Predicting Vaccination Uptake and COVID-19 Cases Using Machine Learning

by Jan Park (jaeunp), Sungmyeon Park (sungmyep), Pleng Witayaweerasak (pwitayaw), Pongsakorn Supakpanichkul (psupak)

1. Introduction

This project aims to develop predictive models for two key COVID-19 outcomes: vaccine uptake and the percentage of positive test cases. These targets are critical for informing public health interventions, resource allocation, and future pandemic preparedness. We will address the following sections through two distinct tasks:

Task 1: Predicting county-level vaccination rates.
Task 2: Predicting county-level COVID-19 positive cases.

The dataset used for this analysis comes from the COVID-19 Trends and Impact Survey (CTIS) conducted by Carnegie Mellon University in partnership with Facebook. It includes 25,626 entries spanning from January 7 to February 12, 2021, representing daily county-level responses from U.S. adults. It contains 19 features covering behaviors (e.g., mask usage, social activity), beliefs (e.g., trust in government or WHO recommendations), and health outcomes (e.g., vaccination status and test positivity rates). The primary target variables are smoothed_wcovid_vaccinated and smoothed_wtested_positive_14d. This dataset enables the exploration of social determinants and temporal patterns influencing pandemic outcomes.

2. Data Analysis

Task 1: Predicting vaccination rate

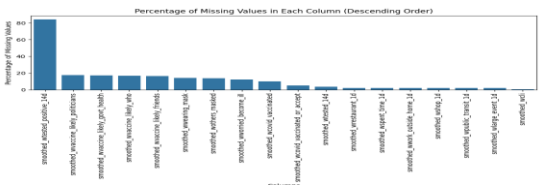
To forecast county-level COVID-19 vaccination rates, we first addressed 2,602 missing values (10%) in smoothed_wcovid_vaccinated by applying forward-fill within geographical groups (geo_value), preserving time-series integrity. No significant outliers were detected outside of 2 SDs (or z = 1.96) based on z-scores from a moving average, and all variables were already on a percentage scale, requiring no further normalization. We enriched the dataset by merging FIPS codes to add state and county names. For temporal integrity, the data was chronologically split into training (pre-Jan 29, 2021), validation (Jan 30–Feb 5), and test sets (Feb 6 onward).

Feature engineering focused on short-term dynamics (3-day lags, rolling averages), behavioral interactions (e.g., 'smoothed_wworried_become_ill' x 'smoothed_wwork_outside_home_1d'), and time-based features like day-of-week. These approaches are well-suited for county-level vaccination prediction as they capture both spatial and temporal patterns, preserve time-series integrity, and reflect behavioral and lagged trends influencing vaccination behavior. Lag features, rolling averages, and interaction terms enhance the model's ability to detect momentum and real-world dynamics.

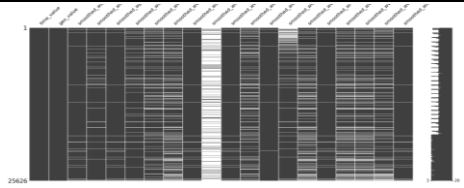
Correlation analysis reveals that the strongest behavioral predictors of vaccination rates include shopping outside (smoothed_wshop_1d, 0.40), working outside the home (smoothed_wwork_outside_home_1d, 0.36), and time spent outside (smoothed_wspent_time_1d, 0.31). These activities likely correlate with higher vaccination uptake because individuals engaging in frequent public interactions may perceive greater personal risk of exposure and thus be more motivated to protect themselves through vaccination.

Task 2: Predicting COVID-19 positive cases

The target variable has 21,632 missing values (84%). After dropping these rows, 3,994 rows remained. For the other columns, we used median imputation, which is a robust approach when the missing data are relatively randomly distributed. Based on the Missingno matrix and heatmap, we did not observe strong patterns or clusters of missingness; therefore, we employed median imputation.



Percentage of Missing Values



Missingno Matrix

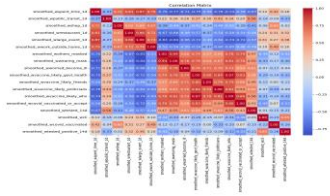
Another version of the dataset was imputed using a moving average. Due to the nature of disease spreading over time, we decided that moving average effectively analyzes COVID-related time series data, preserving smooth trends over time. As a third option, we also explored KNN imputer and tested with the final model, but it underperformed.

After addressing missingness, we conducted a correlation analysis to manage multicollinearity. Highly correlated feature pairs (correlations above 0.8) were identified:

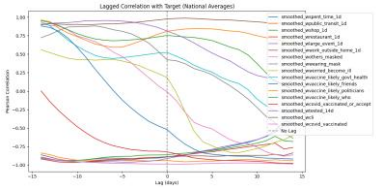
1) smoothed_wvaccine_likely_who and smoothed_wvaccine_likely_govt_health
2) smoothed_wothers_masked and smoothed_wwearing_mask

Among the pair, the features with higher missing values (smoothed wvaccine likely govt health, smoothed wwearing mask) were dropped.

In terms of feature engineering, we used lag Analysis and incorporated state information. Lag analysis revealed that some features, such as public transit use, showed stronger correlation with COVID-19 positivity rates at positive lags, aligning with the intuition that exposure leads to infection after a delay. Based on these results, lagged features were added into the modeling pipeline.



Correlation



Lagged Correlation with Target

Lagged Features	
smoothed_wcli	+2
smoothed_wpublic_transit_1d	+2
smoothed_wcovid_vaccinated	+1
smoothed_wvaccine_likely_friends	+14

To capture spatial correlation, external data for the geo_value was linked to identify the state information for each county. The Spatial Matrix (W) is a row-normalized binary adjacency matrix where counties are considered neighbors if they belong to the same state. This matrix is used in Spatial Lag 1 Regression.

3. Baseline Model and Methods

3.1 Predicting Vaccination Rate

To simulate a real-world scenario, we use chronological train-validation-test split to ensure robust model evaluation and avoid data leakage in this temporal context. The dataset was divided as follows:

- Training Dataset: From January 07, 2021 to January 29, 2021 — used to train predictive models. 16,494 records (64.36%)
- Validation Dataset: From January 30, 2021 to February 05, 2021 — used to tune hyperparameters and compare model performance. 4,653 records (18.15%)
- Testing Dataset: From February 06, 2021 to February 12, 2021 — used to evaluate the final model performance on unseen data. 4,479 records (17.47%)

- **Baseline: Elastic Net Regression without feature selection and engineering**

We began with an Elastic Net regression model as a baseline to evaluate the predictive power of raw features. By combining L1 and L2 penalties, Elastic Net handles multicollinearity and supports sparse feature selection. The model excluded panel data structures and relied on unengineered inputs. To simulate real-world forecasting, we used a fixed temporal split (train: up to Jan 29, 2021; validation: Jan 30–Feb 5; test: after Feb 5). We manually tuned regularization parameters ($\alpha \in \{0.01, 0.1, 1.0, 10\}$, l1_ratio $\in \{0.1, 0.5, 0.9\}$) and selected the best configuration based on validation RMSE. The final model used $\alpha = 0.01$ and l1_ratio = 0.1, offering a simple, interpretable benchmark against which more advanced models were compared.

To reduce redundancy and multicollinearity, we excluded features highly correlated with other. We dropped smoothed_wvaccine_likely_who in favor of smoothed_wvaccine_likely_govt_health which the latter aligns more closely with the messaging from U.S. public health agencies, and replaced smoothed_wwearing_mask with smoothed_wothers_masked, which better captures community behavior.

We also removed smoothed_wcovid_vaccinated_or_accept to avoid mixing vaccination status with intent, as the feature combines individuals who have already been vaccinated with those who merely intend to do so.

- **Regression with spatial temporal features**

To capture spatial and temporal dependencies, we engineered features using a spatial weights matrix that defined neighboring regions by shared state. This matrix was row-normalized and used to compute spatial lag features—weighted averages of neighboring values. In parallel, we created temporal lags and spatial-temporal interaction terms by combining both lags. Applied across all regions and time points, these features enriched the dataset with localized trends and diffusion effects, enabling more accurate modeling in settings with spatial and temporal autocorrelation.

- **Panel Regression with rolling mean, lagging features, and interactive features**

To explicitly model entity-specific heterogeneity and temporal structure, we implemented a fixed-effects panel regression using the PanelOLS framework. This approach assigns each entity (e.g., county) its own intercept, effectively controlling for unobserved, time-invariant characteristics. Our exploratory data analysis (EDA) revealed significant shifts in vaccination rates over time within counties—underscoring the importance of modeling both temporal dynamics and entity-specific effects. To enhance predictive performance and reduce overfitting, we applied recursive feature elimination (RFE) using an Elastic Net estimator ($\alpha = 0.01$, l1_ratio = 0.1, based on baseline tuning results) to identify compact subsets of informative features. For each subset size, we trained a panel regression model with fixed effects, evaluated its performance on the validation set using RMSE, and selected the feature set yielding the lowest error.

- **Random Forest with rolling mean, lagging features, and interactive features**

To capture non-linear relationships and complex interactions among behavioral and belief-based features, we implemented a Random Forest model using ensemble learning. Prior to modeling, we applied recursive feature elimination (RFE) with an Elastic Net estimator to reduce dimensionality and select the 40 most informative predictors. We then conducted an extensive grid search over key hyperparameters—including the number of trees, node depth, and leaf constraints—using a time series cross-validation scheme to ensure temporal integrity in model evaluation. For each parameter combination, the model was trained and validated across multiple rolling time windows, and performance was assessed using average RMSE across folds.

- **Fully Connected Neural Network (FNN) with rolling mean, lagging features, and interactive features**

To model deeper non-linear relationships, we implemented a neural network with two hidden layers of 32 ReLU-activated nodes each. This architecture was selected as a middle ground—complex enough to detect interactions among features, but not so large that it would overfit the training data, which consisted of 8,038 records. In practice, we found this balance was crucial: adding more layers or increasing the number of nodes made the model too flexible, fitting noise in the training data and leading to poor performance on unseen data (overfitting). On the other hand, reducing the size of the network resulted in underfitting—it wasn't expressive enough to capture the underlying relationships in the data.

- **Discarded Model**

We also experimented with a Light Gradient Boosting Machine (LightGBM) to capture complex non-linear interactions through boosting. While powerful, the model consistently overfit the training data despite careful hyperparameter tuning and early stopping. Given its poor generalization performance and instability on our validation set, we discarded LightGBM due to its lack of robustness in this setting.

3.2 Predicting COVID-19 Cases

- **Baseline: Ridge Regression**

We began with a Ridge regression model as the baseline to assess the predictive power of behavioral and belief-based features for county-level COVID-19 positivity rates. Ridge regression extends ordinary least squares by incorporating an L2 penalty, which helps stabilize estimation in the presence of multicollinearity, a common issue in survey-based behavioral datasets. The model does not include spatial

components, making it an interpretable and computationally efficient baseline for comparison with more complex models. To reflect real-world forecasting, we used a temporal split (Train: 80% of earliest time points, Validation: 10% of middle, Test: 10% of most recent). We applied RidgeCV with 5-fold cross-validation to select the optimal regularization strength α. In terms of cross-validation, 5-fold offers a good tradeoff between runtime and estimate reliability and is suitable for medium-sized datasets. This approach was used consistently across models.

- **Multi-Layer Perceptron (MLP)**
  - Model Selection

Given the complex spatio-temporal nature of COVID-19 spread and nonlinear interactions between behavior and health beliefs, we selected a Multi-Layer Perceptron (MLP) as a more flexible alternative. MLPs are capable of capturing nonlinear patterns and feature interactions that Ridge regression cannot.

  - Hyperparameter

Architecture	Optimizer	Loss
· Input layer → 128 → ReLU → Dropout(0.3) · 128 → 128 → ReLU → Dropout(0.3) · 128 → 64 → ReLU → Dropout(0.3) · Output: 1 neuron (regression)	Adam (lr=0.0001, weight_decay=1e-4)	MSELoss
Learning rate scheduler	**Epochs**	**Early stopping:**
StepLR(step_size=100, gamma=0.5)	up to 600	patience = 15 epochs

Dropout and L2 regulization applied for generalization

  - Train/Test Split

The dataset was scaled using StandardScaler, and 5-fold cross-validation was applied using KFold(n_splits=5, shuffle=True, random_state=42) to ensure robustness. Each fold used 80% of the data for training and 20% for validation, maintaining spatial-temporal stratification.

  - Limitations

While the MLP model is capable of modeling nonlinear spatio-temporal patterns, it acts as a black box. The limited interpretability makes them less ideal for policymaking contexts.

- **Spatial Lag 1 Regression ('SLR(1)')**
  - Model Selection

We selected the Spatial Lag 1 Regression model to maintain interpretability and capture spillover effects in COVID-19 positivity rates at the county level. For the same-state adjacency, we reused the previously constructed W matrix to compute the spatial lag Wx1.

  - Hyperparameter

We used L2 regularization to prevent overfitting and deal with multicollinearity. α was selected via RidgeCV, which performs 5-fold cross-validation across a log-scale grid of α values.

  - Train/Test Split

Considering contagiousness over time, we used the same time split as ridge regression.

  - Limitations

1) **Linearity Assumption**: The model assumes additive and linear relationships between features and the outcome, which may not capture the real-world dynamics of disease spread.
2) **Fixed Spatial Effect**: Using only a first-order spatial lag (Wx1) limits the ability to model more diffuse or nonlinear spatial interactions.

- **Random Forest ('RF')**
  - Model Selection

To address the challenges in linearity assumption of SLR and interpretability and the extensive hyperparameter tuning required for MLP, we introduced a Random Forest Regressor as a tree-based ensemble baseline, which requires less hyperparameter tuning and is robust to high-dimensional data.

  - Hyperparameter(5-fold GridSearchCV)

n_estimators	max_depth	min_sample_split	min_samples_leaf	max_features
300	15	10	4	0.3

  - Train/Test Split

Similar to MLP, we used an 80/20 train-test split with random_state=42

4. Results

4.1 Vaccination Rate Prediction

We used Root Mean Squared Error (RMSE) over other metrics because RMSE penalizes larger errors more heavily due to the squaring of residuals. Since large prediction errors in predicting vaccine uptake can lead to disproportionately harmful decisions—such as misallocating resources and failing to identify or prioritize high-risk communities—RMSE is better suited for highlighting and minimizing these high-impact mistakes.

Model	Train RMSE	Validation RMSE	Test RMSE
Panel Regression	1.19	1.53	1.63
Regression with spatial temporal features	1.26	1.56	1.65
Random Forest	1.24	1.79	2.14
Light GBM	0.66	1.82	2.45
Neuron Network	1.99	2.50	2.62
Simple Regression (Baseline)	3.83	6.66	9.32

Panel Regression delivered the best overall performance, achieving the lowest RMSE on the test set (1.63). This result aligned with our expectations, as the model explicitly controls for entity-specific heterogeneity and is well-suited for structured panel data. Its ability to incorporate fixed effects made it particularly effective in capturing underlying regional differences in vaccination behavior.

Random Forest achieved strong performance on the training set but showed increased RMSE on validation and test sets, indicating moderate overfitting. While it successfully modeled non-linear relationships, the model lacked explicit mechanisms to capture spatial and temporal structures, which may have limited its generalization compared to the panel regression.

Neural Network underperformed across the board. The model's limited architecture—constrained to prevent overfitting—struggled to extract meaningful patterns from the moderate-sized dataset (8,038 records). As a result, it failed to outperform simpler models and did not contribute additional predictive value.

Elastic Net (Baseline) served as a useful reference point, but it recorded the highest RMSE on all datasets. Its reliance on unengineered features and lack of panel structure limited its ability to model complex behavioral dynamics, reinforcing the need for more advanced modeling approaches.

Therefore, in predicting vaccine uptake, **Panel Regression** is a highly effective approach because it accounts for **both time-varying trends and entity-specific effects**, such as differences across counties or states. This allows the model to control for unobserved heterogeneity while capturing the temporal dynamics of vaccine behavior. Its strong out-of-sample performance (lowest RMSE on the test set) and high explanatory power ($R^2 = 0.9243$ overall) demonstrate its suitability for panel data.

Interpretation of Panel Regression

- **Vaccine uptake is highly autocorrelated**: The strongest predictor was the 3-day lag of the dependent variable (smoothed_wcovid_vaccinated_lag_3, β = -0.8252), suggesting recent uptake levels heavily influence current behavior.
- **Behavioral patterns matter**: Attendance at large events was negatively associated with vaccine uptake both immediately (smoothed_wlarge_event_1d, β = -0.0602) and with a 3-day lag (smoothed_wlarge_event_1d_lag_3, β = -0.0804), indicating risk-seeking behaviors may delay or reduce uptake.
- **Temporal momentum effect**: The 3-day rolling average of vaccination rates (smoothed_wcovid_vaccinated_rolling_mean_3, β = 1.7814) had a strong positive effect, reinforcing the idea that communities follow established trends or social cues.
- **Work exposure has a small positive link**: Those working outside the home showed slightly higher uptake after a 3-day lag (smoothed_wwork_outside_home_1d_lag_3, β = 0.0145), possibly due to perceived exposure risk.
- **All predictors were statistically significant**, making the model reliable for our analysis and targeted intervention planning.

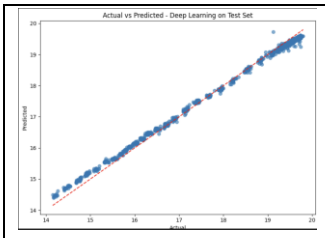
4.2 COVID Case Prediction

To evaluate our models predicting county-level COVID-19 positivity rates, we used R^2, RMSE, and MAE. Since the task involves forecasting a continuous, health-critical outcome, these metrics together capture both goodness of fit (R^2) and the magnitude of prediction errors (RMSE and MAE). In the public health context, minimizing prediction error is critical for timely outbreak detection and resource planning, with RMSE particularly sensitive to large deviations. MAE complements this by offering a straightforward measure of typical error size, supporting reliable, actionable public health predictions.

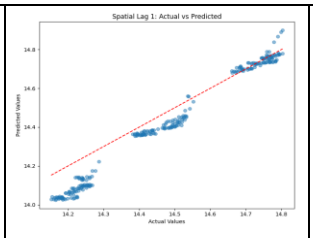
Model	Fold 1-5 Mean			Strengths	Limitation
	R^2	RMSE	MAE		
RR	-0.80	2.40	1.86		
MLP	0.990	0.10	0.07	Captures nonlinearity, generalizes	Requires tuning, less interpretable
SLR(1)	0.999	0.02	0.02	Interpretable, accounts for space	Assumes linearity
RF	0.998	0.09	0.02	Robust, interpretable features	May overfit with small data

Compared to the baseline model, the Random Forest generalized exceptionally well, with minimal overfitting. An important finding was that Random Forest can extract historical patterns through nonlinear interactions, reducing reliance on explicit lag features like the baseline model. The spatial lags ranked lower in importance, and the model performed well without these features, unlike the baseline spatio-linear model, which deemed same-state neighborhood lags crucial. Since Random Forests capture spatio or temporal patterns effectively without explicit lag features, they are suitable for drawing policy implications. Random Forest provides a practical balance of predictive accuracy and feature interpretability, making it highly valuable for informing public health policy.

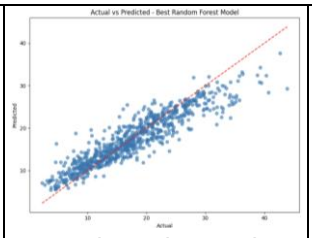
Also, the best model showed that moving average imputation for missing values outperformed the KNN imputer. It better preserves temporal trends than KNN or static methods.



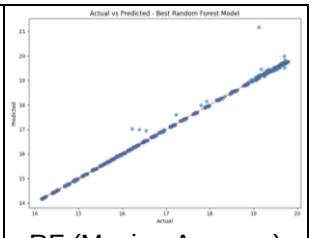
MLP



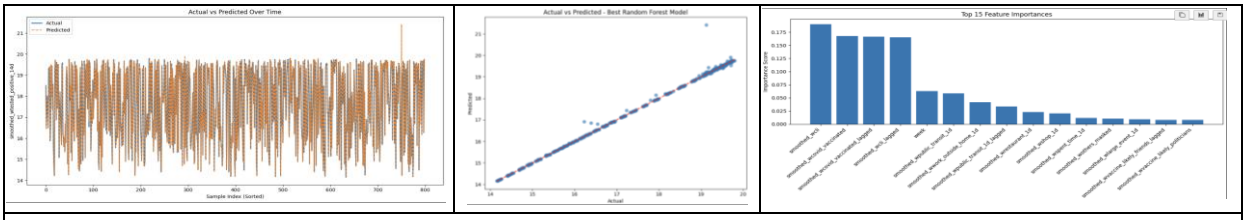
SLR(1)



RF (KNN imputer)



RF (Moving Average)



Random Forest (Excluding spatial lag)

- **Limitations and Future Directions**

Our expectation that the MLP model would capture nonlinear spatio-temporal patterns and generalize effectively was not met, as it did not outperform the Random Forest or baseline model. Random Forests are more robust to noisy data and easier to optimize, making them hard to beat in real-world settings. However, LSTM or Transformer architectures are better suited for temporal sequences. Future work could integrate real-time streams and use LSTM or Transformer to better capture time-evolving relationships.

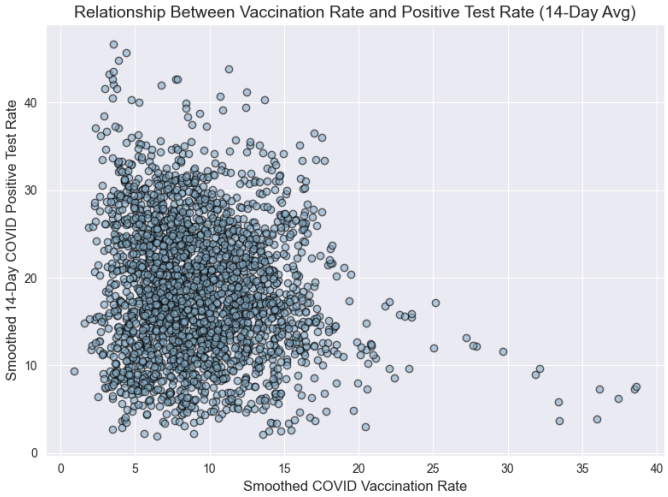
- **Feature Importance**

The Random Forest model achieved near-perfect prediction performance (R^2 = 0.998, RMSE = 0.09), effectively capturing spatial and temporal COVID-19 dynamics. The top predictive features were:

- Vaccination rates (smoothed_wcovid_vaccinated, ..._lagged)
- CLI indicators (COVID-like illness)
- Lagged positivity in same or nearby counties

5. Policy Recommendation & Analysis

Vaccine effectiveness on COVID outcomes



As visualized in the scatter plot, regions with higher smoothed vaccination rates tend to exhibit lower 14-day smoothed positive test rates. This inverse trend suggests that increased vaccination coverage is associated with a reduction in COVID-19 transmission.

Vaccine policy implications

Our modeling demonstrates that both vaccination behavior and COVID-19 test positivity rates are strongly influenced by behavioral patterns and spatial-temporal dynamics. These insights offer actionable guidance for public health interventions:

- **Vaccination Strategy:** Our model shows that vaccine uptake follows strong social momentum—recent local vaccination rates are the most influential predictor, suggesting people are influenced by community trends. Areas with rising uptake should be reinforced with visibility campaigns to sustain momentum, while low-uptake regions may need targeted interventions to break stagnation. Risk-tolerant individuals, such as those attending large events, are less likely to vaccinate and may respond better to autonomy-focused messaging. Meanwhile, those working outside the home show slightly higher uptake, highlighting an opportunity to expand workplace or public-facing vaccination efforts.
- **Prioritize Surveillance of Lagged Indicators:** Early-warning signals for rising positivity rates include increased time spent outside, restaurant visits, and public activity. Lagged COVID-like illness (CLI) and behavioral variables showed strong predictive power, even when testing data were sparse. These metrics should be monitored as leading indicators to trigger preemptive containment efforts.
- **Regional Coordination:** Spatial lags in both tasks confirm that counties are not epidemiologically isolated. Shared policies across state lines, especially in densely populated or highly mobile regions, would enhance the efficiency of interventions. For instance, coordination in vaccination rollout or mask mandates between adjacent counties can reduce overall infection rates more effectively than isolated policies.

While our models performed well, several limitations must be acknowledged. High levels of missing data especially in the infection outcome required imputation, which may have introduced bias or masked important patterns. The county-level aggregation limits granularity, potentially overlooking individual behaviors or intra-county variation. Lagged relationships were modeled based on exploratory analysis, but the true timing of effects may differ by region. Our spatial model assumes uniform influence across counties within the same state, ignoring more nuanced cross-county dynamics. Although Random Forest offers a good balance of predictive performance and interpretability, it does not naturally model sequential data, suggesting the use of models like LSTM or Transformer when richer or longer time-series data is available. Despite these limitations, our models offer valuable insights that can meaningfully inform pandemic response strategies. Future research should explore real-time data integration and advanced temporal modeling (LSTM, Transformer) to enhance forecasting precision and policy applicability.