scientific reports

Check for updates

OPEN

Enhancing COVID-19 forecasting precision through the integration of compartmental models, machine learning and variants

Daniele Baccega[1,3]✉, Paolo Castagno[1], Antonio Fernández Anta[2] & Matteo Sereno[1]

Predicting epidemic evolution is essential for making informed decisions and guiding the implementation of necessary countermeasures. Computational models are vital tools that provide insights into illness progression and enable early detection, proactive intervention, and targeted preventive measures. This paper introduces Sybil, a framework that integrates machine learning and variant-aware compartmental models, leveraging a fusion of data-centric and analytic methodologies. To validate and evaluate Sybil's forecasts, we employed COVID-19 data from several European and U.S. states. The dataset included the number of new and recovered cases, fatalities, and variant presence over time. We evaluate the forecasting precision of Sybil in periods in which there is a change in the trend of the pandemic evolution or a new variant appears. Results demonstrate that Sybil outperforms conventional data-centric approaches, being able to forecast accurately the changes in the trend, the magnitude of these changes, and the future prevalence of new variants.

Keywords Artificial intelligence, Epidemics, Compartment models, Variants, Forecasting, COVID-19

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, highlights the intricate challenges of addressing the most impactful global health crisis of the 21st century. The rapid global spread of the virus has affected nearly every part of the world. Consequently, healthcare systems worldwide are grappling with the significant challenge posed by COVID-19, requiring a continuous COVID-19 monitoring system that includes robust surveillance, widespread testing, and contact tracing, and that can be used to plan and deploy stringent infection control measures.

Establishing a continuous monitoring system aids policymakers in effectively managing the socio-health emergency brought about by the epidemic. Accurate forecasting is a fundamental element of such a system, and is crucial for efficient planning, resource allocation, and decision-making within public health authorities. It facilitates the development of proactive measures, such as vaccination campaigns, travel advisories, and community engagement programs, fostering public awareness and participation in disease control efforts. This proactive approach enhances preparedness and is critical in curbing the spread of infectious diseases and mitigating their impact on communities worldwide. Epidemic forecasting involves a multidisciplinary approach, integrating epidemiology, mathematical modeling, data analysis, and computational methods to gain insights into the future progression of outbreaks.

Numerous methodologies are available for predicting the future trajectory of an epidemic, leveraging diverse modeling approaches[1]. Machine learning (ML) and deep learning (DL) models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) featuring Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) cells, and multivariate CNNs, have emerged as highly prominent approaches for forecasting[2].

In particular, regarding ML models, Alabdulrazzaq et al.[3] investigated and validated the accuracy of an auto-regressive integrated moving average (ARIMA) model over a relatively long period of time using Kuwait COVID-19 data as a case study. Zhao et al.[4] compared the predictions obtained using an ARIMA model, a multiple linear regression (MLR) model, and Prophet[5] with the aim to predict the daily worldwide confirmed cases of COVID-19 during the Omicron outbreak, finding out that ARIMA has superior prediction performance compared to MLR and Prophet models. Rendana et al.[6] used ARIMA to forecast the number of cases, using Spearman correlation

[1]Computer Science Department, Universitá di Torino, Turin, Italy. [2]IMDEA Networks Institute, Madrid, Spain. [3]Laboratorio InfoLife, Consorzio Interuniversitario Nazionale per l'Informatica (CINI), Rome, Italy. ✉email: daniele.baccega@unito.it

to analyze the relationship between B.1.1.7 cases and meteorological variables such as temperature, humidity, rainfall, sunshine, and wind speed. Battineni et al.[7] used Prophet to predict the epidemic trend and derive an epidemic curve in USA, Brazil, India, and Russia, while Wang et al.[8] combined a logistic model with Prophet with the same objective using COVID-19 data for Brazil, Russia, India, Peru, Indonesia, and the whole world. Sardar et al.[9] compared an ARIMA model, Prophet, a generalized linear model elastic net (GLMNet), and a random forest (RF) for predicting the confirmed cases in the South Asian Association for Regional Cooperation (SAARC) countries also finding out that ARIMA is superior to the other models. Finally, a comparison among a support vector machine (SVM), Prophet, and a linear regression (LR) model were used by Gupta et al.[10] to predict the active rate, the death rate, and the cured rate in India showing that Prophet provided the best predictions with vast uncertain and small data sets.

Concerning DL models, Sah et al.[11] compared the results obtained using Prophet, an ARIMA model, and a stacked LSTM-GRU model with the aim to forecast the number of confirmed and active cases in India, revealing that the proposed stacked LSTM-GRU model outperforms all other models. Shahid et al.[12] used an ARIMA model, a support vector regression (SVR) model, an LSTM model, and a bidirectional LSTM (Bi-LSTM) model to make predictions on confirmed cases, deaths, and recoveries in the top-10 countries affected by COVID-19, showing that Bi-LSTM outperforms in terms of endorsed indices. ArunKumar et al.[13] used GRU and LSTM cells, ARIMA, and seasonal ARIMA to forecast the trends in the same countries, finding out that for most of the countries deep learning models outperformed statistical models, but for some countries statistical models outperformed deep learning models. An RNN with LSTM cells and variants such as Deep LSTM, Convolutional LSTM, and Bi-LSTM was used by Arora et al.[14] with the aim to predict the number of positive cases in India, discovering that Bi-LSTM gives the best results. Finally, a comparison among a CNN, a multivariate CNN (MCNN), LSTM, and GRU cells was conducted by Nabi et al.[15] with the aim to predict future outbreak scenarios in Brazil, Russia, and the United Kingdom, showing that uni-variate and multi-variate CNNs outperform RNNs.

Despite their increasing popularity and their capacity to outperform conventional techniques such as auto-regressive models, including ARIMA and SARIMA[16], these data-centric approaches exhibit several notable limitations that warrant attention. First and foremost, the predictions generated by these advanced methods are not easily explainable. The shortage, or even lack, of such a property makes it challenging for users to comprehend the underlying factors and mechanisms driving the forecasts, which can undermine trust and hinder the interpretability of the results. Additionally, these sophisticated approaches often struggle to accurately predict significant changes in trends, such as abrupt peaks in data or the sudden emergence of new variants or outbreaks. Such shortcomings are particularly problematic in scenarios where timely and precise forecasting is critical for effective decision-making and intervention strategies. Consequently, while these data-centric methods hold great promise and have demonstrated superior performance in many instances, their limitations must be carefully considered and addressed to fully leverage their potential in practical applications.

On the other hand, compartmental models are analytic approaches specifically tailored to reproduce the evolution of an infection in a population which take into account different aspects related to an epidemic, like vaccinations and variants[17–19]. In particular, Caldwell et al.[17] provided important insights on the COVID-19 pandemic like calculating the herd immunity threshold, verifying that nascent vaccines can prevent severe disease, determining optimal vaccine allocation strategies, and determining that Variants of Concern (VoC) are more transmissible and lethal than previously circulating strains. Miller et al.[18] demonstrated that systematic properties of population dynamics in infectious diseases can inform the likelihood of the random process of mutation. Gostic et al.[19] provides suggestions for how to mitigate some technical challenges and highlight open problems in effective reproductive number (R_t) estimation. Extended versions of compartmental models were used to improve the accuracy of the predictions in different countries taking into account different age groups, symptoms, hospitalizations, intensive care units admissions, and vaccinations[20–23], undetected infections and human mobility between regions[24], quarantine and governmental intervention measures[22,23,25].

Stochastic transmission models were also used to study the transmission of the COVID-19 disease[26,27]. These models account for the inherent uncertainty in disease transmission by integrating random variables like individual interactions, chance encounters, and variations in infectiousness, offering a more nuanced understanding of how diseases spread and allowing for more robust predictions in complex and dynamic real-world scenarios.

These analytic models incorporate a wide range of factors, including population demographics, rates of infection, recovery, and mortality, thereby providing a comprehensive mathematical representation of epidemic progression[28,29]. This level of transparency simplifies the process of explaining and interpreting the results to various stakeholders, such as policymakers, healthcare professionals, and the general public. Consequently, it fosters a better understanding of the situation and facilitates more informed decision-making. However, despite these advantages, these analytic approaches are not without their challenges. They rely heavily on a set of assumptions about the system being modeled, which may not always hold true in real-world scenarios. Additionally, the process of parameter estimation presents significant challenges, as accurate data collection and validation are crucial for the reliability of the models. Indeed, several COVID-19 forecasting studies incorporate data-centric and analytical approaches, such as a combination of compartmental models and ARIMA with the aim to develop informed mitigation strategies and inform the preparedness of policymakers and health care workers[30], a logistic model and Prophet[8], a compartmental model and a DL model (DNN/RNN-LSTM) to estimate the stochastic transmission parameters of the compartmental model with a discrete time series construct[31], a compartmental model, a machine learning model and a Bayesian model for empirically grounded COVID-19 predictions[32], and several other combinations[33].

Forecasting the epidemic spread aims to accurately predict the percentage of the infected population, the number of fatalities, hospitalizations, and so on, at a given point in the future. However, these metrics result from complex population dynamics that often show nonlinear behaviors, particularly at critical points where there are

changes in the infection course—such as at the peak of the diffusion or when a new variant arises. Conversely, epidemics dynamics are often characterized by widely recognized quantities. The most well-known is the basic reproduction number, R_0, which expresses the number of secondary infections arising from one single infected individual within a population of susceptible individuals. Although R_0 is useful to characterize if and how fast a disease spreads in a population, its time-dependent counterpart, R_t, enables a quantitative evaluation of the infection course. Such indicators, being specific to the disease, tend to be stable. Therefore, their future evolution shows a more predictable behavior.

The joint use of ML and analytical models is gaining momentum in computational epidemiology. Still, to the best of our knowledge, our approach outperforms the available solutions in several aspects, such as its explainable, reproducible and replicable results. For instance, Kiarie et al.[30] used a compartmental model combined with a predictive model of the pandemic to forecast its evolution 60 days in the future, building their approach on the observed data in Kenya. In particular, they propose to estimate the parameters of the compartmental model—precisely the effective reproductive number—using ARIMA and then to use the predicted values to forecast the pandemic with a SEIR compartmental model. Specifically, the authors design a compartmental model accounting for symptomatic and asymptomatic individuals, as well as mild and severe cases.

Watson et al.[32] integrated a Bayesian time series model and a random forest algorithm within an epidemiological model to predict COVID-19 outcomes. The Bayesian model fits location-specific curves to the growth rate of log-transformed cumulative cases, using data from various areas and prior information. This informs the compartmental model, which predicts deaths using a random forest algorithm trained on COVID-19 data. The model provides daily projections and interval estimates for U.S. states, evaluated by measuring 21-day forecast accuracy. This work shares many of the considerations and assumptions that will be made in our paper, making it a natural candidate for comparing (and evaluating) the effectiveness of the choices we will make. Specifically, in the 'Supplementary Material', we will present a comparative analysis of the two methods using data on the spread of COVID-19 in three U.S. states—i.e., Colorado, New York and West Virginia.

Moreover, despite the many similarities with the studies of Kiarie et al.[30] and Deng[31], our approach employs a much simpler compartmental model, which enables us to consider variants and still have a model without significant complexities. Also, using ARIMA makes the approach in[30] less suitable for continuous monitoring platforms. It also makes it harder to apply the same methodology to other scenarios. Eventually, results presented in[30] are not validated against historical data, and the forecasting accuracy is not easy to assess.

Following the approach of combining data-driven approaches with analytical models, we propose **Sybil**, an integrated machine learning and variant-aware compartmental model framework capable of providing improved prediction accuracy and explainability. Sybil exploits the relative stability of disease characteristics indices to project in the future and employs a simple and widely recognized analytical model to draw the infection dynamic. Sybil's strengths mark the difference with approaches present in the literature thanks to *(i)* its capability of providing accurate forecasts, even when there are relevant changes in the diffusion process, and *(ii)* reduced need for training data. Furthermore, the approach offers *(iii)* the possibility to study the evolution of the infection of several variants and *(iv)* the replicability of the results. Additionally, *(v)* the open-source code is available online.

Methods

We propose an integrated framework aimed at providing accurate and explainable predictions for epidemic spreads. Sybil combines a simple compartmental model with a machine learning-based predictive model to forecast the future progression of infection, even in the presence of multiple virus strains. At the core of Sybil, there is a simple analytical model which has a dual functionality. In the first stage of Sybil's operation, the analytical model is used to derive the value of critical model parameters from the surveillance data. Then, these parameters' values—ascribable to the reproductive number over time, R_t—are used as training data for the ML component of Sybil. Based on that knowledge, it predicts the future values for the key parameters, which are then plugged back into the analytical model. Then, Sybil computes the future evolution of daily infections using the analytical model with these key parameters' future values (see Fig. 1-(a) for a schematic representation of Sybil's steps).

The effectiveness and accuracy of Sybil's forecasts are demonstrated through two methods: *(i)* comparing them against real data gathered from active surveillance of the pandemic in several European and U.S. states, and *(ii)* comparing them against forecasts generated by various state-of-the-art approaches, including Prophet[5], ARIMA / SARIMA[16], Neural Prophet[34], LSTM[35] and GRU[36] neural networks, EpiNow2[37], and the framework by Watson et al.[32]. In the latter case, we predict the evolution of different active variants over the specified period.

Compartmental analytical model

The first component of Sybil is the analytical model. In the field of computational epidemiology, compartmental models are widely recognized and employed to study the spreading of an infection in a population. Also, compartmental models are easy to explain: the model contains several compartments, each representing a specific subpopulation, and uses rates to move individuals from one compartment to another. The simplest compartmental model to mimic the evolution of an infection in a population requires only two compartments and one rate; the two compartments represent the only two possible states of the individuals in the population—either susceptible (S) or infected (I)—and one transition rate that represents the pace at which one infected individual infects a susceptible one. Because of the two states characterizing this model, it is commonly known as the Susceptible-Infected (SI) model.

In order to capture some critical features of the SARS-CoV-2 infection, an SI model is not suitable, and a more complex one is required. Specifically, we need to model the end of the infected period explicitly, i.e., when an infected person recovers from the illness. Additionally, fatalities and reinfections must be taken into account.

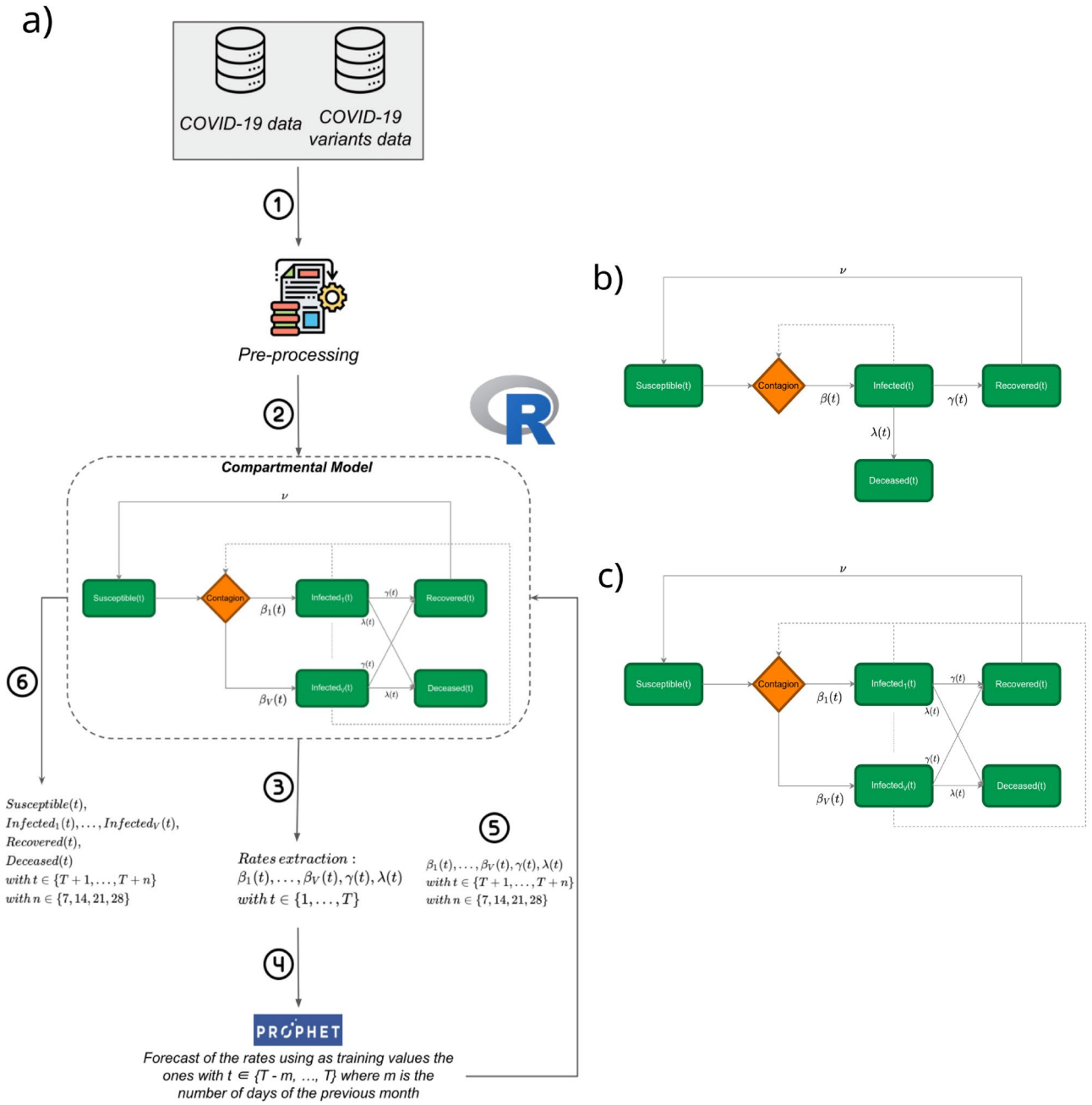

Figure 1. (**a**) Schematic representation of Sybil's steps. Starting from COVID-19 data of a specific country (1) we pre-processed and (2) injected them into the SI^{V}RDS model reconstructing the evolution of the epidemic in the considered country. Then, (3) we extracted the infection rates $\beta_v(t)$ for each variant v, the recovery rates $\gamma(t)$ and the fatality rates $\lambda(t)$ with t ∈ {1, ..., T} (where t = 1 represents the first day with available data and t = T represents the forecasting point) and (4) we applied Prophet[5] using the rates of the previous month (with t ∈ {T - m, ..., T}, where m represents the number of days of the previous month) as training data with the aim to predict the rates one, two, three and four weeks in the future (with t ∈ {T + 1, ..., T + n} and n ∈ {7, 14, 21, 28}). Finally, (5) we injected these predicted rates inside the SI^{V}RDS model to obtain (6) the evolution of Susceptible(t), Infected$_1$(t), ..., Infected$_V$(t), Recovered(t) and Deceased(t) with t ∈ {T + 1, ..., T + n}) (with n ∈ {7, 14, 21, 28}). The R Logo is used under the Creative Commons Attribution-ShareAlike 4.0 International license (CC-BY-SA 4.0), while the Prophet Logo is used under the MIT License. (**b**) SIRDS compartmental model. $\beta(t)$, $\gamma(t)$, $\lambda(t)$ and ν refer to infection, recovery, fatality and end-of-immunization rates, respectively (which are all time dependent, but the end-of-immunization rate). (**c**) SI^{V}RDS compartmental model. $\beta_1(t)$, ..., $\beta_V(t)$ refer to the infection rates of the V considered variants. Dashed lines in (**b**) and (**c**) refer to the fact that an infection is due to contacts among susceptible and infected individuals.

Hence, the compartmental model incorporated in Sybil has two additional compartments, one accounting for individuals recovered from the infection (R) and one for the deceased (D), with the corresponding rates. Specifically, the model used is a Susceptible-Infected-Recovered-Deceased-Susceptible (SIRDS) compartmental model with reinfections—represented in Fig. 1-(b)— that is described by the system of difference equations in Eq. 1. In this system, β represents the infection rate, γ the recovery rate, λ the fatality rate, ν the end-of-immunization rate, and N the total population.

$$\begin{aligned} S(\tilde{t}+1) &= S(\tilde{t}) - \boldsymbol{\beta}(\tilde{\boldsymbol{t}})\frac{S(\tilde{t})I(\tilde{t})}{N} + \nu R(\tilde{t}) \\ I(\tilde{t}+1) &= I(\tilde{t}) + \boldsymbol{\beta}(\tilde{\boldsymbol{t}})\frac{S(\tilde{t})I(\tilde{t})}{N} - \boldsymbol{\gamma}(\tilde{\boldsymbol{t}})I(\tilde{t}) - \boldsymbol{\lambda}(\tilde{\boldsymbol{t}})I(\tilde{t}) \\ R(\tilde{t}+1) &= R(\tilde{t}) + \boldsymbol{\gamma}(\tilde{\boldsymbol{t}})I(\tilde{t}) - \nu R(\tilde{t}) \\ D(\tilde{t}+1) &= D(\tilde{t}) + \boldsymbol{\lambda}(\tilde{\boldsymbol{t}})I(\tilde{t}) \end{aligned} \tag{1}$$

In this model, the rates are time-dependent—meaning that they may vary at each time step, with the time step corresponding to one day. The only exception is end-of-immunization rate ν, which is assumed to be $\nu = \frac{1}{180}$, since on average the immunization due to infection is estimated to be lost after 180 days[38].

Using the surveillance data—possibly after a pre-processing phase (step *1)* of Fig. 1-(a))—, Sybil computes the evolution of the infection process by solving the SIRDS model in Eq. 2, where $\Delta I(\tilde{t})$ represents the new infected at time $\tilde{t}$, $\Delta R(\tilde{t})$ the new recoveries at time $\tilde{t}$ and $\Delta D(\tilde{t})$ the new deceases at time $\tilde{t}$ (with S(0) = N and I(0) = R(0) = D(0) = 0)—step *2)* of Fig. 1-(a).

However, obtaining all the required parameters to solve equations in Eq. 1 is not straightforward. Indeed, surveillance data does not provide the transition rates—namely, the bold elements in the system of difference equations—and hence a further step is required: using Eq. 3 (obtained from Eq. 1), we can estimate the daily infection, recovery, and fatality rates—step *3)* of Fig. 1-(a).

$$\begin{aligned} S(\tilde{t}+1) &= S(\tilde{t}) - \Delta I(\tilde{t}) + \nu R(\tilde{t}) \\ I(\tilde{t}+1) &= I(\tilde{t}) + \Delta I(\tilde{t}) - \Delta R(\tilde{t}) - \Delta D(\tilde{t}) \\ R(\tilde{t}+1) &= R(\tilde{t}) + \Delta R(\tilde{t}) - \nu R(\tilde{t}) \\ D(\tilde{t}+1) &= D(\tilde{t}) + \Delta D(\tilde{t}) \end{aligned} \tag{2}$$

$$\begin{aligned} \lambda(\tilde{t}) &= \frac{D(\tilde{t}+1) - D(\tilde{t})}{I(\tilde{t})} && \text{if } I(\tilde{t}) > 0 \text{ else } \lambda(\tilde{t}) = 0 \\ \gamma(\tilde{t}) &= \frac{R(\tilde{t}+1) - R(\tilde{t}) + \nu R(\tilde{t})}{I(\tilde{t})} && \text{if } I(\tilde{t}) > 0 \text{ else } \gamma(\tilde{t}) = 0 \\ \beta(\tilde{t}) &= \frac{I(\tilde{t}+1) - I(\tilde{t}) + \gamma(\tilde{t})I(\tilde{t}) + \lambda(\tilde{t})I(\tilde{t})}{S(\tilde{t})I(\tilde{t})}N && \text{if } S(\tilde{t})I(\tilde{t}) > 0 \text{ else } \beta(\tilde{t}) = 0 \end{aligned} \tag{3}$$

Bringing in variants into the model of Eq. 1— again, possibly after a pre-processing phase (step *1)* of Fig. 1-(a))—requires introducing one additional compartment to account for infections caused by each virus strain, hence obtaining an $SI^{V}RDS$ model, where the superscript at the I stands for the maximum number of virus strains included in the model (see Fig. 1-(c)). After introducing further compartments, also additional rates are necessary. Specifically, instead of a global infection rate $\beta(\tilde{t})$, there are V different infection rates—one for each variant—for each time step $\tilde{t}$. Making a simplifying assumption, Sybil assumes that the evolution of the $I_v(\tilde{t})$ compartment—for each variant v and for each time step $\tilde{t}$—is computed starting from the evolution of the $I(\tilde{t})$ compartment of the model without variants and the daily proportion of the considered variant; that is,

$$I_v(\tilde{t}) = I(\tilde{t})\pi_v(\tilde{t}) \tag{4}$$

where $\pi_v(\tilde{t})$ is the proportion of infections due to variant v at time $\tilde{t}$. Moreover, we assumed no correlations among variants and that being recovered from a variant v makes a person immune from all variants (and after $\frac{1}{\nu}$ days, in average, a person will be susceptible again).

$$I_v(\tilde{t}+1) = I_v(\tilde{t}) + \boldsymbol{\beta_v}(\tilde{\boldsymbol{t}})\frac{S(\tilde{t})I(\tilde{t})}{N} - \gamma(\tilde{t})I_v(\tilde{t}) - \lambda(\tilde{t})I_v(\tilde{t}) \qquad v \in \{1...V\} \tag{5}$$

$$\beta_v(\tilde{t}) = \frac{I_v(\tilde{t}+1) - I_v(\tilde{t}) + \gamma(\tilde{t})I_v(\tilde{t}) + \lambda(\tilde{t})I_v(\tilde{t})}{S(\tilde{t})I(\tilde{t})}N \qquad v \in \{1...V\} \tag{6}$$

Equation 5 describes the evolution of the I_v compartments—step *2)* of Fig. 1-(a). Here, we used the previously computed global recovery and fatality rates, and $\beta_v(\tilde{t})$ is a column vector with the time-dependent infection rates of V variants—in the 'Supplementary Material' we provide a detailed discussion of these assumptions. Moreover, Eq. 6 (obtained from Eq. 5) describes how to compute the infection rates $\beta_v(\tilde{t})$ for the v variant—step *3)* of Fig. 1-(a).

It is worth noticing that in devising infection and fatality rates from surveillance data, we incorporate all the effects due to the different virus strains and the impact of contention policies, even if such details are not explicitly detailed in the model.

Prophet predictive model

The second component of the Sybil framework is Prophet[5], an open-source framework developed by Facebook for time series forecasting. It is based on an additive model design and it has been conceived to have intuitive parameters that can be adjusted without knowing the details of the underlying model. The modeling approach of Prophet combines the strengths of both statistical modeling and machine learning techniques; it utilizes a generalized additive model that incorporates piece-wise linear trends, nonlinear growth, and seasonality adjustments using a Fourier series. This flexible modeling approach enables Prophet to capture simple and complex data patterns. In particular, it consists of three main model components: trend, seasonality, and holidays combined in the following equation:

$$y(\tilde{t}) = g(\tilde{t}) + s(\tilde{t}) + h(\tilde{t}) + \varepsilon_{\tilde{t}} \tag{7}$$

where $g(\tilde{t})$ is the trend function that models non-periodic changes, $s(\tilde{t})$ represents periodic changes (e.g., daily, weekly, and yearly seasonality), and $h(\tilde{t})$ represents the effects of holidays which occur on potentially irregular schedules over one or more days. The error term $\epsilon_{\tilde{t}}$ represents any characteristic changes that are not accommodated by the model. Furthermore, Prophet makes it possible to estimate uncertainty in trend forecasts. It employs Markov Chain Monte Carlo (MCMC) to generate many plausible future trajectories. The MCMC procedure randomly samples from the posterior distribution of the model parameters, allowing for a range of possible outcomes. These sampled parameter sets are then used to generate multiple forecast trajectories.

As explained in the *Compartmental analytical model* subsection, from Eqs. 3 and 6 we extracted the infection rates $\beta_v(\tilde{t})$ for each variant v, the recovery rates $\gamma(t)$ and the fatality rates $\lambda(t)$ (with $t \in \{1, ..., T\}$, where $t = 1$ represents the first day with available data and $t = T$ represents the forecasting point). Starting from these rates, we applied Prophet to predict the values of the rates one, two, three and four weeks in the future (with $t \in \{T + 1, ..., T + n\}$ and $n \in \{7, 14, 21, 28\}$) using as training data the values of the rates of the previous month (with $t \in \{T - m, ..., T\}$, where m represents the number of days of the previous month)—step *4)* in Fig. 1-(a)—(we also tried to use more months, one year and the whole history as training data, but the performance of Sybil have worsened considerably due to the high variability present over long periods). Finally, we injected these new values into the SI^VRDS model—step *5)* in Fig. 1-(a)—obtaining the evolution of each compartment one, two, three and four weeks in the future—step *6)* in Fig. 1-(a).

Data

All datasets used in this work are open source and publicly available. The surveillance data used for the considered countries—namely, Italy, Austria, Belgium, France, Germany, Sweden and U.S.—in the *Results* section and in the 'Supplementary Material' are available in the COVID19 R library[39,40]—for the results presented for Italy and Austria we used a snapshot dated November 22nd, 2023. These data comprises all the data reported in Figs. 1 to 7 plus other data not relevant to the present study—such as the number of vaccinations, tests, hospitalizations and people in intensive care, information about applied countermeasures, data on mobility, etc. In particular, from the COVID-19 Data Hub[39,40], we used the following data (from February 2020 to May 2023): the cumulative number of cases, the cumulative number of recoveries, the cumulative number of deaths, and the total population. Starting from these, we computed the daily new cases, the daily new recoveries, and the daily new deaths we used in Eq. 2.

Concerning European countries, information about active virus strains comes from the European Center for Disease Control (ECDC)[41,42]—for the results presented for Italy and Austria we used a snapshot dated July 25th, 2023—, while for U.S., this information comes from the Centers for Disease Control and Prevention (CDC)[43]. Here, ECDC and CDC collected the result of serological tests, and there is information about a significant number of variant lineages, such as B.1.1.7, BA.1, BA.2, P.1, XBB, and many others. ECDC data comprises, for each week, the number of sequenced cases (which are a small portion of the new cases) and, for each lineage, the number and the proportion of sequenced cases that had that particular lineage. CDC data comprises, for each week, lineage and Health and Human Services (HHS) U.S. region (and for the whole USA), the proportion of the lineage in the population. To use such data, we aggregated these lineages in four main variant families with the aim of using the WHO variants labels: Alpha, Delta, Omicron, and *Other*[44]. In particular, the *Other* variant comprises the initial SARS-CoV-2 lineage, all the other lineages (e.g., Beta, Gamma, Kappa), and some noisy values in the surveillance data.

Since variants' diffusion data is provided weekly, in a pre-processing phase, we expanded such data to devise approximated daily values. Specifically, we employed splines—piece-wise-defined mathematical functions that use multiple polynomial segments to create a smooth and continuous curve—to obtain daily values. We used these daily values in Eq. 4.

We want to highlight that Sybil strongly depends on the availability of data. In[39,40] there are only few nations with a complete time series (with a daily step) for the used data. For this reason we are extending Sybil to include a pre-processing step (using splines) to fill missing data and to move from a weekly to a daily step in case of availability of data with a weekly step—without alter too much the information present in the time series.

In the following section, we used time-dependent recovery rates thanks to the availability of good reported official data—on cases, recoveries and fatalities—for the selected countries—namely Italy and Austria. In the 'Supplementary Material' we also report results on Belgium and France using time-dependent recovery rates—although the official data were not well reported. Moreover, surveillance data available for many countries worldwide reports incorrect data on recoveries or does not report data to devise daily recovery rates. Hence, to

dispense with data on recoveries, in the 'Supplementary Material' we report results obtained with fixed recovery rate with mean $\frac{1}{\gamma}$ equals to 14 days[45]—same for each variant—considering more European countries—i.e., Italy, Austria, Belgium, France, Germany, and Sweden—and some U.S. states. The results obtained are very good and show the robustness of the Sybil approach.

Results

We test the efficacy and the robustness of Sybil taking into account different European countries—i.e., Italy, Austria, Belgium, France, Germany, and Sweden—and some U.S. states. In this section we focus on surveillance data from Italy and Austria thanks to the availability of good reported official data and for the sake of convenience—the same comments are valid for the other countries, see 'Supplementary Material' for more details. Surveillance data from Italy spanning a period from February 2020 to May 2023 is used. During this period, there are a significant number of lineages, some coexisting and others with an evolutionary advantage taking over. In Fig. 2-(a), lineages are aggregated into the four main variants introduced above.

Accurate forecasts strongly depend on the regularity of the data to predict: it is pretty easy to foresee the future evolution of some dimensions increasing—or equally decreasing—linearly. Conversely, predicting the behavior in the presence of sudden changes in those quantities is much harder. Unfortunately, outbreaks, as well as peak infection fading, exhibit such a behavior. Vertical dashed lines in Fig. 2-(a) mark the time point picked to evaluate the accuracy of Sybil. The first two selected points are placed in the ascending part of an outbreak but close enough to the peak for accurate predictions, Sybil must successfully reproduce a change in the concavity of the function. The third selected point is placed just before the start of Omicron's outbreak to show that Sybil is also able to predict a new emerging variant/outbreak. We compared the predictions obtained using Sybil with the predictions obtained with the plain use of different state-of-the-art approaches—i.e., Prophet[5], ARIMA / SARIMA[16], Neural Prophet[34], LSTM[35] and GRU[36] neural networks, and EpiNow2[37]—to forecast the evolution of the number of daily infections. In the figures we reported the comparison between Sybil and Prophet since the latter is a component of the former, but we computed the Root Mean Squared Error (RMSE) with the standard deviation between the predictions obtained with Sybil and all the considered state-of-the-art approaches and the ground truth—in the figures legends, in Table 1 and in the 'Supplementary Material' we reported these values. Moreover, in the 'Supplementary Material' we conducted a comparative analysis with the study of Watson et al.[32] considering some U.S. states.

The first forecast scenario corresponds to the first infection wave, which in Italy started in February 2020. Only one SARS-CoV-2 strain was detected during this initial wave, namely the *Other* variant of Fig. 2-(a). Figure 2-(b) reports four different forecasts, all starting on April 15th, 2020 and spanning several forecast windows—from one to four weeks. For each forecasting window, we compare predictions obtained using Sybil (green line) and the standard approach used in the literature (red line), which requires using the selected forecasting approach—Prophet in our case—to project the number of infections in the coming weeks. The two approaches use the same period as training data (black line) for a fair comparison. They are compared and contrasted against the surveillance data for the period spanning the forecasting window (blue line). Comparing predictions obtained with the plain use of Prophet to forecast the evolution of the number of daily infections, we notice that it fails to predict the plateau characterizing the highest part of the peak. Considering Sybil's predictions, we see that starting from short-term predictions—i.e., seven days—it catches the decreasing trend. Increasing the forecasting window, Sybil always provides an excellent approximation of the future evolution of the daily infections, but for four weeks. In this latter case, Sybil's predictions diverge in magnitude from the ground truth, but not in trend. It provides a qualitative indication of the future evolution of the pandemic, which eventually fades. Figure 2-(c) shows the infection rate $\beta_v(\tilde{t})$ for the observation period. Here, we clearly see that the values for the infection rate are noisy; what we see is not a smooth line, but saw-tooth function with many peaks and valleys. Nonetheless, the range of variation of that function is limited and with a clear trend, which Sybil easily learns and replicates.

The second scenario considers a period spanning the highest peak in the data set. Here, we consider the period from December 13th, 2021 to January 13th, 2022 as training data, and we forecast the daily infections for the period January-February 2022 (starting from January 14th). In Italy, there were three active variants within this time window: Omicron, Delta, and the *Other* variant. Figure 3-(a) shows the number of daily infections for the three active variants and compares the ground truth (dashed line) with Sybil's forecasts for one to four weeks. Again, forecasts are highly accurate for seven to twenty-one days long predictions and slightly anticipate the peak's descendent phase at four weeks. In Fig. 3-(b), the two approaches are contrasted against the ground truth. Like the first scenario, the Prophet's predictions do not capture the peak. Not only do Prophet's predictions get far from the real data—they grow while the infection fades—but, in this scenario, they do not even provide a valid qualitative prediction.

The third point picked to test Sybil's predictions falls just before the explosion of the largest outbreak, caused by the Omicron variant. In particular, Fig. 4 shows how the one-week forecast changes moving the training window from December 18th, 2021 to December 27th, 2021. In this scenario, the ascending trajectory is very steep, and Sybil finds it harder to calibrate with respect to the previous cases. Here, the Omicron variant is a new emerging variant and Sybil initially foresees a more aggressive exponential growth. Looking at Fig. 4 we can state that Sybil nailed the qualitative prediction, but needed more data to calibrate it. We can also see that Sybil captures well the prediction on the other active variant—the Delta variant. Here we are showing the best results obtained in this scenario. In the 'Supplementary Material' we provide more details on this scenario and show another case where an explosive outbreak occurs, discussing the convenience of establishing a continuous monitoring system.

From the previous scenarios we can observe that Sybil provides a robust, easily portable approach to different use cases without requiring modifications. For instance, in Fig. 5 we apply Sybil to surveillance data from Austria considering three different scenarios. Figure 5-(a) shows the daily infections for the different SARS-CoV-2 strains

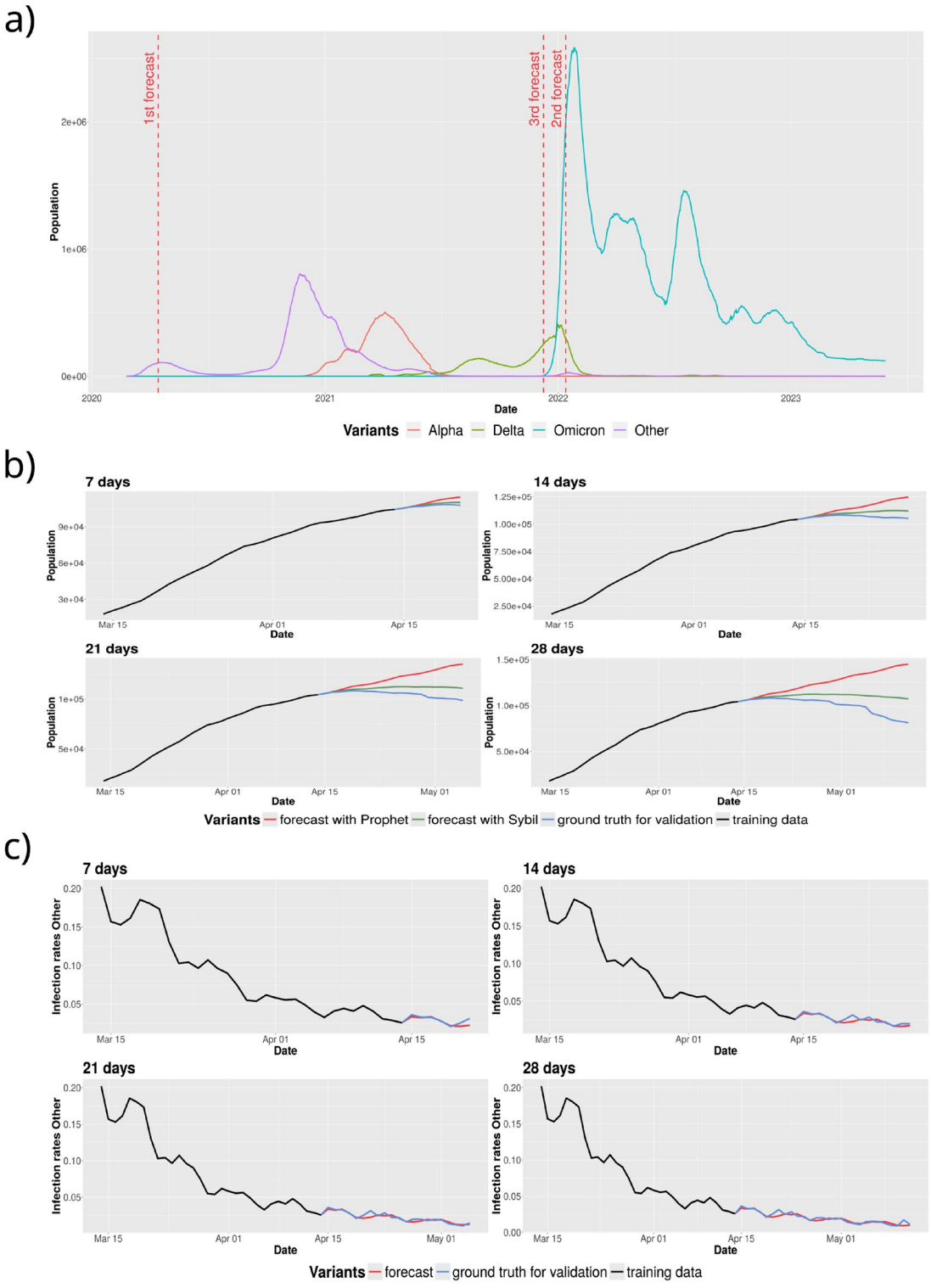

Figure 2. (**a**) Daily active cases in Italy from February 2020 to May 2023 for the four main SARS-CoV-2 strains. Vertical dashed lines mark the selected dates to test Sybil's forecasting. (**b**) Evolution of infections in the first scenario in Italy using Sybil (green line) and Prophet (red line) adopting the same period as training data (black line) comparing and contrasting the predictions against the surveillance data for the period spanning the forecasting window (blue line). We forecast starting from April 14th, 2020. The RMSEs between the ground-truth and the Sybil forecast are: 1381 (1 week), 4141 (2 weeks), 7149 (3 weeks) and 13197 (4 weeks). The RMSEs between the ground-truth and the Prophet forecast are: 3893 (1 week), 10851 (2 weeks), 19615 (3 weeks) and 32593 (4 weeks). (**c**) Sybil applied on the infection rate $\beta_v(\tilde{t})$ of the *Other* variant in the first scenario in Italy (the red line shows the prediction, while black and blue lines represent the training data and the ground-truth values extracted from the surveillance data, respectively).

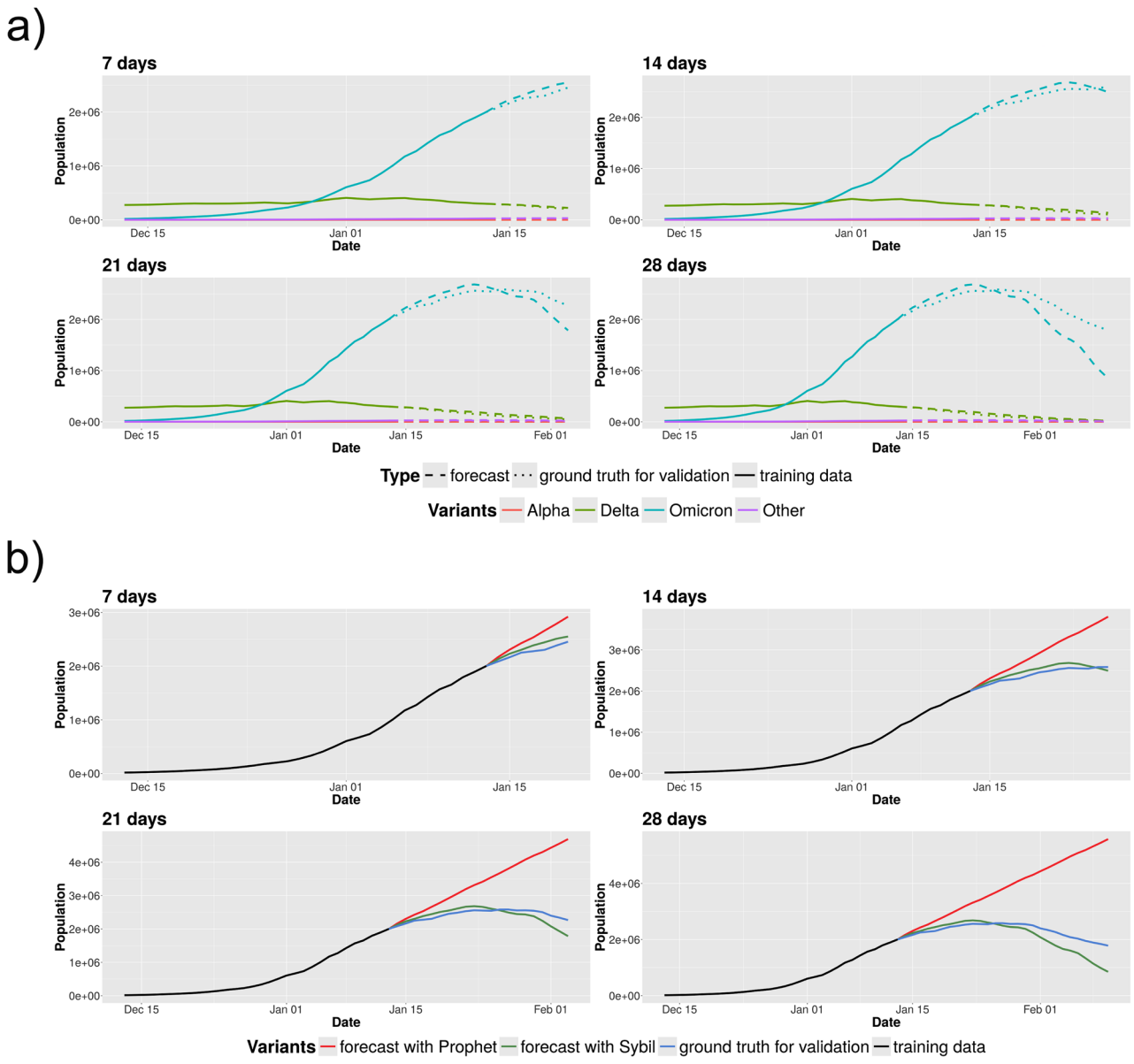

Figure 3. The figures show (**a**) the evolution of infections using Sybil in the second scenario in Italy in which we forecast starting from January 13th, 2022 (the dashed line shows the prediction, while solid and dotted lines represent the training data and the ground-truth values extracted from the surveillance data, respectively) and (**b**) the comparison between Sybil (green line) and Prophet (red line) on the number of infections for the Omicron variant using the same period as training data (black line) comparing and contrasting the predictions against the surveillance data for the period spanning the forecasting window (blue line). The RMSEs between the ground-truth and the Sybil forecast are: 96883 (1 week), 99589 (2 weeks), 190196 (3 weeks) and 380341 (4 weeks). The RMSEs between the ground-truth and the Prophet forecast are: 300266 (1 week), 676381 (2 weeks), 1226726 (3 weeks) and 1946285 (4 weeks).

active in the period from February 2020 to May 2023. As in the case of Italy, lineages are grouped into four main virus strains—Alpha, Delta, Omicron, and the *Other* variant.

Figure 5-(b) refers to the first scenario in Austria, where data from January 1st, 2022 to February 1st, 2022 are employed to train the models, and we compare forecasts obtained for February 2022 (starting from February 2nd). During this period, there were two active variants: Omicron and the *Other* variant, as in the second scenario considered for Italy. The initial point chosen for the forecasts is again placed in the surroundings of the highest peak so that correctly predicting the future evolution of the infection is much more challenging. Nonetheless, Sybil shows high accuracy in predicting the future trajectory of the infection; despite a slightly different evolution,

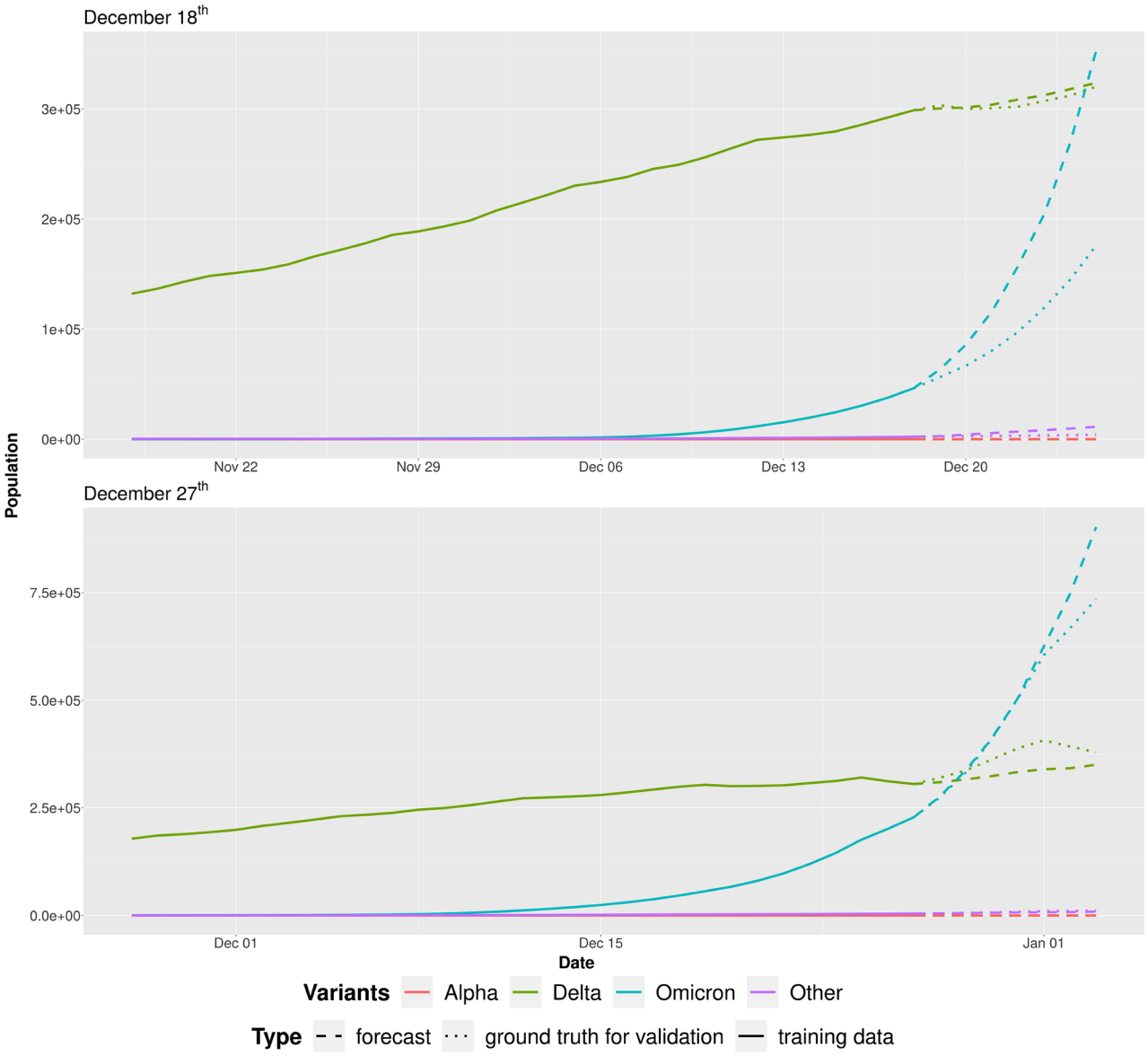

Figure 4. Evolution of infections using Sybil on the third scenario in Italy. In the first plot we forecast starting from December 18th, 2021, while in second one we forecast starting from December 27th, 2021. Both the plots refer to a forecast one week into the future. The dashed line shows the prediction, while solid and dotted lines represent the training data and the ground-truth values extracted from the surveillance data, respectively. In the December 18th, 2021 forecast, the RMSEs between the ground-truth and the Sybil forecast after one week is 89604, while between the ground-truth and the Prophet forecast is 79959. In the December 27th, 2021 forecast, the RMSEs between the ground-truth and the Sybil forecast after one week is 68624, while between the ground-truth and the Prophet forecast is 239562.

Sybil correctly predicts the diffusion slowdown, the successive decreasing phase. Eventually, prediction and surveillance data meet at the end of the four-week-long forecast.

Figure 6-(a,b) refer to the second scenario in Austria, in which we are positioned on the first pandemic peak caused by the *Other* variant. Here, we consider the period from March 3rd, 2020 to April 3rd, 2020 as training data, and we forecast the daily infections for the period April-May 2020 (starting from April 4th). Also in this scenario, Sybil's predictions demonstrate remarkable accuracy in foreseeing the infection's future trajectory, anticipating the deceleration in diffusion and the subsequent decline. Figure 6-(b) shows the comparison between Sybil's and Prophet's predictions.

Finally, Fig. 7-(a,b) refer to the third and last scenario in Austria, in which we are positioned slightly before the Omicron pandemic peak occurred in Summer 2022. Here, we consider the period from June 14th, 2022 to July 14rd, 2022 as training data, and we forecast the daily infections for the period July-August 2022 (starting from July 15th). From these figures, we can see how well Sybil is able to capture the weekly seasonality presents in the data, which is one of Sybil's strengths.

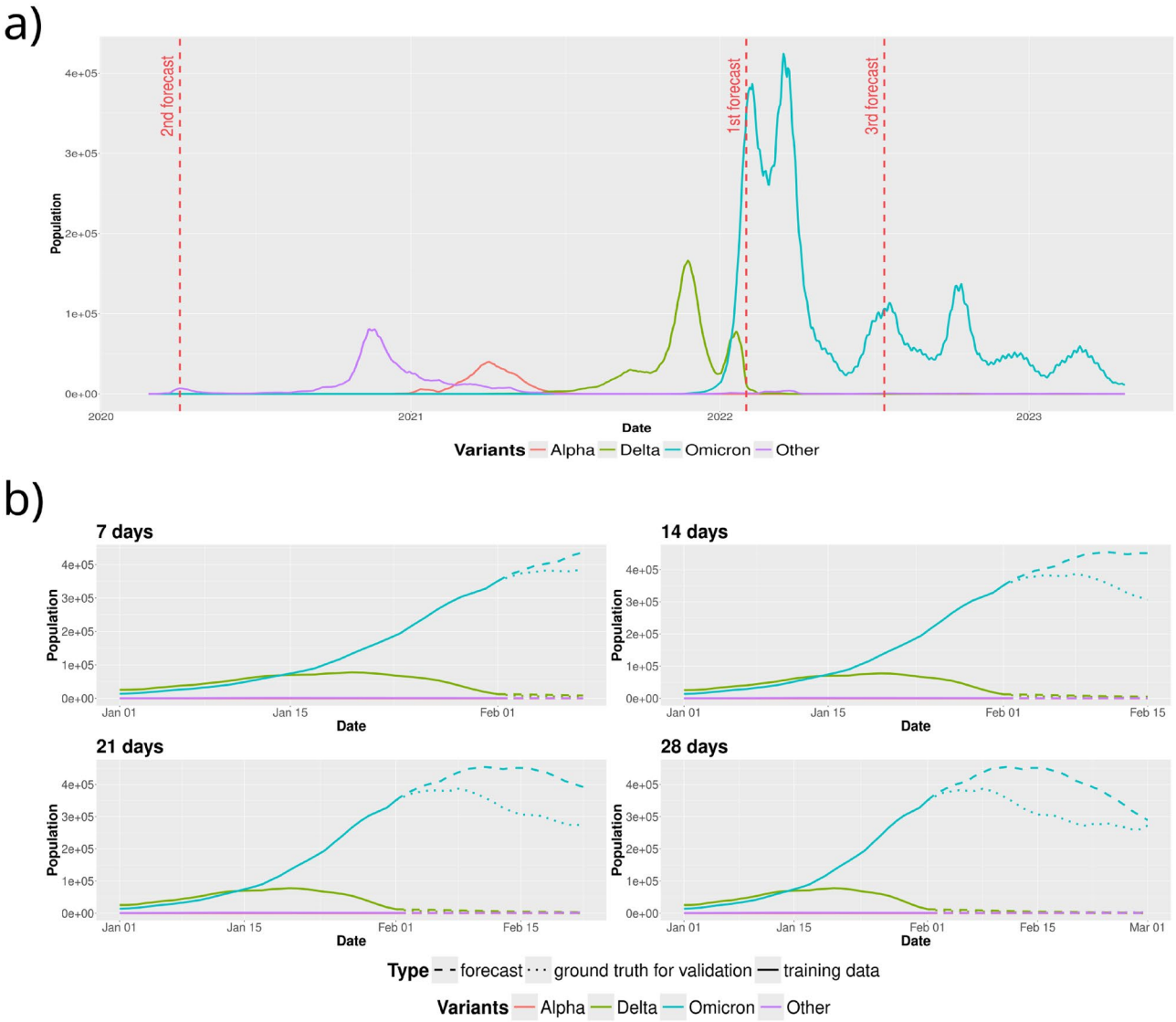

Figure 5. The figures show (**a**) the evolution of infections from February 2020 to May 2023 in Austria with the three considered forecast scenarios and (**b**) the evolution of infections using Sybil on the first considered scenario in Austria in which we forecast starting from February 1st, 2022 (the dashed line shows the prediction, while solid and dotted lines represent the training data and the ground-truth values extracted from the surveillance data, respectively). The RMSEs between the ground-truth and the Sybil forecast are: 30171 (1 week), 80197 (2 weeks), 100774 (3 weeks) and 93937 (4 weeks). The RMSEs between the ground-truth and the Prophet forecast are: 63197 (1 week), 170881 (2 weeks), 278884 (3 weeks) and 373726 (4 weeks).

Discussion

ML approaches have already shown their power when dealing with complexity, thanks to their excellent capabilities in exploiting non-trivial correlations often inaccessible with other tools. Despite the many advantages of employing ML approaches, there are also some caveats worth noting. First, most of these approaches require a tremendous amount of data to learn from, and data available from the surveillance might not be enough for most ML approaches. Sybil faces this issue from two different standpoints; it employs Prophet, a hybrid approach, using ML techniques in combination with simulations—refer to the *Prophet predictive model* subsection for further details. On the other hand, Sybil does not face the challenge of forecasting the virus spread as a single task. First, it tries to predict the rates and then computes the future dynamics using a compartmental model. Therefore, Sybil, combining Prophet and compartmental models, exploits the two approaches in the respective field of application—where they perform the best—and strengthens the individual weaknesses. Specifically, by providing the compartmental model with parameters extracted from the real data or forecasts, there is no need to tune the model and estimate the missing parameters. Estimating model's parameters is a resource-demanding and time-consuming task. Also, it is an activity tightly related to the specific situation or scenario under evaluation. This makes the model almost impossible to apply in a different setting without estimating the parameters again. Hence, Prophet makes Sybil's approach easily deployable in new scenarios without requiring additional tasks,

Scenario	Approach	1 week Mean (± std)	2 weeks Mean (± std)	3 weeks Mean (± std)	4 weeks Mean (± std)
Italy					
1st	Sybil	*1381* (± 747)	**4141** (± 2425)	**7149** (± 4120)	**13197** (± 8408)
	Prophet	3893 (± 2361)	10851 (± 6476)	19615 (± 11653)	32593 (± 20125)
	ARIMA	**778** (± 722)	*4144* (± 3253)	*9736* (± 6885)	*19717*(± 13896)
	SARIMA	**778** (± 722)	*4144* (± 3253)	*9736* (± 6885)	*19717* (± 13896)
	Neural Prophet	4035 (± 2364)	11111 (± 6553)	19983 (± 11791)	33063 (± 20317)
	LSTM	4623 (± 2106)	11411 (± 6192)	19943 (± 11262)	32955 (± 19890)
	GRU	2114 (± 1302)	*6643* (± 4047)	*12339* (± 7448)	*21848* (± 13922)
	EpiNow2	5112 (± 3364)	106799 (± 990)	104885 (± 3054)	100485 (± 8815)
2nd	Sybil	*96883* (± 34583)	**99589** (± 63486)	**190196** (± 186429)	**380341** (± 326962)
	Prophet	300266 (± 133638)	676381 (± 356182)	1226726 (± 703819)	1946285 (± 1158953)
	ARIMA	196298 (± 106323)	502508 (± 290262)	984971(± 602757)	1636012 (± 1022713)
	SARIMA	196298 (± 106323)	502508 (± 290262)	984971(± 602757)	1636012 (± 1022713)
	Neural Prophet	299890 (± 130724)	675966 (± 354417)	1225983 (± 702251)	1945198 (± 1157424)
	LSTM	857350 (± 447682)	–	–	–
	GRU	259534 (± 134814)	768775 (± 470723)	2362125 (± 1714411)	–
	EpiNow2	**64997** (± 63032)	2417232 (± 161043)	2429904 (± 147212)	2322897 (± 245505)
3rd (Dec 18th, 2021)	Sybil	89604 (± 55154)	–	–	–
	Prophet	79959 (± 35214)	–	–	–
	ARIMA	30728 (± 21829)	–	–	–
	SARIMA	30728 (± 21829)	–	–	–
	Neural Prophet	54956 (± 29974)	–	–	–
	LSTM	217615 (± 174845)	–	–	–
	GRU	**7030** (± 3204)	–	–	–
	EpiNow2	58149 (± 41276)	–	–	–
3rd (Dec 27th, 2021)	Sybil	**68624** (± 57142)	–	–	–
	Prophet	239562 (± 117411)	–	–	–
	ARIMA	200101 (± 107520)	–	–	–
	SARIMA	200101 (± 107520)	–	–	–
	Neural Prophet	233592 (± 114532)	–	–	–
	LSTM	316040 (± 270008)	–	–	–
	GRU	124546 (± 61369)	–	–	–
	EpiNow2	213874 (± 114639)	–	–	–
Austria					
1st	Sybil	*30171* (± 15880)	**80197** (± 45647)	**100774** (± 48972)	**93937** (± 45837)
	Prophet	63197 (± 31025)	170881 (± 98196)	278844 (± 155191)	373726 (± 199555)
	ARIMA	75981 (± 39985)	197430 (± 113176)	319520 (± 177087)	428617 (± 228470)
	SARIMA	75981 (± 39985)	197430 (± 113176)	319520 (± 177087)	428617 (± 228470)
	Neural Prophet	62845 (± 30308)	170465 (± 97719)	278293 (± 154718)	373008 (± 199027)
	LSTM	130610 (± 69354)	568862 (± 394725)	–	–
	GRU	57175 (± 32272)	188224 (± 118717)	481200 (± 334214)	–
	EpiNow2	**15423** (± 4989)	361539 (± 25603)	338909 (± 40724)	323021 (± 45713)
2nd	Sybil	**581**(± 244)	**518** (± 267)	**733** (± 733)	**1101** (± 998)
	Prophet	2900 (± 1268)	5698 (± 2757)	8549 (± 4211)	11120 (± 5404)
	ARIMA	1279 (± 654)	2845 (± 1522)	4455 (± 2353)	5785 (± 2929)
	SARIMA	1279 (± 654)	2845 (± 1522)	4455 (± 2353)	5785 (± 2929)
	Neural Prophet	2922 (± 1188)	5667 (± 2672)	8460 (± 4105)	10975 (± 5273)
	LSTM	4950 (± 2260)	18709 (± 12706)	268471 (± 235953)	5911799 (± 5422730)
	GRU	2503 (± 1111)	5289 (± 2702)	10039 (± 5915)	25807 (± 18965)
	EpiNow2	8706 (± 3973)	5752 (± 917)	5039 (± 1441)	4492 (± 1716)
3rd	Sybil	*4383* (± 2052)	**5237** (± 4589)	**10463** (± 9874)	**12605** (± 10515)
	Prophet	*3830* (± 969)	7230 (± 7104)	*21879* (± 17776)	33265 (± 23837)
	ARIMA	8689 (± 2818)	23719 (± 13307)	46740 (± 28039)	66184 (± 38009)
	SARIMA	8689 (± 2818)	23719 (± 13307)	46740 (± 28039)	66184 (± 38009)
	Neural Prophet	*4027* (± 1149)	7145 (± 7055)	*21636* (± 17674)	32931 (± 23698)

Continued

Scenario	Approach	1 week	2 weeks	3 weeks	4 weeks
		Mean (± std)	Mean (± std)	Mean (± std)	Mean (± std)
	LSTM	*3245* (± 2870)	7702 (± 6778)	*22690* (± 17181)	33687 (± 22817)
	GRU	*4434* (± 3173)	*5931* (± 5847)	*19115* (± 15538)	28769 (± 20578)
	EpiNow2	**2696** (± 2695)	105529 (± 5070)	97197 (± 14127)	89894 (± 18865)

Table 1. Root Mean Squared Error (RMSE) with standard deviation between the ground truth used for validation and the obtained forecast with Sybil and the plain use of different state-of-the-art approaches—i.e., Prophet[5], ARIMA / SARIMA[16], Neural Prophet[34], LSTM[35] and GRU[36] neural networks, and EpiNow2[37]—for the considered scenarios in Italy and Austria. For each scenario and forecasting window (one-, two- three- and four-weeks), values in bold represent minimum values, while values in italic represent values close to minimum values. For the third scenario in Italy we do not report the errors for the two-, three- and four-weeks forecast because they are too high—same for LSTM and GRU in the second scenario for two-, three- and four-weeks forecast and four-weeks forecast, respectively, and in the first scenario for three- and four-weeks forecast and four-weeks forecast in Austria, respectively. *Significant values are in bold and italic.

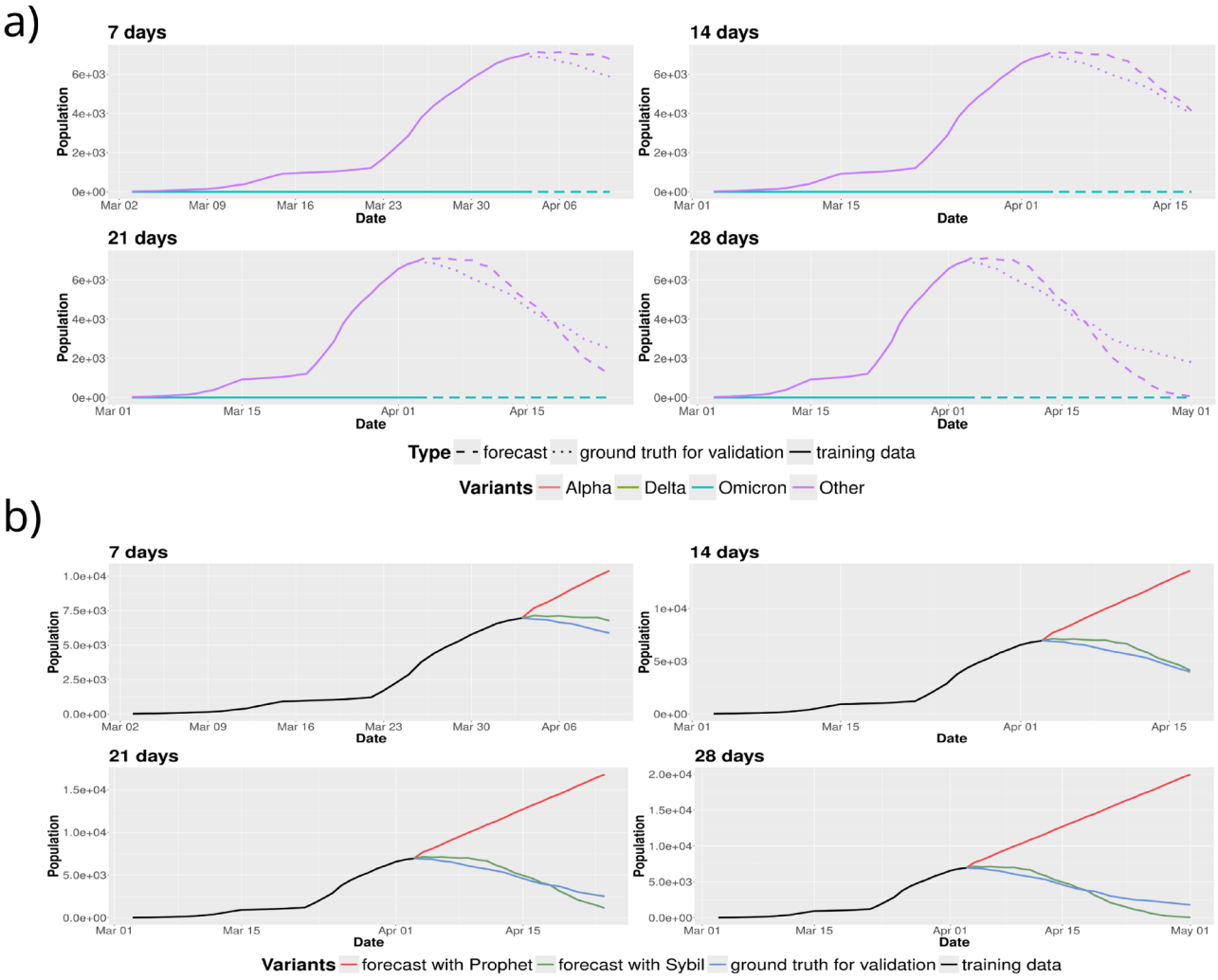

Figure 6. The figures show (**a**) the evolution of the $I_{Other}(\tilde{t})$ compartment using Sybil on the second scenario in Austria in which we forecast starting from April 3rd, 2020 (the dashed line shows the prediction, while solid and dotted lines represent the training data and the ground-truth values extracted from the surveillance data, respectively) and (**b**) the comparison between Sybil (green line) and Prophet (red line) on the number of infections for the *Other* variant using the same period as training data (black line) comparing and contrasting the predictions against the surveillance data for the period spanning the forecasting window (blue line). The RMSEs between the ground-truth and the Sybil forecast are: 581 (1 week), 518 (2 weeks), 733 (3 weeks) and 1101 (4 weeks). The RMSEs between the ground-truth and the Prophet forecast are: 2900 (1 week), 5698 (2 weeks), 8549 (3 weeks) and 11120 (4 weeks).

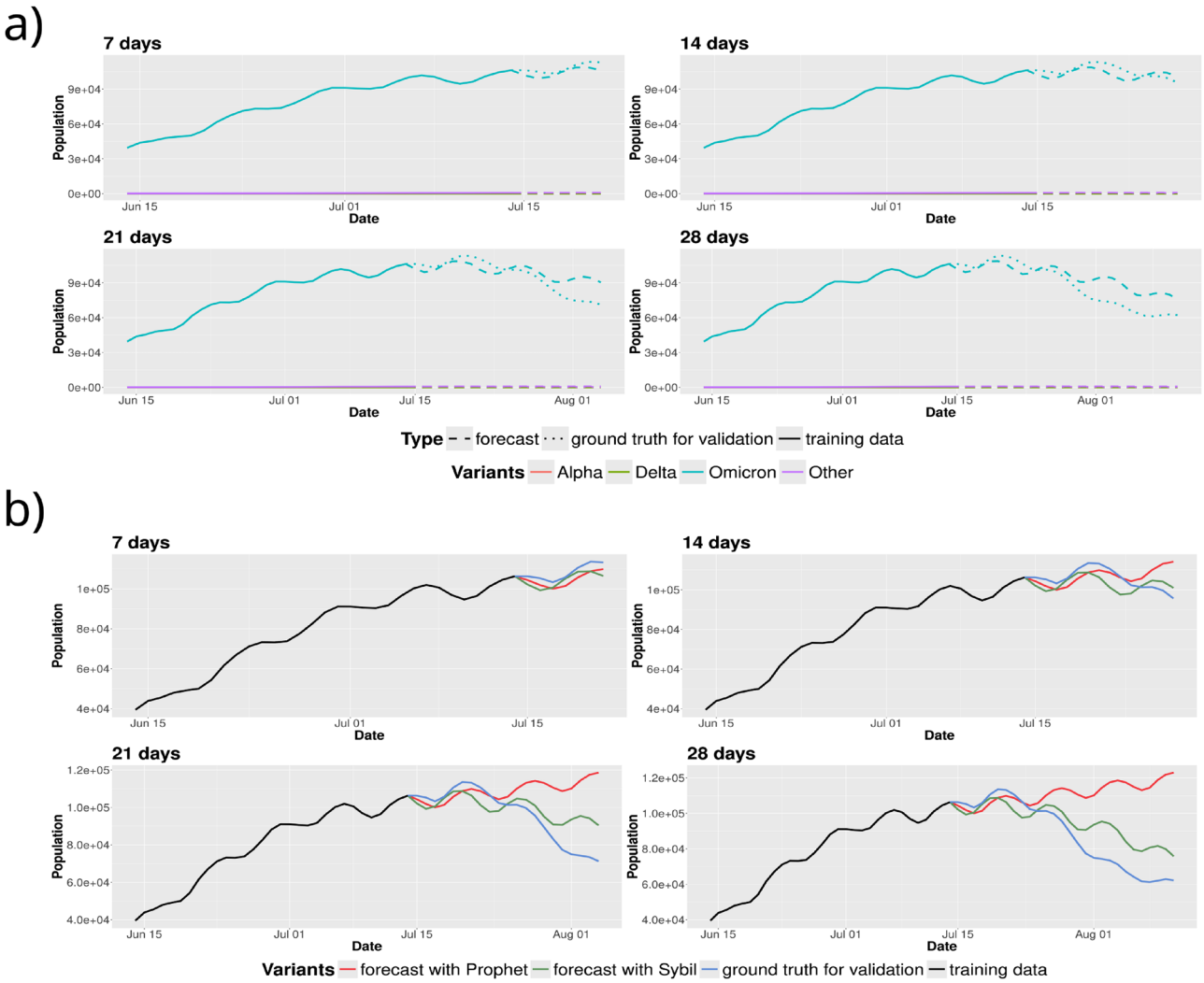

Figure 7. The figures show (**a**) the evolution of the $I_{Omicron}(\tilde{t})$ compartment using Sybil on the third scenario in Austria in which we forecast starting from July 14th, 2022 (the dashed line shows the prediction, while solid and dotted lines represent the training data and the ground-truth values extracted from the surveillance data, respectively) and (**b**) the comparison between Sybil (green line) and Prophet (red line) on the number of infections for the Omicron variant using the same period as training data (black line) comparing and contrasting the predictions against the surveillance data for the period spanning the forecasting window (blue line). The RMSEs between the ground-truth and the Sybil forecast are: 4383 (1 week), 5237 (2 weeks), 10463 (3 weeks) and 12605 (4 weeks). The RMSEs between the ground-truth and the Prophet forecast are: 3830 (1 week), 7230 (2 weeks), 21879 (3 weeks) and 33265 (4 weeks).

but the data must comply with the daily requirement. Likewise, the use of compartmental models in predicting the future trend of the infection makes results more straightforward to explain, for instance, to policymakers, who often do not trust predictions without a robust interpretation.

In the *Results* section and in the 'Supplementary Material', we presented Sybil's forecasts of different lengths for several European states and different periods spanning from February 2020 to May 2023. To showcase Sybil's capabilities, all the forecasting periods cover important changes in the first derivative of the number of daily infections. Then, Sybil's predictions are contrasted against the surveillance data and the plain application of Prophet. Results shows the superiority of Sybil's approach on the plain forecast of the number of daily infections; in particular, Sybil outperforms the plain application of Prophet when primary changes in the virus diffusion happens. For instance, Fig. 3-(b) clearly shows the added value of Sybil approach: it predicts with great accuracy the peak of the current wave while Prophet alone fails to predict the decreasing phase of the infection peak. Further experiments and scenarios are available in the 'Supplementary Material'. Here, for instance, some experiments made choosing forecasting periods far from peaks show that similar and accurate forecasting can be obtained using both approaches and new variants rising or new outbreaks can be predicted accurately using Sybil to set up a continuous monitoring system.

Furthermore, we can assert that Sybil almost always outperforms the other considered state-of-the-art approaches—namely, Prophet[5], ARIMA / SARIMA[16], Neural Prophet[34], LSTM[35], and GRU[36] neural networks,

EpiNow2[37], as well as the framework of Watson et al.[32]—particularly for the two-, three-, and four-weeks forecast periods. In contrast, for the one-week forecast, Sybil provides predictions that are closely aligned with those of the best-performing approach, as detailed in the 'Supplementary Material'. Additionally, when there are minor changes or no changes in the trajectory, Sybil's performance remains robust, effectively matching the performance of the other methods considered.

The presented methodology is very close to being a continuous monitoring system, but strongly depends on the availability of data. In particular, in[39,40] there are few nations with a complete time series (with a daily step) for the data we used (i.e., cumulative number of cases, cumulative number of deceases, cumulative number of recoveries). For this reason, we are working on possible extensions to this methodology. These includes using a fixed recovery rate—same or different for each variant—to dispense with data on recoveries which are often not available—in the 'Supplementary Material' we show some results related to this extension—and adding a pre-processing step using a technique (e.g., splines) to fill missing data and to move from a weekly to a daily step in case of availability of data with a weekly step—without alter too much the information present in the time series. Finally, here we focus on country-level predictions (and state-level predictions for the USA), but Sybil can also be used at regional and city levels.

Conclusion

The global COVID-19 pandemic has brought to light the urgent necessity for sophisticated tools capable of monitoring and predicting the trajectory of infections within the population. This paper introduces a cutting-edge framework designed for continuous monitoring and forecasting, seamlessly integrating machine learning-based predictive models with compartmental models.

Sybil distinguishes itself by delivering forecasts that are not only reliable and replicable but also readily explainable, as evidenced by thorough experimental validation. The adaptability of this innovative approach is clearly demonstrated through its successful application to diverse surveillance data from several European countries—specifically Italy, Austria, Belgium, France, Germany and Sweden—and some U.S. states. In addition, the ability to provide accurate predictions even in the presence of a peak or a new emerging outbreak and the integration of variants are of vital help to policy-makers. Finally, by utilizing only data from the previous month, Sybil significantly reduces the demand for training data, enhancing computational efficiency and enabling effective operation with limited data.

By integrating data-centric and analytic approaches, Sybil effectively addresses the inherent limitations of each method. This amalgamation not only enhances the tool's ability to forecast the evolution of COVID-19 but also positions Sybil as a versatile instrument for predicting the trajectory of various other diseases, thus broadening its scope and impact in the field of infectious disease modeling by helping policy-makers in quickly reacting to a new emerging disease.

Data availibility

COVID-19 data used in this study are available in the COVID19 R library[39,40] and in the ECDC and CDC variants data[41–43]. The framework developed for the analysis presented in this work is available at https://github.com/daniele-baccega/sybil-forecasting. To reproduce the results, see 'Supplementary Material'.

Received: 20 March 2024; Accepted: 7 August 2024
Published online: 19 August 2024

References

1. Gnanvi, J. E., Salako, K. V., Kotanmi, G. B. & Kakaï, R. G. On the reliability of predictions on Covid-19 dynamics: A systematic and critical review of modelling techniques. *Infect. Dis. Model.* **6**, 258–272. https://doi.org/10.1016/j.idm.2020.12.008 (2021).
2. Kamalov, F., Rajab, K., Cherukuri, A. K., Elnagar, A. & Safaraliev, M. Deep learning for Covid-19 forecasting: State-of-the-art review. *Neurocomputing* **511**, 142–154. https://doi.org/10.1016/j.neucom.2022.09.005 (2022).
3. Alabdulrazzaq, H. *et al.* On the accuracy of ARIMA based prediction of COVID-19 spread. *Results Phys.* **27**, 104509. https://doi.org/10.1016/j.rinp.2021.104509 (2021).
4. Zhao, D., Zhang, R., Zhang, H. & He, S. Prediction of global omicron pandemic using ARIMA, MLR, and Prophet models. *Sci. Rep.* **12**, 18138. https://doi.org/10.1038/s41598-022-23154-4 (2022).
5. Taylor, S. J. & Letham, B. Forecasting at scale. *Am. Stat.***72**, 37–45. https://peerj.com/preprints/3190v2/ (2018).
6. Rendana, M. & Idris, W. M. R. New COVID-19 variant (B. 1.1. 7): forecasting the occasion of virus and the related meteorological factors. *J. Infect. Public Health***14**, 1320–1327. https://doi.org/10.1016/j.jiph.2021.05.019 (2021).
7. Battineni, G., Chintalapudi, N. & Amenta, F. Forecasting of COVID-19 epidemic size in four high hitting nations (USA, Brazil, India and Russia) by Fb-Prophet machine learning model. *Appl. Comput. Inf.*https://doi.org/10.1108/ACI-09-2020-0059 *(2020).*
8. Wang, P., Zheng, X., Li, J. & Zhu, B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons Fractals* **139**, 110058. https://doi.org/10.1016/j.chaos.2020.110058 (2020).
9. Sardar, I., Akbar, M. A., Leiva, V., Alsanad, A. & Mishra, P. Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: Methodology, evaluation, and case study in SAARC countries. *Stoch. Env. Res. Risk Assess.* **37**, 345–359. https://doi.org/10.1007/s00477-022-02307-x (2023).
10. Gupta, A. K., Singh, V., Mathur, P. & Travieso-Gonzalez, C. M. Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario. *J. Interdiscip. Math.* **24**, 89–108. https://doi.org/10.1080/09720502.2020.1833458 (2021).
11. Sah, S. *et al.* Forecasting COVID-19 pandemic using Prophet, ARIMA, and hybrid stacked LSTM-GRU models in India. *Comput. Math. Methods Med.***2022**. https://doi.org/10.1155/2022/1556025 (2022).
12. Shahid, F., Zameer, A. & Muneeb, M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons Fractals* **140**, 110212. https://doi.org/10.1016/j.chaos.2020.110212 (2020).
13. ArunKumar, K., Kalaga, D. V., Mohan Sai Kumar, C., Kawaji, M. & Brenza, T. M. Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive

Integrated moving average (SARIMA) for forecasting COVID-19 trends. *Alex. Eng. J.***61**, 7585–7603. https://doi.org/10.1016/j.aej.2022.01.011 (2022).
14. Arora, P., Kumar, H. & Panigrahi, B. K. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons Fractals* **139**, 110017. https://doi.org/10.1016/j.chaos.2020.110017 (2020).
15. Nabi, K. N., Tahmid, M. T., Rafi, A., Kader, M. E. & Haider, M. A. Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks. *Results Phys.* **24**, 104137. https://doi.org/10.1016/j.rinp.2021.104137 (2021).
16. Brockwell, P. J. & Davis, R. A. (eds) *State-Space Models,* 259–316 (Springer, New York, New York, NY, 2002).
17. Caldwell, J. M. *et al.* Vaccines and variants: Modelling insights into emerging issues in COVID-19 epidemiology. *Paediatr. Respir. Rev.* **39**, 32–39. https://doi.org/10.1016/j.prrv.2021.07.002 (2021).
18. Miller, J. K., Elenberg, K. & Dubrawski, A. Forecasting emergence of COVID-19 variants of concern. *PLoS ONE* **17**, e0264198. https://doi.org/10.1371/journal.pone.0264198 (2022).
19. Gostic, K. M. *et al.* Practical considerations for measuring the effective reproductive number. *Rt. PLoS Computat. Biol.* **16**, e1008409. https://doi.org/10.1371/journal.pcbi.1008409 (2020).
20. Fošnarič, M., Kamenšek, T., Žganec Gros, J. & Žibert, J. Extended compartmental model for modeling COVID-19 epidemic in Slovenia. *Sci. Rep.* **12**, 16916. https://doi.org/10.1038/s41598-022-21612-7 (2022).
21. Biala, T. A. & Khaliq, A. A fractional-order compartmental model for the spread of the COVID-19 pandemic. *Commun. Nonlinear Sci. Numer. Simul.* **98**, 105764. https://doi.org/10.1016/j.cnsns.2021.105764 (2021).
22. Nadim, S. S., Ghosh, I. & Chattopadhyay, J. Short-term predictions and prevention strategies for COVID-19: a model-based study. *Appl. Math. Comput.* **404**, 126251. https://doi.org/10.1016/j.amc.2021.126251 (2021).
23. Ramezani, S. B., Amirlatifi, A. & Rahimi, S. A novel compartmental model to capture the nonlinear trend of COVID-19. *Comput. Biol. Med.* **134**, 104421. https://doi.org/10.1016/j.compbiomed.2021.104421 (2021).
24. Zhang, P. *et al.* Usage of Compartmental Models in Predicting COVID-19 Outbreaks. *AAPS J.* **24**, 98. https://doi.org/10.1208/s12248-022-00743-9 (2022).
25. Mandal, M. *et al.* A model based study on the dynamics of COVID-19: Prediction and control. *Chaos, Solitons Fractals* **136**, 109889. https://doi.org/10.1016/j.chaos.2020.109889 (2020).
26. He, S. *et al.* A discrete stochastic model of the COVID-19 outbreak: Forecast and control. *Math. Biosci. Eng.* **17**, 2792–2804. https://doi.org/10.3934/mbe.2020153 (2020).
27. Kucharski, A. J. *et al.* Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet. Infect. Dis* **20**, 553–558. https://doi.org/10.1016/S1473-3099(20)30144-4 (2020).
28. Brauer, F. Compartmental models in epidemiology. *Math. Epidemiol.* 19–79. https://doi.org/10.1007/978-3-540-78911-6_2 (2008).
29. Tolles, J. & Luong, T. Modeling epidemics with compartmental models. *JAMA* **323**, 2515–2516. https://doi.org/10.1001/jama.2020.8420 (2020).
30. Kiarie, J., Mwalili, S. & Mbogo, R. Forecasting the spread of the COVID-19 pandemic in Kenya using SEIR and ARIMA models. *Infect. Dis. Model.* **7**, 179–188. https://doi.org/10.1016/j.idm.2022.05.001 (2022).
31. Deng, Q. Dynamics and development of the COVID-19 epidemic in the United States: a compartmental model enhanced with deep learning techniques. *J. Med. Internet Res.* **22**, e21173. https://doi.org/10.2196/21173 (2020).
32. Watson, G. L. *et al.* Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. *PLoS Comput. Biol.* **17**, e1008837. https://doi.org/10.1371/journal.pcbi.1008837 (2021).
33. Rahimi, I., Chen, F. & Gandomi, A. H. A review on COVID-19 forecasting models. *Neural Comput. Appl.* 1–11. https://doi.org/10.1007/s00521-020-05626-8 (2021).
34. Triebe, O. *et al.* Neuralprophet: Explainable forecasting at scale.https://doi.org/10.48550/arXiv.2111.15397 (2021).
35. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (1997).
36. Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078https://doi.org/10.48550/arXiv.1406.1078 (2014).
37. Sam Abbott *et al. EpiNow2: Estimate real-time case counts and time-varying epidemiological parameters*, https://doi.org/10.5281/zenodo.3957489 (2020).
38. West, J., Everden, S. & Nikitas, N. A case of COVID-19 reinfection in the UK. *Clin. Med.* **21**, e52. https://doi.org/10.7861/clinmed.2020-0912 (2021).
39. Guidotti, E. & Ardia, D. COVID-19 Data Hub. *J. Open Sour. Softw.***5**, 2376, https://doi.org/10.21105/joss.02376 (2020).
40. Guidotti, E. A worldwide epidemiological database for COVID-19 at fine-grained spatial resolution. *Sci. Data* **9**, 112. https://doi.org/10.1038/s41597-022-01245-1 (2022).
41. Data on SARS-CoV-2 variants in the EU/EEA. European Centre for Disease Prevention and Control (ECDC).
42. Khare, S. *et al.* GISAID's role in pandemic response. *China CDC Weekly***3**, 1049. https://doi.org/10.46234/ccdcw2021.255 (2021).
43. Centers for Disease Control and Prevention. COVID Data Tracker. Atlanta, GA: U.S. Department of Health and Human Services, CDC (2024). https://covid.cdc.gov/covid-data-tracker. Accessed: 2024-06-13.
44. Hodcroft, E. B. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. https://covariants.org/ (2021).
45. Byrne, A. W. *et al.* Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open* **10**, e039856. https://doi.org/10.1136/bmjopen-2020-039856 (2020).

Acknowledgements

D.B. is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, health and life sciences course organized by Università Campus Bio-Medico di Roma.

Author contributions

D.B. discussed, developed and tested the framework, conceived and conducted the experiments. P.C., A.F.A. and M.S. discussed and designed the framework, conceived the experiments. All authors reviewed the manuscript.

Funding

This work was supported by grants from "Ripresa delle attività socio-economiche e delle scuole: modelli per la progettazione e supporto di linee guida per la convivenza con il Covid-19" (Cod. ROL 73459, 2020, PI Matteo Sereno), project funded by CRT foundation, and from TED2021-131264B-I00 (SocialProbing) funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-69660-5.

Correspondence and requests for materials should be addressed to D.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.