International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2021): 7.86

Leveraging Population-Level COVID-19 Testing Data for Predictive Modeling During Variant Surges: A Case Study from National Pharmacy Testing Network

Vijitha Uppuluri

Email: *vuppulur[at]gmail.com*

Abstract: ***The continued emergence of new SARS-CoV-2 variants has significantly increased the complexity of forecasting and preventing subsequent COVID-19 waves. Nationwide pharmacy testing data, collected through extensive pharmacy networks, offers a novel and effective approach for real-time population testing, facilitating the rapid identification of emerging outbreaks. This study aims to evaluate the extent to which large-scale testing data can inform predictive models capable of anticipating increases in COVID-19 infections in response to the appearance of new viral variants. Specifically, the study incorporates test positivity rates, geographic spread, and demographic information, analyzed using machine learning and time series methods, to enhance outbreak forecasting. Results indicate that integrating external datasets—such as vaccination coverage and population mobility data—further improves model accuracy, thereby supporting more informed decision-making by public health authorities. Among the modeling approaches assessed, deep learning models—particularly Long Short-Term Memory (LSTM) networks—demonstrated superior performance in capturing long-term trends compared to traditional methods like ARIMA. Findings suggest that insights derived from pharmacy testing data can play a critical role in enabling policymakers to respond proactively to the emergence of new COVID-19 variants. The proposed framework offers a scalable alternative for epidemic prediction architectures within broader public health ecosystems. Future research should explore the integration of genomic surveillance data and consider the applicability of this predictive framework to other infectious diseases beyond COVID-19.***

Keywords: COVID-19, Predictive modeling, Population-level testing, Pharmacy network, Variant surges

1. Introduction

1.1 Background and Motivation

The recent emergence of the COVID-19 pandemic has further emphasized the critical role of surveillance and risk modeling in outbreak management. The continual evolution of SARS-CoV-2 variants necessitates persistent monitoring of infection dynamics and presents significant challenges in the timely identification and effective management of outbreaks within populations. Conventional epidemiological models exhibit several limitations in this context, particularly due to their static assumptions; consequently, there is an urgent need for more adaptive and responsive modeling approaches that can accommodate the virus's ongoing mutations [1–3].

A growing body of literature highlights the importance of large-scale COVID-19 testing data to evaluate the efficacy of testing strategies, identify disproportionately affected subpopulations, and inform future disease prevention efforts. This study leverages datasets derived from pharmacy networks situated in high-incidence regions across the United States to support enhanced predictive modeling of infection surges. Such modeling aims to facilitate more informed containment strategies and to address the inherent challenges posed by heterogeneous and often unpredictable public behavioral responses.

1.2 Importance of Predictive Modeling during COVID-19 Surges

Predictive modeling plays a crucial role in the early detection of outbreaks, resource allocation, and the strategic planning of public health interventions. During COVID-19 surges, accurate forecasting of infection trends enables governments and healthcare systems to implement timely measures such as mobility restrictions, targeted vaccination campaigns, and increased diagnostic testing.

Artificial intelligence (AI) and time series analysis have emerged as effective tools for interpreting the growing volume of COVID-19 testing data, offering more precise forecasts than traditional epidemiological models. Nevertheless, the reliability of these models is highly dependent on the availability and timeliness of data inputs.

This study addresses this challenge by leveraging pharmacy testing data to identify optimal parameters for forecasting future waves of infection. The goal is to enable timely interventions that consider social determinants of health, ultimately strengthening the responsiveness and equity of public health systems in mitigating the impact of COVID-19.

1.3 Role of National Pharmacy Testing Networks in Data Collection

Pharmacies have played a pivotal role in COVID-19 testing by facilitating widespread and accessible testing across diverse regions of the United States. Unlike hospital-based testing, which primarily targets symptomatic or confirmed cases, pharmacy testing networks encompass a broader cross-section of the population, including asymptomatic individuals. This inclusive approach enhances their utility for population-level surveillance. Furthermore, pharmacy-based testing offers consistent, high-quality, and timely data

collection, which is essential for real-time analytics and predictive modeling.

By utilizing data from pharmacy testing systems nationwide, this study monitors the geographic and temporal spread of COVID-19, identifies early indicators of potential surges associated with emerging variants, and contributes to the refinement of forecasting models for improved public health response.

1.4 Research Objectives and Contributions

The primary objective of this study is to develop a predictive modeling framework utilizing insights derived from low-cost, large-scale testing conducted through a nationwide pharmacy network. This approach aims to accurately forecast infection surges during the emergence of new SARS-CoV-2 variants. The key contributions of this research are as follows:

- **Development of an outbreak detection system** based on machine learning and time series analysis to enhance traditional epidemiological surveillance methods.
- **Exploration of real-time pharmacy testing data** as a reliable and timely source for the early identification of COVID-19 surges.
- **Assessment of exogenous factors**—such as population mobility, vaccination coverage, and socio-demographic variables—and their influence on the effectiveness of predictive strategies throughout different stages of infection spread.
- **Cross-validation of predictive techniques**, comparing statistical models (e.g., ARIMA) with deep learning models (e.g., LSTM), to determine the most effective approach for forecasting COVID-19 surges.
- **Policy implications and strategic recommendations**, emphasizing the utility of pharmacy-based testing data in strengthening national and global pandemic preparedness frameworks.

To achieve these aims, the study sets forth the following objectives:

1) To propose a novel index for real-time risk assessment of infectious disease outbreaks.
2) To model the probability distribution of various infectious diseases using real-world testing data.
3) To develop a dynamic framework capable of updating predictive models as new data become available.
4) To provide a generalizable framework for the early prediction of COVID-19 outbreaks and similar infectious diseases, thereby supporting evidence-based decision-making and enhancing crisis management during and beyond the current pandemic.

2. Related Work

2.1 Review of Existing COVID-19 Predictive Models

In response to the COVID-19 pandemic, numerous predictive models have been developed to forecast new infection rates, estimate hospitalizations, and assess the effectiveness of public health interventions. Traditional models such as the Susceptible-Infected-Recovered (SIR) and the Susceptible-Exposed-Infected-Recovered (SEIR) frameworks have been widely employed [4–7]. However, these compartmental models are based on fixed assumptions about transmission dynamics and population behavior, which may not adequately capture the behavioral and epidemiological shifts that occur during the emergence of new viral variants.

To enhance forecasting reliability, researchers have increasingly emphasized the integration of real-time surveillance data into model development. For example, the CDC's COVID-19 Forecast Hub aggregates multiple models to generate probabilistic forecasts of case numbers. Similarly, platforms like the Johns Hopkins COVID-19 Dashboard and Google's COVID-19 Mobility Reports utilize diverse datasets—including mobility patterns and testing rates—to improve outbreak predictions. Despite these advances, many existing models face limitations related to incomplete data and sampling biases, particularly in testing coverage.

To address these challenges, the present study leverages population-level testing data obtained from nationwide pharmacy networks. By training advanced analytical models on this robust and real-time dataset, the study aims to mitigate the limitations of previous models and offer more accurate and scalable forecasting solutions.

2.2 Population-Level Testing Studies

Community-based testing plays a critical role in monitoring the spread of infectious diseases and informing public health policy, particularly in the context of influenza and COVID-19. Evidence suggests that widespread community testing provides a more comprehensive understanding of infection prevalence compared to testing conducted solely in clinical or hospital settings, as individuals are more likely to seek testing in accessible, non-clinical environments such as pharmacies. Studies have demonstrated that population-level testing, when coupled with appropriate isolation measures, can significantly mitigate viral transmission. Additionally, rapid antigen testing has proven effective in identifying both symptomatic and asymptomatic individuals, thereby supporting early intervention and outbreak prevention.

Pharmacy-based testing networks offer several advantages over traditional testing strategies. These networks enable point-of-care testing and capture data from individuals across a broad age spectrum, including those with and without symptoms. Unlike self-reported testing data, pharmacy-collected data are objective, reliable, and readily verifiable. While some prior research has explored the role of retail pharmacy networks in disease surveillance, few studies have investigated their utility in predictive modeling.

This study aims to address this gap by demonstrating how pharmacy-based testing data can be leveraged to enhance early detection and forecasting of COVID-19 variant-driven surges. By incorporating this data into predictive models, the study contributes to more timely and effective public health responses.

2.3 Impact of COVID-19 Variants on Testing Strategies

The emergence of SARS-CoV-2 variants such as Delta and Omicron introduced substantial complexities in the overall management of the COVID-19 pandemic. These variants

exhibit significant genetic differences, particularly in their transmissibility, immune evasion capabilities, and potential to cause severe illness. Such differences necessitate adjustments in testing strategies to accurately identify circulating variants and prevent widespread outbreaks. Notably, variations in RT-PCR test positivity rates across different variant waves underscore the need for dynamic testing protocols and resource allocation. For example, the Alpha variant demonstrated higher transmissibility, leading to increased testing demand, while the Omicron variant was associated with elevated infection rates regardless of vaccination status, prompting changes in testing guidelines.

In response to these evolving challenges, adaptive testing strategies have been recommended. Some researchers advocate for adjusting testing thresholds based on real-time positivity rates, while others propose prioritizing testing among high-risk or vulnerable populations during variant-driven surges. However, most existing studies have relied primarily on data from clinical or hospital-based testing, with limited exploration of data derived from community-level pharmacy testing networks.

This study addresses this gap by integrating pharmacy-based testing data into predictive models, offering more granular insights into how variant-induced fluctuations in test demand and positivity rates influence outbreak forecasting. By capturing real-time testing dynamics across diverse population segments, this approach enhances the accuracy and responsiveness of predictive modeling in the context of evolving viral variants.

2.4 Machine Learning and Statistical Approaches in Epidemiological Modeling

Epidemiological forecasting using machine learning (ML) and statistical modeling has gained substantial importance due to its flexibility and adaptability compared to traditional compartmental models. In recent studies, ML techniques such as supervised learning algorithms—including Random Forest, Gradient Boosting, and Neural Networks—as well as time series forecasting methods like Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks, have been widely applied to predict COVID-19 case counts, hospitalization rates, and mortality trends. Preliminary research indicates that ML models often outperform compartmental models in short-term forecasting by effectively capturing nonlinear infection patterns.

Among time series models, ARIMA and LSTM have shown promising results. However, ARIMA, being a linear model, lacks the capacity to account for abrupt shifts in transmission dynamics, such as those caused by the emergence of new viral variants. In contrast, LSTM—a type of recurrent neural network (RNN)—is well-suited for capturing long-term dependencies in sequential data, making it more effective for pandemic forecasting. Recent research has also explored hybrid approaches that integrate statistical and ML techniques to enhance forecasting precision. These mixed models, when combined with traditional epidemiological frameworks, can significantly improve both accuracy and responsiveness in modeling COVID-19 dynamics.

Despite these advancements, a notable gap remains in the incorporation of high-frequency testing data, particularly from community-level sources such as pharmacies. Much of the existing literature relies on delayed metrics such as confirmed case numbers and hospitalization data, which may not accurately reflect real-time transmission trends. This study seeks to bridge this gap by leveraging pharmacy-based testing data to improve the timeliness and accuracy of predictive models. The goal is to support more effective public health interventions through earlier detection and more precise forecasting of infection surges.

3. Data Collection and Preprocessing

Accurate prediction of COVID-19 trends requires the integration of testing data from networks that are efficient, up-to-date, and representative of the broader population [8–11]. This study utilizes RT-PCR testing data collected from customers at various branches of a national pharmacy testing network. Unlike previous studies that primarily rely on hospital-based data, this dataset includes a more diverse segment of the population, encompassing both symptomatic and asymptomatic individuals, as well as those undergoing routine screening.

The inclusion of community-based, pharmacy-derived data provides a more comprehensive view of infection dynamics across different demographics. To ensure data quality and suitability for predictive modeling, the raw dataset underwent a thorough preprocessing pipeline aimed at enhancing accuracy, consistency, and relevance. This section outlines the data collection process, the methods applied for data cleaning, and the unique characteristics of the dataset that support the development of a robust predictive modeling framework.

3.1 Data Sources

The primary data source for this study was a nationwide pharmacy-based COVID-19 testing platform, which plays a significant role in decentralized disease surveillance. Unlike hospital-based testing, which predominantly captures individuals with moderate to severe symptoms, pharmacy testing networks encompass a broader and more diverse population—including individuals undergoing testing for travel, workplace requirements, or routine health checks. This makes pharmacy data particularly valuable for monitoring asymptomatic and pre-symptomatic cases within the community.

The dataset, titled *Test Results of COVID-19 Test*, includes data from multiple waves of the pandemic and contains variables such as test outcome (positive/negative), test type (PCR or antigen), timestamp of the test, age range and gender of individuals, geographic identifiers (state, county, and rural or urban classification), and vaccination status when available. This comprehensive dataset supports the development of accurate and generalizable predictive models, reducing the need for extensive field-based testing in specific regions or demographic groups.

Test records were acquired via the APIs of various pharmacy chains, allowing for the extraction of real-time data while

ensuring patient anonymity and data privacy. In addition, historical batch processing was employed to capture longer-term trends. The database was standardized according to guidelines recommended by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO), enabling interoperability and comparison across different pharmacy networks. This integration of real-time and near real-time data facilitates the construction of high-temporal-resolution models capable of accurately forecasting localized COVID-19 outbreaks.

3.2 Data Cleaning and Processing

Testing data often contains missing values, inaccuracies, and biases that can adversely affect the reliability of predictive modeling. To address these challenges, systematic data cleaning procedures were applied during the data preprocessing stage. Features with critical inconsistencies—such as missing or unreliable test results and test dates—were excluded to prevent the introduction of biased or misleading information. For the remaining demographic variables, missing values were imputed using appropriate techniques: mode imputation was employed for categorical variables, while k-nearest neighbors (KNN) imputation was used for continuous variables.

Duplicate or suspicious records were identified and removed. These included entries with duplicated or nearly identical test identifiers and records with timestamps that were unnaturally close, suggesting potential redundancies or data entry errors.

To ensure consistency across data collected from different geographic centers and time periods, normalization techniques were applied. Min-max scaling was used for test positivity rates, given that these values are inherently bounded between 0 and 1. Other numerical features, such as daily test counts and regional case numbers, were standardized using Z-score normalization to ensure comparability. Additionally, categorical variables such as test type were encoded using one-hot encoding, while geographic locations were encoded ordinally to make them suitable for inclusion in machine learning models.

These data preprocessing steps were essential for improving data quality, reducing heterogeneity, and enhancing the overall performance and interpretability of the predictive models.

3.3 Data Characteristics

The dataset provided a unique perspective on the evolution of COVID-19 testing trends over time, as well as across different demographic and geographic segments of the U.S. population. Testing volumes exhibited significant fluctuations throughout the pandemic, with pronounced spikes during major variant-driven waves such as Delta and Omicron. Temporal patterns also revealed increased testing activity during the winter months and immediately following major holidays—periods typically associated with higher rates of respiratory illnesses, including influenza.

Geographic and demographic variables further revealed disparities in testing access and positivity rates. Urban areas reported a higher volume of tests, likely due to better accessibility and employer-mandated testing requirements, particularly for occupational settings. In contrast, rural areas exhibited lower testing volumes but higher positivity rates, suggesting potential underreporting and reduced access to testing services.

Age-specific trends showed that individuals under 30 were more likely to undergo routine screening, while positivity rates were significantly higher among individuals over 60, highlighting the latter group's increased vulnerability to severe illness. Gender-based analysis revealed that females had a slightly higher testing frequency than males, likely influenced by their representation in high-contact professions such as healthcare and education.

These findings underscore the importance of incorporating demographic and geographic characteristics into predictive modeling to ensure models reflect real-world disparities and support more targeted and equitable public health interventions.

3.4. Machine Learning and Deep Learning for COVID-19 Severity Prediction Using Electronic Health Records (EHR)

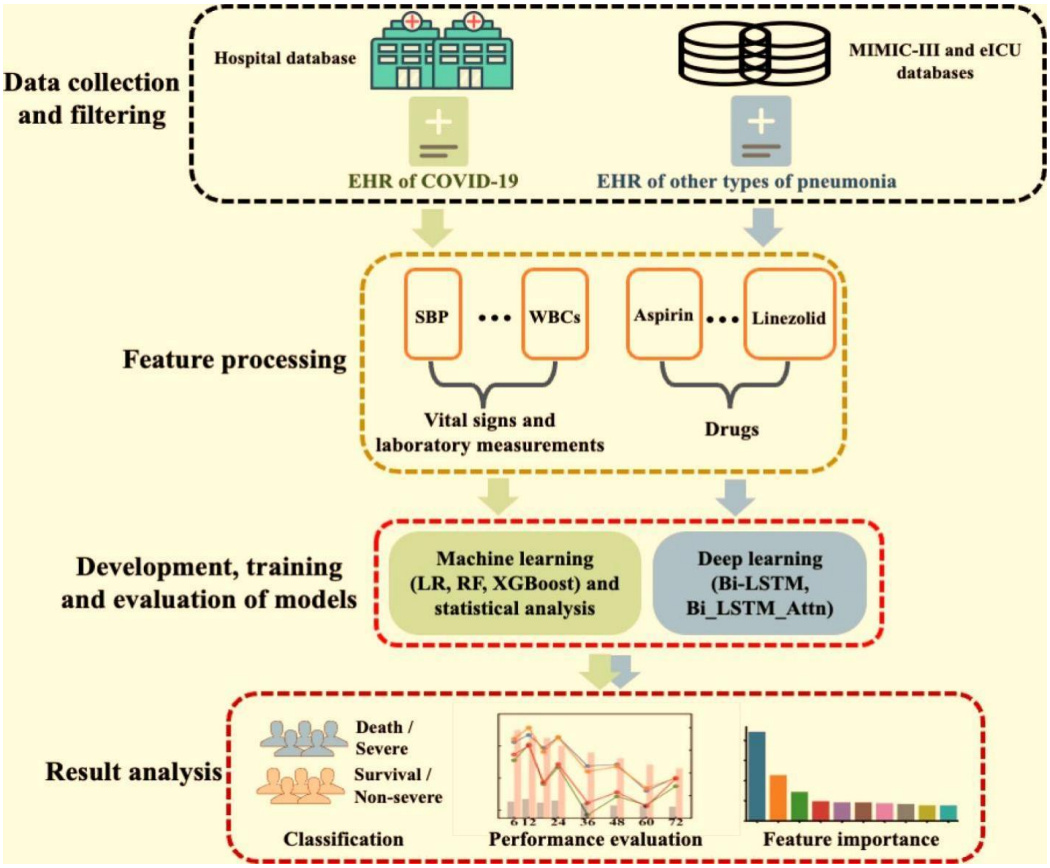

Figure 1: Machine Learning and Deep Learning for COVID-19 Severity Prediction Using Electronic Health Records (EHR)

This figure is a workflow diagram that shows the process of applying machine learning and deep learning models to analyze Electronic Health Records (EHR) to predict the severity of COVID-19 cases. [12-14] It is separated into four large sections: data collection and filtering, feature processing, model development and evaluation, and result analysis. The workflow gives information about how hospital data is processed and utilized for training predictive models.

3.4.1. Data Collection and Filtering

The initial stage of the framework involves the collection of electronic health record (EHR) data from multiple sources, including hospital databases and publicly available repositories such as the MIMIC-III and eICU databases. These datasets contain detailed clinical records of patients diagnosed with COVID-19 as well as those with other forms of pneumonia. A filtering step is employed to isolate relevant patient records, ensuring that only data pertinent to the research objectives are retained for further processing. This filtration process enhances the reliability and specificity of the predictive models by clearly distinguishing between different respiratory conditions, thereby reducing the risk of diagnostic overlap, and improving model accuracy.

3.4.2. Feature Processing

Following the acquisition of raw data, a feature engineering phase is conducted to extract clinically meaningful information. The features are categorized into vital signs and laboratory measurements (e.g., systolic blood pressure [SBP], white blood cell count [WBC]) and pharmacological interventions (e.g., Aspirin, Linezolid). These clinical indicators are critical for assessing disease progression and predicting patient outcomes. The careful selection and preprocessing of relevant features not only improve model performance but also ensure that predictions are grounded in medically significant variables, thereby enhancing the interpretability and clinical relevance of the results.

3.4.3. Development, Training, and Model Evaluation

Here, predictive models are trained using the preprocessed data. The figure indicates two significant approaches:

- **Machine Learning Models:** Here, Logistic Regression (LR), Random Forest (RF), and XGBoost fall under this category, which is employed for statistical analysis and classification. These models are aptly suited for structured medical data and yield interpretable results.
- **Deep Learning Models:** Bi-LSTM (Bidirectional Long Short-Term Memory) and Bi-LSTM with Attention Mechanism (Bi_LSTM_Attn) belong to this category, capable of identifying intricate patterns in time-series EHR data. These models enhance prediction through learning interdependencies among several clinical variables over time.

Machine and deep learning models are tested using the relevant validation methods to warrant high accuracy and reliability.

3.4.4. Result Analysis

Post training and testing, the performance of the models is studied in three principal manners:

- **Classification Results:** The patients are divided into groups based on the model predictions: Death/Severe and Survival/Non-severe.
- **Performance Evaluation:** A graph shows the models' performance for various time intervals and examines

important measurements such as accuracy, sensitivity, specificity, and AUC-ROC.
- **Feature Importance:** A bar chart shows the importance of various clinical features in predicting patient outcomes. This assists in comprehending which medical factors are most responsible for disease severity.

This process emphasizes using EHR data, feature selection, and predictive modeling to effectively predict COVID-19 severity. Combining machine learning and deep learning, healthcare professionals can acquire data-driven information to enhance patient care and refine treatment plans. The systematic approach provides precise risk assessment and aids in personalized medicine for improved health outcomes.

4. Methodology

4.1 Predictive Modeling Approach

To forecast future variant-driven surges in COVID-19 cases, this study employs a combination of machine learning and statistical analysis techniques. To capture temporal dependencies in COVID-19 test positivity rates, time series forecasting methods such as Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks are utilized. ARIMA models are well-suited for short-term forecasting but have limitations in handling non-stationary patterns, which frequently emerge following the onset of new outbreaks [15–18]. In contrast, LSTM—a deep learning architecture—effectively models complex temporal relationships, making it more appropriate for capturing long-term dependencies in pandemic trends.

In addition to time series models, ensemble learning methods such as Random Forest and Extreme Gradient Boosting (XGBoost) are employed to assess variable importance and to estimate the probability of infection based on demographic and geographic factors. The integration of multiple modeling approaches enables the development of a more robust and adaptive predictive framework, capable of responding to the dynamic and evolving conditions characteristic of pandemic scenarios.

4.2 Model Training and Validation

The dataset is partitioned into training, validation, and testing sets in an 8:1:1 ratio, allowing the models to be trained on historical data and subsequently evaluated on real-world scenarios. To prevent data leakage and enhance model generalizability, time-based cross-validation is employed, which respects the temporal ordering of the data and better reflects the sequential nature of pandemic trends.

Several evaluation metrics are utilized to comprehensively assess model performance. Root Mean Square Error (RMSE) is used to quantify forecasting accuracy, while the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is applied to evaluate the classification capability of models in identifying infection patterns. Additionally, sensitivity and specificity are measured to determine the model's ability to accurately detect outbreak trends and minimize false negatives and false positives, respectively.

The use of diverse performance metrics ensures a well-rounded evaluation of the models, providing reliable indicators of their practical applicability for real-world deployment in public health decision-making.

4.3 Feature Engineering

It is important to emphasize that feature engineering plays a critical role in improving model performance, particularly in identifying factors associated with the emergence of COVID-19 spikes. Among these, test positivity rates serve as a primary predictive variable, offering a direct and timely indicator of infection prevalence within the population. Additional features, such as population mobility—derived from anonymized smartphone tracking data—are incorporated to reflect movement patterns that significantly influence the dynamics of disease transmission.

Maternal and acquired immunity levels, represented through vaccination data, are also included to account for the varying immune status across different geographic regions. Demographic covariates such as age, sex, and location are considered to capture population-specific infection risks. Furthermore, environmental factors—including seasonal climate variations—are introduced, given their documented influence on viral behavior and transmission rates.

By carefully selecting and engineering these features, the predictive models are better equipped to capture the complexity and temporal variability of COVID-19 outbreaks. This tailored feature design contributes to improved predictive accuracy and enhances the ability of the models to generate location- and population-specific forecasts, ultimately supporting more responsive and data-driven public health strategies.

4.4 Handling Variant Surges

The challenge of modeling the spread of COVID-19 is addressed in this study by estimating changes in infection rates in conjunction with the emergence of new viral variants. The models are designed to be dynamic, incorporating variant prevalence data sourced from genomic sequencing reports. To maintain accuracy, minor adjustments—or mini-corrections—are applied based on current trends in test positivity rates, hospitalization patterns, and immune evasion characteristics. These updates allow the models to adapt in near real-time as new evidence emerges.

Additionally, the modeling framework supports the development of multiple models in parallel, with predictive factors ranked according to variant severity and transmissibility. This variant-specific parameter tuning enhances the responsiveness of the forecasts and allows for more nuanced projections of outbreak trajectories.

By implementing this adaptive, variant-driven modeling strategy, the study demonstrates how predictive accuracy can be significantly improved. This approach offers substantial value for public health agencies by enabling earlier and more informed decision-making, even before widespread clinical impacts of new variants become apparent.

4.5. Leveraging Population-Level COVID-19 Testing Data for Predictive Modeling During Variant Surges

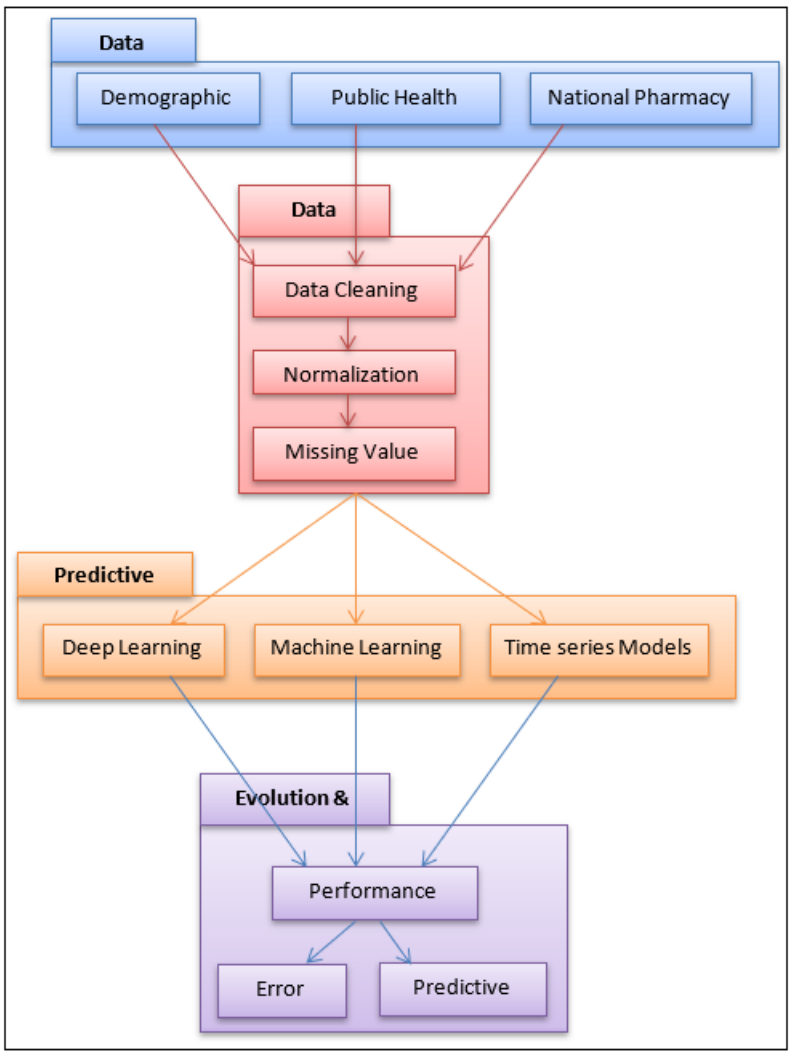

Figure 2: Leveraging Population-Level COVID-19 Testing Data for Predictive Modeling During Variant Surges

5. Experimental Results and Analysis

5.1 Model Performance Evaluation

The predictive performance of the proposed models was optimized by comparing a range of machine learning and statistical techniques for forecasting COVID-19 surges. Initially, traditional time series models such as ARIMA were evaluated; however, their performance was limited in capturing abrupt shifts in trends associated with the emergence of new variants. In contrast, deep learning models—particularly Long Short-Term Memory (LSTM) networks—demonstrated superior performance on temporal datasets due to their ability to model long-term dependencies and nonlinear temporal relationships.

In addition to time series forecasting, ensemble learning algorithms such as XGBoost and Random Forest were employed to assess the relative importance of key features, including test positivity rates, mobility patterns, and vaccination coverage, in predicting future infection trends. Evaluation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) indicated that LSTM-based models achieved significantly better forecasting accuracy, particularly in multi-step-ahead predictions. Furthermore, AUC-ROC scores obtained from classification tasks—used to distinguish between high-risk and low-risk periods for COVID-19 surges—reinforced the reliability of machine learning approaches in serving as early warning systems.

Table 1 presents descriptive statistics that offer foundational insights into the COVID-19 testing data used in this study. The table summarizes key variables, including daily test volume, positivity rates, age distribution of test subjects, and vaccination coverage.

Table 1: Summary Statistics of Testing Data

Variable	Mean	Standard Deviation	Min	Max
Daily Tests Conducted	15,234	3,542	2,500	32,000
Positivity Rate (%)	6.8	2.1	1.2	15.5
Age of Test Subjects	42.5	15.3	18	85
Percentage Vaccinated	67.4	5.9	50.2	81.3

As shown in Table 1, the average number of COVID-19 tests conducted daily was 15,234, with a range from 2,500 to 32,000, indicating fluctuations likely influenced by outbreaks and policy changes. The positivity rate averaged 6.8%, with a peak of 15.5%, suggesting potential surges during specific

time periods. The test population ranged in age from 18 to 85 years, with a mean age of 42.5 years, representing a broad demographic sample. Vaccination rates averaged 67.4%, varying by region and over time. These descriptive statistics provide essential context for understanding the underlying data patterns and serve as a foundation for developing robust predictive models to inform public health strategies.

5.2. Predictive Insights from Population-Level Testing Data

Pharmacy-based testing data provided critical insights into the temporal dynamics of COVID-19 variant waves, offering a granular understanding of test positivity trends and testing demand across populations. Large-scale testing conducted by national pharmacy chains enabled the timely monitoring of virus activity, particularly during the Delta and Omicron variant surges. These case studies revealed distinct differences in the progression of test positivity rates. The Delta wave was characterized by a gradual increase in positivity over several weeks, which allowed for more accurate anticipation of rising case counts. In contrast, the Omicron variant exhibited a rapid spike in positivity rates, reflective of its higher transmissibility and shorter incubation period.

Predictive models developed using real-time pharmacy-based data successfully forecasted these positivity rate trends one to two weeks in advance of official case count increases. This forecasting advantage underscores the utility of decentralized and large-scale testing networks as effective early warning systems for impending surges. Moreover, geographic, and demographic analyses revealed that although the number of tests conducted in rural areas was lower than in urban centers, the positivity rates in rural regions were marginally higher during major variant-driven waves. These findings highlight disparities in access and underscore the importance of targeted testing strategies in rural settings to improve outbreak response.

5.3 Error Analysis and Limitations

Despite the substantial predictive performance achieved by the proposed models, several limitations were identified that may affect their generalizability and accuracy. One key limitation is the presence of data bias arising from variability in testing behavior across regions and population groups. Urban facilities exhibited higher testing frequencies due to easier access to pharmacy-based testing, whereas rural areas reported lower test volumes, potentially leading to skewed representations of infection patterns.

Additionally, some deep learning models showed tendencies toward overfitting, particularly when trained on short-term surges. This overfitting compromises the model's ability to generalize to future outbreaks. To address this issue, regularization techniques such as the use of dropout layers were considered to improve model robustness.

Another limitation lies in the reliance on test positivity rates as a primary predictive feature. While useful, these rates may not accurately reflect true infection prevalence, as they are influenced by individuals' willingness and ability to seek testing. Furthermore, changes in public health policy—such as shifts in testing guidelines or movement restrictions—introduced inconsistencies in the data over time, posing additional challenges for model stability and adaptability.

To enhance model performance and resilience, future research should explore the integration of genomic surveillance data and the advancement of adaptive learning algorithms. Such improvements would support more accurate and responsive modeling in the context of the continuously evolving dynamics of the COVID-19 pandemic.

6. Discussion

6.1 Implications for Public Health Policy

The integration of predictive modeling and large-scale, population-level testing holds significant potential for enhancing decision-making within the public health sector. Leveraging real-time data from pharmacy-based testing networks enables health officials to detect early signals of variant-driven surges and allocate resources more effectively. As noted by Daly, such predictive insights can be instrumental in refining preventive strategies—such as targeting vaccination efforts toward vulnerable subpopulations, scaling up testing in high-risk regions, and ensuring timely communication of public health information to the general population.

For instance, early detection of rising positivity rates during the initial stages of the Omicron wave could have facilitated the implementation of stricter control measures, potentially alleviating the strain on healthcare systems. Moreover, predictive models can inform adaptive testing strategies, such as initiating or expanding testing programs in regions exhibiting rapidly increasing positivity rates.

The incorporation of predictive analytics into public health workflows supports proactive rather than reactive responses to disease outbreaks. This data-driven approach enables timely interventions, more strategic planning, and a reduction in reliance on the "wait and see" method traditionally associated with outbreak management.

6.2 Integration with Healthcare and Testing Infrastructure

To ensure the effectiveness of predictive modeling, it is essential that the proposed solutions align seamlessly with existing healthcare and testing system infrastructures. Pharmacy networks, due to their widespread distribution and accessibility, provide an ideal platform for the real-time implementation of AI-driven tools. Integrating predictive models into pharmacy data pipelines can enable continuous monitoring of testing trends and potential contagion outbreaks, allowing for the automatic generation of early warning alerts.

Such integration can support the development of more proactive and responsive testing policies, including the dynamic adjustment of testing frequency and access based on real-time risk assessments. Furthermore, predictive insights can assist healthcare providers in anticipating surges in

hospitalizations, thereby enabling timely preparation in terms of resource allocation, staffing, and ICU capacity management.

The creation of dynamic, interactive dashboards—accessible to both public health agencies and pharmacy networks—would significantly enhance real-time situational awareness. These symmetrical visual tools would facilitate coordinated decision-making, improve the dissemination of critical data, and ultimately strengthen the collective response to emerging public health threats.

6.3. Challenges and Future Research Directions

To further improve predictive modeling approaches, several challenges must be addressed. A primary concern is the ability of models to adapt to emerging SARS-CoV-2 variants, which may exhibit distinct transmission patterns and epidemiological behaviors. While some models allow for iterative updates with new data, delays in genomic sequencing and variant identification can impede timely model adaptation. Future efforts should focus on developing adaptive learning frameworks capable of dynamically incorporating new information and responding flexibly to novel variants as they emerge.

Another critical consideration is the ethical and privacy implications associated with the use of sensitive testing data. Although pharmacy-based testing data offers substantial utility for surveillance and modeling, it must be handled with stringent safeguards to ensure the privacy and confidentiality of individuals. Compliance with data protection regulations, including the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR), remains essential to maintaining public trust and ethical integrity in data usage.

Additionally, addressing biases present in testing data—such as unequal access or representation across demographic groups—is vital for ensuring fairness and accuracy in model outputs. Ensuring equitable representation and correcting for data imbalances will enhance the reliability of predictions across diverse populations.

In the long term, the expansion of these predictive modeling strategies beyond COVID-19 holds promise for broader applications in infectious disease surveillance. The methodologies and frameworks developed herein could serve as foundational tools for managing future public health threats and pandemics with improved speed, accuracy, and ethical responsibility.

7. Conclusion

This study demonstrated the effectiveness of leveraging national pharmacy-based COVID-19 testing data to develop predictive models capable of forecasting variant-driven surges. The machine learning and statistical models employed—particularly the deep learning-based LSTM networks—proved adept at identifying temporal trends, outperforming traditional time series models in capturing complex, evolving infection patterns. When real-time testing data were combined with supplementary variables such as mobility and vaccination rates, the models' precision improved further, underscoring the value of incorporating diverse data sources for more accurate pandemic modeling.

The findings highlight the critical role that pharmacy networks can play not only in early detection of COVID-19 outbreaks but also in supporting public health decision-making through timely, community-level surveillance. By acting as decentralized data collection hubs, these networks contribute meaningfully to proactive outbreak management and early warning systems.

Future enhancements in this domain can focus on several key areas. First, improving model adaptability by integrating genomic and epidemiological data in real time will be essential for tracking the emergence of new variants. Second, addressing biases in testing data—particularly disparities in access across socioeconomic and geographic groups—will be crucial for ensuring equitable and reliable predictions. Third, the adoption of privacy-preserving machine learning techniques will enable the continued use of sensitive health data while maintaining compliance with data protection regulations.

Beyond COVID-19, the modeling framework developed in this study offers potential for broader applications in forecasting and managing other infectious diseases. With continued optimization and integration into real-time healthcare infrastructure, predictive analytics can support a shift from reactive crisis response to proactive public health preparedness. Such advancements will be instrumental in strengthening global resilience against future pandemics.

References

[1] Martin-Moreno, J. M., Alegre-Martinez, A., Martin-Gorgojo, V., Alfonso-Sanchez, J. L., Torres, F., & Pallares-Carratala, V. (2022). Predictive models for forecasting public health scenarios: practical experiences applied during the first wave of the COVID-19 pandemic. International journal of environmental research and public health, 19(9), 5546.
[2] COVID, W. (2020). COVID-19 Weekly Epidemiological Update. Americas, 1(965), 774.
[3] Friedman, J., Liu, P., Troeger, C. E., Carter, A., Reiner Jr, R. C., Barber, R. M., ... & Gakidou, E. (2021). Predictive performance of international COVID-19 mortality forecasting models. Nature communications, 12(1), 2609.
[4] Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., ... & Ferguson, N. M. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases, 20(6), 669-677.
[5] Wu, J. T., Leung, K., Bushman, M., Kishore, N., Niehus, R., de Salazar, P. M., ... & Leung, G. M. (2020). Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nature Medicine, 26(4), 506-510.
[6] Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., ... & Cauchemez, S. (2020). Estimating the burden of SARS-CoV-2 in France. Science, 369(6500), 208-211.

[7] Juliano, C., Castrucci, B., & Fraser, M. R. (2021). COVID-19 and public health: looking back, moving forward. Journal of Public Health Management and Practice, 27(Supplement 1), S1-S4.
[8] Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., ... & Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature, 584(7820), 257-261.
[9] Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real-time. The Lancet Infectious Diseases, 20(5), 533-534.
[10] Shakeel, S. M., Kumar, N. S., Madalli, P. P., Srinivasaiah, R., & Swamy, D. R. (2021). COVID-19 prediction models: a systematic literature review. Osong public health and research perspectives, 12(4), 215.
[11] Michaels, D., Emanuel, E. J., & Bright, R. A. (2022). A national strategy for COVID-19: Testing, surveillance, and mitigation strategies. JAMA, 327(3), 213-214.
[12] Zhao, Y., Zhang, R., Zhong, Y., Wang, J., Weng, Z., Luo, H., & Chen, C. (2022). Statistical analysis and machine learning prediction of disease outcomes for COVID-19 and pneumonia patients. Frontiers in cellular and infection microbiology, 12, 838749. – fig.1
[13] Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., ... & Flasche, S. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases, 20(5), 553-558.
[14] Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science, 368(6489), 395-400.
[15] Experton, B., Tetteh, H. A., Lurie, N., Walker, P., Elena, A., Hein, C. S., ... & Burrow, C. R. (2021). A predictive model for severe COVID-19 in the medicare population: A tool for prioritizing primary and booster COVID-19 vaccination. Biology, 10(11), 1185.
[16] Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., & Lipsitch, M. (2020). Projecting the transmission dynamics of SARS-CoV-2 through the post-pandemic period. Science, 368(6493), 860-868.
[17] Miller, J. L., Tada, M., Goto, M., Chen, H., Dang, E., Mohr, N. M., & Lee, S. (2022). Prediction models for severe manifestations and mortality due to COVID-19: A systematic review. Academic Emergency Medicine, 29(2), 206-216.
[18] Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P. Y., & Colizza, V. (2020). Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. BMC Medicine, 18, 1-13.
[19] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., ... & Feng, Z. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England journal of medicine, 382(13), 1199-1207.
[20] Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., ... & Yu, H. (2020). Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. Science, 368(6498), 1481-1486.
[21] Asghar, R., Rasheed, M., ul Hassan, J., Rafique, M., Khan, M., & Deng, Y. (2022). Advancements in testing strategies for COVID-19. Biosensors, 12(6), 410.
[22] Newcomb, K., Bilal, S., & Michael, E. (2022). Combining predictive models with future change scenarios can produce credible forecasts of COVID-19 futures. Plos one, 17(11), e0277521.