Predicting COVID-19 Transmission in Southern California with Machine Learning Methods

Han Li
School of Computer Science and Engineering
University of California, Riverside
Riverside, USA
han.li001@email.ucr.edu

Ran Wei
School of Public Policy
University of California, Riverside
Riverside, USA
ranwei@ucr.edu

Wenyu Wang, Nanpeng Yu
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, USA
wwang032@ucr.edu, nyu@ece.ucr.edu

***Abstract*—The COVID-19 pandemic has posed significant challenges to global health, economies, and social structures since its emergence in late 2019. Southern California, a region with unique characteristics and diverse communities, has experienced difficulties in controlling the virus's spread. Understanding the factors that influence COVID-19 transmission in Southern California is crucial for informing public health policies and mitigating the virus's impact on the communities. In this study, we collected weekly zip code level data from March to December 2020 in Southern California and applied various machine learning algorithms, such as graph neural networks (GNN), multi-layer perceptrons (MLP), and gradient boosting trees (XGBoost), to model the spread of COVID-19. Our main contributions include the development of various COVID-19 forecasting models that incorporate socioeconomically significant variables and zip code level mobility data, coupled with variable impact assessments to provide valuable insights for informing future public health policies and strategies. Our predictive model not only attains high levels of accuracy in forecasting but also facilitates the understanding of how each variable contributes to the final outcome. These findings have the potential to enhance our understanding of COVID-19 transmission patterns and inform targeted mitigation strategies in the region.**

***Index Terms*—COVID-19, graph neural networks, machine learning, pandemic modeling, regional mobility data.**

I. Introduction

COVID-19 has profoundly impacted global health, economies, and social structures. The number of reported cases rose to over 104 million, and the number of deaths rose to over 1 million by the end of April 2023 [1]. Southern California, a densely populated region with diverse communities, experienced significant challenges in containing the spread of the virus. The region's health care system was strained, and its economy was also affected by the outbreak [2]. Understanding the factors that influence the transmission of COVID-19 in Southern California is crucial for informing public health policies and mitigating the impact of the virus on communities.

Earlier research on the transmission of COVID-19 employed various epidemiological models, statistical learning methods, and machine learning algorithms to predict the spread of the virus in order to understand its dynamics. However, very little research focuses on the communities in Southern California, which is an important region to study COVID transmission due to its unique characteristics, such as a diverse population, high population density, and significant economic disparities. The role of social determinants of health, such as mobility, socioeconomic status, race, and access to healthcare, has not been comprehensively explored in the context of COVID-19 transmission in this region. It is important to understand these various factors on the spread of COVID-19 at the community level in Southern California to develop targeted strategies for mitigation and public health management in the future.

With this in view, in this paper, we aim to achieve the following goals:

1) Examine the spatial and temporal patterns of COVID-19 transmission in the region to forecast future trends in the area and explore the potential consequences these trends may have on public health policies and interventions.
2) Investigate the impact of some social determinants such as the population of a community, mobility, weekly spatial average of neighboring cases, racial composition, level of employment, and median household income on the spread of the virus in Southern California.

To achieve these goals, our study takes a unique approach by incorporating regional mobility data at the weekly and zip code level into a machine learning model. At such granularity, the data paints a detailed and nuanced picture of the local disease transmission dynamics, thereby filling a crucial gap in the existing literature. We utilize Graph Neural Networks (GNNs) and XGBoost to predict the spread of COVID-19 and interpret the impacts of socioeconomic features and population flows on the spread of the virus.

The rest of this paper is organized as follows. Section II provides an overview of the related work and elaborates on our contribution. Section III provides a detailed description of the data sources collected and pre-processing steps. Section IV describes specific models and algorithms employed in our study. In Section V, we present the results of the models, trends observed in the data, interpretation of the models, and the impact of social determinants on the spread of the virus.

Research reported in this work was supported by National Institute on Minority Health and Health Disparities (NIMHD) of the National Institutes of Health, United States under award number U54MD013368.

Lastly, Section VI summarizes the paper and presents potential future work.

II. Literature Review

In this section, we review the four primary methods for predicting COVID-19 cases: epidemiological models, statistical models, machine learning models, and hybrid models. The classifications of the four methods are shown in Fig 1. We also assess their strengths and weaknesses and examine how these models have been applied in existing research. Furthermore, we outline the contributions of our work and how it fills the gaps of the existing literature.

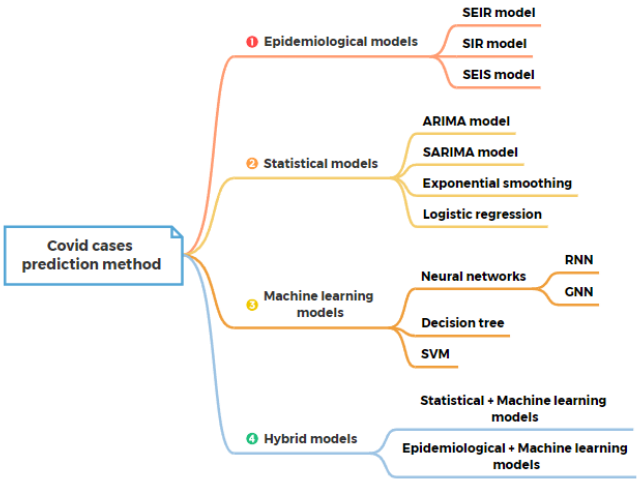

Fig. 1. Current COVID-19 cases prediction methods

A. Epidemiological Models

Epidemiological models attempt to explain the virus transmission process in real-world contexts. One widely used model is the SEIQR model [3], which subdivides human populations into five classes: Susceptible S(t), Exposed E(t), Hospitalized Infected I(t), Quarantine Q(t), and Recovered or Removed R(t) [4]. The sum of these five variables is assumed to be constant. The expression of each variable can be obtained by solving a set of differential equations that describe the transformation between them. After some simplifications, [5] uses four categories: Susceptible S(t), Exposed E(t), Hospitalized Infected I(t), and Recovered R(t) to simulate the process of COVID transmission. Reference [6] and [7] further simplify the model by dividing the population into three groups and use a disease transmission model to characterize the spread of COVID among these three groups.

Despite the usefulness of epidemiological models in understanding infectious diseases and informing health policies, they have limitations. These include the reliance on uncertain or changeable parameters and the tendency to simplify complex real-world scenarios into mathematical equations. They may make simple assumptions that do not hold true for real world diseases.

B. Statistical Models

Statistical models assume that past trends and patterns will continue in the future, with examples of time series analysis, ARIMA models, and exponential smoothing.

Benvenuto [8] uses the Auto Regressive Integrated Moving Average (ARIMA) model on the Johns Hopkins epidemiological data to predict the epidemiological trend of the prevalence and incidence of COVID. Zhao [9] builds a grey Verhulst model using data sequences to predict the daily growth trend of the number of patients confirmed with COVID-19 infection in China. The results show that the grey Verhulst model can predict the S-shaped change characteristics of the cumulative number of confirmed patients. Torrealba-Rodriguez [10] uses a logistic growth model to describe the process of COVID transmission. To solve the differential equations, the nonlinear least square method is used to estimate the parameters in the model. The model exhibits high prediction accuracy for COVID transmission, but estimating the parameters using nonlinear least squares is cumbersome and not universally applicable.

The idea of statistical methods is straightforward. However, these models may not capture the complexity of disease transmission dynamics due to their structure, nor account for changes in underlying population or disease transmission parameters over time.

C. Machine Learning Models

Machine learning models identify patterns and make predictions using large amounts of data. These include decision trees, neural networks, logistic regression, and support vector machines. While these models can detect complex patterns and interactions between variables, their interpretability are often not satisfactory.

In Yang's work [11], the long short-term memory (LSTM) model is used to process and predict the number of new infections over time. They incorporate the COVID epidemiological parameters, such as the probability of transmission, the probability of recovery or death to train the LSTM model. In Chimmula's work [12], a similar method is used to predict COVID transmission. The LSTM network is trained to forecast future COVID cases in Canada.

As another neural network method, Graph Neural Networks (GNNs) excel at capturing the structure of complex networks and modeling intricate relationships, making them effective for predicting the number of virus cases. They can process contextual information and handle temporal dynamics, crucial for accurately modeling the spread of virus.

Fritz [13] presents a multimodal learning framework that amalgamates statistical regression and machine learning models to predict local COVID-19 cases in Germany. Results show that the GNN approach enables the use of a richer collection of data types, including mobility flows and colocation probabilities, thus yielding the lowest mean squared error throughout the observational period in the reported benchmark study. Similar methods can be found in Shah's [14] and Gao's [15] work. Kapoor [16] proposes an approach that learns from a

single large-scale spatiotemporal graph, where nodes represent the region-level human mobility, spatial edges represent the human mobility-based inter-region connectivity, and temporal edges represent node features through time. Evaluations and tests on the US county-level COVID dataset demonstrate that the rich spatial and temporal information leveraged by the graph neural network allows the model to learn complex dynamics.

Machine learning models can detect complex patterns and interactions between variables. One difficulty is that neural networks models may be difficult to interpret. In our work, the data includes many regional socioeconomic features and population flow data between regions. We aim to explore the interpretability of GNN model and identify variables that contribute to the spread of COVID, which are the unique contributions of this paper.

D. Hybrid Models

Hybrid models leverage the strengths of statistical and machine learning models to improve the accuracy of predictions. Gatto [17] combines local SEIR methods with regional mobility data. Compared with the traditional local SEIR model, this regional flow SEIR model is more accurate in prediction, and correspondingly, more parameters are needed. Zheng [18] combines SEIR and RNN on a graph structure to develop a hybrid spatiotemporal model. For node features in graphs, they derive a discrete recursion from SEIR so that the gradient descent method can be applied for optimization. For the edge feature, an RNN model is used to capture the neighboring effect and regularize the landscape of loss function so that local minima are effective and robust for prediction.

These models are beneficial as they integrate the strengths of various models but can be more complex and challenging to implement than single-model approaches.

E. Gaps in the Existing Literature and Contributions

Existing research has applied epidemiological, statistical, machine learning, and hybrid models to predict COVID-19 cases. However, there remain gaps in understanding the impacts of demographic and regional mobility data on COVID transmission by interpreting machine learning models, notably Graph Neural Networks (GNNs).

Recognizing these research gaps, this paper offers the following contributions:

- **Interpretability of GNN Models**: We aim to enhance the interpretability of GNN models. While these models are proficient in detecting complex patterns, their intricacy often results in reduced interpretability. By improving interpretability, we can provide insights into the features deemed significant for prediction by the model. This could be a valuable resource for public health professionals and policy-makers in identifying critical areas to address during a pandemic.
- **Inclusion of Regional Mobility Data, Socioeconomic Features, and Population Flow Data**: Our work stands apart in its inclusion of regional mobility data alongside socioeconomic features and population flow data in the modeling process. This comprehensive approach can provide a more nuanced understanding of disease transmission dynamics, considering not only the health-related aspects but also the influence of social behaviors and population movements.
- **Identification of Key Variables for COVID-19 Spread**: Through our unique modeling approach, we identified key variables that contribute significantly to the spread of COVID-19. This understanding could help policy makers focus in these areas and develop targeted interventions to control the spread of virus.

By addressing these areas, our work seeks to improve the comprehension and effectiveness of machine learning models in predicting COVID-19 cases, thereby informing better strategies for managing and mitigating pandemics.

III. Data overview

The dataset we used in this study spans from March 23, 2020 to December 14, 2020. This dataset comprises weekly COVID-19 case numbers for 556 zip codes across Southern California.

This data reveals two major surges in COVID-19 cases throughout this timeframe. The initial peak appeared in July 2020, with a more pronounced resurgence taking place in December 2020, as depicted in Figure 2. This trend is indicative of the highly contagious nature of the virus and presents a research challenge: deciphering the spatiotemporal pattern of virus transmission using historical data. A key issue we explored in this research is how to utilize the data from March to October to forecast the number of COVID-19 cases for November and December.

To tackle this forecasting problem, we compiled a dataset which includes the number of COVID-19 cases for each zip code on a weekly basis, along with 26 additional features. A detailed overview of these features can be found in Table I. As there may be complex interactions among the input features, all of them are included in the model training.

We have also visualized the progression of the pandemic in Southern California through a heat map (Figure 3), with the scale represented in logarithmic indices. The darker shade corresponds to a higher number of cases. The heat map underlines the regional nature of the COVID-19 spread, highlighting a higher rate of transmission in areas with a larger number of cases.

It is noteworthy that the dataset incorporates zip code level socioeconomic variables, enabling us to analyze potential correlations between these variables and COVID-19 transmission. It stands to reason that wealthier communities, denoted by a higher per capita income, have greater access to medical resources, are more health-conscious, and could be affected differently by the spread of the virus.

In the ensuing section, we delve into our modeling methodology and the specifics of our experiment. This encompasses

TABLE I
VARIABLES NAME AND DESCRIPTION IN OUR DATA SET.

Index	Feature name	Description
1	this-week-case	number of weekly new covid cases
2	last-week-case	number of last week's covid cases
3	access	weekly testing site-to-population ratio
4	median-contacts	estimated weekly sum close contact index
5	pop-flows	estimated weekly sum population flow to this zipcode
6	average-neighbor-case	weekly spatial average of neighboring zipcodes' covid cases, weighted by "pop-flows"
7	total-pop	2020 total population of this zip code
8	pop-density	population density
9	employ-pop	employed population
10	emp-pop-density	employed population density
11	median-household-income	median household income
12	p-unemploy	percentage of population unemployed
13	p-poverty	percentage of population below the poverty level
14	p-24	percentage of age under 24
15	p-25-44	percentage of age 24 to 44
16	p-45-64	percentage of age 44 to 64
17	p-hispanic	percentage of ethnicity-Hispanic or Latino (of any race)
18	p-white	percentage of race-White
19	p-black	percentage of race-Black or African American
20	p-asian	percentage of race-Asian
21	p-car	percentage of commuting to work - car (drove alone + carpooled)
22	p-pubtran	percentage of commuting to work - public transportation (excluding taxi-cab)
23	p-walked	percentage of commuting to work - worked from home
24	rate	percentage increase in the number of cases this week compared to last week
25	week-num	current week number
26	zipcode	zip code number

network construction, hyperparameter selection, and a comparison of our prediction metrics with those of some baseline methods.

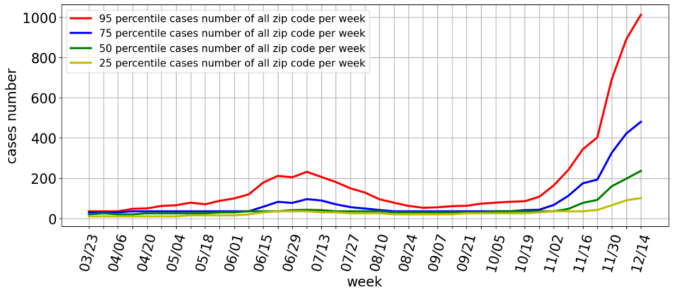

Fig. 2. Weekly COVID-19 cases in Southern California

IV. METHOD

In this project, we use three different machine learning models to predict future COVID-19 cases based on historical infections and input features of each zip code. These models include the GNN model, the multilayer perceptron, and gradient boosting tree models. All models are trained using the spatio-temporal dataset, which includes regional mobility data at a weekly-zip-code level, along with other socioeconomic and demographic variables. The models are evaluated and compared based on their predictive performance, by which we can highlight the strengths and weaknesses of each method.

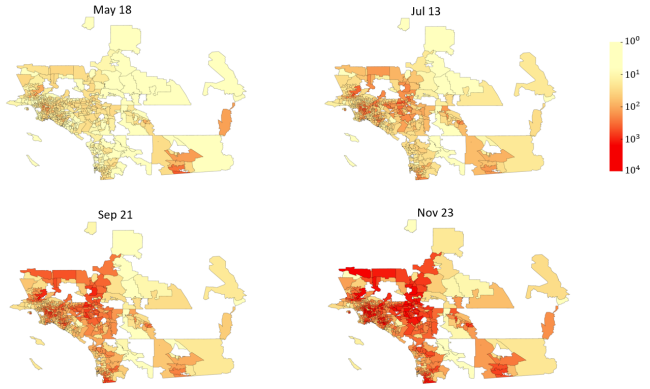

Fig. 3. COVID cases number heat map in Southern California

A. Graph neural network model

1) Problem formulation: GNNs operate on data with graph structures, leveraging the inherent structure and relationships within these graphs to extract valuable information and make predictions. In our study, we employ the GraphSAGE model to predict COVID-19 case numbers at the zip code level across Southern California. Input data includes zip code level historical weekly numbers of COVID-19 cases, population-related statistics, socioeconomic data, and the weekly population flow between zip codes, which provides a measure of regional mobility.

2) Graph construction: Graph nodes: We construct a graph, denoted as $G(\mathcal{V}, \mathcal{E})$, to encapsulate the weekly input data. Each zip code is represented as a graph node, linked with a feature vector that houses both static and dynamic attributes for that specific zip code location. The comprehensive construction encompasses 39 subgraphs and, within each subgraph, we have 556 nodes, indicating the zip-code level granularity of the data.

Graph edges: The graph's edges are designed to account for significant influence of population mobility on disease transmission patterns. We posit a direct correlation between the number of people moving between two areas and the COVID-19 cases in these locations. Therefore, we establish an edge between every pair of areas that have seen population movement. Each edge's weight signifies the number of population movements, enabling us to construct a directed graph that captures the dynamics of regional mobility.

Node features: As for the node features, every node, denoted as n_i, is associated with a 26-dimensional feature vector. This vector includes features such as the current week's COVID-19 case numbers, the previous week's case numbers, weekly spatial average of neighboring zip codes' COVID-19 cases,

population flow, total population, population density, median contacts, accessibility, income, age, race, among others. For a comprehensive breakdown of the node feature variables, please refer to Table I.

B. Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a type of feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. MLP uses a supervised learning technique called back propagation for training. It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. For COVID-19 prediction, an MLP can be trained using the back propagation algorithm, which minimizes the error in the prediction by adjusting the model parameters. MLP is particularly good at capturing complex patterns and interactions in the data, which is useful when dealing with high-dimensional regional mobility data for predicting the spread of COVID-19. The model can learn from temporal and spatial dynamics, and use this information to forecast future outbreaks.

C. XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. XGBoost provides a parallel tree boosting to solve many data science problems in a fast and accurate way. For COVID-19 prediction, we can use XGBoost to capture complex non-linear relationships in our high-dimensional data. The model learns from the trend and pattern of the virus spread in our weekly-zip-code level data and can make accurate predictions of future cases. The most influential features can be identified by analyzing feature importance, which provides insights into the main drivers of COVID-19 transmission.

V. Results and discussion

In this section, we present the results of the study and discuss the implications of the findings for the modeling, prediction, and interpretation of COVID-19 spread in Southern California. In this section, we first describe the experimental settings, which include details of data processing and hyper-parameters selection. Then, we report the performance of the proposed models and the baseline models. Lastly, we discuss the interpretation of the machine learning models.

A. Experimental setup

In this subsection, we describe the process of data pre-processing and transformation, which was essential for the accurate modeling and prediction of COVID-19 in Southern California. The data processing steps included data cleaning and aggregation. Then we discuss the process of hyperparameter selection for our machine-learning models.

1) Data processing: Data cleaning was the first step in the data processing pipeline, ensuring the quality and consistency of the data used in the analysis. We addressed missing values, and outliers by employing the following techniques: for missing values, we imputed missing values using the mean of the respective variable; for outliers, we identified and removed extreme outliers based on the interquartile range (IQR). The dataset included observations at different spatial and temporal scales, requiring aggregation to facilitate the analysis. We aggregated the data at a zip code level for spatial consistency and used a weekly time scale to capture temporal trends. This approach allowed us to reduce noise and improve the robustness of our models training process.

TABLE II
HYPER-PARAMETER COMBINATION IN GNN.

Index	Hyper-parameter name	Value
1	hidden_layer_sizes	[(32,64,128,256),(8,16)]
2	convolutional_layers	Gcn_Conv, Gat_Conv, Sage_Conv, Graph_Conv
3	activation	relu, tanh
4	optimizer	SGD, adam
5	dropout_rate	i/10.0 for i in range(0,10)

TABLE III
HYPER-PARAMETER COMBINATION IN MLP.

Index	Hyper-param name	Value
1	hidden_layer_sizes	[(32,64,128,256),(8,16)]
2	activation	relu, tanh
3	optimizer	SGD, adam
4	dropout_rate	i/10.0 for i in range(0,10)

2) Hyperparameter selection: For GNN model, the following hyper-parameters are adjusted to make the model reach its best performance: *hidden_layer_sizes, convolutional_layers, activation, optimizer*, and *dropout_rate*. The hyperparameter combinations are shown in Table II. It is also worth mentioning that the number of hidden layers in the GNN model is 2.

Grid search is conducted over the space of the above hyper-parameters combination in Table II. This involves specifying a range of values for each hyperparameter and then training and evaluating the model for all possible combinations of these hyperparameters. We implement Grid Search using the existing library scikit-learn's Grid-SearchCV. In the experiment, we find that the best hyperparameter combination is as follows: *hidden_layer_sizes*: [256,8], *convolutional_layers*: Sage_Conv, *activation*: relu, *optimizer*: adam, *dropout_rate*: 0.1.

When the MLP model is used in the experiment, the following parameters are adjusted to make the model reach its best performance: *hidden_layer_sizes, activation, optimizer*, and *dropout_rate*. The parameter combinations are shown in Table III. It is also worth mentioning that the number of hidden layers in the MLP model is also 2. In the experiment, we find that the best hyperparameter combination is as follows: *hidden_layer_sizes*: [256,16], *activation*: relu, *optimizer*: adam, *dropout_rate*: 0.3.

When the XGBoost model is used in the experiment, the following hyperparameters are adjusted to make the model reach its best performance: *n_estimators, min_child_weight, max_depth, gamma*, and *reg_alpha*. The hyperparameter combinations are shown in Table IV. In the experiment, we

TABLE IV
HYPER-PARAMETER COMBINATION IN XGBOOST.

Index	Hyper-param name	Value
1	n_estimators	100, 500
2	max_depth	int(3,10)
3	min_child_weight	int(1,6)
4	gamma	i/10.0 for i in range(0,5)
5	reg_alpha	1e-5, 1e-2, 0.1, 1, 100

find that the best hyperparameters combination is as follows: *n_estimators*: 500, *min_child_weight*: 5, *max_depth*: 4, *gamma*: 0.4, *reg_alpha*: 1e-05.

B. Model performance metrics

The models' performance is evaluated using various metrics, including the mean absolute error (MAE), the root mean squared error (RMSE), and the coefficient of determination (R^2).

(1) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y_i}\right)^2} \tag{1}$$

(2) Mean Absolute Error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widehat{y_i}\right| \tag{2}$$

(3) Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}, \tag{3}$$

where y_i and $\widehat{y_i}$ represent the actual COVID case number and predicted case number of the ith week. N is the number of test data points and $\bar{Y}$ is the average of y_i .

The model performance comparison presented in Table V demonstrates varying degrees of prediction accuracy on the dataset. The three machine models analyzed include MLP, GNN, and XGBoost, with each model yielding different RMSE, MAE, and R^2 values.

The XGBoost model outperforms the other two models, achieving the lowest RMSE of 63.4 and the highest R^2 value of 97.7%, indicating a strong correlation between the predicted and observed values. It also has a relatively low MAE of 35.4, which shows that its predictions are generally close to the true values.

In contrast, the MLP model has the highest RMSE of 77.0 and the lowest R^2 value of 96.5%, which suggests it is less accurate in capturing the underlying patterns in the data compared to the XGBoost model. The GNN model has a slightly better performance than MLP, with an RMSE of 69.9 and an R^2 value of 93.6%, but it is still not as accurate as the XGBoost model.

It is important to note that our models utilize the initial thirty weeks of the dataset for training and validation and the subsequent nine weeks for testing. As observed in Fig 2, the quartile of the number of cases during the first 30 weeks does not surpass 100, whereas the quartile of the number of cases for the following 9 weeks exceeds 500. The XGBoost model achieves a predicted RMSE value close to 60, which demonstrates that it effectively learns the virus transmission trends and the impacts of various factors on the COVID-19 transmission.

TABLE V
MODEL PREDICATION ACCURACY ON OUR DATASET

Method	$RMSE$	MAE	$R^2(\%)$
MLP	77.0	46.4	96.5
GNN	69.9	37.6	93.6
$XGBoost$	**63.4**	35.4	97.7

Based on these results, the XGBoost model demonstrates the best overall performance in predicting the COVID-19 case numbers on our dataset. This superior performance suggests that the XGBoost model is better suited for capturing the complex relationships between the various socioeconomically relevant variables and mobility data in the context of COVID-19 transmission.

C. Model interpretation

1) GNN: In this study, we further enhance the interpretability of the GNN model in predicting COVID-19 cases by employing GNN Explainability [19], a method that allows us to ascertain the importance of node features in the model's predictions. GNN Explainability utilizes the GNNExplainer technique, which can be found in the PyTorch Geometric library.

GNNExplainer, is a model-agnostic method designed to provide explanations for the decisions made by GNN models. It works by learning a soft mask over the input node features and edges to determine the most relevant subgraph structure that contributes to the model's prediction results. This approach enables us to identify important features and connections within the graph that influence the spread of COVID-19, ultimately leading to a better understanding of the factors affecting virus transmission dynamics.

Based on the GNN Explainability analysis, we have identified the top 10 node features that significantly influence the model's predictions for COVID-19 cases. Fig. 4 presents these features along with their corresponding feature importance scores.

Fig. 4 suggests that a region's total and employed population, the proportion of the Hispanic population, and median household income are among the most critical factors affecting the spread of COVID-19. Furthermore, our model also emphasizes the importance of recent trends in case numbers (both in the current week and the previous week), population density, car access, age distribution, and neighboring area case counts.

By pinpointing these key factors, our study can aid public health officials and policymakers in implementing targeted interventions that address these specific aspects of the population. For instance, efforts to reduce the spread of COVID-19

can be concentrated in areas with high population density and high numbers of employed individuals. Additionally, the findings can help direct resources and support toward communities with lower median household incomes and higher proportions of Hispanic residents, who may be more vulnerable to the virus.

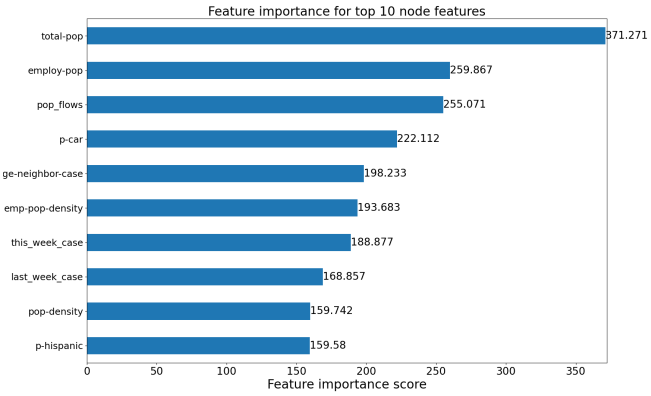

Fig. 4. Values for feature importance in the GNN model

2) XGBoost: In order to better understand the contribution of each feature in the XGBoost model, we employ the concept of Shapley Values, which originated from cooperative game theory. Shapley Values provide a fair distribution of the contributions of each player (in this case, features) in a cooperative game. In the context of machine learning models, Shapley Values help us to quantify the contribution of each feature to model's prediction for COVID cases. This approach provides valuable insights into the importance of individual features and their effects on the model's predictions [20]. Fig. 5 reveals that the following features have the most significant impact on the model's predictions: total-pop, average neighborhood case, p-hispanic, employ-pop, median-contacts, and median-household-income.

We will analyze the Shapley Values of these variables separately to find out their impacts on COVID spreading. We divide time into four windows based on the different phases of virus spread in the data. The first stage is the first six weeks before the virus first appeared, during which everyone is advised to stay at home; the second stage is from week seven to week twelve, corresponding to the period of reopening; the third stage is from thirteen to thirty weeks in the data set, during which time people are advised to stay at home to reduce the spread of the virus; the fourth stage is from the thirty-first week to the thirty-ninth week. During this period of reopening, the virus spread reached a new peak.

Firstly, we analyze the impact of this week's cases and last week's cases, and draw a dependence plot, as shown in Fig 6. The first subplot shows the impact of this week's cases, we can see that when this week's cases are greater than a certain value (e.g. 50), it has a positive impact on the COVID spread, and vice versa. We can also find a similar pattern in the second

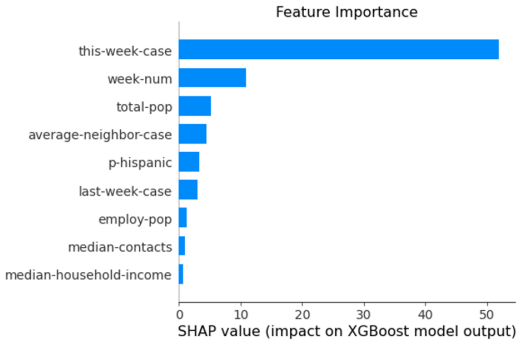

Fig. 5. Shapley values for feature importance in the XGBoost model

subplot. This shows that it is much easier to control the spread of the virus when it is spreading early than it is to take action after the virus has spreaded.

Then, we analyze the impact of social-economic variables and draw a dependence in Fig. 7. The first subplot shows the impact of p-hispanic. From the subplot, we can find that the areas where Hispanics account for less than 30% are not conducive to COVID transmission, and the areas where Hispanics account for more than 70% are conducive to COVID transmission. This phenomenon may be related to the living habits of this ethnic group. Frequent communication between neighbors and living together with huge families may lead to the acceleration of the spread of the COVID virus. In addition, areas with a Hispanic proportion of more than 70% are conducive to the spread of the virus in the four stages of virus transmission. It can be seen that areas with a Hispanic proportion of more than 70% significantly impact the spread of the virus during the COVID-19 outbreak. We also analyzed some of the outlier points with high SHAP values. These areas exhibit strong clustering characteristics, which means the areas with high proportions of Hispanics are often close to each other. This provides convenience for frequent visits between families. In addition, Hispanics tend to live with big families, thus the virus spreads faster in these areas. We look into the areas with low p-hispanic, which are scattered and not connected to each other. People in these areas communicate less with other areas during the pandemic, which also explains to some extent why these areas are less affected by the virus.

The second subplot shows the impact of employ-population. Areas with fewer than 5,000 working people are a barrier to virus spread. In these areas, fewer people go out, these areas have less contact with the outside world, thus the spread of the virus is hindered. On the contrary, areas with more than 25,000 people with jobs can increase the spread of COVID. The more people go out to work, the more exchanges with densely populated areas there are. This phenomenon also confirms the effectiveness of staying at home and reducing going out to block the spread of the virus. In addition, in the initial stage of virus transmission, regardless of the number of employees, each region has a tendency to spread the virus, indicating that everyone's normal living will spread the virus, and protective

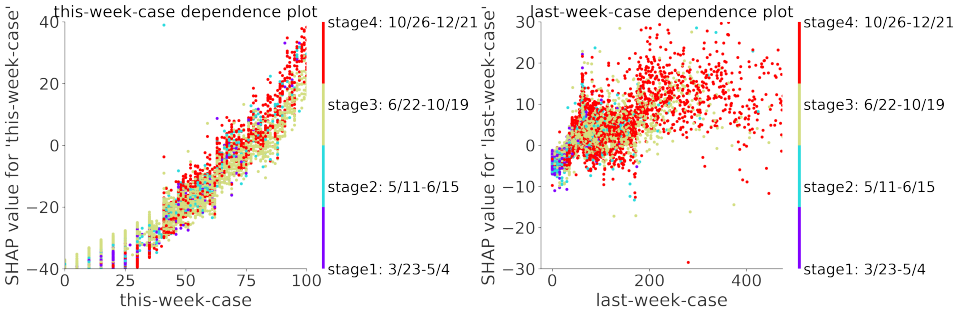

Fig. 6. Shapley values for cases variables in XGBoost model

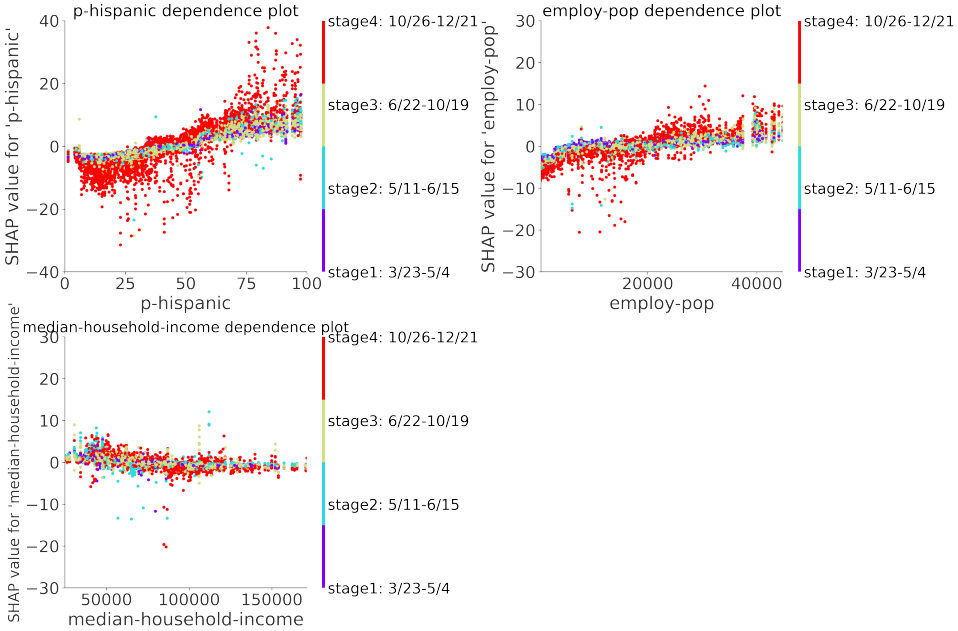

Fig. 7. Shapley values for social-economic variables in the XGBoost model

measures need to be taken.

The impact of median household income is shown in the third subplot. In this subplot, we find that the higher the average income, the greater the obstacles encountered in the spread of the virus, and this phenomenon becomes especially obvious in the later stages. Areas with an annual income of more than $120,000 are basically not affected by the spread of the virus at all stages. However, areas with an annual income of less than $40,000 are greatly affected by the spread of the virus. From Fig. 4, we can see that the spread of the virus in areas near the Pacific Ocean is generally slower than inland areas. To further study this effect, we take a closer look at Fig. 7 and evaluate some of the outlier points which have rather low SHAP values. Some of these points correspond to zip codes such as 90731, 90740, and 91911, which are regions close to the ocean. These beach cities are spots where people go for vacation. However, When the stay-at-home regulations were implemented, people were asked to go out less, and the number of people in these beach cities was greatly reduced, so they were much less affected by the virus than inland cities. In addition, the per-household income of seaside communities is higher and such families may have more resources to pay for personal protection and family health.

The impacts of mobility variables are shown in Fig. 8. We can tell from the first subplot that the lower average-neighbor-case has a hindering effect on the spread of the virus. The higher average-neighbor-case leads to faster spread of the virus. However, after the average-neighbor-case reaches a certain threshold, the spread of the virus gradually reaches saturation. From the perspective of controlling the spread of the virus, it is very effective to block the spread of the virus in the early stage, but it is difficult to control the virus in the later stage.

The second subplot and the third subplot also show the impact of pop flow and p-car on COVID cases. From subplot 2, we can not identify the direct impact of pop-flow on cases, but some cases still show that pop-flows exhibit the characteristics of promoting the spread of virus. In subplot 3, the positive

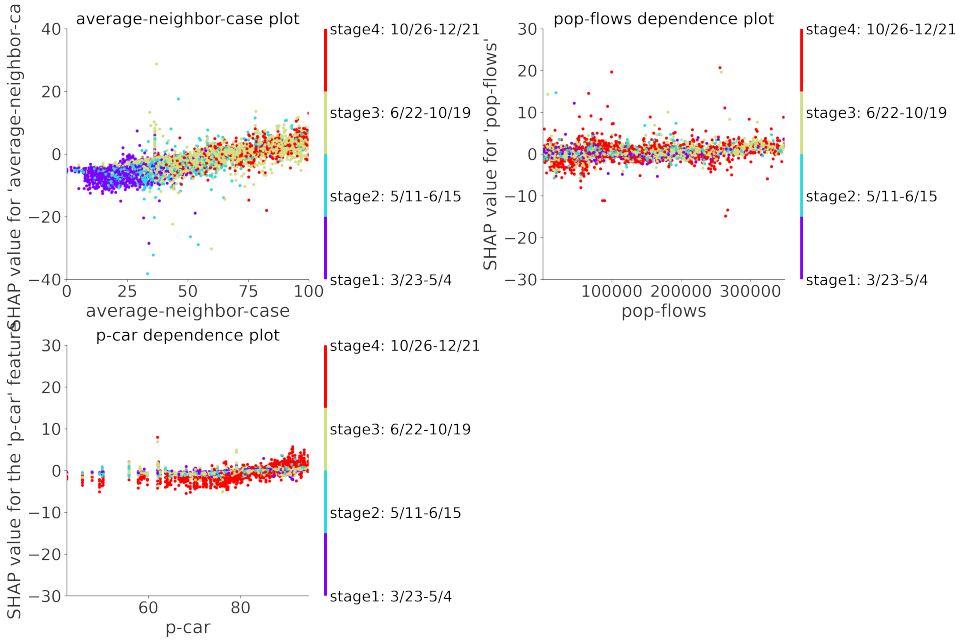

Fig. 8. Shapley values for mobility variables in the XGBoost model

impact of p-car on COVID cases is much more obvious. Especially when the time is between the end of October and the end of December, p-car has a positive impact on COVID cases.

These findings suggest that areas with higher populations, and lower socioeconomic status may be more vulnerable to the spread of COVID-19. Moreover, the racial and ethnic composition and reliance on public transportation also appear to play a role in the transmission dynamics in Southern California.

VI. Conclusion

In conclusion, this paper presents a comprehensive analysis of modeling, forecasting, and interpreting COVID-19 transmission in Southern California using various machine-learning models. Our study highlights the importance of several features in determining the virus's transmission dynamics, including the total population, weekly spatial average of neighboring cases, percentage of Hispanic ethnicity, employed population, estimated weekly close contacts, and median household income.

Our predictive framework showcases high prediction accuracy that can be used in practice while allowing interpretation of variable impacts. We employ Shapley Values for feature interpretation, revealing valuable insights into the traits of regions more prone to COVID-19 transmission. Our results emphasize that areas with larger Hispanic populations, lower socioeconomic status, and higher population density are more susceptible to the virus's spread. This information can guide public health officials and policymakers in designing targeted interventions, such as enhanced community outreach, more accessible testing, and increased resources for vulnerable communities.

Additionally, our study reinforces the effectiveness of stay-at-home measures and limited social interactions in curbing COVID-19 transmission. The relationship between employ-pop and virus transmission demonstrates the importance of minimizing contact, particularly in densely populated areas, to control outbreaks. By understanding the factors influencing transmission dynamics, we can better prepare ourselves for future public health challenges and strive to create a more resilient and equitable society in the face of pandemics.

References

[1] CDC, "COVID Data Tracker," https://covid.cdc.gov/covid-data-tracker, Mar. 2023.

[2] M. A. Achterberg, B. Prasse, L. Ma, S. Trajanovski, M. Kitsak, and P. Van Mieghem, "Comparing the accuracy of several network-based COVID-19 prediction algorithms," *International Journal of Forecasting*, vol. 38, no. 2, pp. 489–504, Apr. 2022.

[3] M. Mandal, S. Jana, S. K. Nandi, A. Khatua, S. Adak, and T. K. Kar, "A model based study on the dynamics of COVID-19: Prediction and control," *Chaos, Solitons & Fractals*, vol. 136, p. 109889, Jul. 2020.

[4] K. Chatterjee, K. Chatterjee, A. Kumar, and S. Shankar, "Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model," *Medical Journal Armed Forces India*, vol. 76, no. 2, pp. 147–155, Apr. 2020.

[5] S. He, Y. Peng, and K. Sun, "SEIR modeling of the COVID-19 and its dynamics," *Nonlinear Dynamics*, vol. 101, no. 3, pp. 1667–1680, Aug. 2020.

[6] P. Martínez-Fernández, Z. Fernández-Muñiz, A. Cernea, J. L. Fernández-Martínez, and A. Kloczkowski, "Three mathematical models for COVID-19 prediction," *Mathematics*, vol. 11, no. 3, p. 506, Jan. 2023.

[7] M. Siwiak, P. Szczesny, and M. Siwiak, "From the index case to global spread: The global mobility based modelling of the COVID-19 pandemic implies higher infection rate and lower detection ratio than current estimates," *PeerJ*, vol. 8, p. e9548, Jul. 2020.

[8] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, "Application of the ARIMA model on the COVID-2019 epidemic dataset," *Data in Brief*, vol. 29, p. 105340, Apr. 2020.

[9] Y.-F. Zhao, M.-H. Shou, and Z.-X. Wang, "Prediction of the number of patients infected with COVID-19 based on rolling grey verhulst models," *International Journal of Environmental Research and Public Health*, vol. 17, no. 12, p. 4582, Jan. 2020.

[10] O. Torrealba-Rodriguez, R. A. Conde-Gutiérrez, and A. L. Hernández-Javier, "Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models," *Chaos, Solitons & Fractals*, vol. 138, p. 109946, Sep. 2020.

[11] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, and J. He, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, pp. 165–174, Mar. 2020.

[12] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals*, vol. 135, p. 109864, Jun. 2020.

[13] C. Fritz, E. Dorigatti, and D. Rügamer, "Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany," *Scientific Reports*, vol. 12, no. 1, p. 3930, Mar. 2022.

[14] C. Shah, N. Dehmamy, N. Perra, M. Chinazzi, A.-L. Barabási, A. Vespignani, and R. Yu, "Finding patient zero: Learning contagion source with graph neural networks," *arxiv*, Jun. 2020.

[15] J. Gao, R. Sharma, C. Qian, L. M. Glass, J. Spaeder, J. Romberg, J. Sun, and C. Xiao, "STAN: Spatio-temporal attention network for pandemic prediction using real-world evidence," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 733–743, Apr. 2021.

[16] A. Kapoor, X. Ben, L. Liu, B. Perozzi, M. Barnes, M. Blais, and S. O'Banion, "Examining COVID-19 forecasting using spatio-temporal graph neural networks," in *16th International Workshop on Mining and Learning with Graphs*, Aug. 2020.

[17] M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, and A. Rinaldo, "Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures," *Proceedings of the National Academy of Sciences*, vol. 117, no. 19, pp. 10 484–10 491, May 2020.

[18] Y. Zheng, Z. Li, J. Xin, and G. Zhou, "A spatial-temporal graph based hybrid infectious disease model with application to COVID-19," in *The 10th International Conference on Pattern Recognition Applications and Method*, 2020, pp. 357–364.

[19] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Annual Conference on Neural Information Processing Systems*, Dec. 2019, pp. 9240–9251.

[20] W. Wang, N. Yu, J. Shi, and N. Navarro, "Diversity factor prediction for distribution feeders with interpretable machine learning algorithms," in *IEEE Power & Energy Society General Meeting (PESGM)*, 2020, pp. 1–5.