst., Man, Cybern., vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[59] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3D classification and segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 652–660.

[60] S. Ravvanbakhsh, J. Schneider, and B. Poczos, “Equivariance through parameter-sharing,” in Proc. 34th Int. Conf. Mach. Learn., vol. 70, 2017, pp. 2892–2901.

[61] J. N. Roemmich, L. H. Epstein, S. Raja, L. Yin, J. Robinson, and D. Winiewicz, “Association of access to parks and recreational facilities with the physical activity of young children,” Preventive Med., vol. 43, no. 6, pp. 437–441, 2006.

[62] S. Ruder, “An overview of multi-task learning in deep neural networks,” 2017, arXiv:1706.05098. [Online]. Available: https://arxiv.org/abs/1706.05098

[63] A. G. Rundle, M. D. M. Bader, C. A. Richards, K. M. Neckerman, and J. O. Teitler, “Using Google street view to audit neighborhood environments,” Amer. J. Preventive Med., vol. 40, no. 1, pp. 94–100, 2011.

[64] A. G. Rundle and S. B. Heymsfield, “Can walkable urban design play a role in reducing the incidence of obesity-related conditions?” J. Amer. Med. Assoc., vol. 315, no. 20, pp. 2175–2177, 2016.

[65] T. E. Seeman, B. H. Singer, J. W. Rowe, R. I. Horwitz, and B. S. McEwen, “Price of adaptation allostatic load and its health consequences: MacArthur studies of successful aging,” Arch. Internal Med., vol. 157, no. 19, pp. 2259–2268, 1997.

[66] K. Simoyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556

[67] C. M. Thornton, T. L. Conway, K. L. Cain, K. A. Gavand, B. E. Saclens, L. D. Frank, C. M. Geremia, K. Glanz, A. C. King, and J. F. Sallis, “Disparities in pedestrian streetscape environments by income and race/ethnicity,” SSN-Population Health, vol. 2, pp. 206–216, 2016.

[68] P. Townsend, P. Phillimore, and A. Beattie, Health and Deprivation: Inequality and the North. Evanston, IL, USA: Routledge, 1988.

[69] K. D. Truong and S. Ma, “A systematic review of relations between neighborhoods and mental health,” J. Mental Health Policy Econ., vol. 9, no. 3, pp. 137–154, 2006.

[70] H. Tyroler, S. Wing, and M. Knowles, “Increasing inequality in coronary heart disease mortality in relation to educational achievement: Profile of places of residence, United States, 1962 to 1987,” Ann Epidemiol, vol. 3, pp. S51–S54, 1993.

[71] [Online]. Available: https://www.cdc.gov/50cities/index.html

[72] O. Vinyals, S. Bengio, and M. Kudlur, “Order matters: Sequence to sequence for sets,” 2015, arXiv:1511.06391. [Online]. Available: https://arxiv.org/abs/1511.06391

[73] M. C. Wang, S. Kim, A. A. Gonzalez, K. E. MacLeod, and M. A. Winkleby, “Socioeconomic and food-related physical characteristics of the neighborhood environment are associated with body mass index,” J. Epidemiol. Community Health, vol. 61, no. 6, pp. 491–498, 2007.

[74] J. S. Wilson, C. M. Kelly, M. Schootman, E. A. Baker, A. Banerjee, M. Clemmip, and D. K. Miller, “Assessing the built environment using omnidirectional imagery,” Amer. J. Preventive Med., vol. 42, no. 2, pp. 193–199, 2012.

[75] S. Wing, E. Barnett, M. Casper, and H. Tyroler, “Geographic and socioeconomic variation in the onset of decline of coronary heart disease mortality in white women,” Amer. J. Public Health, vol. 82, no. 2, pp. 204–209, 1992.

[76] D. M. Woebbecke, G. E. Meyer, K. Von Bargen, and D. A. Mortensen, “Shape features for identifying young weeds using image analysis,” Trans. Amer. Soc. Agricult. Eng., vol. 38, no. 1, pp. 271–281, 1995.

[77] L. Yin, Q. Cheng, Z. Wang, and Z. Shao, “Big data for pedestrian volume: Exploring the use of Google street view images for pedestrian counts,” Appl. Geography, vol. 63, pp. 337–345, Sep. 2015.

[78] L. Yin and Z. Wang, “Measuring visual enclosure for street walkability: Using machine learning algorithms and Google street view imagery,” Appl. Geography, vol. 76, pp. 147–153, 2016.

[79] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, “Deep sets,” in Proc. Adv. neural Inf. Process. Syst., 2017, pp. 3391–3401.

[80] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 2881–2890.