Forecasting High-risk Areas of COVID-19 Infection Through Socioeconomic and Static Spatial Analysis

Abdulaziz Alhamadani*, Shailik Sarkar†, Lei Zhang‡, Lulwah Alkulaib§, and Chang-Tien Lu¶
Department of Computer Science, Virginia Tech
Falls Church, VA
Email: *hamdani@vt.edu, †shailik@vt.edu, ‡Lei@vt.edu, §lulwah@vt.edu, ¶ctlu@vt.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671896

***Abstract*—Existing COVID-19 prediction models focus on studying the dynamic nature of the virus spread by using pandemic-related temporal data. In this paper, we present a work that exclusively uses comprehensive socioeconomic factors to predict the high risk areas of COVID-19 infection based on fine-grained static spatial analysis. Moreover, the most and least influential socioeconomic factors on COVID-19 spread are identified. This paper uses a uniquely built dataset by combining local states' cumulative COVID-19 statistics and their associated socioeconomic features on the zip code level. Further, the work solves the lack of data by augmentation. To evaluate the work, four case studies are conducted on Florida, Illinois, Minnesota, and Virginia. Experimental results show that the study provides accurate predictions with respect to ground truth data. By identifying high risk areas and socioeconomic factors, policymakers can use this study to take necessary measures to help disadvantaged communities.**

***Index Terms*—COVID-19, Socioeconomic, forecast, static spatial analysis, Data Augmentation**

I. Introduction

Towards the end of 2019, the novel coronavirus was first identified in Wuhan, China [20]. The virus was named Sars-Cov-2, and the associated acronym COVID-19 was issued by the World Health Organization (WHO). The transmission rate of COVID-19 was extremely high that by March 11th, 2020 a pandemic was declared by WHO. At the beginning of 2020, the virus has spread globally without any known treatment or approved vaccines. WHO recommended that people follow guidelines to prevent contracting COVID-19. Since there was no known timeline on when vaccines would be widely available for the public, we needed understand the social and economic determinants of COVID-19 and identify the spread in high risk geographical regions for any future pandemic such as this.

Most of the existing works on pandemic modeling focus on using historical data for forecasting the dynamic trends [14], [24], [32], [41]. Very few works focus on the relatively static characteristics of the spatial areas that play an essential role in the disease spread. Therefore, this research takes a different direction to forecast based on socioeconomic and static spatial analysis. Throughout history, socioeconomic status has been linked to health. Individuals higher in the social hierarchy typically enjoy better health than those below; socioeconomic status differences are found to increase rates of mortality and morbidity in almost every disease and condition [1].

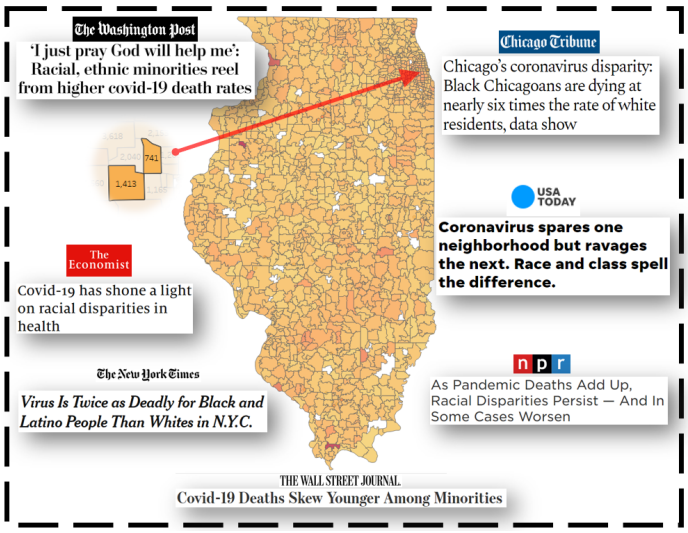

Fig. 1: Newspapers linking racial minorities with COVID-19 spread. A zoomed example from Chicago of two neighboring areas with disparate rates of race and income. Although they are neighboring areas, one has more cases than the other.

Socioeconomic factors impact all facets of human functioning, including health and quality of life.

While the impact of socioeconomic status on the pandemic has been studied in previous work [18], there was limited quantitative analysis available. The correlation between socioeconomic factors and the affected regions is yet to be exhaustively investigated. Studies have shown the disparity in the spread of the virus among different demographic groups [39], see "Fig. 1". Conclusions were made based on this superficial observation in both mainstream media and academic research about how people from certain racial backgrounds were more susceptible to the virus, but the key element behind that peculiar observation is often the underlying socioeconomic factors in their residential regions. Therefore, it is important to identify those factors so that policymakers can locate more vulnerable geographical areas and take necessary measures to help these communities.

In this paper, we focus solely on static socioeconomic features collected from Census data to train our model to predict areas that are most and least vulnerable to COVID-19 or similar future pandemics. Since our goal does not in-

volve future predictions, we avoid using time-series forecasts. Instead, we are more interested in answering the following research questions:

- Can a model predict high risk areas of COVID-19 infection in an area based on socioeconomic features alone?
- Can we accurately identify which socioeconomic features have the most and least effect on the spread?

Hence, we use machine learning models along with a GCN model to predict the high risk areas of COVID-19 spread based on zip codes using data collected from the U.S. Census and local states health departments. Then, we rank the zip codes according to the severity in infection ratio and analyze the performance of our methodology to identify those places.We also provide 4 case studies on four U.S. state along with quantitative and quantitative analysis in hopes that policymakers would explore the impacts of socioeconomic disparities on COVID-19 spread in our society and allocate needed resources ahead of time in similar future situations.

The main contributions of this work are summarized as follows:

- We present a model that exclusively uses comprehensive socioeconomic factors to predict the high-risk areas of COVID-19 infection based on fine-grained static spatial analysis.
- Identify the socioeconomic factors that have the most and least impact on the spread of COVID-19 in a community.
- We build a dataset combining local states' COVID-19 statistics and their associated socioeconomic features on the zip code level.

II. Related Works

Previous research in this area focused on COVID-19 spread forecasting and distribution. Furthermore, papers that examine how socioeconomic factors affect individuals' health fell short in the number of factors they consider. We discuss them in detail below.

COVID-19 modeling and forecasting: The importance and scale of the pandemic globally have attracted researchers to work on modeling the spread of COVID-19. Fong et al. [14] propose an optimized forecasting model which utilizes a polynomial neural network with corrective feedback to forecast the outbreak of COVID-19 using a small dataset. Kumar et al. [24] proposes a forecasting model using ARIMA and Prophet time series model. Moreover, Melin et al. [32] present a multiple ensemble neural network model with fuzzy response aggregation for COVID-19. The model was used to predict COVID-19 time series in Mexico on state and country levels. Ramchandani et al. [41] proposed a Deep Learning model using a combination of temporally static(e.g., census characteristics) and dynamic features to predict future data about the pandemic. Most of the work in this regard uses a combination of dynamic features (e.g., mobility flow) and historical pandemic data to make forecasting for a certain number of days into the future. Therefore, the effect of static features of a particular region on the spread of the disease has not been explored. Our work addresses this issue by considering a set of static socioeconomic features that have the most proven impact on the spread of COVID-19 see Fig. 2.

Socioeconomic factors and health: Previous research has explored the relationship between socioeconomic factors and health-related issues. Sethi et al. [47] propose a model that integrates geographical, socioeconomic, behavioral, demographic, and healthcare indices on a county-level resolution to discover factors of the longevity gap in the United States. Babar et al. [3] study three disease outbreaks in Pakistan to better understand environmental and socioeconomic factors impact on them. They perform independent factor analysis using decision tree and logistic regression to show the association of factors with disease outbreaks. However, they do not leverage the spatial dependency of the affected areas for their predictive model. When it comes to COVID-19 spread, researchers have explored static features like comorbidity and ethnicity along with some socioeconomic factors [38]. However, we focus solely on socioeconomic factors as the use of ethnicity can often be misleading due to the underlying socioeconomic conditions of the localities where people of a certain race are more in number. Furthermore, existing research solely explored the effect of socioeconomic factors on the severity of COVID-19 outbreak [29], [33], but their works are insufficient since they do not consider the various categories of socioeconomic factors that can affect the spread. These works depend on a factor called Socioeconomic Status (SES) derived using a predefined formula that takes select few factors. Our work differs significantly, as it focuses on exploring how each of the different socioeconomic factors correlates to the spread and how it can be leveraged to build a predictive model that learns the most important features and the spatial dependency among multiple different regions. Hence, we address the gap by determining more than 30 socioeconomic factors correlated with COVID-19 spread. We use machine learning models and a Graph Convolutional network (GCN) to predict the infection ratio of the areas at risk of COVID-19 in fine-grained spatial analysis.

III. Determinants of the COVID-19 Cases (Features)

Socioeconomic features in our research are used to determine COVID-19 high risk regions. Those features have been selected using the knowledge of a domain expert and backed with studies from the medical and social science fields (see sec II. Previous research focused on race and ethnicity in relation to the spread of COVID-19 due to higher infection rates in regions where minority groups resided [9], [46]. We believe that there has been a misconception in the exclusive association between race and increased positive cases of COVID-19, and the socioeconomic variables shared among those regions should be investigated. In this study, we group the investigated socioeconomic factors into categories. For each category, detailed features are used to describe how each factor is surveyed. Our resulting categories include: education,

employment, income, home/rent values, house characteristics, healthcare, transportation, and area characteristics.

The socioeconomic factors can disallow a large portion of people from practicing the safety precautions of COVID-19 and avoid environments with a high probability of infection. The level of education in a community has been overlooked in previous studies. Although higher education is not the only characteristic required to obtain a job to enable employees to carry out their jobs remotely during the pandemic, it is one of the leading means. Moreover, education levels play a role in helping people think critically about social matters in general and conspiracy theories in particular [50]. Believing in conspiracy theories that promote misinformation like "COVID-19 is a hoax" may result in serious implications [21], [50]. In this work, three traits were selected to represent the education level in a community: the number of people over 18 with less than high school degree, people over 25 with a minimum of bachelor's degree, and without a bachelor degree. In addition, lower education levels have been associated with higher COVID-19 mortality rates [17]. It is not necessary that lower education levels could be the source factor in the spread, but it is crucial to include the category in the study.

Retail employees, security personnel, farmers, and other workers from the service industry are essential workers. Their jobs compel them to be at risk of exposure to COVID-19 more than others. Similarly, healthcare workers and protective service workers are also at a higher risk of unavoidable exposure [4]. These front-line jobs require working onsite, which makes physical distancing impossible due to frequent interactions with others. Besides, some workers who experience some COVID-19 symptoms try to avoid taking tests because they may be afraid to lose their jobs, have no insurance, or may not be able to have paid sick leaves. As a result, they must be physically present at work to fight for unemployment and risk others the exposure to COVID-19 infection. According to Pew research center, 25% of U.S. adults report that they or someone in their household was laid off during the pandemic [36]. Thus, employees in similar situations have high chances of being exposed or being carriers of COVID-19 [16]. Our work added the following traits related to employment category for each zip-code in the states of IL, MN, FL, and VA: *number of insured people, number of people in service jobs, production/transportation/material jobs, private/wage salary jobs, management/business/science/arts jobs, government jobs, and public administration jobs*.

The spread of COVID-19 cycle continues for those who cannot afford to choose quarantine over the risk of exposure. Blue-collar workers generally depend on public transportation to get to their jobs [10]. No matter how a workplace can afford to apply safety measurements to mitigate the exposure of COVID-19 in work environments, those employees who cannot afford to have car and use public transportation can have a high risk of exposure to COVID-19. For example, a person travelling through public transportation (a bus) can be a super spreader source of COVID-19 [30]. Therefore, the number of *houses with no vehicles available and houses with a vehicle or more* is considered in the socioeconomic traits.

A study conducted in the U.K. concluded that those with the lowest household income were six times less likely to be able to work from home and three times less likely to self-isolate [2].Consequently, essential workers not only risk themselves but also their family members of exposure to COVID-19, unless their residence is large enough to isolate and conform to social distancing measures. Unfortunately, not all workers can afford bigger residences such as one room for each person. Therefore, this work considers the households' characteristics and home/rent values and their impact on the spread of COVID-19. For instance, a worker comes back from his job using public transportation to their family of 4 members living in a 2-bedroom apartment. If the person is exposed to COVID-19, it would be hard for him/her to isolate without risking the rest of the family members. Thus, overcrowding or poor housing conditions increase the chances of spreading the virus. Household crowding is a situation where "the number of occupants exceeds the capacity of the dwelling space available" [35]. Multiple studies have concluded that there is a strong association between household crowding and the high risk of COVID-19 infection [7], [8], [13], [40]. Our study considers household characteristics traits through the following statistics for each zip-code in the four U.S. states such as: *number of houses with 4 or more people occupied household, 1 or more occupants per room, 1 or less occupants per room, 3 or fewer people occupied household, the average population per house, total housing units with less than or equal to 4 rooms, total housing units with less than or equal to 5 rooms, total housing units* . Additionally, studies show that homeowners are more likely to work from home and self-isolate than those who rent or live in shared apartments [2]. Hence, we also considered some owned homes and rent value traits in our study: *total occupants per house units, total occupied units paying rent, total owner-occupied units, home values above 200k, and total occupied units paying rent over 3k*.

There are other socioeconomic factors connected to the spread of COVID-19 that we included in our study. Poverty levels soar in disadvantaged communities, especially in situations where area deprivation is prominent. Such areas are described as areas where people are clustered with limited possibilities for choosing residence destinations [37]. People are more likely to be infected with COVID-19 in those areas [13], [31], [34]. The poverty socioeconomic traits are included as well; *population density, the persons below 50%, 100%, and 150% of the poverty level*. Distance of food access in an area is one of the socioeconomic determinants that may oblige some people who seek safety from COVID-19 to leave their homes to obtain their necessities. For instance, someone who lives in a secluded area (safe from the spread of COVID-19) with a family of 5 may leave to the nearest food access (supermarket), which is located 10 miles away. That person has a high risk of infection due to that visit to shop for groceries [25]. Thus, we include *areas with food access of greater than 5 miles* in our analysis.

TABLE I: Chosen Socioeconomic factors taxonomy and examples after testing the hypothesis that there is a correlation between each socioeconomic factor and the cumulative number of COVID-19 cases

Category	Feature	Cor	95% CI
Area characteristics	Poverty $< 150\%$	0.188	(0.157 to 0.212)
	Total housing units	0.154	(0.123 to 0.185)
	Total housing units ≤ 4 rooms	0.215	(0.184 to 0.245)
	Food Access	0.075	(0.043 to 0.106)
	Population Density	0.233	(0.203 to 0.263)
Education	25+ years with less than highschool degree	0.237	(0.206 to 0.267)
	25+ years with minimum bachelor's degree	-0.06	(-0.094 to -0.03)
Employment	Service jobs	0.134	(0.102 to 0.165)
	Production/transportation/material job	0.061	(0.029 to 0.093)
	Private wage salary job	0.106	(0.073 to 0.137)
	Population not working from home	0.117	(-0.085 to -0.15)
Healthcare	Uninsured Civilian noninstitutionalized	0.187	(0.155 to 0.217)
Home/rent values	Total occupied housing units	0.158	(0.123 to 0.189)
	Owner occupied housing units	-0.29	(-0.322 to -0.26)
	Total occupied units paying rent	0.302	(0.272 to 0.330)
House characteristics	1 or more occupants per room	0.283	(0.253 to 0.312)
	Average population per house	0.209	(0.179 to 0.234)
	Houses with no vehicles available	0.238	(0.207 to 0.267)

The socioeconomic factors investigated in this paper have been carefully selected by a domain expert. We provide the dataset for reproducibility and explain each factor in detail[1]. A subset of these factors is shown in TableI.

IV. Methods

In this section, we describe the methodology. Our objective is to predict the ratio of the population infected in a zipcode given a set of static socioeconomic features discussed in Section III. First, we use three traditional regression models running in the transductive manner. To tackle the issue of data paucity at the zipcode level we enhance the data with our proposed novel data augmentation method. Then we introduce a Graph Neural Network model under inductive learning settings to leverage the spatial correlation among different zipcodes.

A. Pseudo-Zipcode Data Augmentation

Fine-grained zip code level predictions of the infection ratio are highly desired for risk evaluation of spatial areas. Unfortunately, not all states report COVID-19 data on the zip code level. Moreover, the number of zip codes in each state is often insufficient for training. To solve this problem, data augmentation was used to overcome this limitation. *Data augmentation* is an approach to increase the diversity and size of training samples from the original data. Data augmentation makes collecting new data dispensable when not enough data is available. Moreover, data augmentation can increase the models' prediction accuracies and enhance the robustness of a model toward overfitting [11], [12], [22], [26], [28], [49]. We applied *permutation* [28], [49] and *shuffling* [48] a combined pseudo-zipcode data augmentation methods that have enhanced accuracies of other models and which can be further extended for any level of a geographical area such as county, tract or census block). Given the data of two locations, it creates a new pseudo data point where the mapping between the features and the label still makes sense.

Um et al. [49] proposed permutation as a method for data augmentation, and shuffling was applied by [48]. Permutation rearranges all elements of the dataset, dividing them into equal segments and permute each segment. Shuffling randomly rearranges all elements of the dataset instead of dividing the data into segments. In our data augmentation, we applied permutation and shuffling. To describe how the data is augmented, a description of the structure of the collected data from Florida state is shown below. SE_f represents the full data structure in the shape of a matrix. The columns of the matrix represent the socioeconomic features, and the last column is the cumulative COVID-19 cases. The rows denoted as n are all the zip codes in Florida state.

$$SE_f = \begin{bmatrix} a_1^f & b_1^f & \dots & z_1^f \\ a_2^f & b_2^f & \dots & z_2^f \\ \vdots & \vdots & \vdots & \vdots \\ a_n^f & b_n^f & \dots & z_n^f \end{bmatrix}$$

Permutation is implemented by dividing the rows of matrix SE_f into s equal sized segments, and the size of each segment is n/s. The matrix SE_f becomes the segmented matrix $SE_f^{'}$

$$SE_f^{'} = \begin{bmatrix} SE_f(1) \\ SE_f(2) \\ \vdots \\ SE_f(S) \end{bmatrix}$$

where

[1]github link hidden

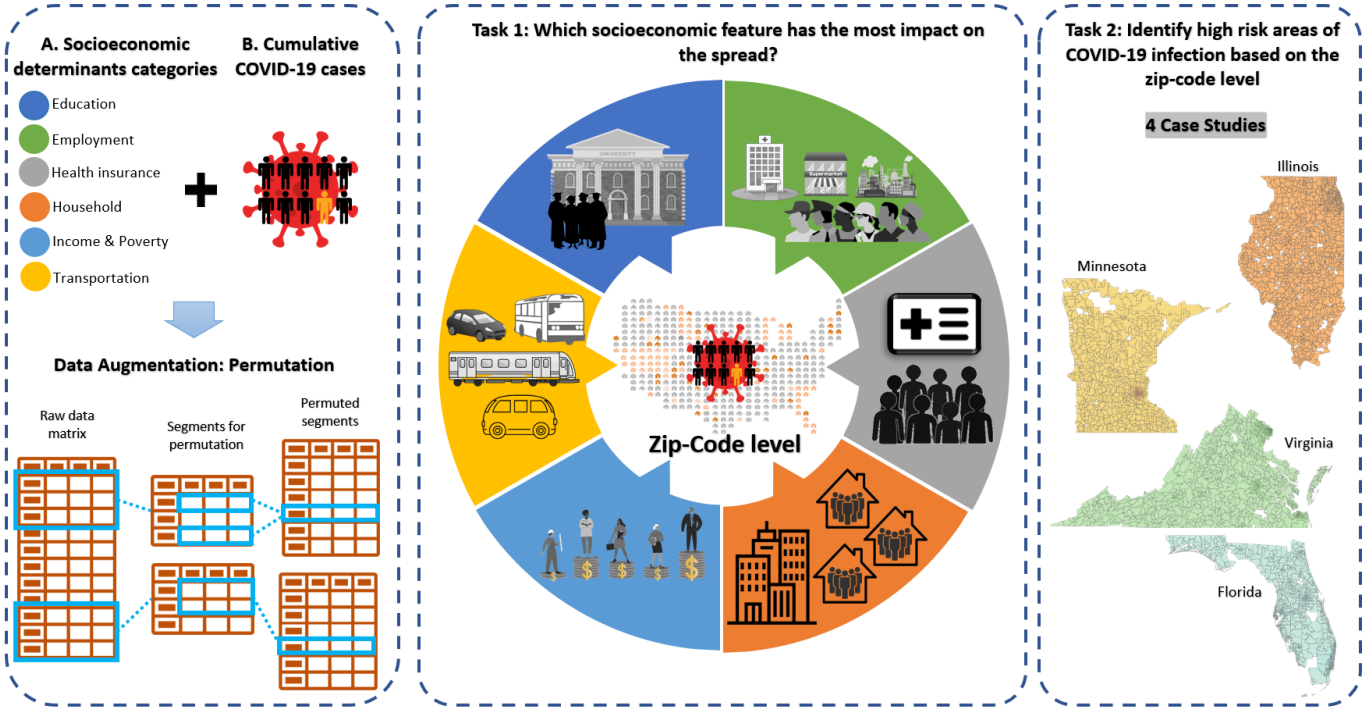

Fig. 2: The figure explains the overview of the work. The top left section shows that collected and merged dataset from socioeconomic factors and cumulative COVID-19 cases for each zip code. The bottom left explains data augmentation by applying permutation. The middle section defines the first task which is identifying the socioeconomic features of most impact on COVID-19 spread. Each part of the circle represents a socioeconomic category that is related to the spread of COVID-19. The middle map is a U.S. map. The right section shows that four case studies are conducted to show the effectiveness of the work.

$$SE_f(i) = \begin{bmatrix} a^f_{(i-1)n/s+1} & b^f_{(i-1)n/s+1} & \cdots & z^f_{(i-1)n/s+1} \\ a^f_{(i-1)n/s+2} & b^f_{(i-1)n/s+2} & \vdots & z^f_{(i-1)n/s+2} \\ \vdots & \vdots & \cdots & \vdots \\ a^f_{in/s} & b^f_{in/s} & \cdots & z^f_{in/s} \end{bmatrix}$$

for each $SE_f(i)$ $(i \in [1, s], s = 10)$, the rows are permuted, and each permutation involves two rows (zip codes') features to augment the data. During permutation and shuffling the features of each zip code is merged using a merging function depending on their type.

Feature and label merging: Given the features of zip code $a^f_{(i-1)n/s+1}$ and $a^f_{(i-1)n/s+2}$, the feature merging is defined as $x_{i-j} = F(x_i, x_j)$ where F is the merging function (e.g. average, maximum). The merging function must be defined in respect with the specific features and label.Sometimes additional information are needed for the merging. That additional values are different for different categories of feature. Let a be the feature. The merging function can be expressed as:

$$F(x_i, x_j)[a] = \frac{x_i[a] * I(x_i) + x_j[a] * I(x_j)}{I(x_i) + I(x_j)]} \quad (1)$$

Where I(x) is the additional information corresponding to the zipcode x. I(x) can be different based on the categories of features as described in Table I.

- if a is **Population Desnity** then then I(x)=area in square miles.
- With the exception of Population Desnity, if a belongs to the category of **Area Characteristics** or **Healthcare** feature as mentioned in sections III then I(x)=Population
- if a belongs to the Employment category of features then I(x)=Number of Workers over 16
- if a belongs to **"Home/Rent Values"** or **"Housing Characteristics"** then I(x)=Number of Housing Units

In practice, we only use state-level 2nd order Pseudo-Zipcode data augmentation. For every state with N number of zip codes, we create n number of bins such that $n = N/10$. Then, within each of those bins, we take every possible combination of two zip codes to do the merging task. The data can be augmented from $O(n)$ to $O(n^2)$ for each state. This method can also be applied in higher orders by changing the value of n in order to generate more data if needed.
Along with the Pseudo-Zipcode Data Augmentation method, we propose to use the following 3 models for the risk evaluation task in the experiments. Following are the models used on both unaugmented and augmented data to predict the ratio of infected people.

a) Ridge Regression: Ridge with data augmentation (Ridge-DA) [19]. Ridge is the linear model with L2 regularization which theoretically improves the basic linear regression model. L2 regularization helps in giving larger co-efficient to the most important features.

b) SVR: This is a version of SVM proposed for regression analysis. SVM classifier focuses on building a decision boundary even on a higher dimension for non-linearly separable data. Hence the model produced by the classification algorithm focuses on a subset of the training sample, ignoring the training points lying beyond the margin. Similarly, the regression model depends only on a subset of the training sample, ignoring any training data close to the model prediction for the cost function. As the proposed problem has a multitude of features that may not be linearly separable in a lower dimension, one needs to consider SVR for a solution to the regression problem.

c) LightGBM: It is a subtype of Gradient Boosting Decision Tree, which is efficient and robust for large datasets. We have an augmented dataset with 21000 data points in the experimental setup, making lightGBM an ideal candidate for the problem. The two novel technique introduced by Ke et al. [23] which makes it unique from other Gradient Boosting algorithms are Gradient based One Sided Sampling(GOSS)and Exclusive Feature Bundling (EFB). GOSS helps reduce the number of training data by excluding the instances with smaller gradients, and EFB helps reduce the dimension by Bundling the mutually exclusive features

B. Graph Neural Network Model

According to [51], the degree of risk has strong spatial dependencies, which, if utilized, could help make more accurate estimations on the risk associated with the zip codes. Graph Convolutional Network (GCN) is a recently emerging technique that integrates graph structure and node attributes. It has been proven, that GCN is effective for handling spatial dependencies. [44] We formulate the problem into an inductive semi-supervised regression problem on an undirected graph. We consider the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with n nodes. The nodes $\mathcal{V}$ represent the zipcodes, and the edges $\mathcal{E}$ represent the spatial dependency relationships. The edges can be represented in the adjacency matrix: $\mathcal{A} = [A_{ij}] \in \{0,1\}^{n \times M}$. We hypothesize that the closer a zipcode is to another, the higher the spatial dependency will be. Hence, edges are defined according to the distance between two zip-codes. The distance is calculated according to haversine. We try to determine a threshold d where $e_{(i,j)} = 1$ when $distance(i,j) < d$. While increasing d, we calculate the number of connected components ($\mathcal{N}$) in the graph. In the beginning, $\mathcal{N} = n$ because d is too small to connect any nodes. We stop increasing d when the $\mathcal{N}$ reaches a certain level that is determined by domain experts. $X \in \mathcal{R}^{n \times M}$ represents the M features described in section III.

After the construction of the graph, the GCN model enables the node-level information to propagate according to the neighborhood relationships. One layer of the propagation rule is defined as the following convolution:

$$g_\theta * X \approx (\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}) X \Theta \tag{2}$$

This is parameterized with Θ, with $\tilde{A} = A + I$, where I is the identity matrix and $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$. By stacking two layers of the graph convolutions, our final model is defined as follows:

$$Y = F(g_{\theta_2} * \sigma(g_{\theta_1} * X)) \tag{3}$$

Here, F is the linear layer for the regression task, and σ is the activation function.

V. Experiment

This section describes data collection, experimental settings, and the subsequent result and analysis.

A. Data Collection

The analysis in this research is based on two types of data: cumulative COVID-19 cases and combined socioeconomic statistics. The data has been collected for four states FL, IL, MN, and VA, because of their availability based on the zip code level. Here are more details about the datasets:

COVID-19 cases Dataset: For each of the four states, the number of cases has been collected, individually, from the official sources, which are the Department of Health of each state, as shown in Table II. Each dataset includes the cumulative (total) number of COVID-19 cases for each zip code from the beginning of the pandemic in the U.S. until the data of the collection 11/09/2020.

Socioeconomic Dataset: While other works (see sec II) have focused on individual groups socioeconomic factors to find the association between disadvantaged communities and the spread of COVID-19 cases, we utilize a COVID-19 related comprehensive group of socioeconomic dataset from multiple resources combined by the zip code to give more accurate predictions of the high risk areas of infections. In Table II, 6 datasets are taken from the U.S. Census Bureau and one from the U.S. Department of Agriculture. We use subsets of these datasets to extract more than 50 distinct socioeconomic features that are used in our experiments. [2] The food access datasets were initially on a census tract level. We used tract to zip conversion method to calculate it for all the zip codes.

B. Experiment Settings

For transductive learning, we implement two different experimental settings. The first one does not use the augmented data and splits the dataset with an 80:20 ratio into the training and testing sets. To see if we can tackle the lack of data and make predictions better. In the second setting, we use the augmented dataset for training and the original dataset for testing. In both settings, the data is standardized, and we tune the hyperparameters using grid search method [27]. Furthermore, to test whether incorporating spatial dependency

[2]Hidden github link

TABLE II: Data Sources

Data	Source	Details
COVID-19 Dataset		
Florida COVID-19 cases	Florida Department of Health	Feb-11/9/2020
Minnesota COVID-19 weekly report	Minnesota Department of Health	Feb-11/9/2020
Illinois COVID-19 statistics	Illinois Department of Health	Feb-11/9/2020
Virginia COVID-19 Cases & Tests	Virginia Department of Health	Feb-11/9/2020
Socioeconomic Dataset		
Health Insurance Coverage	U.S. Census Bureau (ACS)	2019
Housing Characteristics	U.S. Census Bureau (ACS)	2019
Occupancy Characteristics	U.S. Census Bureau (ACS)	5-year 2018
Poverty Status	U.S. Census Bureau (ACS)	2019
Educational Attainment	U.S. Census Bureau (ACS)	2019
Means of Transportation to Work	U.S. Census Bureau (ACS)	2019
Food Access Research Atlas	U.S. Department of Agriculture	2015

will improve the performance, we use GCN in the inductive learning setting. We run our experiment on GCN by using 1,3,10,30, and 80 percent of nodes for training, respectively.

As the study aims to investigate predicting the ratio of the population infected (COVID-19 cases ratio) based entirely on static socioeconomic factors, we lack similar works for baseline comparison. Most of the related literature to COVID-19 modeling focuses on investigating the dynamic nature of the spread rather than the static nature of the vulnerability of a population. Some works have chosen very few of these factors or sometimes, even a single factor for their feature-set, but their objective was different [5], [42], [43]. Similarly, Mena et al. [33] utilizes a term called socioeconomic status(SES) which is a predefined formula using select few features and their goal was not to predict the ratio of population infected like ours is. Hence, we use the following as the baseline methods for comparison:

- Population-based: a non-socioeconomic feature (population density)
- Population-income-based: a socioeconomic feature and population density
- Single-category-based: uses a single category of socioeconomic features(Income and Poverty rate)

As we are predicting COVID-19 ratio, metrics like MAE or RMSE will not be meaningful as they will be minimal and do not make for a good comparison as it denotes an exact value and does not provide a comparative score more useful. However, Pearson's r [6] can be used as a metric to give a more standardized understanding of the performance of different methods as it is also more robust to outliers [45].

C. Experimental Results

The results for Transductive Learning models are given in Table III. From the comprehensive set of features' categories mentioned in Table I, if one uses the features or category of features used in related work in isolation, the prediction of COVID-19 infection ratio will not be satisfactory as per our experiments. Hence, we perform Lasso Feature Selection [15] to select the most and least impactful features from the comprehensive list of 50 socioeconomic features. Fig 3 displays the most and least influential features. The values in

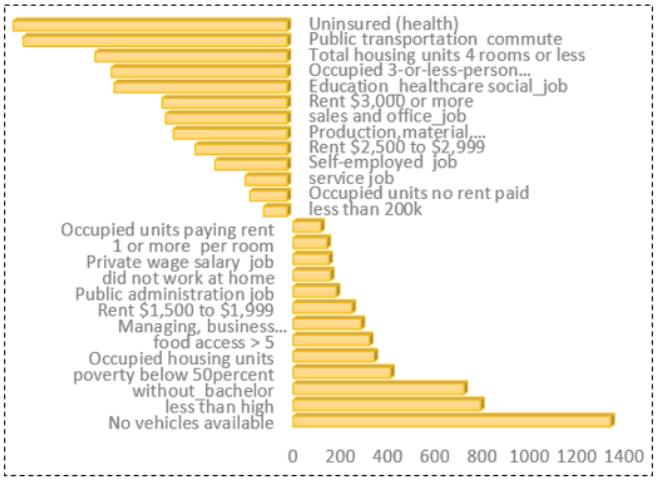

Fig. 3: The figure displays the top socioeconomic features that have the most and least impact on COVID-19 spread ranked according to their coefficient in lasso. The bars to the left shows the ones with the negative correlation on the spread, and the bars to the right shows the ones with the most impact on the spread (bottom is the most).

the scale denote the coefficient assigned to each of the selected features. As we can see, the most positively correlated features are *population with no vehicles available* and population with *education level less than high school* or *without a bachelor's degree*. In addition, the areas that suffer from poverty were mostly affected by the spread. Followed by the overcrowded places, areas that have no food access within a radius of 5 miles, and other factors related to jobs and income. However, surprisingly, we also see a strong negative correlation with people availing public transport and the uninsured population. Given that the feature called *no vehicle available* has a strong positive correlation with the infection ratio, one could think that *workers availing public transportation* will have a similar effect. However, in the census data, *no vehicle available* corresponds to the households with no vehicles available. That indicates more towards a subsection of people who avails public transportation to go to work. As public transports

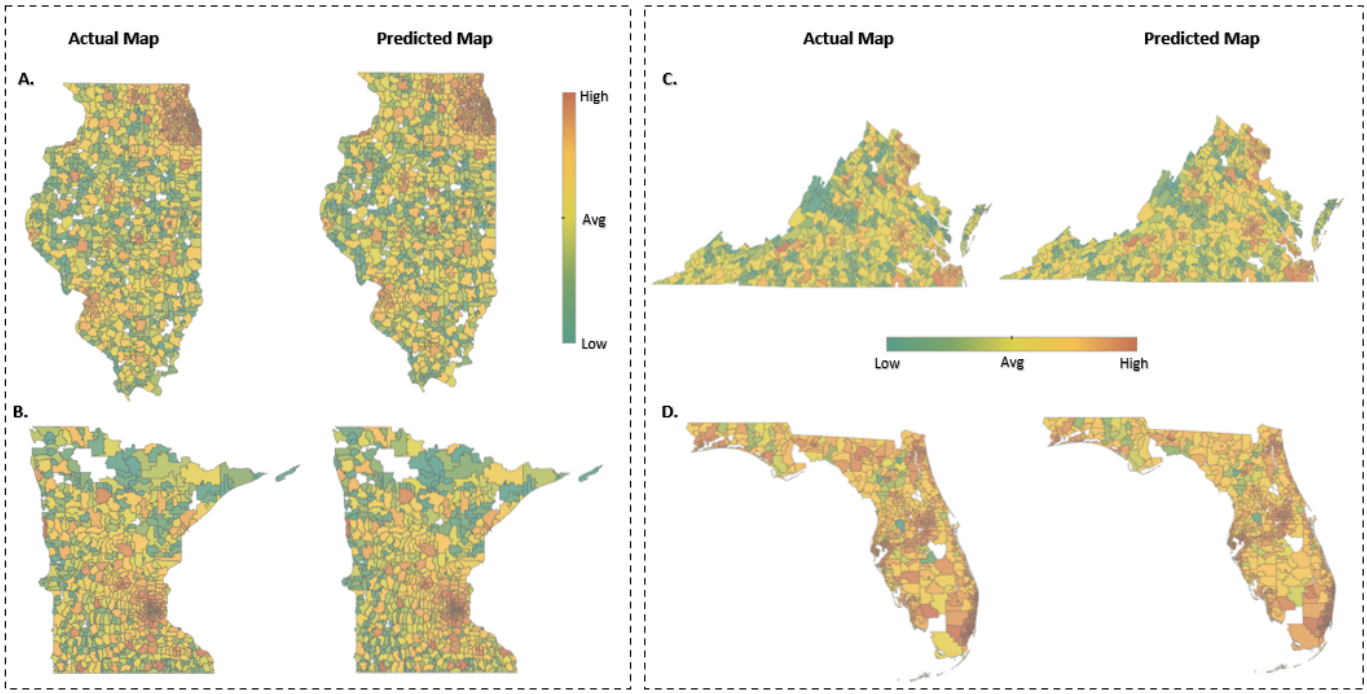

Fig. 4: The figure displays the results of 4 case studies of the following U.S. states: (A) Illinois, (B) Minnesota, (C) Virginia, (D) Florida. For each state, the map on the left is the actual map of the zip codes that have COVID-19 cases from the highest to lowest zip code areas, and the map on the right is the predicted map based on our experiment. High risk areas are colored in red, and low risk areas are colored in green. The results are based on the data collected till November 2020

were mostly made infrequently and due to social distancing protocol, a possible explanation for the observation could be that a lot of the people availing subway or buses started working from home or started availing cabs or ride-share services where there is less chance of infection. In addition, another unexpected result that appeared as one of the most negatively correlated features toward the spread of COVID-19 is the population without health insurance. However, it is evidence of the fact that uninsured people might be wary of going to the testing center due to their worries of being laid off of work [36]. Another reason for that may also be because of a lack of information about the nature of the healthcare service.

The performance with the three base machine learning models described in Section IV is better but due to the paucity of zip code level data, the performance is still unsatisfactory. Hence, we implement our data augmentation method and as shown in Table III, a significant improvement in performance is observed while using the Augmented Dataset(Pearson's r 0.546). Also, the Gradient Boosting Decision Tree-based algorithm LightGBM-DA outperforms the linear and kernel-based models. This is a potential solution for COVID-19 prediction on a smaller geographical region when challenged with small dataset.

The results using GCN models are shown in Table IV. GCN(α) indicates the GCN model with α ratio of the training set. For example, GCN(80%) is the GCN model that used 80% of the nodes as a training set, 10% as the validation set (which remains fixed), and the rest 10% as the testing set. GCN(80%) can achieve almost 70% Pearson's r. It is not surprising that GCN outperforms the other models because GCN models are inductive models trained in a semi-supervised manner. Even though GCN used additional information, the coordinates, inductive learning by nature is easier than transductive learning. The drawback of GCN model is that it cannot be applied to new data points. However, even with only 10% of labeled data in the training set, the GCN model can achieve better results

TABLE III: Transductive learning Model Comparison. The top third of the table describes the baselines using the set of features used in related work. The models Ridge, SVR and LightGBM shows the results of the respective models without the use of data augmentation. Ridge-DA, SVR-DA and LightGBM-DA indicates the same models used on augmented data.

Model	Pearson's r
Population-based	0.24
Population-Income-based	017
Single-Category-based	0239
Ridge	0.249
SVR	0.285
LightGBM	0.332
Ridge-DA	0.271
SVR-DA	0.326
LightGBM-DA	**0.546**

TABLE IV: Results of Inductive Learning with GCN. Following are 5 experimental settings of GCN based on the percentage of nodes used as training data

Model	Pearson's r
GCN(80%)	0.695
GCN(30%)	0.586
GCN(10%)	0.58
GCN(3%)	0.490
GCN(1%)	0.297

(0.58 Pearson's r) than all the transductive learning models in Table III. Our preliminary finding will enable researchers to extend this research to much larger geographical areas, which will benefit policymakers in identifying high risk areas (zip codes in our case).

D. Qualitative Analysis

We show the effect of using a comprehensive set of socioeconomic features to predict the high risk areas of COVID-19 spread on a fine-grained spatial region by evaluating the performance of our proposed methods. For a given set of zip codes of any area and a set of socioeconomic features, our work exploits the COVID-19 predicted infection ratio to identify the severity of infection in each of the given zip codes. We perform four cases (four U.S. states) to evaluate the method.

Case study 1 (Illinois): We perform a case study on the state of (IL), the most impacted city Chicago and its suburban areas to validate the merit of our proposed method. Fig 4 shows a side-by-side comparison of the ground truth and the predicted results of more than 100 zip codes. The colors in the maps indicate the severity of infection of each zip code based on the predicted infection ratio. We grouped the first top ten high risk zip codes (highest cases of COVID-19) from IL state and found that our method has successfully predicted those areas as well. Those zip codes are 60629, 60639, 60632, 60804, 60623, 60085, 60634, 60402, 60505, and 60641. The maps (A) in Fig 4 show the accuracy of the predicted map in comparison to the actual map.

Case study 2 (Minnesota): For the case study of Minnesota, our work identifies the hot spots of COVID-19 infections based on the socioeconomic features for the zip-code level. The map in Fig 4 shows a barely different map of the actual and the predicted map of the most and least cases of COVID-19. The work identified the following zip codes as the top ten areas COVID-19 has affected: 55106, 56187, 56560, 56001, 55407, 55117, 56301, 55404, and 55443. The top ten of our predicted zip codes are among the top 15 on the ground truth map.

Case study 3 (Virginia): In the state of Virginia, COVID-19 spread is clustered in some areas in the south east and north east, but there are some difficult zip code areas that have been highly infected by COVID-19 and our work has successfully identified those areas as seen in the maps (C) in Fig 4. The following zip codes are identified as high risk areas of COVID-19 spread and are in the top 10 areas of ground truth map: 22191, 22193, 24060, 20110, 20164, 23234, 22003, and 22204.

Case study 4 (Florida): The fourth case study is for the state of Florida. As displayed in the left ground truth map in (D) in Fig 4, Florida is one of the most highly affected states in the U.S. because many of its areas are affected by the spread, and many out-of-state visitors are present in the area for a short period. Even though it is a harder to predict the hot spots of the spread compared to the other 3 states, our work has predicted 6 out of the top 10 hot spots in FL. The identified zip codes are 33125, 33012, 33126, 33015, 33142, and 33165.

VI. Conclusion

We propose a work that exclusively uses comprehensive socioeconomic factors to predict the high risk areas of COVID-19 spread based on finely-grained static spatial analysis. We then rank zip codes according to the severity of the infection ratio and test the accuracy of our model's ability to identify those places. Further, this work identifies the most and least influential socioeconomic factors on COVID-19 spread in a community. Extensive experiments show that our methods accurately predict high risk infection areas based on stable features. Our work can be utilized by authorities to predict COVID-19-like high risk areas in the future to take proper precautions.

References

[1] Aaron Antonovsky. Social class, life expectancy and overall mortality. *The Milbank Memorial Fund Quarterly*, 45(2):31–73, 1967.

[2] Christina J Atchison, Leigh Bowman, Charlotte Vrinten, Rozlyn Redd, Philippa Pristera, Jeffrey W Eaton, and Helen Ward. Perceptions and behavioural responses of the general public during the covid-19 pandemic: A cross-sectional survey of uk adults. *medRxiv*, 2020.

[3] Z Babar, A Mannan, F Kamiran, and A Karim. Understanding the impact of Socio-Economic and environmental factors for disease outbreak in developing countries. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 124–131, November 2015.

[4] Marissa G Baker, Trevor K Peckham, and Noah S Seixas. Estimating the burden of united states workers exposed to infection or disease: a key factor in containing risk of covid-19 infection. *PLoS One*, 15(4):e0232452, 2020.

[5] Olivier Bargain and A Ulugbek. Poverty and covid-19 in developing countries. *Bordeaux University*, 2020.

[6] Kenneth A Bollen and Kenney H Barb. Pearson's r and coarsely categorized measures. *American Sociological Review*, pages 232–239, 1981.

[7] Kevin A Brown, Aaron Jones, Nick Daneman, Adrienne K Chan, Kevin L Schwartz, Gary E Garber, Andrew P Costa, and Nathan M Stall. Association between nursing home crowding and covid-19 infection and mortality in ontario, canada. *JAMA internal medicine*, 2020.

[8] Jarvis T Chen and Nancy Krieger. Revealing the unequal burden of covid-19 by income, race/ethnicity, and household crowding: Us county versus zip code analyses. *Journal of Public Health Management and Practice*, 27:S43–S56, 2020.

[9] Merlin Chowkwanyun and Adolph L Reed Jr. Racial health disparities and covid-19—caution and context. *New England Journal of Medicine*, 2020.

[10] Madelaine Criden. The stranded poor: Recognizing the importance of public transportation for low-income households. *National Association for*, 2008.

[11] Sumeyra Demir, Krystof Mincev, Koen Kok, and Nikolaos G Paterakis. Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. *Applied Energy*, 304:117695, 2021.

[12] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.

[13] Ukachi N Emeruwa, Samsiya Ona, Jeffrey L Shaman, Amy Turitz, Jason D Wright, Cynthia Gyamfi-Bannerman, and Alexander Melamed. Associations between built environment, neighborhood socioeconomic status, and sars-cov-2 infection among pregnant women in new york city. *Jama*, 324(4):390–392, 2020.

[14] Simon James Fong, Gloria Li, Nilanjan Dey, Rubén Gonzalez-Crespo, and Enrique Herrera-Viedma. Finding an accurate early forecasting model from small dataset: A case of 2019-nCoV novel coronavirus outbreak. *IJIMAI*, 6(1):132, 2020.

[15] Valeria Fonti and Eduard Belitser. Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, 30:1–25, 2017.

[16] Elise Gould and Valerie Wilson. Black workers face two of the most lethal preexisting conditions for coronavirus—racism and economic inequality. *Economic Policy Institute*, 1, 2020.

[17] R.B. Hawkins, E.J. Charles, and J.H. Mehaffey. Socio-economic status and covid-19–related cases and fatalities. *Public Health*, 189:129 – 134, 2020.

[18] Robert B Hawkins, Eric J Charles, and J Hunter Mehaffey. Socioeconomic status and coronavirus disease 2019 (covid-19) related cases and fatalities. *Public health*, 2020.

[19] Arthur E Hoerl, Robert W Kannard, and Kent F Baldwin. Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2):105–123, 1975.

[20] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.

[21] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621–1629, 2020.

[22] Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7):e0254841, 2021.

[23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[24] N. Kumar and S. Susan. Covid-19 pandemic prediction using time series forecasting models. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2020.

[25] Fan-Yun Lan, Christian Suharlim, Stefanos N Kales, and Justin Yang. Association between sars-cov-2 infection, exposure risk and mental health among a cohort of essential retail workers in the usa. *Occupational and environmental medicine*, 78(4):237–243, 2021.

[26] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. Data augmentation for time series classification using convolutional neural networks. In *ECML/PKDD workshop on advanced analytics and learning on temporal data*, 2016.

[27] PM Lerman. Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(1):77–84, 1980.

[28] Yantao Li, Hailong Hu, and Gang Zhou. Using data augmentation in continuous authentication on smartphones. *IEEE Internet of Things Journal*, 6(1):628–640, 2018.

[29] Christine Little, Mathilda Alsen, Joshua Barlow, Leonard Naymagon, Douglas Tremblay, Eric Genden, Samuel Trosman, Laura Iavicoli, and Maaike van Gerwen. The impact of socioeconomic status on the clinical outcomes of covid-19; a retrospective cohort study. *Journal of community health*, pages 1–9, 2021.

[30] Kaiwei Luo, Zhao Lei, Zheng Hai, Shanliang Xiao, Jia Rui, Hao Yang, Xinping Jing, Hui Wang, Zhengshen Xie, Ping Luo, et al. Transmission of sars-cov-2 in public transportation vehicles: a case study in hunan province, china. In *Open Forum Infectious Diseases*, volume 7, page ofaa430. Oxford University Press US, 2020.

[31] KC Madhav, Evrim Oral, Susanne Straif-Bourgeois, Ariane L Rung, and Edward S Peters. The effect of area deprivation on covid-19 risk in louisiana. *medRxiv*, 2020.

[32] Patricia Melin, Julio Cesar Monica, Daniela Sanchez, and Oscar Castillo. Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: The case of mexico. *Healthcare (Basel)*, 8(2), June 2020.

[33] Gonzalo E Mena, Pamela P Martinez, Ayesha S Mahmud, Pablo A Marquet, Caroline O Buckee, and Mauricio Santillana. Socioeconomic status determines covid-19 incidence and related mortality in santiago, chile. *Science*, 372(6545), 2021.

[34] Los Angeles County Department of Public Health. Report on la county covid-19 disaggregated by race/ethnicity and socioeconomic status. *http://publichealth.lacounty.gov/docs/RacialEthnicSocioeconomicDataCOVID19.pdf*, apr 2020.

[35] World Health Organization et al. Who housing and health guidelines. In *WHO housing and health guidelines*. World Health Organization, 2018.

[36] Kim Parker, Rachel Minkin, and Jesse Bennett. Economic fallout from covid-19 continues to hit lower-income americans the hardest. *Pew Research Center*, 21, 2020.

[37] Fredrik Niclas Piro, Øyvind Næss, and Bjørgulf Claussen. Area deprivation and its association with health in a cross-sectional study: are the results biased by recent migration? *International Journal for Equity in Health*, 6(1):10, 2007.

[38] Albert Prats-Uribe, Roger Paredes, and Daniel Prieto-Alhambra. Ethnicity, comorbidity, socioeconomic status, and their associations with covid-19 infection in england: a cohort analysis of uk biobank data. *medRxiv*, 2020.

[39] Eboni G Price-Haywood, Jeffrey Burton, Daniel Fort, and Leonardo Seoane. Hospitalization and mortality among black patients and white patients with covid-19. *New England Journal of Medicine*, 2020.

[40] Benjamin Rader, Samuel Scarpino, Anjalika Nande, Alison Hill, Robert Reiner, David Pigott, Bernardo Gutierrez, Munik Shrestha, John Brownstein, Marcia Castro, et al. Crowding and the epidemic intensity of covid-19 transmission. *medRxiv*, 2020.

[41] A. Ramchandani, C. Fan, and A. Mostafavi. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *IEEE Access*, 8:159915–159930, 2020.

[42] Santanu Roy, Gouri Sankar Bhunia, and Pravat Kumar Shit. Spatial prediction of covid-19 epidemic using arima techniques in india. *Modeling Earth Systems and Environment*, pages 1–7, 2020.

[43] Srikanta Sannigrahi, Francesco Pilla, Bidroha Basu, Arunima Sarkar Basu, and Anna Molter. Examining the association between socio-demographic composition and covid-19 fatalities in the european region using spatial regression approach. *Sustainable Cities and Society*, 62:102418, 2020.

[44] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[45] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, 2018.

[46] Thomas M Selden and Terceira A Berdahl. Covid-19 and racial/ethnic disparities in health risk, employment, and household composition: Study examines potential explanations for racial-ethnic disparities in covid-19 hospitalizations and mortality. *Health Affairs*, 39(9):1624–1632, 2020.

[47] Tavpritesh Sethi, Anant Mittal, Shubham Maheshwari, and Samarth Chugh. Learning to address health inequality in the united states with a bayesian decision network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 710–717, 2019.

[48] Odongo Steven Eyobu and Dong Seog Han. Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. *Sensors*, 18(9):2892, 2018.

[49] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 216–220, 2017.

[50] Jan-Willem van Prooijen. Why education predicts decreased belief in conspiracy theories. *Applied cognitive psychology*, 31(1):50–58, 2017.

[51] Liang Zhao, Jiangzhuo Chen, Feng Chen, Fang Jin, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. Online flu epidemiological deep modeling on disease contact network. *GeoInformatica*, 24(2):443–475, 2020.