Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A machine learning-driven spatio-temporal vulnerability appraisal based on socio-economic data for COVID-19 impact prevention in the U.S. counties

Mohammad Moosazadeh a, Pouya Ifaei a, Amir Saman Tayerani Charmchi a, Somayeh Asadi b, ChangKyoo Yoo a,*

a Department of Environmental Science and Engineering, Center for Environmental Studies, College of Engineering, Kyung Hee University, Seocheon-dong 1, Gihang-gu, Yongin-Si, Gyeonggi-Do 446-701, South Korea
b Department of Architectural Engineering, Pennsylvania State University, 213 Engineering Unit, University Park, PA 16802, United States

ARTICLE INFO

Keywords:
COVID-19 control policies
Data analysis
Machine learning
Social vulnerability
The U.S. cities

ABSTRACT

A mature and hybrid machine-learning model is verified by mature empirical analysis to measure county-level COVID-19 vulnerability and track the impact of the imposition of pandemic control policies in the U.S. A total of 30 county-level social, economic, and medical variables and a timeline of the imposed policies constitutes a COVID-19 database. A hybrid feature-selection model composed of four machine-learning algorithms is developed to emphasize the regional impact of community features on the case fatality rate (CFR). A COVID-19 vulnerability index (COVULin) is proposed to measure the county’s vulnerability, the effects of model’s parameters on mortality, and the efficiency of control policies. The results showed that the dense counties in which minority groups represent more than 45% of the population and those with poverty rates greater than 24% were the most vulnerable counties during the first and the last pandemic peaks, respectively. Highly-correlated CFR and COVULin scores indicated a close agreement between the model outcomes and COVID-19 impacts. Counties with higher poverty and uninsured rates were the most resistant to government intervention. It is anticipated that the proposed model can play an essential role in identifying vulnerable communities and help reduce damages during long-term alike disasters.

1. Introduction

The worst pandemic of the last century was first identified in Wuhan, China, in December 2019. Coronavirus disease 2019 (COVID-19) rapidly became the fifth most fatal pandemic in history. Despite initial countermeasures, the pandemic infected more than 300 million people within 24 months. According to the Johns Hopkins Coronavirus Resource Center (2021), 5.5 million COVID-19–related deaths were officially recorded in 224 countries through January 2022. In January 2020, the World Health Organization (WHO) declared the outbreak as a Public Health Emergency of International Concern (WHO, 2020). In addition to the loss of life, the global gross domestic product (GDP) contracted by 3.5% in 2020 (International Monetary Fund, 2021). Government interventions such as travel restrictions, social distancing protocols, quarantines for COVID-19 patients, and lockdown policies increased the risks of socio-psychological disorders (Xiao, 2020). Due to travel restrictions, the international tourism sector may lose more than $4 trillion over the years 2020 and 2021 (UNCTAD, 2021). The COVID-19 crisis also led to dramatic swings in household spending. Retail sales decreased by 8.7% from February to March 2020, the greatest month-to-month decline since the Census Bureau began tracking the data (U.S. Census Bureau, 2020). The COVID-19 pandemic has been particularly damaging for small businesses, which represent the majority of businesses in the country and employ nearly half the private-sector labor force (Bartik et al., 2020). These and other setbacks led to dramatic increases in unemployment in many regions, threatening even the governance of some countries (OECD, 2020). Considering the continuing pandemic-related economic damage; there is a need to investigate multiple aspects of the COVID-19 crisis, acknowledge socio-economic weaknesses in responses, and minimize potential damage from future disasters.

Since the first description of a COVID-19–related pneumonia outbreak, numerous efforts have been made to analyze the effects of the pandemic on human health and society. Some studies have proposed solutions to mitigate damage and draw a picture of the post-pandemic world and life. Clinical (Pourhomayoun & Shakibi, 2021), virologic...
(Cevik et al., 2020), epidemiological (Cooper et al., 2020), socio-economic (Josephson et al., 2021), medical (Ibrahimagic et al., 2020), and other perspectives have been applied to describe the challenge. The pattern of COVID-19 was studied as a new pandemic from the first days. Epidemiologists effectively used data-driven strategies to combat the spread of infectious diseases and used them in response to the COVID-19 (Hale et al., 2021). These emergency responses included detection, prevention, and control of the disease itself and encouraged recovery from global economic disruption (Verschuur et al., 2021). Following the initial peaks of the pandemic, data-driven models were extended to inspect other aspects of the pandemic beyond laboratory-scale medical studies.

Data-driven tools are essential to monitor, control, mitigate, and prevent spread of disease during a pandemic (Kashem et al., 2021). Mature data analytics tools can test available hypotheses and differentiate between scientific and conspiratorial theories. Hence, much research has been devoted to exploring the relationships between COVID-19 morbidity and mortality rates in different communities using statistical analyses (Chen et al., 2021). This category of the COVID-19 literature includes social variables, background risk factors of illness, and the effectiveness of restrictive policies. Ethnicity and racial background (Maiti et al., 2021), population density (Aral & Bakir, 2022; Lulbadda et al., 2021), poverty (Millán-Guerrero et al., 2020), community networks (Seto et al., 2020), and mask-wearing (Zhang et al., 2020) have been confirmed to be influential social variables, while diabetes (Corrales-Reyes et al., 2021) and asthma (Mendes et al., 2021) are recognized as background risks of illnesses. The effects of specific interventions such as stay-home (Fu & Zhai, 2021), mask mandates (Van Dyke et al., 2020), digital contact tracing data (Grekusis & Liu, 2021), social distancing (Su et al., 2021), and travel restrictions (Chinazzi et al., 2020) have been extensively discussed in different communities, and the vulnerability of poor, elderly, and racially heterogeneous populations have been already studied (Khavarian-Garmsir et al., 2021; Wadhera et al., 2020; Hu et al., 2022). Spatio econometric analysis is widely used to examine the impact of COVID-19 on mortality (Shobande & Ogbeifun, 2020). However, these methods will restrict the model by utilizing various sorts of variables. Some of the selected features, such as income, poverty, care facilities, and education could not be translated into geometric or spatio-econometric properties. Grekusis et al. (2021) showed a lower education level and annual income led the U.S counties to a more vulnerable area. Moreover, accurate forecasts of the number of infected and deceased individuals have proven to be extremely useful (Paiva et al., 2020). Mollalo et al. (2021) examined the association of case fatality rate of COVID-19 with 20 underlying diseases and showed Asthma, Hepatitis and Leukemia have positive associations with COVID-19 CFR. Furthermore, spatiotemporal analysis widely investigated the relationship between risk factors and COVID-19 mortality. Pan et al. (2021) developed a spatiotemporal analysis using Random Forest (R.F.) regression to use data-driven models in fighting the global pandemic and showed R.F. can perform acceptable outcomes in forecasting COVID-19 mortality. However, the number of features was limited, and the increasing number of features caused a decrease in the model’s accuracy. Luo et al. (2021) proposed random forest regression model to estimate the relationship between COVID-19 death rate and 47 risk factors and proved a high correlation between risk factors and COVID-19 related mortality rate. Zhou et al. (2021) showed spatiotemporal distribution and their changes could provide a beneficial support for the COVID-19 pandemic and implementation of resumption plans for sustainable development.

The previous studies suffer from five shortcomings. First, they could not monitor the long-term effects of the pandemic because they were limited to a short time, and the employed data were mainly from the first year of the pandemic. Second, they had eliminated many influential factors such as hospitalized individuals, financial status, race, poverty, and other social features that could affect the severity of a pandemic. Third, they bundled features with similar effects on mortality and analyzed the outcomes of the data bundles instead of each feature. This caused neglecting the independent impact of each variable on the vulnerability. Finally, they turned the data into a binary set by assigning 1 flag to the top 10% of the counties with respect to each feature, and no flags (0) to the rest of them. These assumptions were too simplistic, and the model’s continuity was disrupted, leading to unrealistic results. Therefore, it is essential to develop a comprehensive model to evaluate multiple consequences of the pandemic. Advanced artificial intelligence models can overcome these problems if mature multivariate statistical analyses validate them. However, comprehensive COVID-19 models with multiple variables, a wide geographical distribution, and a sufficient temporal period are scarce in the literature.

The present study contributes to four research topics concerning the COVID-19 pandemic using data-assisted smart models and empirical analyses. U.S. selected as a case study since it is the most affected country. First, we assembled a large dataset for 30 features in 3142 counties of 50 states to extract the influential factors in the spread of COVID-19 in the U.S. This dataset can be used in vulnerability assessment and for the management of and response to disasters. Second, several machine-learning feature-selection methods were used in a new platform to quantitatively highlight the regional impact of each feature on case fatality rates (CFRs) and reduce model complexity by eliminating irrelevant features. This method removed redundant parameters and increased model accuracy. Comparing the proposed multi-step method with other methods shows more comprehensive and reliable outcomes by employing communities’ variables and government responses. Third, the selected features were classified using machine-learning algorithms to create a more realistic smart model that mimics vulnerability patterns in which different numbers and kinds of input variables can provide reliable results. Fourth, county vulnerability classes were compared at two peaks of the pandemic to analyze regional responses to pandemic control policies.

2. Case study

On April 11, 2020, about weeks after the first reported case of COVID-19 in the U.S. (which occurred in Washington state on January 20, 2020) (Link, 2020), COVID-19 cases were confirmed in all 50 states (JHU, 2020). Despite all efforts, the first peak occurred in June of the same year. Eighteen months later, COVID-19 had killed more than 600,000 people in the U.S., accounting for 18% of the worldwide total (JHU, 2020) and putting the U.S. at the top of the list of countries with the most confirmed cases and the highest official mortality count (JHU, 2020). Shortly after confirmation of the pandemic, many states imposed policies to curb its spread. Puerto Rico issued the first territorial stay-at-home order on March 15, 2020, and the first state order was issued in California four days later (Moreland et al., 2020). Forty-four states issued a stay-at-home order through the end of June 2020, although details of the execution of these orders varied. Mask-mandate policies were applied broadly to control the spread of SARS-CoV-2. The state of New Jersey issued the first mask mandate in April 2020, and Wyoming was the last among 40 states to mandate the wearing of face coverings in January 2021 (Fischer et al., 2021). Fig. 1 presents the number of confirmed cases and deaths for two peaks (July 15, 2020, and January 8, 2021) in all U.S. counties.

Given the potential power of multivariate data-driven models to analyze the COVID-19 pandemic, the U.S. Centers for Disease Control and Prevention (CDC) and the Surgo Foundation compiled a COVID-19 Community Vulnerability Index for each census tract or county in the U.S. (Surgo Foundation, 2020). The study was an attempt to identify vulnerable regions and help reduce the risk of mortality through the efficient allocation of resources. The index was a simple algebraic

1. The states of Arkansas, Iowa, Nebraska, North Dakota, Utah, and Wyoming and the District of Columbia are not included in this list.
summation of six social features and ignored the weight of each feature. Following this project, Acharya et al. (2020) conducted a COVID-19 risk and vulnerability assessment in India. The results confirmed the significance of geography in population vulnerability and concentration of COVID-19 cases but neglected the role of policies in controlling the spread. Snyder & Parks (2020) developed a county-level hierarchical socio-ecological vulnerability index four dimensions of ecological, social, health, and economic and identified a high COVID-19 vulnerability in the Southeast U.S during the first COVID-19 wave. Tiwari et al. (2021) employed a machine-learning algorithm to classify counties in the U.S according to vulnerability to the CFR and the six social themes identified by the CDC. However, the index considered data gathered only through the first peak of the pandemic. Mollalo et al. (2020), who used GIS spatial modeling of the COVID-19 incidence rate in the U.S., found that spatial variability of disease incidence could be explained at the county level by using environmental, socioeconomic, topographic, and demographic variables. Bosancianu et al. (2020) reported that a county’s political conditions and local government decisions affected the COVID-19 mortality rate. However, each of these studies applied simple linear and equal-weighting methods to each feature, leading to

Fig. 1. County-level COVID-19 mortality and morbidity in the U.S. (a) Confirmed cases during the first peak. (b) Confirmed cases during the last peak. (c) Mortality during the first peak. (d) Mortality during the last peak. (e) Confirmed and mortality rates during the first and second pandemic peaks.
uncertain results. To overcome these problems, comprehensive and multi-aspect research on various pandemic-related datasets, including community features, government interventions, and responses to the policies, is needed.

The primary purpose of this study was to develop a hybrid machine-learning, and multivariate model that considered community variables and government interventions to assess county-level vulnerability to the COVID-19 pandemic. We developed advanced artificial intelligence models to overcome the unseen problems in previous studies that were built on limited data from the first months of the pandemic, which are insufficient to analyze long-term and multi-aspect disasters such as the pandemic. However, using equal-weighted values for the different types of community features can make outcomes unreliable and unrealistic when it comes to crafting effective public health policies. We therefore collected multi-aspect parameters, including data from the CDC data and mandate policies, and applied a multi-step machine-learning algorithm. After cleaning and normalizing the data (described in Section 3.1), a peak-detection method was used to slice the time-series data into comparative periods (Section 3.1). The most important features determining how different communities managed their pandemic responses were identified by developing a hybrid feature-selection algorithm (Section 3.2).

Furthermore, we used an unsupervised clustering method to classify counties based on COVID-19–related features. Both clustering method and machine-learning classification algorithm were used to distinguish which regions were more in danger during the pandemic (Section 3.3). Finally, a novel vulnerability index (COVULin) was developed to investigate the effect of specific policies in different counties. By analyzing the variables that had the greatest effect on COVID-19 mortality, the potential obstacles to reducing vulnerability were examined using COVULin scores (Section 3.4).

3. Materials and methods

The methods of exploring the relationship between county variables as well as local policies and their effect on the community vulnerability are discussed in this section. The methods are shown graphically in Fig. 2–(a). First, county- and national-level, publicly available social and non-social data were collected to develop a comprehensive database of COVID-19 impacts on each county. The data were transformed using a power transformer method into a unified shape. To build a COVID-19 vulnerability model, county-level COVID-19–related mortality and morbidity were obtained and transformed to a CFR. Using a peak-detection method, COVID-19 mortality time-series were sliced according to each county’s pandemic peak. In the next step, machine-learning algorithms were employed in three steps: feature selection, clustering, and classification of counties based on the CFR. A hybrid feature-selection algorithm was used to identify the most important COVID-19 variables, including those with the strongest relationship with the mortality rate. Factor analysis was used to validate the feature-selection results. Unsupervised clustering was applied to feature-selection outcomes to divide counties into the sub-similar groups with the same pandemic-related challenges. In the final step of artificial intelligence modeling, multiple machine-learning algorithms classified counties according to vulnerability. The COVULin served as an indicator to rank counties. Finally, all five steps were replicated for two COVID-19 peaks. The results were used to investigate which communities improved their situation through the imposition of pandemic-control policies and which ones saw conditions worsen. These steps are described in the following sections.

3.1. Case selection

A large dataset was prepared at the county level by assimilating, processing, and merging relevant data from publicly available datasets. County-level population statistics and demographics as well as 30 socioeconomic, topographic, health-care system, housing type, and transportation variables were collected and used as explanatory variables to stratify 3142 U.S. counties (CDC, 2018; U.S. Census Bureau, 2019). Flanagan et al. (2020) used 4 domains of socioeconomic data, household composition and disability, minority status and language, and housing and transportation, including 16 predictors. They tried to develop a social vulnerability model of disasters. In this study, health care facilities and high-risk environments are added to the dataset to build a more specific COVID-19 vulnerability model. Furthermore, the relationship between Riots and COVID-19 pandemic has already been studied (Bloem & Salemi, 2021), but it was substituted with other variables in our model that can have inherent effects on riots, including poverty, race, income and etc. In addition, variants could effect on the severity of the pandemic. However, this study focused on the community variables and obtaining the effect of the strain variants on the pandemic severity needs a separate study.

All 30 data features are summarized in Table 1. These features were reported in the literature to be correlated with COVID-19 spread. Data for confirmed COVID-19 cases in the U.S. from January 22, 2020, to July 15, 2021, including all reported infections and mortality, were obtained from Johns Hopkins University (COVID C, 2020). The Federal Information Processing Standard (FIPS) code was used to merge the data into a single dataset. Control policies such as mandates to quarantine infected cities and the requirement of wearing a mask in public are hypothesized to affect COVID-19 infection rates (Liu et al., 2020). Data on stay-at-home orders (Boston University, 2021; Moreland et al., 2020) and mask mandates (Ballotpedia, 2021) were gathered using state-and-county-level public data, and the policies’ timelines were extracted for 50 states. Only counties with at least one reported case at each peak were selected for analysis to ensure input data quality. A total of 2510 and 3040 counties were selected in models for the first and last peaks, respectively.

Because the features of the constructed dataset exhibited different variances and shapes, they were scaled to facilitate machine learning. Power transformers were used to normalize data and enhance machine learning. Power transformers are a family of parametric, monotonic transformations that turn data into semi-Gaussian patterns. This transformation is crucial when modeling data with significant heteroscedasticity or non-normal patterns in multiple fields such as statistical data analysis, medical research, modeling of physical processes (Gluzman & Yukalov, 2006) and many other clinical, environmental, and social research areas. Here, a power transformer with a mean of zero and unit-variance normalization was imported from the Scikit-Learn (SkLearn) library of Python. The normalized data are supplied in the Supplementary files. Mortality and morbidity data should be normalized so that two counties with different populations and infection rates but similar infection curves can be treated comparably. In addition, both confirmed and death indicators depend on many criteria including population, population density and type of the spread variant. The CFR was selected as an epidemiological factor to report the severity of the pandemic from possible indices. This index incorporates the impacts of infrastructure, policies, and other variables to detect severity of the COVID-19 (Ahmed et al., 2020; Amram et al., 2020; Nayak et al., 2020b; Onder et al., 2020; Pinato et al., 2021). In addition, the relationship between a COVID-19 CFR and a policy of austerity (health expenditure cuts) and open testing has already been investigated (Cao et al., 2020; Sherpa, 2020). The CFR was obtained using Eq. (1):

$$CFR = \frac{CD}{CDC} \times \Delta T$$  

(1)

where $CD$ is the number of confirmed deaths, $CDC$ is the number of confirmed diagnosed cases, and $\Delta T$ is a given time interval. Due to the nature of a pandemic, although parameters themselves are fixed, their effect on long-term disaster may vary on mortality. Based on the available data, communities with higher minorities and old people have a
higher risk of mortality in the first and last peaks of the pandemic, respectively. Hence, a peak-to-peak analysis was employed in this study to identify the impacts of the community variables on the CFR. The time-series data were sliced based on the mortality peak in each county to measure the effect of dynamic variables, including local policies on the pandemic pattern. A peak-finding module from the SciPy library was used to identify the peak of a time series. This module has been used to detect pattern peaks in chest X-ray images of COVID-19 patients, breathing and heart rates, and traffic monitoring systems (Chan et al., 2021; Evteeva et al., 2019; López-Reyes et al., 2021). The function was tuned for the prominence of the peaks in the mortality time-series data to control its sensitivity and avoid an excessive number of peaks. A 60-day window was applied to each peak to avoid reporting faulty peaks. This means that a peak was defined as the greatest daily death rate one month prior and one month after a given date.

Fig. 2. A representation of employed methods and outcomes. (a) A flow diagram of the process and methods from data collection to the comparison of outcomes for 2 different peaks. (b) A graphical view of the expected outcomes at each step of the research.
3.2. Selection of indicators

The presence of a relatively large number of potentially relevant variables may result in a theoretical discrepancy that decreases model accuracy. Feature-selection methods are widely used in machine-learning models to reduce complexity, decrease the number of irrelevant and redundant features, and enhance prediction accuracy (Hancer et al., 2015). A typical method of estimating variable contribution is to rank the features and select the best subset, respectively. A hybrid feature-selection algorithm was then employed to identify the most important features among the 30 variables that can affect the COVID-19 CFR. First, a Pearson correlation is used to measure multicollinearity between variables. This method identifies the dependency between variables based on which the variables with similar effects on the mortality rate can be removed. Then, a combination of four mature explanatory variables and policies used in this study together with definitions and sources. Table 1

| Category                         | Variable name                                                                 | Source                                             |
|----------------------------------|-------------------------------------------------------------------------------|----------------------------------------------------|
| 1 Socioeconomic status           | 1 Percentage of population below a poverty level                               | U.S census tract (2018)                           |
|                                  | 2 Percentage of the unemployed population age 16 and older                    | U.S census tract (2018)                           |
|                                  | 3 Estimated income per capita                                                 | U.S census tract (2018)                           |
|                                  | 4 Percentage of persons age 25+ with no high school diploma                  | U.S census tract (2018)                           |
| 2 Household composition/disability| 5 Percentage of persons aged 65 and older                                     | U.S census tract (2018)                           |
|                                  | 6 Percentage of persons aged 17 and younger                                   | U.S census tract (2018)                           |
| 3 Minority status/language       | 7 Percentage of civilian noninstitutionalized population age 5 and older with a disability | U.S census tract (2018)                           |
|                                  | 8 Percentage of single-parent households with children under 18              | U.S census tract (2018)                           |
| 4 Housing/transportation         | 9 Estimated minority percentage of total population (all persons except white, non-Hispanic) | U.S census tract (2018)                           |
|                                  | 10 Percentage of persons age 5 and older who speak English "less than well"  | U.S census tract (2018)                           |
|                                  | 11 Percentage of housing in structures with 10 or more units                  | U.S census tract (2018)                           |
|                                  | 12 Percentage of mobile homes                                                 | U.S census tract (2018)                           |
|                                  | 13 Percentage of occupied housing units with more people than rooms          | U.S census tract (2018)                           |
|                                  | 14 Percentage of households with no vehicle available                         | U.S census tract (2018)                           |
|                                  | 15 Percentage of persons in institutionalized group quarter                  | U.S census tract (2018)                           |
| 5 Health-care system factors     | 16 Percentage of the population uninsured                                      | U.S census tract (2018)                           |
|                                  | 17 Percentage of households without access to indoor plumbing                 | U.S census tract (2018)                           |
|                                  | 18 Intensive care unit beds per 100,000 people                                | U.S census tract (2018)                           |
|                                  | 19 Hospital beds per 100,000 people                                          | U.S census tract (2018)                           |
|                                  | 20 Agency for Healthcare Research and Quality - Prevention Quality Indicator Overall Composite: admission rates for preventable conditions (via good outpatient care) adjusted per population | U.S census tract (2018)                           |
|                                  | 21 Emergency services per 100,000 people (includes emergency and relief services and freestanding ambulatory surgical and emergency centers) | U.S census tract (2018)                           |
|                                  | 22 Epidemiologists per 100,000 people                                        | U.S census tract (2018)                           |
|                                  | 23 Health lab per 100,000 people                                             | U.S census tract (2018)                           |
|                                  | 24 Health spending per capita                                                 | U.S census tract (2018)                           |
|                                  | 25 Total Public Health Emergency Preparedness funding per capita              | U.S census tract (2018)                           |
| 6 High-risk environment          | 26 Long-term care (nursing homes, assisted living, and care homes) residents per 100,000 | U.S census tract (2018)                           |
| 7 Population                     | 27 Percentage of the population employed in a high-risk industry (includes employees in farming, manufacturing, printing, and related support activities and textile North American Industry Classification System subsectors) | U.S census tract (2018)                           |
| 8 Pandemic control policies      | 28 Prisons population per 100,000                                            | U.S census tract (2018)                           |
|                                  | 29 County level population density (person/km²)                              | U.S census tract (2018)                           |
|                                  | 30 County-level population                                                   | U.S census tract (2018)                           |
|                                  | 31 Including state-level stay-at-home orders, travel restrictions, etc.       | U.S census tract (2018)                           |
|                                  | 32 Mask-mandate order (state- and county-level mask mandates)                 | U.S census tract (2018)                           |

3.3. K-means-MOGA clustering and county classification

Considering the socio-economic variety of U.S. counties, they were labeled based on their specifications to develop an efficient model. No predetermined labels such as rich-poor classes or share of minorities were assigned to the data to avoid bias while making decisions about resource allocation during the pandemic. Instead, the counties were clustered using an unsupervised machine-learning technique based on their COVID-19 features. This unsupervised clustering method separates counties based on their features similarity and geographic data and facilitates the learning procedure. The unsupervised clustering approach helps find the geological relationship based on selected predictors and group counties to similar ones to overcome the geo-spatial complexity.
Nayak et al. (2020a) used a county-level clustering algorithm to identify social vulnerability in the U.S. and showed that the clustering algorithm can help build a vulnerability assessment model. The outcomes will help develop a more accurate model for each group of similar counties. A centroid-based k-means algorithm grouped heterogeneous elements into homogenous clusters based on data structure and analysis targets. Unsupervised k-means clustering has already provided strong and sharp differences among clusters by identifying optimal parameter settings (Carrillo-Larco & Castillo-Cara, 2020; Chandu, 2020). The distance from each node to the centroid of the cluster \( c_k \) and a silhouette test were used to assess model performance. A genetic algorithm was employed to find the optimal parameter options as a population-based hyper-parameter-tuning approach. The objective functions and decision variables were defined as follows:

\[
\min_{p_1 \neq p_2} \sum_{i=1}^{n} d(i, c_k) \quad \text{where} \quad c_k = \left( \frac{1}{n_k} \right) \sum_{i=1}^{n_k} i^2 \]  

(2)

\[
\max_{p_i \in P} S(i) = \begin{cases} 
\frac{b(i) - a(i)}{\max\{a(i), b(i)\}} & \text{if } |C_i| > 1 \\
0 & \text{if } |C_i| = 1
\end{cases}
\]  

(3)

\[
\begin{align*}
a(i) &= \frac{1}{|C_i|} \sum_{j \in C_i, i \neq j} d(i, j) \\
\text{where } b(i) &= \min_{i \neq j} \frac{1}{|C_i|} \sum_{j \in C_i} d(i, j)
\end{align*}
\]  

(4)

**Fig. 3.** A hybrid feature-selection method to determine the variables most related to COVID-19 CFRs.

Given \( K, DM, N_k \)
with respect to \( k \in \{2..20\} \),
\[
DM \in \{\text{sqeuclidean', 'cityblock', 'cosine', 'correlation'}\},
\]
\[
N_{k} \in \{\{2..20\}\}.
\] (5)

where \( k\text{means} \) is a k-means clustering function adopted from Matlab built-in functions, \( K \) is the number of clusters, \( DM \) represents distance metric for \( d(.) \) as the distance function, \( N_{k} \) is the number of repetitions of clustering using new initial cluster centroid positions, \( S(i) \) is the silhouette function, \( |C| \) and \( |C_{k}| \) denote the number of points belonging to cluster \( i \) and \( k \) such that \( C_{i} \neq C_{k} \). The knee point of the Pareto front was selected as a solution with the maximum marginal utility.

After clustering counties, several machine-learning algorithms were used to classify counties according to vulnerability and provide a priority list for counties that need special attention with respect to the pandemic experience. The classification models are developed for each cluster separately and based on each region’s selected variables and their importance. Five classes are defined based on the CFR value which shows the impact of the pandemic in each cluster. Here, the Decision Tree classifier (DTC), Random Forest classifier (RFC), and a neural network (N.N.) with a multi-layer perceptron (MLP) classifier and three hidden layers were used to classify counties in each cluster based on their selected variables and CFR. The results defined a vulnerability class for each county. Because the performance of machine-learning algorithms depends on hyperparameters, a grid search tool was used to determine the optimal hyperparameters. Table S.1 summarizes the tuned hyperparameters for each model.

Data from all counties were divided into training (70%) and test (30%) datasets to inhibit overfitting problems. The models were evaluated using 10-fold random cross-validation with no overlap or replacement. A receiver operating characteristic (ROC) and an overall accuracy metric were used to compare the performance of each classification approach identifying vulnerable regions. A ROC curve is a graph of the performance of a classification model at all classification thresholds (Aurélien Géron, 2019). ROC curves were generated for every algorithm, and the area under the curve (AUC) was calculated to reveal how accurate the results of each model are. The AUC measures the entire two-dimensional area beneath the ROC curve from 0 to 1. A higher ROC-AUC measures the trustworthiness of the results from a machine-learning approach, given the accuracy of the model.

3.4. Vulnerability and policy analysis

Social vulnerability refers to the socio-economic and demographic factors that impact community resilience, while population vulnerability is a subset of geographical vulnerability. However, COVID-19 vulnerability is more than simply facing a disease; it is assumed as a dynamic concept in which an individual or a group may not be vulnerable at the start of the epidemic, but depending on how the government responds, they may become vulnerable afterward (Lancet, 2020). In addition to the epidemiologically vulnerable groups such as elder people and those with the underlying disease, people from diverse socio-economic backgrounds such as poor communities or minorities might be vulnerable to the pandemic (Acharya & Porwal, 2020). Thus, it is crucial to obtain social and community vulnerabilities to manage the COVID-19 pandemic.

A county-level COVID-19 vulnerability index is proposed as a comparative indicator to detect the vulnerable regions during different pandemic periods. This index is developed according to the social and non-social variables of the communities that quantifies the pandemic vulnerability in each county in the United States. The index can be calculated after restricting the data samples to the counties with confirmed cases, selecting the most important features with respect to the CFR, and using the importance values of the selected features, as given in Eq. (6).

\[
COVULin = \sum_{i=1}^{n} (w_{i} \times P_{i})
\] (6)

Here, \( COVULin \) is the vulnerability index, \( w \) is the importance factor of each feature (from the feature-selection method) and \( P \) is the feature’s value. Compared with the traditional method of assigning a feature’s value with hard boundaries based on the minimum and maximum values (0 or 100) used in the CDC’s social vulnerability index (SVI), this research represents a new method based on a \( P \)-value evaluation that ranges from 0 to 100. Instead of using the original value for each feature according to the different effects of each feature on the CFR trend, a fuzzification method is used to evaluate the \( P \)-value for each feature. Based on the correlation factor between the variables and CFR, all variables were categorized into two classes:

1. Those with a positive correlation with the CFR, such as poverty, minority, and people above the age of 65. For this category, Eq. (7) is used to obtain the \( P \)-value.

\[
P(x) = \begin{cases} 
0 & x \leq a \\
\frac{100(x - a)}{b - a} & a < x \leq b \\
100 & x > b 
\end{cases}
\] (7)

2. Those with a negative correlation with the CFR, such as hospital beds per capita and household income. For these variables, the \( P \)-value is calculated using Eq. (8).

\[
P(x) = \begin{cases} 
0 & x > a \\
\frac{100(d - x)}{d - c} & c < x \leq d \\
100 & x \leq c 
\end{cases}
\] (8)

Here, \( x \) is the feature’s normalized value from the CDC, \( a \) and \( c \) are the mean values, \( b \) is equal to 0.9 of the maximum value, and \( d \) is equal to a 0.1 maximum value. \( COVULin \) scores were estimated for and rank each county. According to the definition, the variables with positive correlations with CFR can increase the COVULin, which addresses the higher vulnerability. On the other hand, the variables with negative correlations decrease the COVULin. In this regard, a higher COVULin score indicates a greater vulnerability in a region. Flanagan et al. (2020) confirmed that a higher vulnerable index in a community accompanies a higher risk of mortality by disasters. To divide counties based on their COVULin scores and compare them with the impact assessment results, 5 levels of vulnerability were defined as below:

✓ High vulnerability 0.8 < COVULin < 1
✓ Most vulnerable 0.6 < COVULin < 0.8
✓ Moderate vulnerable 0.4 < COVULin < 0.6
✓ Low vulnerable 0.2 < COVULin < 0.4
✓ Least Vulnerable 0 < COVULin < 0.2

All the analyses were performed on two detected COVID-19 peaks to compare the changes in the vulnerability and most important features in each county. The changes in the vulnerability can come from different imposed policies, portion of people who are following the policies, kind of the virus variant, or other factors. Although vulnerability index does not represent each county’s policy responses, the changes in vulnerability index can represent the responses to the same policies in different counties (Davradakis et al., 2020). Each county’s vulnerability class and COVULin score were compared during these peaks to determine which communities experienced improved conditions following the introduction of those control policies and which ones were less responsive to the
mandates. Finally, based on the fixed features in the counties, the model was developed based on a static statistical model for two different pandemic peaks, and the autocorrelation approach was not used in its development.

A categorical regressor technique is used to quantify the impact of the policies (features 31 and 32 in Table 1) and integrate them into the model. First, the start and end dates of the imposed policies were determined. Then, two county-level Boolean variables were chosen for the imposed policies: mask mandate and stay-at-home (or extremely restricted policies). These Boolean variables are assigned to the counties during two different peaks. If a county spends more than 80% of the time under the restricted policy in each pandemic peak, the Boolean variable is set to True. Otherwise, it will be False.

\[ x = \begin{cases} \text{False} & \text{if a county spends less than 80\% of the period under the restricted policy} \\ \text{True} & \text{if a county spends more than 80\% of the period under the restricted policy} \end{cases} \]

Fig. 4 illustrates the timeline of the mask mandate policy in different counties. Counties A1, C1, C2 and C3 are designated as True in the first peak because of spending more than 80% of the period under the mask requirement policy. In the last peak, all counties except B1, D3 and D4 were designated as True. Hence, Boolean variables were developed out of the data to quantify the imposed policies considering the following simplifying assumptions:

1- It was assumed that everyone followed the restrictions in a county if formally imposed. However, the people’s responses to the mask mandate policies and stay-home orders could vary.
2- Assuming the independency of the socio-economic variables from the imposed policies through time, the intercorrelation between the policies and other variables was not considered. However, a policy might affect social and economic situation in real world.
3- Only two policies were assumed to be implemented in the U.S. However, various pandemic-controlling policies were employed in the U.S, including money distribution, rapid testing, masking mandate, stay-home, etc. Here, the latest two were considered to be disruptive policies.
4- The restriction policies were assumed to be followed by all counties of a state if there were no explicit county-scale data available. However, the CDC did not report a particular policy timeline for each and every county in the U.S.

This comparison was used to identify the obstacles to improving resiliency while imposing pandemic control policies. We used a multistep method to provide a comprehensive view of the impacts of the COVID-19 pandemic on different communities and then investigated how community characteristics can decrease the effects of applied policies in different regions. Fig. 2-(b) depicts the expected outcomes at each study step.

4. Results and discussion

The unique characteristics of the proposed nationwide COVID-19 model are discussed using the analytical results in this section. County-level data for confirmed cases, deaths, and CFRs were used to assess the pandemic from January 22, 2020, to July 15, 2021. Results are detailed in the following sections.

4.1. Peak detection

A find_peaks algorithm was imported from the Sklearn library to detect pandemic peaks using sliced county-level mortality data. The peak-detection results are summarized in Table S.2, using 60-day windows for peak detection. Counties with no confirmed cases were removed from the analysis of each peak. According to Table S.2, the first and last pandemic periods are assumed to occur from January to August 2020 and November 2020 to January 2021, respectively. A total of 2486 counties in the first peak and 3044 counties in the last pandemic peak were selected to analyze COVID-19 impacts. All county-level mortality and confirmed cases during the first and the last days of the peaks were selected to calculate each peak’s CFR. For the first peak, all mortality and confirmed cases since the first report day until the peak day were used to calculate the CFR. The starting days of the last peak in the county-level analysis were detected by a local minimum in the time-series data using the find_peaks method.

As can be seen in Table S.2, the U.S. experienced a lower CFR in the last peak compared with the first, despite a greater number of deaths and confirmed cases. According to the Johns Hopkins Coronavirus Resource Center (2021), the highest daily rate of COVID-19 tests per population was twice as great in the last peak compared with the first, resulting in faster diagnosis and treatment. The high testing rate caused a dramatic decline in the CFR during the last peak of the pandemic. The respective net mortality and morbidity numbers in the U.S. were 4.2 and 2.6 times higher in the last peak than in the first peak, but the CFR was half as large. However, two other influential factors, in addition to rapid
diagnosis and intensity, must be considered during a long-term crisis such as the COVID-19 pandemic: community specifications and government interventions, as described in the following sections.

4.2. Feature selection and model validation

A multi-step procedure was followed to choose the best subset of important variables in the communities. The feature-selection procedure with respect to model accuracy and the RFECV is detailed in Fig. 3. The best results of the machine-learning algorithm were obtained by combining the first 16 variables in each pandemic peak (out of the 30 initial variables), which were hypothetically related to the COVID-19 CFR. These features included population density, percentage of members of minority groups in the population, households without a personal vehicle, health-care residents per 100,000 people, prisons population per 100,000, percentage of persons aged 65 and older, prevention quality index (pqi), percentage of the population employed in a high-risk industry, people living in mobile homes, housing with multiple units, population of uninsured people, hospital beds per 100,000 people, disabled people, poverty, unemployment, per capita income (PCI). The values for the selected features in each feature-selection method, as well as the results of the hybrid method, are shown in Fig. 5. The two different pandemic period are considered to show how a community variable can affect pandemic’s severity differently in different pandemic time. The outcomes reveal that target vulnerable communities can be changed during the pandemic. As illustrated in Fig. 5(a), strong correlations are evident among population density, the minorities proportion of the population, people without any vehicle access, and the CFR during the first pandemic peak. However, one year after the beginning of the pandemic, other variables, such as population density, the portion of people older than 65, and poverty, were the significant features associated with an increased mortality rate, as shown in Fig. 5(b). Although the same 16 important features were found as the most important factors in both peaks, a meaningful difference appeared in the importance of the

![Figure 5](https://example.com/figure5.png)

**Fig. 5.** (a) The most important features related to COVID-19 CFR in the first peak, and (b) the most important features related to CFR in the last peak. (c) A comparison of the top 10% of counties with the highest value of each feature among 200 counties with the highest mortality in each peak.
effective variables. For example, the proportion of minorities was the second important variable associated with counties with a high mortality risk during the first peak, but the significance of the relationship was considerably weaker during the last peak. This may be a result of many factors, including government intervention and public awareness (Chaudhry et al., 2020).

As can be seen in Fig. 5(a) and (b), counties with high population densities were more vulnerable during both peaks, and government interventions, such as lockdowns and distancing restrictions in public places were required. Of the 200 counties with the highest mortality rates during the first peak, 134 (67%) were among the top 10% of counties with a high population density, as shown in Fig. 5(c). During the last peak, this number rose to 169 counties. An important variable that has been frequently addressed in the literature is the proportion of minorities such as Hispanics and African-Americans in a county (Karaye & Horney, 2020). According to Fig. 5(a), this portion was the second most important variable associated with an increased mortality rate during the first peak, but ethnic minorities were not as threatened as poor communities during the last peak. This finding reveals the importance of studying peaks individually during the pandemic. As shown in Fig. 5(c), 54 of 200 counties with the highest mortality rate were in the top 10% of cities with the highest proportion of minorities during the first peak. This number fell to 31 counties in the last peak. The poor communities were at greater risk during the last peak compared with the

Fig. 6. Clustering results in (a) a GIS map of counties and (b) a significance diagram of clustered features.
first; the number of poor communities among the counties with the highest mortality increased from 15 in the first peak to 28 in the last one. The outcomes from the feature-selection method confirmed the previous finding that the proportion of older people in a county may have affected the mortality rate during both peaks (Mallapaty, 2020).

4.3. County clustering and classification

An unsupervised k-means clustering algorithm accompanied by a multi-objective genetic algorithm (MOGA) optimizer was used to group the COVID-19 data, excluding mortality and morbidity, based on the similarities. The clustering algorithm was run on all 3142 counties with 16 significant COVID-19 features and included longitude and latitude data. To reach the most stable and robust clusters, a MOGA algorithm tried to maximize summation of Silhouette function while minimize the summation of distance metric by three decision variables including: Number of clusters, Distance Metric functions and Number of replication. The best results obtained from knee point of the Pareto front as a solution with the maximum marginal utility. Table S.6 provides the optimization results.

With this method, four clusters were obtained based on their silhouette and centroid-distance values to obtain the best results from k-means clustering, Fig. 6(a) is a GIS map of the clustered counties based on the 16 features most related to COVID-19 mortality. Standardized values of the features are supplied in Fig. 6(b) to highlight the significant features in each cluster, illustrating the importance of features in each region and their differences from the nationwide analysis.

Although the unsupervised method weighed all features equally to find the best clusters, a meaningful pattern for the distribution of counties was evident in four clusters. As can be seen in Fig. 6(a), most of the counties in clusters 1 and 3 come from southern states. These clusters have relatively high population densities compared with other clusters. Clustering results show that capitals and larger cities are grouped in these clusters. Of 50 state capitals, 34 are categorized into these clusters. In contrast, clusters 0 and 2 contain counties from northern and central states (with the exception of major metropolitan areas such as New York City). These clusters have a relatively high number of rural areas and small counties.

According to Fig. 6(b), the population of old people is the most important feature in cluster 0. The proportion of people older than 65 in cluster 0 is 25.9%, the highest value among other clusters (Table S.3). According to the feature importance’s result in Fig. 5, the proportion of older people in a community is a variable that warrants special attention. In contrast, as can be seen in Table S.3, this cluster has relatively low values for three important variables that can increase COVID-19 mortality, including population density (24.35 persons/km²), minority populations (18.9%), and people without access to vehicles (4%). Lower values for these important variables led counties in cluster 0 to lower mortality rates compared with other clusters.

With the highest population densities, the counties in cluster 1 had the highest COVID-19 confirmed infection and death rates during the pandemic’s first and last peaks. As can be seen in Table S.3 the numbers of people living in multiple-unit buildings and the population densities are considerably greater in cluster 1 than in other clusters. Almost 1000 people live in each km² in this cluster, potentially making these counties one of the regions with the highest vulnerability. This population density is made possible by multiple-unit structures such as apartments and condominiums. In cluster 1, more than 11% of the people live in multiple-unit structures, which is four times the share in cluster 2. The proportion of minority groups is the third important variable in this cluster, with 33.5% of the population categorized as members of minority communities, including Hispanics and African-Americans. As is evident in Fig. 5(a), the proportion of minorities groups was one of the most important features affecting mortality during the first peak of the pandemic. Counties of cluster 1 contain the highest values for critical features, including population density (999 persons/km²), people who live in multiple-unit structures (11.26%) and the proportion of minority groups (33.5%), which put them in the most vulnerable area. As shown in Table S.3, the number of long-term health-care residents (including those in nursing homes, assisted living facilities, and care homes) per 100,000 people were 1046.7 in cluster 3, the highest among all clusters. This cluster is also the second oldest, with the senior community accounting for 21.5% of the population. As can be seen in Fig. 5, this variable increases the mortality rate of a community. This cluster is the only one not to see a CFR reduction in the last peak compared to the first peak because it had an older community that retained its vulnerable status.

As Fig. 6 makes clear, poverty and mobile homes are two significant variables in cluster 3. Counties in this cluster have the greatest share of people who live below the poverty level (21.7%) compared with other clusters. In addition, this cluster has the highest proportion of uninsured, disabled, and unemployed of all clusters. With these variables, this cluster saw the highest number of deaths per population during the first peak and second highest during the last peak. To describe the impact of COVID-19 on the counties, 5 classes of COVID-19 vulnerabilities were defined based on the CFR as follows:

1. Least impacted counties; 0 < CFR < 0.9
2. Low impacted counties; 0.9 < CFR < 1.8
3. Moderate impacted counties; 1.8 < CFR < 2.7
4. High impacted counties; 2.7 < CFR < 3.6
5. Most impacted counties; 3.6 < CFR

Using these categories, an impact label was assigned to each county for the pandemic’s first and last peaks. Fig. 7 shows the number of counties with COVID-19 mortality reported for each peak (2486 counties in the first peak and 2099 counties in the last peak) and their intensity of impact was labeled for each cluster. According to this system, 509, 296, 464, 692, and 505 counties were categorized as the most, highly, moderately, low, and least impacted classes in the first peak, respectively. In addition, 228, 201, 516, 1216, and 838 counties were classified in the same groups for the last peak. Clusters 1 and 0 had the highest and lowest populations in both peaks. Clusters 1 and 3 had almost 69% of counties with high and most impacted situations. However, the percentage of counties experiencing a ‘low and least impact’ in cluster 0 (52%) and cluster 2 (63%) were higher than in cluster 1 (40%) and cluster 3 (41%). Counties in clusters 1 and 3 were therefore at higher risk of being impacted by COVID-19 during the first peak.

Although the unsupervised clustering method was developed without considering mortality, a nationwide analysis revealed an interesting pattern. A strong relationship was found between COVID-19 mortality and both population density and minority group representation during the first peak. As can be seen in Fig. 7(a) and Table S.3, the majority of the population living in the most impacted area are in cluster 1 during the first peak. In this peak, 154 counties out of the first 509 with a CFR greater than 3.6 (high impacted areas) belong to cluster 1. The highest average CFR (2.13) occurred in this cluster. In addition, 56.7 million people of 69.5 million living in high-impacted counties live in these counties. The greatest number of deaths per population (0.3 per 1000 people) and the second greatest CFR (2) occurred in cluster 3, with a high proportion of minorities, poor communities, and a number of people without vehicle access. As Fig. 5(b) shows, poorer and older communities were significantly impacted by the pandemic along with the counties with higher population densities during the last peak of the pandemic. As the data in Table S.3 indicate, cluster 3, with the largest senior population and the second highest population density, contained 42% of all counties in the most and high-impacted situations, and cluster 0 had the second greatest number of clusters with the highest proportion of older people with the most and high impact situations.

Various machine-learning algorithms assigned counties into the most-to-least impact categories and provided a preliminary list of counties which impacted by the pandemic, including 5 types of impacts.
The DTC, RFC, and N.N. methods were used to classify counties in each cluster based on selected variables and CFRs. In addition to accuracy, Receiver Operating Characteristic (ROC) and Area Under Curve (AUC) were calculated for each algorithm. The ROC-AUC technique was used to validate model prediction accuracy and choose the best model for each cluster. The ROC curve was created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings in which a higher ROC-AUC value indicated a higher model’s accuracy. Seventy percent of the data was used for training and 30% to test for four clusters in two different peaks of the pandemic. Fig. 7(b), which depicts the ROC-AUC results, indicates that RFC provides accurate results, with ROC-AUCs of between 0.95 and 0.96 for clusters in the first peak and 0.95–0.97 in the last peak. The trained N.N. with three hidden layers produced ROS-AUCs of between 0.88 and 0.91 for clusters in the first peak and 0.90–0.91 for the last peak. Finally, DTC produced ROS-AUCs of between 0.86 and 0.89 for clusters in both peaks. Detailed results are provided in the Supplementary files. Overall, validation and reliability results indicate that an RFC model with a high degree of training and testing accuracy provides a high-quality COVID-19 vulnerability class during both pandemic peaks.

4.4. Vulnerability index and policy analyses

The COVULin was developed to gauge county-level vulnerability during the pandemic as well as the effect of mandates on county vulnerability. This index is based on each cluster and their feature importance values to increase the model’s accuracy and facilitate the learning procedure. Four separate and accurate models were developed for 4 clusters in this case. The index varies between 0 and 1; the higher the score the more vulnerable the region. Fig. 8(a) and (b) provide GIS maps of vulnerability in the first and last COVID-19 peaks. Most of the vulnerable counties in the first peak are in the southern part of the U.S. There is close agreement between COVID-19 mortality and COVULin scores, as shown in Fig. 1(c). The vulnerability of the southern counties
decreased during the last peak, but southwestern counties remained highly vulnerable.

Fig. 8(c) depicts the distribution of counties with different vulnerabilities in four clusters. The COVULin data reveal that clusters 1 and 3 contained a higher number of vulnerable counties during the first pandemic peak. As shown in Table S.3, these clusters have greater population densities, a higher proportion of minority groups, and more people with no vehicle access. As a result of applying the feature-selection method, these parameters can increase the CFR, as can be seen in Fig. 5(a); counties in these clusters were associated with a higher vulnerability score. However, Fig. 8(c) shows that cluster 3 had the largest proportions of poor and unemployed residents and was second in population density, with the highest number of counties in the most and highly vulnerable classes. This cluster contained 44% of all counties with a COVULin score greater than 0.6 (most and highly vulnerable areas) in the last peak.

New York City (CFR = 3.46) from cluster 1 had the highest COVULin score in this peak. The city had the greatest population density, the most multiple-unit housing structures, and the second highest number of people with no vehicle access. It was also among the top 10% of counties with a high percentage of minority residents among all counties. During the last peak, in the Louisiana county of East Carroll from cluster 3, in which 48.6% of the population lived below the poverty line (ranking fourth of all counties), 71.1% of its population were members of minority groups (ranking 96th), 18.9% had no vehicles access (ranking 33rd), the proportion of its population that was incarcerated was the third largest, had the highest COVULin score. East Carroll’s CFR of 2.87 placed it among the highest-impact counties. The compatibility between COVULin scores and impact assessment results confirmed that the index could be used to describe the vulnerability of counties facing pandemic pressures. The index can also be used for regions with faulty or missed data to predict conditions in comparison with other counties as it is based on community features and their values.

Once the relationship between infection rates and county-level variables was established, the policy’s timeline for each county was used to measure the impact on COVID-19 mortality rates. Among all policies,
two controlling policies were identified as disruptive factors: stay-at-home orders (or travel restrictions) and mask mandates. Based on the Fig. 4, the Boolean variables are assigned to the counties for two different policies. In the first pandemic peak, seven states used mask mandate and 10 applied travel restrictions for short periods which just 76 counties have been assigned True for both policies. Hence, It was assumed that no effective policies were implemented during the first peak of the pandemic. However, among all states and federal districts, just eight states did not impose travel restrictions and stay-at-home orders (439 counties were assigned False for travel restrictions policy) and 12 states and one federal district did not use state or territorial mask-mandate policies (905 counties were assigned False for mask-mandate policy) through the end of the last peak. The last peak was characterized by high levels of government intervention. Among all counties, only 258 did not use any local and federal restrictions during the pandemic (False for both Boolean variables) and 1848 imposed both travel restrictions and mask mandates (True for both Boolean variables). To determine the impact of policies, counties with strong interventions (both kinds of policies) and counties without government intervention (without either kind of policy) were chosen. Four indicators were employed to investigate the effect on the counties, including death rates, confirmed cases, CFR, and the COVULin scores.

Fig. 8. (continued).

| Least Vulnerable | Low Vulnerable | Moderate Vulnerable | High Vulnerable | Most Vulnerable |
|------------------|---------------|---------------------|-----------------|-----------------|
| First peak       |               |                     |                 |                 |
| Clusters         |               |                     |                 |                 |
| 0                | 5.8           | 235.2               | 23.2            | 30.5            | 6.9           | 246.7        | 31.3           | 35             |
| Counties         | 360           | 622                 | 668             | 816             | 416           | 710           | 903            | 970            |
| Most Vulnerable  |               |                     |                 |                 |               |               |                 |                 |
| 1                | 41            | 210                 | 98              | 223             | 76            | 22            | 41             | 96             |
| Counties         | 92            | 106                 | 110             | 154             | 71            | 68            | 243            | 224            |
| Moderate Vulnerable |           |                     |                 |                 |               |               |                 |                 |
| 2                | 94            | 109                 | 214             | 149             | 89            | 280           | 359            | 351            |
| Counties         | 104           | 84                  | 186             | 146             | 142           | 314           | 217            | 198            |
| Least Vulnerable |               |                     |                 |                 |               |               |                 |                 |
| 3                |               |                     |                 |                 |               |               |                 |                 |
| Counties         |               |                     |                 |                 |               |               |                 |                 |

5. Discussion

The data-assisted smart models were employed to estimate county-level vulnerability indices for disaster management based on multi-aspect features during a long-term pandemic. Consequently, 30 county-level social and non-social features were assembled and fed to several machine learning algorithms to highlight the regional impact of each feature on case fatality rates (CFRs). The feature selection results showed strong correlations between population density and increased risk of COVID-19 mortality during both pandemic peaks. 67% (and 86%) of the counties with the greatest mortality rates during the first (and the last) pandemic peak were clustered as crowded counties. The feature importance analysis highlighted the significance of population density in the urban vulnerability against public disasters. The population density can be directly and indirectly correlated with the pandemic. According to the literature, public transport enhanced virus spread and increased the infection probability in crowded areas (Sy et al., 2021). The clustering analysis showed that the most vulnerable counties appeared in cluster 1 in the first peak. Cluster 1 mainly consisted of state capitals and large cities with the highest population density, percentage of people without cars (8%), and use of public transportation. However, a lower population density (i.e. rural areas) is not necessarily associated with lower COVID-19 mortality. Rural areas may become vulnerable to morbidity and death due to higher proportions of older people, more underlying diseases, and limited access to care facilities, including ICU.
beds and ventilators. The cluster 0 included counties with the lowest population density (24.35 persons/km$^2$), the largest percentage of elderly people (25.9%), and the fewest hospital beds per capita. These counties had high death rates per capita in spite of low population. Despite the positive correlation, other factors should be considered when evaluating the effect of population density on fatality rates. In other words, the population density is an essential (but not sufficient) variable to describe the vulnerability. Overall, the counties with higher population densities were more vulnerable than others, and government interventions, such as lockdowns and distancing restrictions in public places, were required at both pandemic peaks. The peak-to-peak analysis revealed that government interventions were more successful in populated counties by reducing their vulnerability index. Although the local governments successfully controlled the disaster and reduced damages in urban areas, they were unable to diminish the vulnerability index in rural areas.

The racial minorities were the second vulnerable community in the first pandemic peak, which was already approved in the literature (Karaye & Horney, 2020). According to the COVID-19 data, the minorities had greater COVID-19 hospitalization and mortality rates than the White race in the States (Centers for Disease Control & Prevention, 2022). The vulnerability of racial and ethnic minorities against the COVID-19 might be due to the previous inequalities in the American societies (Kim & Bostwick, 2020). The racial and ethnic minorities had higher rates of diabetes, hypertension, obesity, asthma, and heart diseases compared to other communities. In parallel with health parameters, lower education, household income, and employment rate are the economic factors associated with racial minorities that may prevent easy imposing of the restrictions among them. Noting the federal intervention by money allocation among the people, this feature lost its initial importance in the multi-variate model one year after the pandemic break. The comparative analysis among communities showed minorities and elder communities were more responsive to pandemic-related policies than poor communities. The COVULin score did not decrease after implementing the social restrictions among the poor communities indicating the ineffectiveness of governmental restrictions in poorer communities.

Fig. 9. Comparing counties in two pandemic peaks, considering the policies regarding (a) the number of counties reporting an increase or decrease in indicators (deaths, confirmed cases, CFR, and COVULin score) in the last peak compared with the first, and (b) four important variables in counties that imposed restrictive policies and reported an increase (red) or decrease (blue) in COVULin scores during the last peak (compare to the first peak).
counties. As the epidemic spread, the increasing unemployment rate and food insecurity were significant barriers to poor classes in complying with public health guidelines (Wolfson & Leung, 2020). These factors caused lower-income people to encounter the disease while becoming more vulnerable.

5.1. Limitations

The present study is built on some assumptions and limitations. First, although the county-level features are used for the model, these variables are for 2018. Some parameters such as poverty, minimum wage, and health facilities could be changed during the pandemic. Second, based on the limited data on the policies’ impact on the pandemic severity, the assumptions are infant, and a model based on the real impact of each policy is suggested for new research. Third, this research developed a static COVULin based on the features. However, coupling this method to dynamic data makes it possible to announce a dynamic vulnerability index and monitor the vulnerable areas during a disaster.

6. Conclusion

We developed and applied an innovative and multi-step model to detect regions in the U.S. that were most vulnerable to the COVID19 pandemic. This approach integrated reliable and high-functioning domains of machine learning as well as numerical models. A combination of multiple powerful algorithm is employed to obtain the most effective feature on the Covid-19 mortality. Beside, a COVID-19 vulnerability index (COVULin) was developed to describe county-level conditions. The outcomes show that a population density, portion of minority, poverty rate, and portion of people without access to the vehicle are the most influential parameters on the COVID-19 CFR. Furthermore, the outcomes showed the effectiveness of the mandate policies in the counties with a higher portion of poverty and uninsured people are lower than other counties and couldn’t decrease the COVID-19 CFR properly.

The feature-selection algorithm found that a higher population density was the most important feature associated with counties in the vulnerable category. Of counties with the highest mortality rates during the first and last pandemic peaks, 67% and 86%, respectively, were considered crowded. In addition, counties with a higher portion of minorities and people living below the poverty line needed more attention during the first and last pandemic peaks, respectively. These results confirm that a COVULin model based on social and non-social variables can provide reliable predictions of community vulnerability. Compared with previous investigations of a relatively brief period of the pandemic, this research used a multi-aspect and long-term view of the pandemic. Our analysis of the effectiveness of imposed policies showed they failed to decrease vulnerability in a third of the counties. This counties had a greater proportion of poor and uninsured inhabitants by 5% and 4%, respectively. These findings suggest that in the event of a long-term disaster, impoverished communities are more resistant to policy control and require more resources and attention to reduce their susceptibility. Although this approach proved effective during the COVID-19 pandemic in the U.S., it can be extended to other case studies and more vulnerable areas and help communities prepare for, mitigate, respond to, and recover from disasters.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1A2C2007838) and the Korea Ministry of Environment (MOE) as a Graduate School specialized in Climate Change.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.scsoce.2022.103990.

References

Acharya, R., & Porwal, A. (2020). A vulnerability index for the management of and response to the COVID-19 epidemic in India: An ecological study. The Lancet Global Health, 8(9), e1142–e1151. https://doi.org/10.1016/S2214-109X(20)30300-4
Ahmad, R., Williamson, M., Hamid, M. A., & Ashraf, N. (2020). United States county-level Covid-19 death rates and case fatality rates vary by region and urban status. Healthcare (Switzerland), 8(3), 330. https://doi.org/10.3390/healthcare8030330
Amram, O., Amiri, S., Lutz, R. B., Rajan, B., & Monsivais, P. (2020). Development of a vulnerability index for diagnosis with the novel coronavirus, COVID-19, in Washington State, USA. Health and Place, 64, Article 102377. https://doi.org/10.1016/j.healthplace.2020.102377
Aral, N., & Bakir, H. (2022). Spatiotemporal analysis of COVID-19 in Turkey. Sustainable Cities and Society, 76. https://doi.org/10.1016/j.scs.2021.103421
Aurelien Geron. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: Concepts, tools, and techniques to build intelligent systems. In O’ReillyMedia. https://www.oreilly.com/library/view/hands-on-machine-learnin g/9781492026322/
Ballotpedia. (2021). State-level mask requirements in response to the coronavirus (COVID-19) pandemic, 2020–2021 - Ballotpedia. https://ballotpedia.org/State-level_mask_requirements_in_response_to_the_coronavirus_(COVID-19)_pandemic, 2020–2021#Map of mask orders.
Barik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., & Stanton, C. (2020). The impact of COVID-19 on small business outcomes and expectations. Proceedings of the National Academy of Sciences of the United States of America, 117(30), 17656–17666. https://doi.org/10.1073/pnas.2006991117, National Academy of Sciences.
Boscaniaccio, C.M., Dione, K.Y., Hilbig, H., Humphreys, M., kc, S., Lieber, N. et al. (2020). Political and social correlates of Covid-19 mortality. 10.31235/osf.io/uh2zd. Boston University. (2021). Stay-at-Home order – CUSP. https://statepolicies.com/data/graphs/stay-at-home-order/.
Bloom, J. R., & Salemi, C. (2021). COVID-19 and conflict. World Development. https://doi.org/10.1016/j.worlddev.2020.105294
Cao, Y., Hiroya, A., & Montgomery, S. (2020). COVID-19 case-fatality rate and demographic and socioeconomic influencers: Worldwide spatial regression analysis based on country-level data. BMU Open, 10(11), Article e043560. https://doi.org/10.1136/bmjopen-2020-043560
Carrillo-Larco, R. M., & Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Welcome Open Research, 5(56). https://doi.org/10.26688/welcomereprints.15819.1
CDC. (2018). Centers for disease control and prevention/agency for toxic substances and disease registry/ geospatial research, analysis, and services program. CDC/ATSDR social vulnerability index [2018]Database [US]. https://www.atstd.cdc.gov/placecen d/builder/svs/data_documentation_download.html.
Johns Hopkins Coronavirus Resource Center. (2021). COVID-19 Map - Johns Hopkins Coronavirus Resource Center. In Johns Hopkins Coronavirus Resource Center (p. 1). https://coronavirus.jhu.edu/map.html.
Johns Hopkins Coronavirus Resource Center. (2021). COVID-19 Map - Johns Hopkins Coronavirus Resource Center. In John Hopkins Coronavirus Resource Center (p. 1). https://coronavirus.jhu.edu/map.html.
Johns Hopkins Coronavirus Resource Center. (2021). COVID-19 Map - Johns Hopkins Coronavirus Resource Center. In Johns Hopkins Coronavirus Resource Center (p. 1). https://coronavirus.jhu.edu/map.html.
Johns Hopkins Coronavirus Resource Center. (2021). COVID-19 Map - Johns Hopkins Coronavirus Resource Center. In Johns Hopkins Coronavirus Resource Center (p. 1). https://coronavirus.jhu.edu/map.html.
Centers for Disease Control and Prevention. (2022). Impact of racism on our nation’s health | Health Equity. Minority Health and Health Equity. https://www.cdc.gov/h ealthequity/racism-disparities/impact-of-racism.html.
Cevik, M., Kuppalli, K., Kindrachuk, J., & Peris, M. (2020). Virology, transmission, and pathogenesis of SARS-CoV-2. The BMJ, 371, 10.1136/bmj.m3862.
Chan, B., Korfanzy, E.R., Nataj, S., Park, J., Peng, L., & Zhang, J. (2021). Traffic monitoring with distributed acoustic sensing (pp. 1–11).
Chandu, V. C. (2020). Identification of spatial variations in COVID-19 epidemiological data using K-Means clustering algorithm: A global perspective. MedRxiv, 2020.06.03.20121194. 10.1101/2020.06.03.20121194.
Chaudhry, R., Dranitsaris, G., Mbushir, T., Bartoszkju, J., & Riaz, S. (2020). A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. E|ClinicalMedicine, 25, https://doi.org/10.1016/j.eclinm.2020.100464.
Sy, K. T. L., White, L. F., & Nichols, B. E. (2021). Population density and basic reproductive number of COVID-19 across United States counties. *PloS One*, 16, Article e0249271. https://doi.org/10.1371/journal.pone.0249271, 4 April.

The Lancet. (2020). Redefining vulnerability in the era of COVID-19. In *The Lancet* (Vol. 395, Issue 10230), p. 1089. Elsevier. 10.1016/S0140-6736(20)30757-1.

Tiwari, A., Dadhania, A. V., Ranganathrao, V. A. B., & Oliveira, E. R. A. (2021). Using machine learning to develop a novel COVID-19 vulnerability index (C19VI). *Science of the Total Environment*, 773, Article 145650. https://doi.org/10.1016/J.SCITOTENV.2021.145650

US Census Bureau. (2019). Index of programs-surveys/popest/datasets/2010-2018/counties/asrh/. https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/counties/asrh/.

US Census Bureau. (2020). Week 12 household pulse survey: July 16 - July 21. https://www.census.gov/data/tables/2020/demo/hhp/hhp12.html.

Van Dyke, M. E., Rogers, T. M., Pezvner, E., Satterwhite, C. L., Shah, H. B., Beckman, W. J., et al. (2020). Trends in county-level COVID-19 incidence in counties with and without a mask mandate — Kansas, June 1–August 23, 2020. *MMWR Morbidity and Mortality Weekly Report*, 69(47), 1777–1781. https://doi.org/10.15585/mmwr.mm6947e2

Verschuur, J., Koks, E. E., & Hall, J. W. (2021). Observed impacts of the COVID-19 pandemic on global trade. *Nature Human Behaviour*, 5(3), 305–307. https://doi.org/10.1038/s41562-021-01060-5. Nature Publishing Group.

Wadhera, R. K., Wadhera, P., Gaba, P., Figueroa, J. F., Joynt Maddox, K. E., Yeh, R. W., et al. (2020). Variation in COVID-19 hospitalizations and deaths across New York City Boroughs. *JAMA - Journal of the American Medical Association*, 323(21), 2192–2195. https://doi.org/10.1001/jama.2020.7197, American Medical Association.

Wolfson, J. A., & Leung, C. W. (2020). Food insecurity and COVID-19: Disparities in early effects for US adults. *Nutrients*, 12(6), 1648. https://doi.org/10.3390/nu12061648

World Health Organisation. (2020). Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-Emergency-Committee-Regarding-the-outbreak-of-Novel-Coronavirus-(2019-nCoV).

Xiao, C. (2020). A novel approach of consultation on 2019 novel coronavirus (COVID-19)-related psychological and mental problems: Structured letter therapy. *Psychiatry Investigation*, 17(2), 175–176. https://doi.org/10.30773/pi.2020.0047

Zhang, K., Vilches, T. N., Tariq, M., Galvani, A. P., & Moghadas, S. M. (2020). The impact of mask-wearing and shelter-in-place on COVID-19 outbreaks in the United States. *International Journal of Infectious Diseases*, 101, 334–341. https://doi.org/10.1016/j.ijid.2020.10.002

Zhou, Y., Feng, L., Zhang, X., Wang, Y., Wang, S., & Wu, T. (2021). Spatiotemporal patterns of the COVID-19 control measures impact on industrial production in Wuhan using time-series earth observation data. *Sustainable Cities and Society*, 75. https://doi.org/10.1016/j.scs.2021.103388