

Machine Learning for Infectious Disease Risk Prediction: A Survey

Mutong LIU ^a, Yang LIU *^a, and Jiming LIU ^a

^aDepartment of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

Abstract

Infectious diseases, either emerging or long-lasting, place numerous people at risk and bring heavy public health burdens worldwide. In the process against infectious diseases, predicting the epidemic risk by modeling the disease transmission plays an essential role in assisting with preventing and controlling disease transmission in a more effective way. In this paper, we systematically describe how machine learning can play an essential role in quantitatively characterizing disease transmission patterns and accurately predicting infectious disease risks. First, we introduce the background and motivation of using machine learning for infectious disease risk prediction. Next, we describe the development and components of various machine learning models for infectious disease risk prediction. Specifically, existing models fall into three categories: Statistical prediction, data-driven machine learning, and epidemiology-inspired machine learning. Subsequently, we discuss challenges encountered when dealing with model inputs, designing task-oriented objectives, and conducting performance evaluation. Finally, we conclude with a discussion of open questions and future directions.

→ also predict infectious dis. w/ a ut.

Keywords: Machine Learning, Data-driven Modeling, Epidemiology-inspired Learning, Infectious Disease Risk Prediction, Transmission Dynamics Characterization

*Corresponding author: csygliu@comp.hkbu.edu.hk

Contents

1	Introduction	3
2	Machine learning for infectious disease risk prediction	5
2.1	Problem statement	5
2.2	Statistical prediction	6
2.3	Data-driven machine learning	7
2.3.1	Traditional machine learning models	7
2.3.2	Deep learning models	10
2.3.3	Discussions: advantages and limitations	15
2.4	Epidemiology-inspired machine learning	15
2.4.1	Introduction to epidemiological models	17
2.4.2	Epidemiological parameterization	18
2.4.3	Epidemiology-embedded learning	21
2.4.4	Discussions: advantages and limitations	23
3	Challenges	23
3.1	Data challenges	23
3.1.1	Heterogeneous risk-related data	24
3.1.2	Multiple-resolution/granularity	27
3.1.3	Data latency	28
3.2	Task challenges	28
3.2.1	Characterizing spatiotemporal patterns	28
3.2.2	Distribution shift	30
3.2.3	Uncertainty quantification	30
3.2.4	Model interpretability	32
3.3	Evaluation challenges	32
3.3.1	Measurements of disease severity	32
3.3.2	Evaluation of model performance	33
4	Conclusions and future directions	34

1 Introduction

The propagation of infectious diseases, whether emergent (e.g., coronavirus disease 2019 (COVID-19), which is responsible for the ongoing pandemic and has caused nearly 7 million deaths worldwide so far ¹) or long-standing (e.g., malaria, which has an ancient history and still causes more than 600 thousand deaths every year [1]), significantly affects human well-being and social development on a global scale [1, 2]. Thus, the battle against infectious disease is never-ending. Humankind's development of countermeasures to various diseases has been based on conceptual innovation together with scientific development in multiple disciplines, from the use of vaccination to eradicate smallpox, a high-mortality disease, to the use of combinations of multiple interventions to contain the transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes COVID-19.

In recent decades, machine learning has been widely and successfully applied in many fields, e.g., natural language processing and computer vision, to perform various tasks, e.g., regression and classification. Inevitably, due to the ability of machine learning to deal with large and heterogeneous data and capture complex patterns, it has also been employed in infectious disease research [3, 4]. However, the focus of coping with disease risks varies across the different stages of infectious disease progression, due to the different goals of public health strategies (e.g., the prevention, mitigation, or containment of disease transmission), and thus the goal of disease modeling for prediction of disease risk also varies across these stages. With reference to [5], the developmental process of infectious disease risk can be divided into three phases: the watch phase, the warning phase, and the emergency phase. In the watch phase, an infectious disease has not yet occurred in humans but possibly exists in the environment surrounding the areas where humans live. Thus, in this phase, data-driven modeling is used to investigate hosts (e.g., wild animals) that may carry pathogens (e.g., microorganisms that could cause infectious diseases, such as certain species of bacteria, viruses, protozoa, fungi, and prions), as this is essential to prevent transmission of pathogens from hosts to humans and thereby prevent outbreaks in human populations. If these pathogens come into contact with humans under specific conditions, human infection can occur; thus, once human cases have been verified, the level of infectious disease risk is upgraded to the warning phase. In this phase, it is crucial for disease transmission to be understood to allow public health agencies to appropriately respond. The importance of applying phylogenetic and phylodynamic modeling techniques in this phase was highlighted, as these techniques can help us to understand the properties and potential of disease transmission by offering informative predictions in situations in which spatiotemporal disease transmission data are scarce [6]. If an epidemic is not efficiently brought under control or the efforts of interventions are overwhelmed due to the rapid spread of a disease, the epidemic can spread over such a large geographical range that it becomes a pandemic, leading to high morbidity and mortality, such as the COVID-19 pandemic. In such situations, the level of infectious risk enters the emergency phase. As a result, data-informed modeling and prediction of disease risk and severity (such as infection size, scope, and duration), which are recognized as the main focus in the influenza season Challenges held by the Centers for Disease Control (CDC) in the United States [7], and exploration of the effect of available intervention strategies [8], are more urgent in the emergency phase than in previous phases, as such investigations are needed to inform decision-makers on how to take action to achieve the goal of reducing dam-

¹<https://covid19.who.int/>. Accessed April 29, 2023

age to human health.

In this paper, we focus on machine learning approaches for infectious disease risk prediction in the emergency phase. Disease risk prediction can provide insights and quantitative information for decision-making processes on how to contain disease transmission and mitigate loss, so it is considered an important aid for the formulation of public health responses [9]. Specifically, accurate prediction of epidemic or pandemic trends can provide advance warning of potential outbreaks and thus allow timely action to be taken, such as the allocation of anti-disease resources to regions with urgent needs or the adjustment of quarantine policies [10] to prevent an outbreak as quickly as possible. Furthermore, retrospective analysis of disease trends based on prediction models can reveal the transmission patterns underlying observations and enable future outbreaks to be dealt with more effectively than the current outbreak [11]. Given that big data technologies and machine learning models have been successfully applied in many fields, increasing numbers of researchers in the fields of machine learning and statistics are considering how to utilize large-scale available data and the capacity of machine learning for data representation and data fitting to accurately predict disease risks. Numerous novel infectious disease risk prediction models have been devised, with a range of goals. Initially, the main focus of most studies was to develop statistical prediction models and data-driven machine learning models by designing various model structures to automatically capture implicit dependencies based on observed data and by minimizing prediction errors. However, when such models are used in practice, accurate prediction based on statistical relationships is not the ultimate goal, and questions continually emerge. For example, how do we know we can trust the predictions of a data-driven model? What information can be provided by predictions? These questions reflect the importance of accurate, informative, and interpretable predictions, as they provide reliable and valid information for disease prevention and control. In recent studies, this has been achieved by integrating prior knowledge of epidemiological models with data-driven models to create epidemiology-inspired machine learning models.

Over the past two decades, several authors have summarized progress in the development of infectious disease models. For instance, Grassly and Fraser [12] examined multiple topics on the linking of mathematical hypotheses and modeling to the process of infectious disease transmission, and summarized studies that have devised mathematical models of infectious disease transmission. In recent years, some surveys have been performed on models for specific infectious diseases, such as malaria [13], dengue [14], influenza [15, 16], and COVID-19 [17, 18]. [13] and [14] have introduced and summarized the development of determinate or stochastic mathematical approaches for modeling malaria and dengue transmission, and the evolution of elaborate hypotheses on these processes. These surveys have covered several areas, such as population-level compartmental models (also known as mass-action compartmental models), structured metapopulation models, and agent-based models, but have not paid much attention to machine learning models, which have undergone continual development over the past decade. [15] and [16] have surveyed studies on mechanism-based models and data-driven models for influenza forecasting. The two most recent published surveys are those of [17] and [18] and these have examined many state-of-the-art deep learning models. For example, [17] sorted mathematical models of COVID-19 into three categories: statistical, mechanistic, and hybrid models. In contrast, [18] sorted computational models of COVID-19 transmission and diagnosis more finely, to give five categories: compartmental, statistical, data-driven, machine learning- and deep learning-based, and mixed models. However, al-

though the categories in [17] and [18] encompass epidemiologically inspired models, these surveys mainly focused on mechanism-based models and statistical models, and therefore did not cover some recently developed epidemiologically inspired machine learning models.

In this paper, we categorize, summarize, and discuss machine learning methods for infectious disease risk prediction. To this end, we first provide an overview of the previously developed machine learning models by sorting them into three categories: (1) statistical prediction models (Section 2.2), (2) data-driven machine learning models (Section 2.3), and (3) epidemiology-inspired machine learning models (Section 2.4). Next, we briefly introduce a selection of related methods in each category to show their innovations, differences, and similarities. Subsequently, we discuss in three sections the three kinds of challenges commonly faced when predicting epidemic risk: (1) data-related challenges (Section 3.1); (2) task-related challenges (Section 3.2); and (3) evaluation-related challenges (Section 3.3). In each section, we cover a series of sub-topics and introduce the techniques that have typically been used to address these challenges.

2 Machine learning for infectious disease risk prediction

As mentioned, we divide current machine learning approaches for infectious disease risk prediction into three categories. The first category is statistical prediction models and contains studies that have mostly treated epidemic prediction as a time-series prediction problem to analyze the statistical characteristics of trends; the second category is data-driven machine learning models and contains many studies that have presumed that there are implicit and unknown disease propagation patterns (such as spatiotemporal transmission networks) that can be captured by various model structures, whose parameters can be inferred from available data; and the third category is epidemiology-inspired machine learning models and comprises studies that have incorporated prior knowledge of epidemiological models with the inferential ability of data-driven machine learning to better characterize disease transmission than previous approaches.

In this section, we provide a brief introduction to the models in each category and then subdivide these categories to show their differences and relationships. The specific structure of the taxonomy we employ to classify the models is given in Fig. 1.

2.1 Problem statement

In the task of predicting infectious disease risks, given observations of disease dynamics-related data \mathbf{x} , the goal of machine learning approaches is to train a model f to accurately predict the future disease dynamics y in one location or multiple locations using historical data:

$$y = f(\mathbf{x}). \quad (1)$$

Usually, \mathbf{x} denotes the model input, and it could be the historical data of the indicator of disease risks, such as the disease case number or disease prevalence. In some circumstances, it could also include other risk-related data, such as climate data, mobility data, and population data, to enhance the prediction performance. y is the model output, i.e., the future risks to be predicted, which is usually the indicator of infectious disease risks. In general, in the temporal dimension, the input and output could cover one time step or multiple time steps;

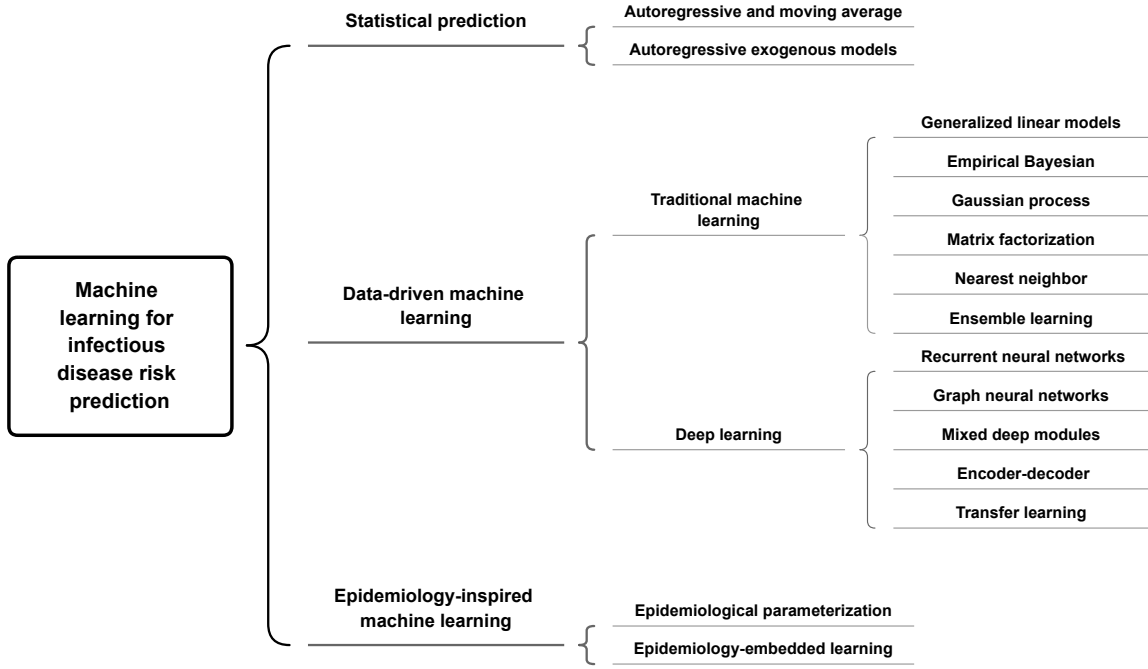


Figure 1: Categories of machine learning models that have been developed for infectious disease risk prediction: (1) Statistical prediction models (Section 2.2); (2) data-driven machine learning models (Section 2.3); and (3) epidemiology-inspired machine learning models (Section 2.4).

in the spatial dimension, the input and output could cover one location or multiple locations. In the following content, we use x to denote the model input, y to denote the model output, and we do not make assumptions on their dimensions. We use f to denote the general format of model's prediction functions; the specific formulations of them depend on the used model structures in different works.

2.2 Statistical prediction

Because epidemic or pandemic data are usually presented in time-series form, the task of infectious disease risk prediction can usually be treated as a time-series prediction problem. Therefore, many statistical models are applied to epidemic or pandemic prediction. Some classic statistical models are based on linear model structures, such as **autoregressive** (AR) models, which linearly combine past observations within a time window p (i.e., $x_{t-1}, x_{t-2}, \dots, x_{t-p}$) with the disturbance term E_t (the AR model equation is shown in Eq. 2 [19]); **moving average** (MA) models, which complement AR models by linearly combining disturbance terms within a time window q (i.e., $E_t, E_{t-1}, \dots, E_{t-q}$) (the MA model equation is shown in Eq. 3 [19]); and their combination, known as autoregressive moving average (ARMA) models (the model equation is shown in Eq. 4 [19]), which characterize the features of single time-series dynamics to forecast future trends [19, 20].

$$y \text{ or } x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \dots + \alpha_p x_{t-p} + E_t \quad (2)$$

$$y \text{ or } x_t = E_t + \beta_1 \cdot E_{t-1} + \dots + \beta_q \cdot E_{t-q} \quad (3)$$

$$y \text{ or } x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + E_t + \beta_1 E_{t-1} + \cdots + \beta_q E_{t-q} \quad (4)$$

However, the above-mentioned models can only be applied to time series that are stationary, and so cannot be applied in many situations, as time series are often non-stationary due to the effects of seasonal factors, persistent interventions, or other determining factors. Therefore, variations of the above-mentioned models have been proposed to cope with these situations. For example, **autoregressive integrated moving average (ARIMA)** models [19] remove obvious trends (such as upward or downward trends)—which are caused by determining factors—by using d -order differencing processes. This affords stationary time series, to which ARMA models can be applied. Seasonal ARIMA models [19] remove the effects of seasonality by performing lagged differencing processes with a period s . **Autoregressive exogenous (ARX)** models also use disease dynamics data to make predictions but also take other risk-related factors into account; these are denoted extra or exogenous variables and formulated as a weighted sum item that is in addition to the original weighted sum item. Wang et al. developed a variation of the standard ARX model that they denoted the **dynamic Poisson autoregressive model with exogenous inputs variables (DPARX)** model, whose parameters dynamically change over time [21]. That is, the DPARX model learns a set of parameters for a prediction at each time step, which results in many parameters to learn. To effectively do so and avoid overfitting, Wang et al. assumed that some models at different time steps—and thus also their parameters—are similar to each other. They provided three types of prior constraints with graphical structures (i.e., fully connected, nearest neighbors connected, and seasonal nearest neighbors connected structures) to depict this kind of structural similarity between models at different time steps.

2.3 Data-driven machine learning

2.3.1 Traditional machine learning models

In addition to a series of statistical autoregressive models that consider time dependency, many machine learning models have been developed that use more flexible model structures to predict disease risk. A summary of the data-driven machine learning models is given in Table 1.

Generalized linear models Some studies have used **generalized linear models (GLMs)** to predict disease risks in a single location or multiple locations. The general formulation of GLMs is given as follows [42]:

$$p(y|\mathbf{x}, \mathbf{w}, \sigma^2) = \exp \left[\frac{y\mathbf{w}^T \mathbf{x} - A(\mathbf{w}^T \mathbf{x})}{\sigma^2} + \log h(y, \sigma^2) \right], \quad (5)$$

where y denotes the response variable to be predicted, \mathbf{x} denotes the input feature vector used to predict y , \mathbf{w} denotes the weighting parameter vector on \mathbf{x} , σ^2 represents the dispersion term, $A(\cdot)$ represents the log normalizer, $h(\cdot, \cdot)$ represents the base measure, and $\exp[\cdot]$ denotes a given probability distribution from the exponential family. Furthermore, there is a mean function g^{-1} that maps $\mathbf{w}^T \mathbf{x}$ to the mean value of the response variable: $\mu = g^{-1}(\mathbf{w}^T \mathbf{x})$ [42]. In different instances, different probability distributions are used to model disease risks. For example, in [22], Zhang et al. used the Poisson distribution to model case numbers as integer values. In [23, 24], Pei et al. used the Gaussian distribution to model the disease case

Table 1: Summary of data-driven machine learning models.

Categories		References	Data-driven components
Traditional machine learning	Generalized linear models	[22]	Poisson regression
		[23, 24]	Gaussian regression for continuous variables, and Bernoulli regression for discrete variables
	Empirical Bayesian	[25]	Empirical Bayes framework
	Gaussian process	[26]	Gaussian process with spatial and temporal kernel for spatiotemporal dependency
		[27]	Gaussian processes for temporal dependency
	Matrix factorization and nearest neighbor	[28]	Matrix factorization based regression using nearest neighbor embedding
	Ensemble learning	[28]	Fusion of multiple models trained by different data sources
		[27]	Fusion of multiple regression models trained by different features of data
Deep learning	RNN	[29]	Hierarchically stacked RNN
		[30]	Sequentially stacked LSTM
		[31]	Independent two LSTM layers and a fusion layer
	GNN	[32]	GNNs with spatial and temporal connections
		[33]	GCN combining with GFT and DFT
	Mixed deep modules	[34]	CNN and RNN for spatial and temporal dependency respectively
		[35]	Dilated convolution and RNN for temporal dependency and GNN with attention matrix for spatial dependency
		[36]	LSTNet/N-Beats and GAT for temporal and spatial dependency respectively
		[37]	Functional neural process
		[38]	GAT for spatial dependency, GRU and dilated CNN for temporal dependency
	Encoder-decoder	[39]	LSTM and IDEC to encode and cluster the temporal dependency respectively
		[40]	Multi-channel CNNs to encode temporal dependency and GRE for encode spatiotemporal dependency
	Transfer learning	[41]	EpiDeep as source model and CAEM to encode spatiotemporal features

numbers as continuous values; they also used the Bernoulli distribution to model the status of getting infected. Based on the distribution selected for modeling the data, the mean function varies accordingly, such as $\mu = \exp(\mathbf{w}^T \mathbf{x})$ for Poisson distribution, $\mu = \mathbf{w}^T \mathbf{x}$ for Gaussian distribution, and $\mu = \frac{1}{1+e^{-\mathbf{w}^T \mathbf{x}}}$ for Bernoulli distribution. Although GLMs have a similar regression framework to generate predictions, their specific structures and the corresponding inference methods in different works are different, as they are elaborately designed based on various assumptions to reflect specific disease transmission processes. For example, Zhang et al. incorporated the effects of intra-regional, inter-regional, and external factors on disease risk into a unified Poisson regression-based framework to model epidemic diffusion between multiple locations [22]. In their approach, information from prior knowledge about disease transmission is taken into consideration to specify what heterogeneous risk-related factors are involved in three aspects of disease transmission: intra-transmission, inter-transmission, and external transmission. Specifically, in the intra-transmission part of their framework, climate data (e.g., temperature, rainfall), geographical data (e.g., elevation), and demographic data (population) are combined to predict self-infections within a region; in the inter-transmission part of their framework, a diffusion matrix with a prior structure constrained by the transportation network within a region is used to describe disease transmission between locations; and in the external-transmission part of their framework, a quadratic function that has uni-modal patterns is employed to model the effect of seasonal imported cases of the disease. Similarly, Pei et al. have utilized a multivariate regression model denoted the **group sparse Bayesian learning model (GSBL)**, which is based on a transmission network, to predict disease dynamics [23, 24]. In contrast to [22], they focused on the sentinel selection problem, which arises due to limited disease surveillance resources. Sentinel selection, which is closely related to the concept of active surveillance in the domain of public health, is the selection of representative locations from all targeted locations at which to conduct disease surveillance. Pei et al. have formulated this problem as a learning process of a row-sparse disease transmission network, in which sentinels are indicated by the non-zero rows [23, 24]. Thus, their model uses the disease data of these sentinels and an inferred network to recover or predict the global dynamics of all target locations. In addition, the model does not use other prior knowledge about disease transmission and only uses historical case number data to infer a transmission network, so it can be more easily applied than the above Poisson regression model to various diseases and other domains.

Empirical Bayesian In contrast to GLMs, which assign a pre-defined prior distribution to the parameters of a model (i.e., a distribution that is irrelevant to the observational data), empirical Bayesian models usually estimate prior distributions from historical observations. A typical example is the semiparametric empirical Bayes framework for epidemic modeling proposed by Brooks et al. [25]. This framework first estimates the prior, i.e., the shape of the influenza-like illness (ILI) curve, the noise, the peak height, the peak week, and the pacing, using a set of uniform distributions over the historical observations. It then generates the underlying ILI curve of the current ILI season by linearly adjusting the piecewise quadratic curves of historical seasons using the current year’s CDC baseline weekly ILI level.

Gaussian process To predict disease risks, **Gaussian process (GP)** models assume that the random variable $f(\mathbf{x}_i)$ in continuous domains (e.g., time or space) follow a Gaussian distribution with the mean $\mu = m(\mathbf{x}_i)$ and the variance σ_i , and that the joint distribution of a finite set

of these variables $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_M)]$ will follow the multivariate Gaussian distribution with the mean $\boldsymbol{\mu} = [m(\mathbf{x}_1), \dots, m(\mathbf{x}_M)]$ and the covariance $\Sigma_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ ($i, j = 1, \dots, M$), where the \mathcal{K} is the kernel function and M is the number of observations [43, 42]. In GP models, given the training set $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$, we have $\mathbf{f}_X \sim \mathcal{N}(\boldsymbol{\mu}_X, \mathbf{K}_{X,X})$, where $\mathbf{K}_{X,X}$ is the $M \times M$ covariance matrix of the data set \mathbf{X} . Given a test set \mathbf{X}_* , the joint distribution $p(\mathbf{f}_X, \mathbf{f}_{X_*} | \mathbf{X}, \mathbf{X}_*)$ is represented as follows:

$$\begin{pmatrix} \mathbf{f}_X \\ \mathbf{f}_{X_*} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_{X_*} \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{X,X} & \mathbf{K}_{X,X_*} \\ \mathbf{K}_{X,X_*}^\top & \mathbf{K}_{X_*,X_*} \end{pmatrix} \right). \quad (6)$$

The covariance of these variables is calculated by choosing the appropriate kernel function $\mathcal{K}(\cdot, \cdot)$ and is used to describe the characteristics of processes. Due to the inherent ability of a covariance matrix to model the similarity between data points, conventional GP models are generally used as interpolation models. However, some recent studies have extended their use by applying them to epidemic prediction tasks. For instance, Senanayake et al. proposed a model based on GP regression that predicts influenza cases by capturing the spatiotemporal dependency of data [26]. They constructed a non-linear kernel with both spatial and temporal components, and spatiotemporal covariance components, to address the challenges associated with the complicated characteristics of disease dynamics, such as temporal characteristics (i.e., periodicity, non-stationarity, and short- and long-term dependency) and spatial characteristics (i.e., the distance between locations and morphology of a region). Zimmer and Yaesoubi proposed a GP-based framework to forecast seasonal epidemics [27]. In contrast to Senanayake et al. [26], they did not design kernels to represent the dependencies between spatial locations but rather focused on exploring the temporal dependency between within-seasonal and between-seasonal time series.

Matrix factorization and nearest neighbor These methods, which are popular in the field of recommender systems [44, 45], are also used to predict disease risks. For instance, Chakraborty et al. proposed matrix factorization with nearest-neighbor regression (MFN), which incorporates MF regression and nearest-neighbor-based regression, for ILI count prediction [28]. In their MFN model, they integrated disease-related features, historical disease dynamics, and the disease dynamics to be predicted across time into a prediction matrix. Then, they factorized the prediction matrix as a factor-feature matrix and a factor-prediction matrix, such that the prediction matrix could be reconstructed by multiplying the factor-feature matrix and factor-prediction matrix. Subsequently, they incorporated nearest neighbor regression to correct the reconstructed prediction matrix with the K nearest samples.

Ensemble learning Some studies have improved the robustness of predictions by generating them from ensemble models rather than from a single model. For instance, [28] used an ensemble model derived from the fusion of outputs of models trained on data from various sources. Specifically, they trained multiple MFN models on data whose effects on disease dynamics had been previously studied and combined their results to give the final prediction. Similarly, [27] trained various GP regression models on different features of disease-related data and aggregated these models' results to generate the final predictions.

2.3.2 Deep learning models

Due to the excellent ability to represent high-dimensional features in latent space and capture

complex dependencies, deep learning has been widely explored and applied in the task of disease risk prediction. Many sophisticated structures of **deep neural network** (DNN) models—e.g., **convolutional neural networks** (CNNs), **recurrent neural networks** (RNNs), and **graph neural networks** (GNNs)—have been fully explored as a means to capture the non-linear relationships and spatiotemporal patterns of disease transmission and thereby achieve good predictive performance. In our survey, for the simplicity of notations, we use $\hat{y}_d = f_d(\mathbf{x}, \boldsymbol{\theta}_d)$ to represent the prediction of y_d from the input feature \mathbf{x} via a non-linear function f_d . Here the f_d could be specified by any DNN model structure, and the $\boldsymbol{\theta}_d$ denotes the corresponding model parameters. Generally, $\boldsymbol{\theta}_d$ is optimized (over the parameter space Θ_d) using the following loss function:

$$\arg \min_{\boldsymbol{\theta}_d \in \Theta_d} \mathcal{L}_d(\boldsymbol{\theta}_d) \quad (7)$$

$$\mathcal{L}_d(\boldsymbol{\theta}_d) = \ell(\hat{y}_d, y_d) = \ell(f_d(\mathbf{x}, \boldsymbol{\theta}_d), y_d), \quad (8)$$

where \mathcal{L}_d is the predictive loss, quantified by the difference between the model prediction \hat{y}_d and the ground truth label y_d . In the disease risk prediction task, various distance metrics could be used to measure the difference, such as the ℓ_1 -norm loss (mean absolute error) or ℓ_2 -norm loss (mean squared error).

Recurrent neural networks RNNs are widely used to model the temporal dependency of time series data, such as voice or text data. RNN modules are formed based on the assumption that the current output is not only related to the current input but also depends on the previous inputs. Thus, as infectious disease dynamics are a type of time series data, they can also be modeled by RNNs. For example, the **interactively and integratively connected deep recurrent neural network** model (I²DRNN) [29] uses stacked RNN modules to capture spatiotemporal dependencies from heterogeneous and multiple-scale risk-related data. The model structure contains three components: (1) an input module, which is used to integrate heterogeneous (i.e., fine-, coarse-, and same-scale) data; (2) a hidden module, which is designed as a hierarchical structure to extract dependencies from different locations and the heterogeneous factors of different scales; (3) and an output module, which is used to generate final predictions based on extracted hidden features. RNN architectures based on a gating mechanism, such as a **long short-term memory** (LSTM) network [46], have also been used in recent studies for disease risk prediction, due to their ability to preserve the long-term information of data sequences. Venna et al. proposed an LSTM-based deep learning model that consists of multiple LSTM cells sequentially stacked over time [30]. In the sequential structure, every LSTM cell takes two inputs (except the first cell, which only takes the dynamic data as input): the dynamic data at a single time point and the output of the previous cell, to generate a prediction for the next time step. That study also examined the effects of climate variables by applying the symbolic time-series approach and the effects of regions with geographical proximity by applying weighted summation to adjust the output of an LSTM to generate the final prediction. Volkova et al. adopted two LSTM layers to learn temporal dependencies from the ILI case number and social media data, respectively, and merged their outputs via a fully connected layer to generate predicted ILI proportions [31].

Graph neural networks In contrast to RNN models, which capture the temporal dependency of sequential data, GNN models can deal with data with graphical structures [47]. That is, a GNN model is based on a graph with a given structure and encodes structural

information by passing messages between nodes of the graph. Due to this ability of GNN models to capture characteristics in graph structures, they are naturally used to capture and represent spatial patterns of disease dynamics, which can be regarded as driven by a disease transmission network. For example, the **spatio-temporal graph neural network** (STGNN) [32] utilizes daily mobility data from Google to construct the structure of time-varying disease transmission networks. Based on their constructed network, Kapoor et al. designed two types of edges, i.e. edges between nodes within the network at the same time, and edges between nodes within the network at the current time and nodes within the network at the previous time, to characterize varying spatiotemporal dependencies driven by cross-regional human mobility and the effect of historical risk trends, respectively [32]. Moreover, GNNs are not limited to modeling intuitive spatial relationships by delineating the network structure between locations; they can also be used to model the dependency between extracted features. For instance, the **spectral temporal graph neural network** (StemGNN) [33] uses a **graph convolutional network** (GCN) structure to model temporal dependency to predict newly confirmed COVID-19 cases. Specifically, instead of modeling the time series in the time domain, it utilizes the **graph Fourier transform** (GFT) to model inter-series correlations within the spectral domain and the **discrete Fourier transform** (DFT) to model intra-series temporal correlations within the frequency domain, and then feeds the representation of correlations into a GNN.

Mixed deep modules More recent studies have applied combinations of multiple neural network structures to model complex spatiotemporal patterns of disease transmission. These models with mixed neural network structures make full use of the aforementioned common neural network structures, i.e., RNNs, GNNs, and CNNs, to design new composite architectures serving various purposes. Usually, these architectures contain two separate modules—i.e., a spatial module and a temporal module—that are connected to form an integrated model that is subsequently optimized in an end-to-end manner to capture and model spatial and temporal dependencies simultaneously. For example, Wu et al. proposed a model which incorporates CNN, RNN, and residual structures, named CNNRNN-Res, to capture spatiotemporal dependencies in historical disease dynamics [34]. The CNN module uses an adjacency CNN filter to represent the adjacent graph of different regions, which is employed to integrate the information from neighbors. The RNN module uses a **gated recurrent unit** (GRU) to capture the temporal correlation in data. To solve the problem of overfitting, a sparse residual link is used to skip connections with some previous layers. The **Cross-location attention-based GNN** (ColaGNN) [35] is designed for long-term ILI prediction. It uses location-aware attention to infer the spatial influence between different regions from learned hidden features (denoting temporal dependencies), which are extracted from an RNN module. In addition, it employs a dilated convolution module to learn attributes for each node from historical disease trends and thereby capture multiple-scale local temporal dependency. Based on the above-mentioned network structure and node attributes, a graph message-passing mechanism is used to integrate the spatiotemporal information, which is then used to generate ILI predictions. The **Hierarchical spatial-temporal framework** (HierST) [36] includes a temporal module that combines two time-series architectures—the long- and short-term time series network (LSTNet) and the neural basis expansion analysis for time series (N-BEATS)—to model temporal dependency; and a spatial module that contains the gated edgeGNN, which adaptively adjusts the connections of edges, and the nodeGNN, which learns the representa-

tion of node features. The novelty of this approach is also reflected by the introduction of prior knowledge of common sense to constrain the model inference. Specifically, given that the predictions for different administrative levels (i.e., country, state, and county) should be close to each other, [36] designed a consistency optimization objective that includes items representing the difference between predictions at different spatial scales in addition to the difference between ground truth and predictions. The **epidemic forecasting model based on functional neural process (EPIFNP)** proposed by Kamarthi et al. [37] also includes temporal and spatial modules, which are implemented by a probabilistic neural sequence encoder and a stochastic correlation graph, respectively. Instead of generating point estimates of forecast value, the EPIFNP model generates the probability distribution of prediction via a probabilistic generative process model to evaluate the uncertainty of prediction. The population-level disease prediction model (named PopNet) proposed by Gao et al. assumes that an undirected disease transmission network drives disease dynamics [38]. PopNet learns the connection structure of this network by using population and geographical distance to calculate the similarity of each pair of locations. Then, based on the learned network structure, PopNet uses two graph attention networks (GATs) to obtain node embedding from real-time disease data and updated disease data (i.e., revised data released after the initial release), respectively. Subsequently, PopNet fuses these two kinds of node embeddings by **spatial latency-aware attention (S-LAtt)** and **temporal latency-aware attention (T-LAtt)**, sequentially. S-LAtt uses a feature similarity-based attention mechanism and considers the marginal effects of time latency on final predictions to learn the edge weights between pairs of nodes together to update node embedding. T-LAtt uses GRU networks to learn temporal dependency. Finally, PopNet concatenates the learned node embeddings to generate final predictions.

Encoder–decoder The encoder–decoder framework is a general framework that consists of different deep modules to manage sequential data and was initially used in machine translation [48]. As trends in disease dynamics exhibit a sequential dependency that is similar to that of sentences of text, an encoder–decoder framework with an RNN structure can also be used to predict epidemic trends within a given time period [39, 40]. Adhikari et al. proposed EpiDeep to predict weighted ILI (wILI) using an encoder–decoder framework, together with deep clustering components [39]. EpiDeep uses an LSTM-based encoder to encode an input influenza sequence as latent variables that contain temporal information, and a deep clustering component (an **improved deep embedded clustering (IDEC)** module [49]) to learn the embedding of the existing observed epidemic trend in the current season whose trend is to be predicted, and then clusters this embedding with the most similar epidemic trends in historical seasons. EpiDeep also uses this approach to learn and cluster the embedding of full-length historical trends. Next, it learns a mapping function to map the embedding of the incomplete sequence to the space of the full-length sequence. Finally, EpiDeep uses a decoder to predict the future sequence of the epidemic trend in the current season by taking the mapped clustering embedding and the encoded trend (both are in the current season) as inputs. Cui et al. also used an encoder–decoder framework to predict the dynamics of COVID-19 pandemic [40]. The encoder component employs several CNN modules with different kernel sizes to extract temporal features within multiple time ranges from the data of case numbers and regional visitor counts. They designed a graph-based module that characterizes spatial patterns by modeling human mobility and infection processes in each range in which graph structure is learned from data by the attention mechanism, and then fuses features learned from each

range into one feature via a multi-headed self-attention mechanism. The decoder component employs a temporal embedding module to embed the case numbers and death numbers, and the obtained embedding is fed into the multi-head attention layers with the output of the encoder. Finally, the features obtained from the encoder and decoder are passed through a multilayer perceptron (MLP) to generate a final prediction.

Transfer learning However, knowledge on the properties and transmission of emerging diseases may be ambiguous, and there may be a paucity of observational data on epidemic trends. Additionally, the data for regions with inadequate systems for surveillance and reporting of infectious diseases may be incomplete and noisy. Therefore, it is of interest to determine how to use abundant and high-quality data on similar diseases to the disease of interest or for regions with similar characteristics to the region of interest to facilitate the prediction of a given disease epidemic. One approach used is transfer learning (TL) architecture, which consists of one source task and one target task, and is designed to transfer knowledge learned from the previous model to enable the learning of the target task [50]. Similarly, the COVID augmented ILI deep network (CALI-NET) is a heterogeneous transfer learning (HTL) framework for COVID-ILI forecasting [41] that applies the EpiDeep [39] model as the source model to learn representations of temporal dependency from historical wILI data. [41] designed the COVID-augmented exogenous model (CAEM) to encode representations of spatiotemporal features of exogenous data signals of COVID-19 by Laplacian regularization of a geographical adjacent matrix and a GRU module, and used these encoded representations in the target model. They also designed a knowledge distillation (KD) loss, which consists of the hint loss between the mapped representations from source and target models, and the imitation loss between source predictions and ground truth to ensure effective knowledge transfer.

The above-described studies demonstrate that previously developed deep learning-based models use spatial and temporal modules, which are designed to capture disease patterns over space and time, respectively. The relationship between spatial information is typically modeled by a GNN module. However, using a GNN module is challenging, due to its unknown network structure. Some approaches pre-build connections in a network by using proxy data, such as human mobility data [32], population and geographical distance data [38], and regional demarcation data [41], and then feed the network into a GNN module or constrain model learning by applying graph Laplacian regularization. If there is a lack of related data, a graph structure can be constructed by using the attention mechanism to infer the weights of edges in a network structure during model optimization [35, 36, 38, 40]. Temporal patterns are modeled by mining the temporal dependency of time series data using various RNN structures, such as a classic RNN [35], a GRU [34, 41, 37, 38], LSTM [39], or an LSTNet/N-BEATS [36]. However, in addition to using an RNN and GNN to characterize temporal and spatial patterns, respectively, these neural network modules can be flexibly designed to serve these purposes. For instance, in addition to RNN modules being used to extract temporal dependencies, CNN modules (such as temporal convolution modules [40] or dilated CNNs [35, 38]) and GNN modules (such as spectral temporal GNNs [33]) can also be used for this purpose; and in addition to GNN modules being used to capture spatial dependencies, a modified CNN module can also be used for this purpose [34].

2.3.3 Discussions: advantages and limitations

The above-mentioned studies show that many data-driven machine learning methods have been developed for infectious disease risk prediction. Their excellent predictive performance may be due to the following aspects.

- The sophisticated modules designed for spatial and temporal characterization can extract complex hidden representations and learn complex non-linear relationships from abundant data sources to capture spatiotemporal disease transmission patterns.
- The above-described data-driven models are based on supervised learning algorithms that enable the inference of model parameters that fit well with the data.

However, most of the above-described advanced machine-learning methods are not designed for the modeling of diseases with complex transmission environments and conditions, such as malaria. Thus, these methods may not perform well when they are applied to model such diseases because they do not fully consider and utilize disease-related knowledge on a given disease. Furthermore, it is difficult to quantify the effects of various risk-related factors on predicted transmission intensity or risk in an interpretable way, so modeling results may not provide clear guidance for decision-makers on implementing time/location/factor-specific measures in response to potential risks or outbreaks.

2.4 Epidemiology-inspired machine learning

Although data-driven machine learning approaches greatly improve the accuracy of disease risk prediction, they still struggle to provide insights to facilitate disease control. To overcome this drawback, epidemiological models have been re-visited and integrated with machine learning methods. Epidemiological models mathematically depict a disease transmission process based on domain experts' understanding of the disease's biological characteristics. Thus, each parameter and the overall structures of epidemiological models have a clear epidemiological meaning. However, the structures of epidemiological models are typically based on relatively simplified assumptions, so these models may struggle to provide sufficiently accurate predictions. Conversely, data-driven machine learning models can fit training data very well and generate accurate predictions, but in some cases the physical meaning of learned patterns is ambiguous and thus cannot effectively support public health decision-making. Therefore, a key question in disease risk prediction modeling is how to exploit the complementary strengths of data-driven models and epidemiological models to obtain modest explanatory power while utilizing their strong representation ability to determine complex dependencies. Driven by this question, a large body of literature has investigated the potential of combinations of epidemiological models and data-driven machine-learning models. In this paper, we denote this type of model "epidemiology-inspired machine learning" and we divide it into two classes: (1) epidemiological parameterization and (2) epidemiology-embedded learning. In the following, we first provide some preliminary information on epidemiological models. Then, we introduce previous studies by categorizing their epidemiology-inspired machine learning models and describing how they combine epidemiological prior knowledge with machine learning methods. A summary of the epidemiology-inspired machine learning models is given in Table 2.

Table 2: Summary of epidemiology-inspired machine learning models.

Categories	How to combine	Ref.	Targeted diseases	Epidemiological components	Data-driven components
Epidemiological parameterization	Inferring epidemiological parameters from data	[11, 51]	Influenza	Humidity-driven SIRS model	EAKF/PF (Data assimilation)
		[52]	Influenza	Four types of compartmental models	EAKF/PF (Data assimilation)
		[53]	Influenza	Metapopulation compartmental model	EAKF/PF (Data assimilation)
		[54]	A/H1N1	GLEaM	Monte Carlo maximum likelihood analysis
		[55]	Influenza	GLEaM	Monte Carlo maximum likelihood analysis
		[56]	COVID-19	SuEIR model	Loss function with logarithmic-type MSE
		[57]	COVID-19	Spatiotemporal-SuEIR	AutoODE
		[58]	Airborne disease	Metapopulation SIR model	Non-negative network inference model with power-law distribution and data priori
	Modeling epidemiological parameters	[59]	COVID-19	Improved SEIR model	Generalized additive model
		[60]	COVID-19	Stochastic SIR process	Mixed effects model
Epidemiology-embedded learning	Epidemiological guides	[61]	Malaria	EIR/VCAP	Nonlinear stochastic model
		[62]	Malaria	NGM	Multivariate regression with non-linear parameters
	Epidemiological regularization and constraints	[63]	COVID-19	SEIR model	Social media based simulation model
		[64]	COVID-19	SIR model	Spatiotemporal tensor factorization
		[65]	Influenza	SIR model	Dynamic Bayesian
		[66]	COVID-19	SIR model	GAT and GRU
		[67]	COVID-19	SIRD model	Dynamic attention-based GCN
		[68, 69]	Influenza	SEIR model	LSTM

2.4.1 Introduction to epidemiological models

In the 20th century, many epidemiological models were developed to mathematically depict the process of infectious disease transmission based on the understanding and knowledge of disease characteristics and transmission modes. These models are also known as mechanism-based models, compartmental models, or physics-based models. In these models, a studied population is usually divided into several compartments representing different disease statuses, and a set of rules is designed to describe the transition between these statuses. These models can be subdivided based on the granularity level of modeling (from coarse to fine) into three types [70, 17]: (1) compartmental models at the population level; (2) compartmental models at the meta-population level; and (3) agent-based at the individual level. Compartmental models at the population level usually include a set of differential equations (DEs) that depict the dynamics of state variables and thereby quantitatively represent disease risk. These models comprise a representative and classical group of epidemic models that are used to mathematically depict disease transmission. Various DEs have been developed for a wide range of infectious diseases, such as influenza, malaria, dengue, Aids, and COVID-19. These models assume that a disease transmission environment is homogeneous, i.e., individuals are mixed evenly within the environment and those with the same status have the same probability of moving from their current status to another status. In the following section, we introduce various compartmental models at the population level that have been constructed for diseases that propagate in different ways, e.g., respiratory diseases and vector-borne diseases.

The susceptible–infected–recovered (SIR) model is a classic compartmental model [71]. It has a simple structure with three statuses (susceptible, infected, and recovered) and two parameters (effective contact rate β and recovery rate γ), and is widely used to model the dynamics of infectious diseases [72], especially respiratory diseases, such as influenza, ILI, and COVID-19. Following the development of the SIR model, many other compartmental models with more sophisticated structures were designed to describe more complex scenarios, such as susceptible–exposed–infected–recovered (SEIR) models, which consider the latent period of a disease [73, 74] and use the parameter β to represent the probability of an individual entering the incubation period after being in contact with an infectious individual, and the parameter α to represent the probability of an individual leaving the latent period; and the susceptible–exposed–infected–recovered–death (SEIRD) model [75, 76], which considers deaths due to disease. Many variations of compartmental models have been developed for vector-borne diseases, such as malaria and dengue, to depict disease transmission between vectors (e.g., mosquitoes) and humans [13, 14]. For instance, the Ross model [77] is the most fundamental model to describe vector-borne diseases, while the Macdonald model [78] is based on the Ross model but also considers the latent status of vectors. In addition to considering the latent status of vectors, the Anderson and May model [79] considers the latent status of humans, and susceptible–latent–infected–recovered (SLIR) models [80, 81] consider the recovered status of humans in terms of acquired immunity.

However, sometimes the homogeneous-mixing assumption does not accurately reflect the real situation of disease transmission because individuals in a host group may have different characteristics, such as different susceptibilities to disease and abilities to recover from infection. These characteristics significantly influence disease spread throughout a population and also determine how epidemic interventions should be enacted. Therefore, in addition to models based on the assumption that disease spreads occurs in a homogeneous environment (i.e., that individuals have the same probability of coming into contact with each other and of

moving from one status to another), many models—i.e., compartmental models at the meta-population level—have been developed that are not based on this assumption; instead, they are based (to some extent) on a heterogeneous assumption. Studies have divided populations into subgroups and designed model structures according to different population properties, such as age structures [82, 83], geographical distributions [84, 85, 86], and human behavioral patterns [87].

As mentioned, mechanistic models at the meta-population level consider the heterogeneity of subgroups of a whole population. However, their characterization of the heterogeneity of population traits is still limited because of the low resolution of subgroup partitions. Contact between hosts and hosts, or between hosts and vectors, is the natural way in which infectious diseases are transmitted in the real world. Thus, agent-based models are usually built on a network constructed at the individual level and simulate interactions between individuals, such that they model disease transmission in more realistically than mechanistic models at the meta-population level. EpiSims, proposed by Eubank et al., is an agent-based simulation tool for modeling disease spread caused by human mobility [88]. EpiSims simulates the physical contact patterns of humans by constructing a bipartite social contact network that consists of two types of vertices: individual vertices and location vertices. Compared with the results of compartmental models at the population and meta-population levels, the simulations generated by agent-based models are closer to the real-world situation because the characteristics of these models' networks are similar to those of real networks. Similarly, Hoertel et al. developed a stochastic agent-based microsimulation model for modeling the COVID-19 epidemic in France [89]. The two above-described studies show that fine-grain agent-based models enable the flexible setting of interventions and can help to reveal potentially effective intervention strategies.

2.4.2 Epidemiological parameterization

Epidemiological parameterization uses or improves existing mechanism-based models (e.g., SIR and SEIR models) as the basis for predicting disease dynamics. The initial values and epidemiological parameters of compartmental models that are described by ordinary DEs (ODEs) are crucial for determining disease dynamic. In contrast to traditional mechanism-based models, which use given or fixed epidemiological parameters, epidemiological parameterization models use disease-related data to estimate model variables and epidemiological parameters in compartmental models. Many machine learning approaches are used to infer model variables and epidemiological parameters. The relatively simplified structures (compared with real systems) and inaccurate parameterization of initial values and parameters in mechanism-based or compartmental models may cause them to generate biased predictions (i.e., predictions that do not reflect the real situation), so inference approaches try to compensate for biased predictions to some extent by balancing model predictions with observed data.

Inferring epidemiological parameters from data Data assimilation techniques, which are widely applied in atmospheric and oceanic sciences and in numerical weather forecasting [90], aim to utilize observations to optimize mechanism-based models. Thus, they have also been applied in disease dynamic prediction [11, 51, 52, 53]. In a set of data assimilation models used for epidemic prediction, the **Kalman filter** (KF) and its variants [91, 92] and

particle filter (PF) [93] methods have been used to estimate model statuses. For instance, Shaman and Karspeck applied data-assimilation techniques to the problem of influenza forecasting and generated retrospective ensemble forecasts of influenza seasons from 2003 to 2008 in New York City, USA [11]. They proposed the SIRS–EAKF framework, which uses the ensemble adjustment Kalman filter (EAKF) [92] and a PF [93] to assimilate the observations of infections (i.e., estimates of influenza infections from Google Flu Trends) into the susceptible–infectious–recovered–susceptible (SIRS) model [94], which is a humidity-forced compartmental model. The SIRS–EAKF framework can estimate the posterior of probabilistic distributions of system states (i.e., susceptible populations S_t and infected populations I_t) and epidemiological parameters (e.g., the mean infectious period D , the average duration of immunity L , and the maximum and minimum of daily basic reproductive number R_{0max} and R_{0min}) in the used SIRS model. In [94], Shama et al. represented model states and epidemiological parameters by a set of variables $Z_t = (S_t, I_t, R_{0max}, R_{0min}, L, D)$. Then the posterior of Z_t can be represented as follows:

$$p(Z_t|y_t, y_{t-1}, \dots) \propto p(y_t|Z_t)p(Z_t|y_{t-1}, \dots), \quad (9)$$

where the first term on the right-hand side is the likelihood of observational disease risk given states and parameters, while the second term is the prior distribution of the states and parameters. For KF, these two terms are assumed to be Gaussian distributions; in contrast, for PF, these two terms are not under these assumptions. Subsequently, Shaman et al. used similar data-assimilation techniques to generate weekly influenza forecasts for the influenza season in 2012 and 2013 in 108 cities in the USA [51]. Yang et al. tested four types of compartmental models and two types of filter models in their model-data assimilation framework and analyzed the epidemiological characteristics of influenza dynamics from the 2003–2004 season to the 2012–2013 season in 115 cities in the USA [52]. Pei et al. developed a model-data assimilation framework based on a metapopulation compartmental model to accurately predict the spatial spread of influenza [53]. In this metapopulation compartmental model, which is based on a humidity-driven SIRS model [94] that they had used in their previous studies [11, 51, 52], they divided a population into different groups in terms of geographical locations (i.e., different states), and incorporated two types of human mobility (i.e., fixed commuting flows and irregular movement of visitors).

Balcan et al. proposed the **global epidemic and mobility** (GLEaM) computational model for simulating infectious disease transmission [95, 96]. The GLEaM model is a global model based on a stochastic compartmental model at the meta-population level (i.e., a SLIR model considering both symptomatic and asymptomatic infections) and incorporates multiscale human mobility (short-range commuting and long-range airline flows) to effectively capture disease transmission patterns. Tizzoni et al. used the GLEaM model to model the disease transmission of 2009 H1N1 influenza and utilized the Monte Carlo maximum likelihood method to estimate some parameters [54]. Similarly, Zhang et al. proposed an epidemic computational framework based on the GLEaM model [55]. The computation is performed via three steps: (1) microblogging data from Twitter and surveillance data are used to estimate initial infections; (2) epidemiological parameters are searched in four-dimensional space by running Monte Carlo simulations with selected sampling points, and the GLEaM model is used to generate simulations; (3) a set of best-fit models is selected by using a multi-model information approach, which minimizes the loss of information (which is calculated by the Akaike information criterion (AIC)).

Aside from data-assimilation methods and simulation-based methods, some machine learning approaches are proposed to estimate the model states and epidemiological parameters. In these works, the loss function is generally formulated as the difference between states simulated using epidemiological models and the ground truth of these states. The general formulation of such loss function can be represented as follows:

$$\mathcal{L}_e(\boldsymbol{\theta}_e) = \ell(\hat{\mathbf{y}}_e, \mathbf{y}_e) = \ell(f_e(\boldsymbol{\theta}_e, \mathbf{s}_0), \mathbf{y}_e), \quad (10)$$

where $\boldsymbol{\theta}_e$ (in the parameter space Θ_e) denotes epidemiological parameters (e.g., contact rate and recovery rate) in a given epidemiological model, \mathbf{y}_e denotes the ground truth of the target variable (usually are model states with records, e.g., infected case number and death number), $\hat{\mathbf{y}}_e$ denotes predictions on the target variable, \mathbf{s}_0 denotes the initial value of model states, and f_e denote the given epidemiological model, which is generally described by an ODE. With such an ODE representation, $\hat{\mathbf{y}}_e$ can be calculated using model parameters and initial values. Given the above loss function, the optimal model parameters can be inferred by minimizing the loss:

$$\hat{\boldsymbol{\theta}}_e = \arg \min_{\boldsymbol{\theta}_e \in \Theta_e} \mathcal{L}_e(\boldsymbol{\theta}_e). \quad (11)$$

For example, Zou et al. formulated a loss function with a logarithmic-type mean square error (MSE) [56]. Based on this loss function, parameters can be optimized by the general gradient-based optimizer. Moreover, Zou et al. also developed a novel compartmental model, named the SuEIR model—an improved **SEIR** model that considers a scenario of **un**tested or **un**reported cases of COVID-19—and trained it with their machine learning approach. Wang et al. formulated a similar loss based on MSE to estimate the parameters of epidemiological models. Based on this loss function, they formalized the learning procedure as the AutoODE algorithm, which infers the model parameters of mechanism-based models by an automatic differentiation method. In addition, based on a case study on the forecasting of COVID-19 dynamics, they proposed the spatiotemporal SuEIR model, which is an extension of the SuEIR model [56] that better models spatiotemporal patterns of COVID-19 spread. The power-law degree and data priori jointly regularized non-negative network inference (D^2PRI) approach of Wang et al. [58] is based on a SIR model at the meta-population level. As this model regards the infectious interactions between individuals at different locations as a transmission process in a disease propagation network, Wang et al. [58] formulated the parameter inference of edge weights in the network and disease transmission rate in the SIR model as an integrated network inference problem. Moreover, according to prior knowledge of network structure, i.e., the power-law distribution of node degree and the features extracted from mobility-related data, they designed corresponding regularization items to constrain the parameter inference.

Modeling epidemiological parameters In contrast to models that infer values or probabilistic distributions of model parameters from observations, other models estimate the variation of epidemiological parameters and formulate them as functions of covariates. For instance, Arik et al. proposed the use of time-varying functions to model parameters [59]. That is, they used an improved compartmental model that is based on the SEIR model: instead of using the static epidemiological parameters in the traditional compartmental model, they used learnable functions to estimate parameter values from various covariates, which enable parameter values to vary over time. Specifically, they used the generalized additive model

to encode the effects of covariates on epidemiological parameters. Baek et al. predicted the disease dynamics of multiple regions by using a stochastic SIR model [60]. This stochastic model employs a mixed-effects model that incorporates a random-effects term within each region and a fixed-effects term between different regions to encode the effects of static and time-varying covariates on the disease transmission rate.

2.4.3 Epidemiology-embedded learning

In contrast to epidemiological parameterization, epidemiology-embedded learning focuses on using machine learning models to predict disease dynamics directly, while using mechanism-based models to guide, regularize, or constrain the machine learning models.

Epidemiological guides Some studies have utilized the epidemiological concept to guide the construction of model structures. For instance, rather than directly describing and predicting malaria dynamics by using ODEs, Shi et al. constructed a nonlinear stochastic model based on the formula for vectorial capacity (VCAP) and entomological inoculation rate (EIR), which are epidemiological concepts derived from corresponding ODEs that describe malaria transmission [61]. VCAP is defined as the daily rate of future inoculations from mosquitoes to humans caused by a currently infected human case [97], whereas EIR is defined as the number of infectious bites received from mosquitoes per day by a human [97]. Based on the epidemiological meaning of local transmission, the local infections at time t can be formulated as a function involving the VCAP/EIR and the infections at time $t - 1$. Then, by considering the effects of cross-regional transmission, Shi et al. used periodic function modeling to depict the periodic transmission patterns [61]. Thus, their nonlinear stochastic model consists of the items of local infections and imported infections. Liu et al. developed a multivariate regression model based on a next-generation matrix of a meta-population vector-human compartmental model to predict the malaria risk in multiple locations [62]. The next-generation matrix [98, 99], consisting of the non-linear relationships between epidemiological parameters that can be derived from a compartmental model, represents the change in model variables from one time step to the next time step, thereby enabling the prediction of disease dynamics.

Epidemiological regularization and constraints Some studies have added epidemiological constraints and regularizations, which are derived from compartmental models, to standard objective functions of supervised machine learning models to aid model parameter optimization. Hua et al. proposed the social media based simulation (SMS) model for influenza dynamics prediction [63]. This model incorporates two learning spaces: the social media space, which is designed to identify individuals' health statuses from social media posts; and the epidemiological simulation space, in which a transmission network is built to simulate disease propagation between individuals. These two spaces are linked by minimizing the loss in terms of the inconsistency between the health status at the population level, which is obtained from the social media space and the simulation space. Kargas et al. applied epidemiological constraints in tensor factorization approaches to predict disease dynamics [64] by devising spatio-temporal tensor factorization with epidemiological regularization (STELAR). A tensor is an intuitive and natural structure used to represent and preserve the complex structure of high-dimensional data, especially spatiotemporal data with multiple risk-related factors.

Tensor factorization is usually employed for dimensionality reduction and data decomposition rather than to predict disease transmission dynamics. STELAR enables the prediction of long-term epidemic trends by the addition of the latent epidemiological regularization of the SIR model into a standard tensor factorization method, i.e. **c**anonical **p**olyadic **d**ecomposition (CPD). Osthus et al. proposed a **d**ynamic **B**ayesian (DB) influenza forecasting approach that models discrepancies between mechanistic model-generated simulations and observations [65]. This approach assumes that the uncertainty of prediction cannot be fully explained by observational noise and therefore models a wILI as the sum of three items: the logit of the infections that are described by the SIR model, a common discrepancy item for all influenza seasons, and a specific discrepancy item for each influenza season.

Some deep learning models, such as GNNs [66, 67] and RNNs [68, 69], also incorporate mechanistic models to constrain the learning of model structures and parameters, such that they effectively fit the realistic situation of disease transmission. In addition to the prediction loss, some of the aforementioned methods further introduce the epidemiological loss to ensure that the prediction of deep learning models is consistent with the dynamics described by the epidemiological model. The general formulation of the loss function is shown as follows:

$$\mathcal{L} = \mathcal{L}_d(\boldsymbol{\theta}_d) + \mathcal{L}_e(\boldsymbol{\theta}_e) = \ell(\hat{\mathbf{y}}_d, \mathbf{y}_d) + \ell(\hat{\mathbf{y}}_e, \mathbf{y}_e), \quad (12)$$

where \mathcal{L}_d denotes the prediction loss to ensure the prediction accuracy of the deep learning model, and \mathcal{L}_e denotes the epidemiological-constrained loss. The approaches to introduce \mathcal{L}_e vary in different works. For example, **s**patio-**t**emporal **a**ttention **n**etwork (STAN) is a GAT model with epidemiological constraints that is designed for long-term pandemic prediction [66]. In the network, nodes denote different locations and have both static features (latitude, longitude, population size, and population density) and dynamic features (case number, and the information on hospitalizations across all timestamps), whereas the edges denote disease transmission. The weights of edges in the STAN are calculated in terms of geographical proximity and population size. After constructing the network structure of the STAN, the graph attention mechanism is used to update node attributes and feed them to a GRU to extract temporal features. Epidemiological constraints are incorporated into STAN learning and prediction, as it generates two types of output by multitasking prediction settings: (1) epidemiological parameter predictions (i.e., the transmission rate and recovery rate); and (2) disease dynamics predictions (i.e., the increases in the infected and recovered case numbers). Gao et al. [66] also designed the loss function for model optimization based on the above-mentioned two kinds of outputs. In their loss function, the first item is the prediction loss which captures short-term trends by calculating errors between the dynamics predicted by the deep modules and real case numbers, the second item is the epidemiological loss which captures long-term trends by calculating the errors between the disease dynamics simulated with the SIR model and real case numbers. The causal-based graph neural network (CausalGNN) model proposed by Wang et al. is another framework that constrains the dynamic attention-based GNN module with an epidemiological model (i.e., the SIRD model) [67]. Similar to [66], the CausalGNN model generates two types of outputs: data-driven predictions of case numbers; and the epidemiological parameters in the SIRD model, which are used to simulate predictions of case numbers. The loss function in this model consists of two l_1 -norm items, which are errors between the ground truth of the case number and the two types of predictions. Unlike the model developed by Gao et al. [66], the CausalGNN model also feeds the simulations obtained from the SIRD model together with the input features into the data-driven model to

generate the model outputs. Wang et al. [68, 69] proposed an epidemic prediction framework, named **deep learning based epidemic forecasting with synthetic information (DEFSI)**, to conduct short-term and high-resolution ILI incidence prediction. The novelty of DEFSI is that it generates fine-scale ILI incidence data from an agent-based simulator (EpiFast) of an SEIR model, whose transmission parameters are estimated from the surveillance data. The obtained synthetic data are used to train a two-branch LSTM model to capture the within-season and between-season temporal dependencies of the incidence trends, and the model outputs are merged to generate final predictions.

2.4.4 Discussions: advantages and limitations

As aforementioned, epidemiology-inspired machine learning exploits the advantages of epidemiological models and data-driven machine learning models to construct models that are more interpretable than models built in a totally black-box manner while preserving representation and fitting capacity. Thus, epidemiology-inspired machine learning can infer disease transmission patterns from available data to generate a model that is more consistent with epidemiological constraints than black-box models. Based on the results of epidemiology-inspired machine learning models, estimated disease patterns can be further analyzed to provide more potential information hidden in data. However, most existing epidemiology-inspired machine learning models, especially those based on neural networks, incorporate a compartmental model in a serial or parallel manner, such as by utilizing a compartmental model to generate data that are fed into the learning architecture or by predicting the parameters of epidemiological models and then calculating the output of the compartmental model as a constraint. As such, their model structures are still somewhat black-box-like and thus are insufficiently interpretable. Consequently, there is a need for the development of methods that can naturally and intrinsically integrate knowledge of epidemiological models into structures of data-driven machine learning models.

3 Challenges

In Section 2, we introduced machine-learning methods and models that are currently used for infectious disease risk prediction and categorized the models by their structures. The plurality of model structures and their integrations that can be used to explicitly model disease transmission in different ways or directly predict disease dynamics are evident from the abundant previous literature. However, in practice, the designs of various models could benefit from specific challenges being addressed, and these have not been summarized thoroughly in previous surveys. Furthermore, there is no unique way to comprehensively classify prediction models; however, their various advantages are revealed by classifying these models in different ways. Therefore, in this section and Fig. 2, we summarize the above-mentioned models in terms of input data, task nature, and output evaluation, to present the challenges that are met during the prediction of disease risk.

3.1 Data challenges

The availability and limitations of disease-related data must be carefully examined to enable appropriate algorithms to be designed, i.e., algorithms that utilize or manage abundant data.

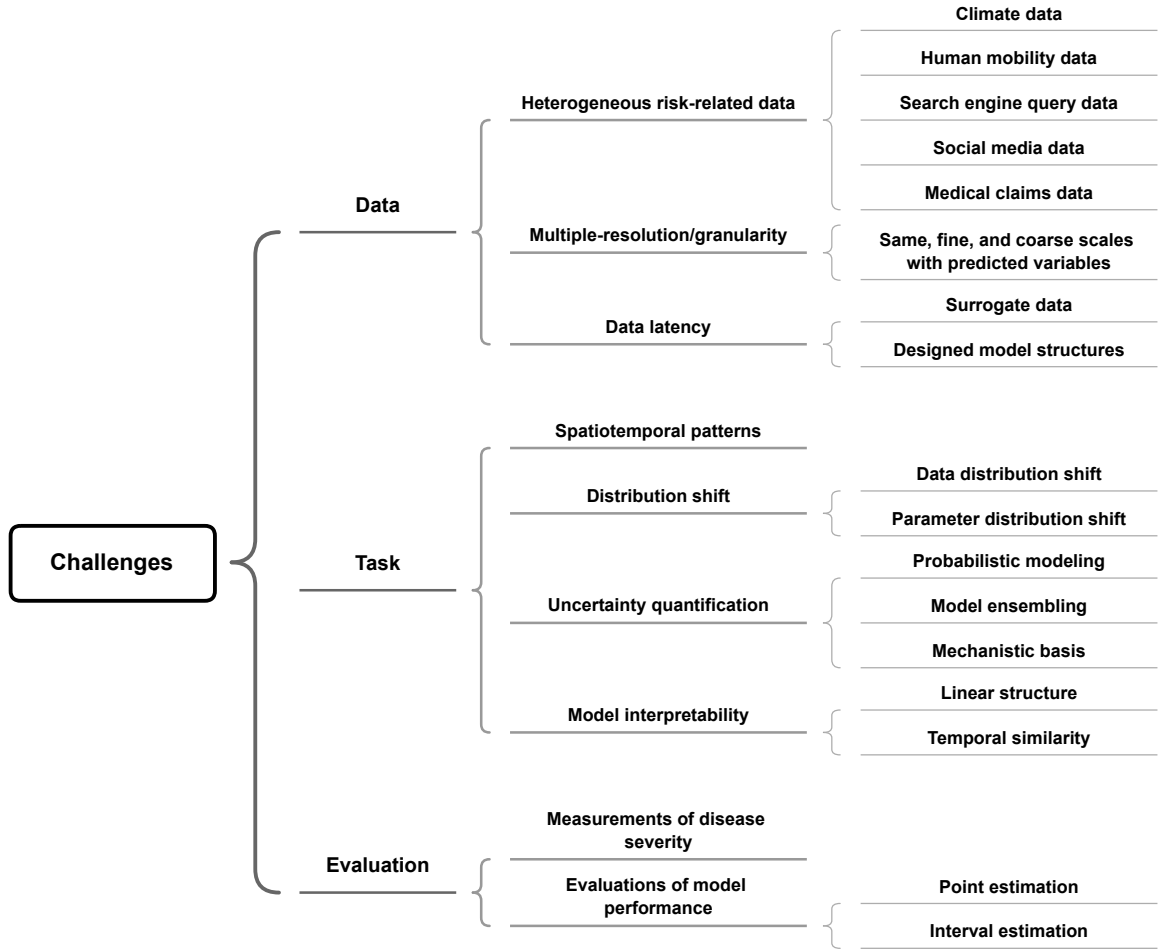


Figure 2: Three main challenges in machine learning for infectious disease risk prediction: (1) data-related challenges (Section 3.1); (2) task-related challenges (Section 3.2); and (3) evaluation-related challenges (Section 3.3).

3.1.1 Heterogeneous risk-related data

Because disease propagation is closely related to interactions between humans, the environment, and pathogens, data are commonly collected from the physical world. However, with the advent of the Internet and social media, non-physical types of information, such as social interactions, are increasingly reflecting transmission patterns between humans. Thus, many researchers have used abundant information from different sources by exploring and exploiting data of various risk-related factors to comprehensively characterize the patterns of disease spread. Some quantitative relationships between risk-related factors and disease dynamics have been explored and defined in empirical studies, but most of the causal links and correlations between these factors and disease dynamics are complex and non-intuitive. Therefore, when using these rich related data in prediction, many studies have either directly utilized known formulations or automatically discovered the statistical relationship between various risk-related factors to improve epidemic prediction.

Climate data One of the most significant features of the dynamics of diseases such as influenza and malaria, which are triggered or influenced by the climate, is their obvious correlation with trends in climate conditions. For example, Jeffrey and Melvin reanalyzed data obtained in laboratory experiments on guinea pigs [100] and demonstrated that absolute humidity (AH) greatly affected the influenza virus transmission (IVT) and influenza virus survival (IVS) in temperate regions [101]. On this basis, Shaman et al. further examined these findings at the human population level and proposed an AH-forced SIRS model, which quantitatively defines the relationship between AH and basic reproduction number, to simulate the seasonal patterns of influenza-related deaths [94]. In their subsequent studies, Shaman et al. used an AH-driven SIRS model that has a similar structure to their SIRS model [94] to incorporate AH data into their data-assimilation framework and generate influenza predictions for cities in the USA [11, 51]. In regions with non-temperate climate patterns, such as subtropical and tropical regions, the relationship between AH and influenza can be quite different. In contrast to the above-mentioned studies, which have used compartmental models to incorporate risk-related factors, Tamerius et al. collected data from numerous sites across a global range and explored the relationship between several climate variables (i.e., temperature, solar radiation, specific humidity, and precipitation) and the influenza peaks by rank order analysis. They also established univariate and multivariate logistic regression models to assess this relationship [102]. Venna et al. used a symbolic time-series approach to model the nonlinear relationship between climate variables (i.e., precipitation, temperature, and sun exposure) and dynamics of influenza [30].

Vector-borne diseases, such as malaria, are greatly affected by climate conditions, as these influence the biological features of vectors and pathogens, such as the survival rate of mosquitoes and the incubation period of *Plasmodium*. The quantitative relationship between temperature or rainfall and VCAP was defined by Ceccato et al. [103]. Shi et al. utilized temperature and rainfall data to estimate some epidemiological parameters in their model, which incorporates VCAP, to assist with the prediction of malaria dynamics which show obvious seasonal fluctuations [61]. Zhang et al. used temperature and rainfall and other disease-related data to form the feature vector of locations to learn their impact on intra-regional transmission risk with the Poisson regression model [22].

Human mobility data Human behavior plays an important role in infectious disease transmission [104], and human mobility is a crucial factor affecting the range and distribution of disease risk [105]. Several studies have examined the effects of different human mobility patterns on disease transmission. For instance, Balcan et al. studied two kinds of mobility patterns—short-range commuting flows and long-range flights—and explored their effects on the spatiotemporal dynamics of infectious disease by integrating a mobility network into the GLEaM model [95]. Pei et al. used commuter data from the US census survey² to support the prediction of influenza spread [53]. In their model, recurrent commuters and random visitors are considered to compose the connections between locations in the meta-population compartmental model. Kapoor et al. utilized human mobility data from two Google mobility datasets (the Google COVID-19 Aggregated Mobility Research Dataset and Google Community Mobility Reports³) to construct a human mobility network at the daily scale [32] for studying the COVID-19 pandemic. In their network, nodes represent locations, and

²<https://www.census.gov/topics/employment/commuting.html>. Cannot be accessed on March 25, 2023

³<https://www.google.com/covid19/mobility/>. Accessed March 25, 2023

spatial and temporal edges denote human movement between neighborhoods and connections between historical days and the current day, respectively. Cui et al. used visitor counts in the census block group provided by SafeGraph⁴ to model human mobility [40]. Unlike the method of Kapoor et al. [32], which directly constructs the edges of a human mobility network from movement data, the method of Cui et al. [40] treats visitor counts and other disease-related factors as features of each location and uses the attention-based mechanism to learn the structure of a disease transmission network.

Search-engine query data In addition to physical movement, the online behavior of individuals can partly reflect the patterns of disease risk. For example, some researchers have assumed that there is a correlation between changes in ILI levels and online search activities. For example, Ginsberg et al. developed Google Flu Trends, which uses the proportion of ILI-related search queries to overall search queries from Google as an explanatory variable for predicting ILI physician visits (the outcome) by fitting a simple linear model [106]. They focused on designing an automated method to identify the ILI-related search queries from a huge number of search records, which contribute most to the model when fitting with ILI physician-visit data. However, maintenance of the website for generating Google Flu Trends stopped in August 2015⁵. Nevertheless, the Google Extended Trends (GET) application programming interface remains accessible and provides the statistics of online search trends at various temporal and geographical granularities that researchers can use to train their models [107]. Yang et al. utilized GET data to devise the autoregression with Google search data (ARGO) method for influenza epidemic estimation [107]. They obtained the search terms that are strongly correlated with the ILI from Google Correlate⁶ (which stopped providing data after March 28, 2015) and search trends from Google Trends⁷, and then used these data as input for an ARX model. Kandula et al. also used Web-based search activity data from GET to nowcast ILI dynamics at subregional geographic scales (i.e., state-level) [108]. They formalized strongly correlated query terms for a specific region in a period as explanatory variables and used them to train an ARIMA model to predict the response variable (i.e., ILI observations). Then, they treated the forecast of the ARIMA model as an additional explanatory variable and used it with the original explanatory variables to train a random forest model for making final predictions of the response variable. Similarly, search query data from Baidu⁸, one of the biggest search engines in China, were used by Yuan et al. to monitor influenza case counts in China [109]. They also used a multiple regression model to predict case counts based on the previous case count and a composite index of searches.

Social media data Other important online behaviors of individuals that reflect the health status and contact patterns of humans are social media posts and virtual interactions. Although these data do not directly quantitatively reveal the extent and patterns of disease transmission, they can be mined to obtain some useful information. For example, Achrekar et al. [110] calculated Pearson correlation coefficients and fitted a regression model, finding a linear correlation between the number of Twitter users who tweeted influenza-related posts

⁴<https://docs.safegraph.com/docs/places#section-patterns>. Accessed March 25, 2023

⁵<https://ai.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html>. Accessed March 25, 2023

⁶www.google.com/trends/correlate/. Cannot be accessed on March 25, 2023

⁷<https://trends.google.com/trends/>. Accessed March 25, 2023

⁸<https://index.baidu.com/>. Accessed March 25, 2023

and the percentage of ILI physician visits. Based on this evidence, they proposed the **social network enabled flu trends (SNEFT)** framework, which is based on an ARX model, to predict ILI case numbers from ILI physician visits and collected Twitter data. Similarly, Zhang et al. counted ILI-related tweets to estimate the number of influenza infections in a given week and in given locations, but without filtering retweets and successive posts [55].

However, in [110, 55], only the number of users who post influenza-related information or the number of tweets relevant to the influenza keywords were used to estimate the number of influenza infections; other features of textual information were not utilized. In contrast, Volkova et al. extracted detailed linguistic features and communication patterns as latent embeddings from social media posts of a Twitter dataset and fed these data, together with ILI data, into a joint neural network model based on LSTM modules to predict ILI dynamics [31]. Hua et al. used a Bayesian model to identify individual health states (i.e., healthy, exposed, and infectious) from Twitter posts, aggregated these states at the population level, and then employed these states to inform the epidemiological parameters of a simulation model [63].

Medical claims data In addition to data that describe disease severity and transmission in a whole population (i.e., incidence rates or case numbers), there are abundant individual-patient data that can provide insights into disease severity and medical resource utilization. Gao et al. extracted daily data on hospitalizations, intensive care unit stays, and frequency of diagnosis codes from IQVIA’s medical claim data⁹ [66] and combined these data with the number of COVID-19 cases (active, confirmed, and death cases) provided by Johns Hopkins University¹⁰ to serve as the dynamic features of nodes in a graph neural network to facilitate the characterization of spatiotemporal patterns. In [38], Gao et al. extracted the statistics of disease-related features from the same IQVIA claim dataset to facilitate the training and prediction of a downstream deep learning model.

3.1.2 Multiple-resolution/granularity

One of the most common problems with using data from heterogeneous sources is that data on different risk factors have different spatial and temporal resolutions. The common spatial resolutions for disease-related data, in order from coarse to fine, are region, country, province/state, county, and village, whereas temporal resolutions are year, month, week, and day. In [29], Tan et al. considered data with three types of resolution: (1) the same scale as the predicted variables (same-scale data); (2) a scale that is finer than the predicted variable (fine-scale data); (3) a scale that is coarser than the predicted variable (coarse-scale data). They designed an input module to integrate the data from heterogeneous data sources with the above-described scales as a vector and treat the vector as the input of a hierarchical RNN model. Specifically, at each time step (at the same resolution as the target variables), the fine-scale data are encoded as a vector representation by an encoder structure based on an RNN and then concatenated with the same-scale data and coarse-scale data to give an integrated vector.

⁹<https://www.iqvia.com/solutions/real-world-evidence/real-world-data-and-insights>. Accessed March 25, 2023

¹⁰<https://github.com/CSSEGISandData/COVID-19>. Accessed March 25, 2023

3.1.3 Data latency

Traditional surveillance data (e.g., reports of case numbers from government institutions) on infectious diseases usually take several weeks to be published because of the processes of data collection, organization, and revision [28, 38]. One of the ways used to manage this challenge in recent studies of real-time prediction approaches is to use surrogate or proxy data with high timeliness [28], such as online search data (e.g., Google Trends¹¹), social media posts (e.g., Twitter) [110], news (e.g., influenza-related news stories), and meteorological data (e.g., temperature, rainfall, and humidity), to make predictions at the same time. Another way is to use a designed model structure to deal with the latency within data. For instance, Gao et al. designed two attention mechanisms: S-LAtt and T-LAtt [38]. In the S-LAtt module, the spatial embedding of updated data with latency is fused with the real-time data and its embedding. To obtain the embeddings of real-time data and updated data, they designed a dual GAT to learn these embeddings separately. To obtain the embedding of updated data, the attention weights are calculated by considering the spatial similarity and the degressive marginal influence of latency. In the T-LAtt module, the temporal embeddings from real-time data and updated data are fused. When the embedding of updated data is calculated, the learning of attention weights is constrained by the time latency between the real-time data and the updated data.

3.2 Task challenges

In addition to the challenges associated with the heterogeneity of data, the task of infectious disease risk prediction faces many other challenges related to modeling disease transmission. These include how to capture the spatiotemporal patterns of disease transmission, how to cope with the problems of distribution shift and uncertainty, and how model structures and predictions should be interpreted. In this section, we summarize several concerns about task-related challenges and introduce how they are addressed by various models.

3.2.1 Characterizing spatiotemporal patterns

Characterizing spatial and temporal patterns is one of the most important tasks required for determining infectious disease propagation. Herein, we discuss how temporal and spatial patterns are incorporated, respectively, into models in previous studies.

Temporal modeling Naturally, current approaches assume that disease dynamics are related to historical dynamics, and this assumption is obviously consistent with the course of disease transmission. Some models assume that the dynamics at the current time step develop from the dynamics at the immediately previous time step, such as the ODEs of the SIR model and other similar compartmental models [71, 73], or that they are related to the dynamics of multiple time steps ahead, due to the existence of disease incubation, such as autoregressive-type models [19, 20], and many deep learning models based on RNN modules [31, 30, 34, 36, 37, 41, 38], CNN modules [35, 40], or GNN modules [33]. Specifically, unlike traditional RNN modules, which can naturally consider multiple time steps, some CNN-based deep learning models capture the temporal patterns at different time scales by

¹¹<https://trends.google.com/trends/>. Accessed March 25, 2023

conducting dilated convolution across time [35] or encode temporal patterns in various time windows by using kernels of different sizes [40]. There is also a GNN-based model [33] that uses the GFT to capture the inter-series temporal dependency in the spectral domain, and uses the DFT to learn the intra-series in the frequency domain.

Spatial modeling Similarly, the spatial pattern is also a necessary characteristic in the description of disease propagation. Some studies have focused on the modeling of disease transmission in a specific location, such as compartmental models at the population level [71, 73] and machine learning models for single locations [31, 30, 27], whereas many other studies have predicted disease dynamics at multiples locations simultaneously. Some of these studies have assumed that spatial patterns can be reflected by the statistical correlation of data between different locations; that is, locations with similar feature conditions have similar disease severities. For instance, the GP regression model [26] uses designed kernels to calculate the covariance matrix, which reflects spatial similarity. Other studies have modeled disease transmission between locations by assuming that a disease transmission network exists, such as compartmental models at the meta-population level [86] and the individual level [89], a multivariate regression model [22, 23], and deep learning models based on a CNN module [34] or a GNN module [32, 35, 36, 38]. The nodes of disease transmission networks typically consist of locations or individuals of interest, whereas edges consist of the interactions or relationships between locations or individuals. Although network-based methods aim to achieve the same goal—i.e., generate epidemic predictions for multiple locations by considering their relationships—they have different focuses on adopted network hypotheses, based on which they can be distinguished. However, although such methods assume that a disease transmission network exists, the interactions in a transmission network cannot be observed directly in the real world, which means there is no ground truth for a disease transmission network. Therefore, some ODE-based models construct a transmission network based on physical contact patterns, and some machine learning models infer network structure from data during model optimization. In doing so, various hypotheses are made regarding network properties to approximate the real situation. In the following, we introduce three types of network structure hypotheses that have often been used in previous studies.

The first common type of setting of network properties is whether edges in a network are undirected or directed, where an undirected edge and a directed edge denote a symmetrical and an asymmetrical correlation or dependency between different locations, respectively. In studies that have assumed that a network is undirected (e.g., [38]), some symmetrical similarity measurements of features have been considered as the quantitative relations between different locations. This implies that the disease dynamic of locations with similar conditions can be used to improve the prediction in a target location. In other studies, a network has been assumed to be directed [22, 23, 34, 32, 35, 36] which implies that there are asymmetrical interactions between different locations. Some of these studies have incorporated prior knowledge and data on the drivers of disease transmission, such as human mobility, which causes cross-regional disease transmission, to assist with network inference [22, 32]. Other studies have directly inferred a transmission network from disease dynamic data [23, 34, 24, 35, 36].

The second common type of setting of network properties is whether the network structure is static or dynamic. Some studies have used a fixed network to predict disease risk [34, 23, 22], which means that after a model is trained on historical data, an inferred transmission network remains constant during the whole prediction. However, in the real world, the

structure of a disease transmission network changes over time, due to changes in risk-related factors (e.g., climate, socio-economic indicators, and human behaviors) and interactions between these factors. Therefore, many studies have tried to characterize dynamic disease transmission networks [32, 62].

In addition to basic network properties, more advanced network characteristics such as the power law of degree distribution are ubiquitous in real-world networks, such as the World Wide Web, worldwide airline networks, and interurban commuting networks [111, 112, 113]. Some studies have used this power law to constrain network structure learning for disease transmission. Specifically, Wang et al. [58] adopted a metapopulation SIR model to model disease propagation between different locations and thereby construct an infection network. That is, by assuming that the node degrees of the disease transmission network followed the power-law distribution, they formalized the prior probability of the whole network and incorporated it into their Bayesian framework of network inference. However, in recent studies that have inferred disease transmission networks, this aspect has not been fully explored.

3.2.2 Distribution shift

Generally, machine learning methods that are trained based on empirical risk minimization face an inherent issue: the generalization ability. A model requires good ability of generalization to make accurate predictions when receiving inputs that it has never seen. However, machine learning models for epidemic prediction struggle to achieve good generalization ability. Moreover, as epidemic trends can change quickly in a short period due to complex interactions between multiple factors, such as intervention strategies and climate conditions, the problem of distribution shift arises [114, 115]. A few studies have examined distribution shifts as part of the topic of epidemic prediction. Wang et al. [57] investigated two distribution shift scenarios: data distribution shift and parameter distribution shift. For each scenario, they studied interpolation and extrapolation tasks via machine learning. The extrapolation task can be regarded as model learning with distribution shift, which means that the distribution of the data or system parameters that need to be predicted is different from the distribution of the data or system parameters that are used for model training. The interpolation task is associated with a situation without distribution shift, which most current machine learning models can handle well. Wang et al. [57] showed that physics-based mechanistic models outperform deep learning models in both of the above-mentioned scenarios, which suggests that it is possible to improve the generalization ability of deep learning models by introducing the inductive bias of mechanism-based models.

3.2.3 Uncertainty quantification

As epidemic predictions are closely related to the development and establishment of public-health intervention strategies, predictions must be both accurate and reliable to enable decision-makers to make good decisions. Usually, point estimation is used to represent a model's output and assess the model's accuracy. However, although calculating the errors between point estimations and observed data is a good way to determine a model's performance, it is insufficient to enable the development of good intervention strategies. That is, when applying epidemic prediction models in the real world, flawed data, incomplete understanding of disease transmission, unknown future potential changes, and even model design bring significant uncertainty into the results and model parameterization [116]. Therefore, many studies

have generated interval estimates, which provide not only estimated values but also their confidence intervals for model outputs or model parameters.

Generally, uncertainty quantification methods can be classified into two categories: intrinsic and extrinsic methods [117]. Intrinsic methods generate predictions and uncertainty estimates simultaneously. Extrinsic methods train auxiliary or meta-models to give confidence estimates in a post-hoc manner. Current models for epidemic prediction tend to generate uncertainty measurements in an intrinsic way. Based on the ways in which they incorporate uncertainty, models can be divided into two classes: Bayesian learning-based models and ensembling models.

Probabilistic modeling The first category of model is based on probabilistic modeling, which takes uncertainty into account by incorporating the probabilistic distribution of parameters or functions. Usually, these models first assign a prior distribution for targeted variables, and then use Bayes’ theorem to calculate its posterior distribution. This category can be subdivided into two classes. The first class assumes that model parameters follow a probabilistic distribution. For instance, the empirical Bayes framework proposed by Brooks et al. to predict ILI trends uses historical data to estimate the prior distribution of model parameters and produces the posterior distribution of epidemic curves [25]. Other Bayesian inference methods, including the Kalman filtering method and its variants [92], and PFs [93], have also been used in disease dynamic prediction [11, 54, 51, 52, 55, 53]. Rather than assuming that the model parameters follow a probability distribution, stochastic process-based models (e.g., the GP model) define the probability distribution over functions [43]. Some representative studies have used the GP model to predict disease dynamics and provide the uncertainty of results. For instance, Senanayake et al. used the Gaussian process regression model with designed spatiotemporal kernels to model the spatiotemporal characteristics of influenza transmission and used variational inference to optimize the model to adapt to a large dataset [26]; Zimmer and Yaesoubi trained an independent GP model for each location to capture the correlation between historical data in the previous season and predict the dynamics t weeks ahead based on available data for the current year [27]. Furthermore, some studies have used deep learning models to perform stochastic processes; this type of model is called **neural processes** (NPs) [118]. Its extensions, such as the **functional neural process** (FNP) [119] and the **recurrent neural process** (RNP) [120, 121], have also been developed to capture complex dependency. For instance, the EPIFNP is a neural process model that incorporates RNP and FNP modules for disease dynamics prediction [37]. It uses a probabilistic deep-sequence model to encode the sequence data of influenza outbreaks as latent variables, and uses a stochastic correlation graph to learn the correlations between these variables. Finally, it uses an MLP module to parameterize the predictive distribution of model outcomes.

Model ensembling Another way is to use ensemble approaches, which collect a set of predicted values from multiple trained models and calculate their probabilistic distributions to enhance the robustness of predictions. For instance, the COVID-19 Forecast Hub ¹² ensembles the forecasts from different models from various institutions to generate US COVID-19 death data [122]; and [123] evaluates the uncertainty from data by using bootstrapping, which is a data sampling technique, to re-sample an entire training dataset to multiple subsets and

¹²<https://covid19forecasthub.org>. Accessed March 25, 2023

then uses the subsets to train a set of models. Based on the trained models, multiple prediction results are produced and can be used to calculate the confidence interval.

3.2.4 Model interpretability

In addition to the broad practical applications of deep learning models in other domains closely related to human well-being, such as healthcare [124], the interpretability of deep learning models has been examined, and its rigorous definition and associated challenges have been widely discussed [125, 126]. Moreover, the interpretability of epidemic prediction models is pivotal, because inappropriate interpretation of models and incomplete understanding of results may lead to unsound decisions, which could adversely affect human well-being and waste precious anti-epidemic resources. To avoid this kind of loss as much possible, researchers need to be cautious when they interpret models and obtain results.

Two widely known and intuitive interpretable models for disease transmission are the classic mechanism-based models (e.g., the SIR model [71] and the SEIR model [73, 74]), which explicitly describe and simulate the disease transmission process by using ODEs in which the relationships between various variables are mathematically defined.

Recently, some machine learning models have also explored interpretability. In particular, some models incorporate machine learning methods to infer the epidemiological parameters of compartmental models [59], whereas other models use a linear model structure, such as AR and MA-based models [19, 20], which assume that a prediction is the weighted sum of historical dynamics. Deep learning models are usually treated as black-box models because the relationships between input and output are highly non-linear and are implicitly encoded by the model structure and learned parameters. However, recently, many researchers have explored the possibility of incorporating explainable elements into deep learning model structures. Thus, some studies have used the similarity of time series to explain predictions. For example, [39] assumed that the current (to be predicted) season is similar to some historical seasons and that this similarity can be used to aid the prediction of the incidence curve of the current season. Based on this assumption, the deep learning modules are first used to learn the similarity between historical trends by clustering and then the incomplete data of the current season are mapped to the closest historical season in the latent space. The approach in [37] is based on similar assumptions and uses a functional neural process module to learn the correlation between the predicted season and past seasons.

3.3 Evaluation challenges

When constructing models and evaluating their performance, many different types of outcomes and evaluation measurements can be involved, depending on data availability and practical needs. How to identify and use appropriate measurements is thus a challenge to the modeling task. In the following section, we summarize the outcomes of prediction models and measurements that have been widely used to evaluate these outcomes.

3.3.1 Measurements of disease severity

When constructing an epidemic prediction model, one of the most important tasks is to determine the model outcome, which is usually a measurement of disease severity in the target population. In general, the choice of predicted variables is based on the goals of pub-

lic health policy and on data availability. Various indicators of disease severity have been used. The most commonly used indicators include disease incidence [68, 123], case numbers [23, 22, 26, 32, 40, 35, 34, 36], death counts [32, 40, 36], patient visit counts related to the disease [35, 37], and disease activity levels [35, 34]. In addition, some specialized indicators have been used to describe the seasonal outbreak of influenza, such as peak intensity, peak time, final epidemic size, onset time, and duration of outbreaks [55, 27, 25].

3.3.2 Evaluation of model performance

The above introduction to previous studies shows that some models generate point estimations that can be directly compared with observed data. Some researchers have considered uncertainty in their model designs, and have therefore presented their predictions in interval/quantile-based format due to the requirements of practical use (e.g., [127]). This accounts for the different evaluation methods that have been used for these two kinds of output formats: point estimation and interval estimation.

Point estimation The most common methods used to evaluate the accuracy of point estimation include the **root-mean-square error** (RMSE), as shown in Eq.(13); the **mean absolute error** (MAE), as shown in Eq.(14); the **mean absolute percentage error** (MAPE), as shown in Eq.(15); and the **root mean squared percent error** (RMSPE), as shown in Eq.(16). These indicators all calculate the deviation of predicted values Y^* from ground truth Y , where Y_t^* and Y_t denote the predicted value and the ground truth at time step t , respectively. The **Pearson correlation coefficient** (CORR), as shown in Eq.(17), is used to evaluate the correlation between predicted trend Y^* and real trend Y ; where \bar{Y}^* and \bar{Y} denote the mean value of the predicted trend and the real trend, respectively, during the time slot from 1 to T .

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (Y_t - Y_t^*)^2} \quad (13)$$

$$MAE = \frac{1}{T} \sum_{t=1}^T |Y_t - Y_t^*| \quad (14)$$

$$MAPE = \left(\max_t \frac{|Y_t - Y_t^*|}{Y_t} \right) * 100 \quad (15)$$

$$RMSPE = \sqrt{\frac{1}{T} \sum_{t=1}^T \left(\frac{Y_t - Y_t^*}{Y_t} \right)^2} \quad (16)$$

$$CORR = \frac{\sum_{t=1}^T (Y_t - \bar{Y})(Y_t^* - \bar{Y}^*)}{\sqrt{\sum_{t=1}^T (Y_t - \bar{Y})^2} \sqrt{\sum_{t=1}^T (Y_t^* - \bar{Y}^*)^2}} \quad (17)$$

Interval estimation Some indicators are widely used for interval estimation [127]. For instance, prediction interval coverage, denoted as $k_M(c)$, calculates the percentage of observed values falling into the c (i.e., 50% or 95%) confidence interval of predicted distributions of M [122]; calibration score is the integral of $\|k_M(c) - c\|$ over c from 0, \dots , 1, as shown in Eq.(18) [37], and a calibration plot shows the relationship between c and $k_M(c)$ [37]. Another indicator is the logarithmic score, also called the log-score, which is used in the CDC's

influenza prediction Challenge in the US. It is calculated as follows: given the predicted distribution of the outcome, first calculate the sum of probability of bins within a given interval around the true value, and then determine the natural logarithm of the calculated sum, which represents the final score [128, 27].

$$CS(M) = \int_0^1 \|k_M(c) - c\| \approx 0.01 \sum_{c \in \{0, 0.01, \dots, 1\}} \|k_M(c) - c\| \quad (18)$$

4 Conclusions and future directions

In this paper, we review the development of machine learning in the realm of infectious disease risk prediction. First, we introduce previous studies that have explored three types of approaches to risk prediction: statistical prediction, data-driven machine learning, and epidemiology-inspired machine learning. In each category, we depict the relationship and differences between different approaches by subdividing them according to their structural characteristics. In addition, we summarize common challenges encountered when dealing with inputs, designing prediction algorithms, and conducting performance evaluation, and we also discuss several related challenges and provide examples of how previous studies have coped with these challenges.

In research on the fusion of epidemiological models with data-driven models to aid in disease risk prediction, some directions that have been exploited are worthy of further exploration. With the increasing use of approaches that employ sophisticated deep learning structures to infer the spatiotemporal patterns of disease transmission, the problems of model overfitting and poor generalization are highlighted. Many studies of epidemiology-inspired machine learning have attempted to combine disease transmission mechanisms based on domain knowledge with data-driven models to guide the structure construction of models, regularize model parameters, and constrain model learning, and thereby generate rational model structures and meaningful results.

Specifically, epidemiological parameterization and epidemiology-embedded learning both endow models with the power of expert knowledge and data but in different ways. Currently, most models primarily comprise a white-box structure (such as epidemiological models) or a black-box structure (such as deep learning models), with the corresponding black- or white-box structure playing a subsidiary role. Therefore, it remains to be explored how to more naturally and intrinsically merge these two types of structures to enhance the interpretability of models and results.

Another important point is that the mechanism that guides data-driven machine learning models is not and should not be limited to domain knowledge of the biological process in disease transmission but should rather be based on prior knowledge from multidisciplinary fields. For example, when modeling a disease transmission network, network science theory can be applied to constrain model learning to force the learned patterns to follow physical laws. Similarly, in learning settings, the constraints of network properties for a specific network can be linked to the network properties that are empirically validated by real networks. However, there has been very little work in this direction. Therefore, improving the learning and optimization of disease transmission networks by incorporating more network properties

than have previously been incorporated could be a promising future direction. Furthermore, because disease transmission networks are unobservable, analyzing the traits of inferred network structures with tools from the field of complex network research to verify these structures' rationality and provide feedback into the model could be another promising future direction.

References

- [1] World Health Organization et al. *World Malaria Report 2022*. World Health Organization, 2022.
- [2] Indranil Chakraborty and Prasenjit Maity. Covid-19 outbreak: Migration, effects on society, global environment and prevention. *Science of the Total Environment*, 728:138882, 2020.
- [3] Said Agrebi and Anis Larbi. Use of artificial intelligence in infectious diseases. In *Artificial Intelligence in Precision Health*, pages 415–438. Elsevier, 2020.
- [4] Omar Enzo Santangelo, Vito Gentile, Stefano Pizzo, Domiziana Giordano, and Fabrizio Cedrone. Machine learning and prediction of infectious diseases: A systematic review. *Machine Learning and Knowledge Extraction*, 5(1):175–198, 2023.
- [5] Barbara A Han and John M Drake. Future directions in analytics for infectious disease intelligence: Toward an integrated warning system for emerging pathogens. *EMBO Reports*, 17(6):785–789, 2016.
- [6] Stephen W Attwood, Sarah C Hill, David M Aanensen, Thomas R Connor, and Oliver G Pybus. Phylogenetic and phylodynamic approaches to understanding and combating the early sars-cov-2 pandemic. *Nature Reviews Genetics*, 23(9):547–562, 2022.
- [7] Matthew Biggerstaff, David Alper, Mark Dredze, Spencer Fox, Isaac Chun-Hai Fung, Kyle S Hickmann, Bryan Lewis, Roni Rosenfeld, Jeffrey Shaman, Ming-Hsiang Tsou, et al. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. *BMC Infectious Diseases*, 16(1):1–10, 2016.
- [8] Kiesha Prem, Yang Liu, Timothy W Russell, Adam J Kucharski, Rosalind M Eggo, Nicholas Davies, Stefan Flasche, Samuel Clifford, Carl AB Pearson, James D Munday, et al. The effect of control strategies to reduce social mixing on outcomes of the covid-19 epidemic in wuhan, china: a modelling study. *The Lancet Public Health*, 5(5):e261–e270, 2020.
- [9] Chelsea S Lutz, Mimi P Huynh, Monica Schroeder, Sophia Anyatonwu, F Scott Dahlgren, Gregory Danyluk, Danielle Fernandez, Sharon K Greene, Nodar Kipshidze, Leann Liu, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19(1):1–12, 2019.

- [10] Nakul Chitnis, Allan Schpira, David Smith, Simon I Hay, Thomas Smith, Richard Steketee, et al. *Mathematical Modelling to Support Malaria Control and Elimination*. World Health Organization, 2010.
- [11] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.
- [12] Nicholas C Grassly and Christophe Fraser. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487, 2008.
- [13] Sandip Mandal, Ram Rup Sarkar, and Somdatta Sinha. Mathematical models of malaria-a review. *Malaria Journal*, 10(1):1–19, 2011.
- [14] Mathieu Andraud, Niel Hens, Christiaan Marais, and Philippe Beutels. Dynamic epidemiological models for dengue transmission: a systematic review of structural approaches. *PloS One*, 7(11):e49085, 2012.
- [15] Elaine O Nsoesie, John S Brownstein, Naren Ramakrishnan, and Madhav V Marathe. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and Other Respiratory Viruses*, 8(3):309–316, 2014.
- [16] Jean-Paul Chretien, Dylan George, Jeffrey Shaman, Rohit A Chitale, and F Ellis McKenzie. Influenza forecasting in human populations: a scoping review. *PloS One*, 9(4):e94130, 2014.
- [17] Aniruddha Adiga, Devdatt Dubhashi, Bryan Lewis, Madhav Marathe, Srinivasan Venkatramanan, and Anil Vullikanti. Mathematical models for covid-19 pandemic: a comparative analysis. *Journal of the Indian Institute of Science*, 100(4):793–807, 2020.
- [18] J Christopher Clement, VijayaKumar Ponnusamy, KC Sriharipriya, and R Nandakumar. A survey on mathematical, machine learning and deep learning models for covid-19 transmission and diagnosis. *IEEE Reviews in Biomedical Engineering*, 15:325–340, 2021.
- [19] Marcel Dettling. Applied time series analysis. lecture notes. *Zurich: Zurich University of Applied Sciences*, 2013.
- [20] Helmut Lütkepohl and Markus Krätzig. *Applied Time Series Econometrics*. Cambridge university press, 2004.
- [21] Zheng Wang, Prithwish Chakraborty, Sumiko R Mekaru, John S Brownstein, Jieping Ye, and Naren Ramakrishnan. Dynamic poisson autoregression for influenza-like-illness case count prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2015.
- [22] Yu Zhang, William K Cheung, and Jiming Liu. A unified framework for epidemic prediction based on poisson regression. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2878–2892, 2015.

- [23] Hongbin Pei, Bo Yang, Jiming Liu, and Lei Dong. Group sparse bayesian learning for active surveillance on epidemic dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 800–807, 2018.
- [24] Hongbin Pei, Bo Yang, Jiming Liu, and Kevin Chen-Chuan Chang. Active surveillance via group sparse bayesian learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1133–1148, 2022.
- [25] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. Flexible modeling of epidemics with an empirical bayes framework. *PLoS Computational Biology*, 11(8):e1004382, 2015.
- [26] Ransalu Senanayake, Simon O’Callaghan, and Fabio Ramos. Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3901–3907, 2016.
- [27] Christoph Zimmer and Reza Yaesoubi. Influenza forecasting framework based on gaussian processes. In *Proceedings of International Conference on Machine Learning*, pages 11671–11679, 2020.
- [28] Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O Nsoesie, Sumiko R Mekaru, John S Brownstein, Madhav V Marathe, et al. Forecasting a moving target: Ensemble models for ili case count predictions. In *Proceedings of the SIAM International Conference on Data Mining*, pages 262–270, 2014.
- [29] Qi Tan, Yang Liu, and Jiming Liu. Demystifying deep learning in predictive spatiotemporal analytics: An information-theoretic framework. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3538–3552, 2020.
- [30] Siva R Venna, Amirhossein Tavaneai, Raju N Gottumukkala, Vijay V Raghavan, Anthony S Maida, and Stephen Nichols. A novel data-driven model for real-time influenza forecasting. *IEEE Access*, 7:7691–7701, 2018.
- [31] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS One*, 12(12):e0188941, 2017.
- [32] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O’Banion. Examining covid-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113*, 2020.
- [33] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems*, 33:17766–17778, 2020.
- [34] Yuexin Wu, Yiming Yang, Hiroshi Nishiura, and Masaya Saitoh. Deep learning for epidemiological predictions. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1085–1088, 2018.

- [35] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pages 245–254, 2020.
- [36] Shun Zheng, Zhifeng Gao, Wei Cao, Jiang Bian, and Tie-Yan Liu. Hierst: A unified hierarchical spatial-temporal framework for covid-19 trend forecasting. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pages 4383–4392, 2021.
- [37] Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B Aditya Prakash. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. *Advances in Neural Information Processing Systems*, 34:19796–19807, 2021.
- [38] Junyi Gao, Cao Xiao, Lucas M Glass, and Jimeng Sun. Popnet: Real-time population-level disease prediction with data latency. In *Proceedings of the ACM Web Conference*, pages 2552–2562, 2022.
- [39] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B Aditya Prakash. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 577–586, 2019.
- [40] Yue Cui, Chen Zhu, Guanyu Ye, Ziwei Wang, and Kai Zheng. Into the unobservables: A multi-range encoder-decoder framework for covid-19 prediction. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pages 292–301, 2021.
- [41] Alexander Rodríguez, Nikhil Muralidhar, Bijaya Adhikari, Anika Tabassum, Naren Ramakrishnan, and B Aditya Prakash. Steering a historical disease forecasting model under a pandemic: Case of flu and covid-19. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4855–4863, 2021.
- [42] Kevin P Murphy. *Probabilistic Machine Learning: an Introduction*. MIT press, 2022.
- [43] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [44] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
- [45] Badrul M Sarwar, George Karypis, Joseph Konstan, and John Riedl. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the International Conference on Computer and Information Technology*, volume 1, pages 291–324, 2002.
- [46] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

- [47] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.
- [48] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [49] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 1753–1759, 2017.
- [50] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.
- [51] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, 4(1):1–10, 2013.
- [52] Wan Yang, Marc Lipsitch, and Jeffrey Shaman. Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*, 112(9):2723–2728, 2015.
- [53] Sen Pei, Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Forecasting the spatial transmission of influenza in the united states. *Proceedings of the National Academy of Sciences*, 115(11):2752–2757, 2018.
- [54] Michele Tizzoni, Paolo Bajardi, Chiara Poletto, José J Ramasco, Duygu Balcan, Bruno Gonçalves, Nicola Perra, Vittoria Colizza, and Alessandro Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC Medicine*, 10(1):1–31, 2012.
- [55] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the International Conference on World Wide Web*, pages 311–319, 2017.
- [56] Difan Zou, Lingxiao Wang, Pan Xu, Jinghui Chen, Weitong Zhang, and Quanquan Gu. Epidemic model guided machine learning for covid-19 forecasts in the united states. *medRxiv*, 2020.
- [57] Rui Wang, Danielle Maddix, Christos Faloutsos, Yuyang Wang, and Rose Yu. Bridging physics-based and data-driven modeling for learning dynamical systems. In *Learning for Dynamics and Control*, pages 385–398. PMLR, 2021.
- [58] Jingyuan Wang, Xiaojian Wang, and Junjie Wu. Inferring metapopulation propagation network for intra-city epidemic control and prevention. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 830–838, 2018.

- [59] Sercan Arik, Chun-Liang Li, Jinsung Yoon, Rajarishi Sinha, Arkady Epshteyn, Long Le, Vikas Menon, Shashank Singh, Leyou Zhang, Martin Nikoltchev, et al. Interpretable sequence learning for covid-19 forecasting. *Advances in Neural Information Processing Systems*, 33:18807–18818, 2020.
- [60] Jackie Baek, Vivek F Farias, Andreea Georgescu, Retsef Levi, Tianyi Peng, Deeksha Sinha, Joshua Wilde, and Andrew Zheng. The limits to learning an sir process: granular forecasting for covid-19. *arXiv preprint arXiv:2006.06373*, 2020.
- [61] Benyun Shi, Shan Lin, Qi Tan, Jie Cao, Xiaohong Zhou, Shang Xia, Xiao-Nong Zhou, and Jiming Liu. Inference and prediction of malaria transmission dynamics using time series data. *Infectious Diseases of Poverty*, 9(1):1–13, 2020.
- [62] Mutong Liu, Yang Liu, Ly Po, Shang Xia, Rekol Huy, Xiao-Nong Zhou, and Jiming Liu. Assessing the spatiotemporal malaria transmission intensity with heterogeneous risk factors: A modeling study in cambodia. *Infectious Disease Modelling*, 8(1):253–269, 2023.
- [63] Ting Hua, Chandan K Reddy, Lei Zhang, Lijing Wang, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. Social media based simulation models for understanding disease dynamics. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3797–3804, 2018.
- [64] Nikos Kargas, Cheng Qian, Nicholas D Sidiropoulos, Cao Xiao, Lucas M Glass, and Jimeng Sun. Stelar: Spatio-temporal tensor factorization with latent epidemiological regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4830–4837, 2021.
- [65] Dave Osthus, James Gattiker, Reid Priedhorsky, and Sara Y Del Valle. Dynamic bayesian influenza forecasting in the united states with hierarchical discrepancy (with discussion). *Bayesian Analysis*, 14(1):261–312, 2019.
- [66] Junyi Gao, Rakshith Sharma, Cheng Qian, Lucas M Glass, Jeffrey Spaeder, Justin Romberg, Jimeng Sun, and Cao Xiao. Stan: spatio-temporal attention network for pandemic prediction using real-world evidence. *Journal of the American Medical Informatics Association*, 28(4):733–743, 2021.
- [67] Lijing Wang, Aniruddha Adiga, Jiangzhuo Chen, Adam Sadilek, Srinivasan Venkattraman, and Madhav Marathe. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12191–12199, 2022.
- [68] Lijing Wang, Jiangzhuo Chen, and Madhav Marathe. Defsi: Deep learning based epidemic forecasting with synthetic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9607–9612, 2019.
- [69] Lijing Wang, Jiangzhuo Chen, and Madhav Marathe. Tdefsi: Theory-guided deep learning-based epidemic forecasting with synthetic information. *ACM Transactions on Spatial Algorithms and Systems*, pages 1–39, 2020.

- [70] Jiming Liu and X Shang. *Computational Epidemiology*. Springer, 2020.
- [71] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical*, 115(772):700–721, 1927.
- [72] Juliana Tolles and ThaiBinh Luong. Modeling epidemics with compartmental models. *The Journal of the American Medical Association*, 323(24):2515–2516, 2020.
- [73] Herbert W Hethcote and P Van den Driessche. Some epidemiological models with nonlinear incidence. *Journal of Mathematical Biology*, 29(3):271–287, 1991.
- [74] Michael Y Li and James S Muldowney. Global stability for the seir model in epidemiology. *Mathematical Biosciences*, 125(2):155–164, 1995.
- [75] Andrea Maugeri, Martina Barchitta, Sebastiano Battiato, and Antonella Agodi. Estimation of unreported novel coronavirus (sars-cov-2) infections from reported deaths: a susceptible–exposed–infectious–recovered–dead model. *Journal of Clinical Medicine*, 9(5):1350, 2020.
- [76] Elena Loli Piccolomini and Fabiana Zama. Preliminary analysis of covid-19 spread in italy with an adaptive seird model. *arXiv preprint arXiv:2003.09909*, 2020.
- [77] Ronald Ross. *Report on the Prevention of Malaria in Mauritius*. J. & A. Churchill, 1908.
- [78] George Macdonald et al. *The Epidemiology and Control of Malaria*. Oxford University Press, London, 1957.
- [79] Roy M Anderson and Robert M May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, London, 1992.
- [80] Hyun M Yang and Marcelo U Ferreira. Assessing the effects of global warming and local social and economic conditions on the malaria transmission. *Revista De Saude Publica*, 34(3):214–222, 2000.
- [81] Hyun M Yang. Malaria transmission model for different levels of acquired immunity and temperature-dependent parameters (vector). *Revista De Saude Publica*, 34(3):223–231, 2000.
- [82] Carlos Castillo-Chavez, Herbert W Hethcote, Viggo Andreasen, Simon A Levin, and Wei M Liu. Epidemiological models with age structure, proportionate mixing, and cross-immunity. *Journal of Mathematical Biology*, 27(3):233–258, 1989.
- [83] João A N Filipe, Eleanor M Riley, Christopher J Drakeley, Colin J Sutherland, and Azra C Ghani. Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. *PLoS Computational Biology*, 3(12):e255, 2007.

- [84] Chris Cosner, John C Beier, Robert Stephen Cantrell, D Impoinvil, Lev Kapitanski, Matthew David Potts, A Troyo, and Shigui Ruan. The effects of human movement on the persistence of vector-borne diseases. *Journal of Theoretical Biology*, 258(4):550–560, 2009.
- [85] Olivia Prosper, Nick Ruktanonchai, and Maia Martcheva. Assessing the role of spatial heterogeneity and human movement in malaria dynamics and control. *Journal of Theoretical Biology*, 303:1–14, 2012.
- [86] Sen Pei and Jeffrey Shaman. Initial simulation of sars-cov2 spread and intervention effects in the continental us. *MedRxiv*, 2020.
- [87] John A Jacquez, Carl P Simon, James Koopman, Lisa Sattenspiel, and Timothy Perry. Modeling and analyzing hiv transmission: the effect of contact patterns. *Mathematical Biosciences*, 92(2):119–199, 1988.
- [88] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.
- [89] Nicolas Hoertel, Martin Blachier, Carlos Blanco, Mark Olfson, Marc Massetti, Marina Sánchez Rico, Frédéric Limosin, and Henri Leleu. A stochastic agent-based model of the sars-cov-2 epidemic in france. *Nature Medicine*, 26(9):1417–1421, 2020.
- [90] Bin Wang, Xiaolei Zou, and Jiang Zhu. Data assimilation and its applications. *Proceedings of the National Academy of Sciences*, 97(21):11143–11144, 2000.
- [91] Greg Welch, Gary Bishop, et al. An introduction to the kalman filter. Technical report, Chapel Hill, NC, USA, 1995.
- [92] Jeffrey L Anderson. An ensemble adjustment kalman filter for data assimilation. *Monthly Weather Review*, 129(12):2884–2903, 2001.
- [93] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.
- [94] Jeffrey Shaman, Virginia E Pitzer, Cécile Viboud, Bryan T Grenfell, and Marc Lipsitch. Absolute humidity and the seasonal onset of influenza in the continental united states. *PLoS Biology*, 8(2):e1000316, 2010.
- [95] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.
- [96] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J Ramasco, Vittoria Colizza, and Alessandro Vespignani. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science*, 1(3):132–145, 2010.

- [97] David L Smith and F Ellis McKenzie. Statics and dynamics of malaria infection in anopheles mosquitoes. *Malaria Journal*, 3(1):1–14, 2004.
- [98] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan AJ Metz. On the definition and the computation of the basic reproduction ratio r_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28:365–382, 1990.
- [99] Pauline Van den Driessche and James Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(1-2):29–48, 2002.
- [100] Anice C Lowen, Samira Mubareka, John Steel, and Peter Palese. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathogens*, 3(10):e151, 2007.
- [101] Jeffrey Shaman and Melvin Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248, 2009.
- [102] James D Tamerius, Jeffrey Shaman, Wladimir J Alonso, Kimberly Bloom-Feshbach, Christopher K Uejio, Andrew Comrie, and Cécile Viboud. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathogens*, 9(3):e1003194, 2013.
- [103] Pietro Ceccato, Christelle Vancutsem, Robert Klaver, James Rowland, and Stephen J Connor. A vectorial capacity product to monitor changing malaria transmission potential in epidemic regions of africa. *Journal of Tropical Medicine*, 2012, 2012.
- [104] Sebastian Funk, Marcel Salathé, and Vincent AA Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of the Royal Society Interface*, 7(50):1247–1256, 2010.
- [105] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.
- [106] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.
- [107] Shihao Yang, Mauricio Santillana, and Samuel C Kou. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.
- [108] Sasikiran Kandula, Daniel Hsu, Jeffrey Shaman, et al. Subregional nowcasts of seasonal influenza using search trends. *Journal of Medical Internet Research*, 19(11):e7486, 2017.

- [109] Qingyu Yuan, Elaine O Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S Brownstein. Monitoring influenza epidemics in china with search query from baidu. *PloS One*, 8(5):e64323, 2013.
- [110] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *Proceedings of the IEEE Conference on Computer Communications Workshops*, pages 702–707, 2011.
- [111] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge university press, 2008.
- [112] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.
- [113] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.
- [114] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- [115] Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.
- [116] Inga Holmdahl and Caroline Buckee. Wrong but useful—what covid-19 epidemiologic models can and cannot tell us. *New England Journal of Medicine*, 383(4):303–305, 2020.
- [117] Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. *arXiv preprint arXiv:2212.07359*, 2022.
- [118] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.
- [119] Christos Louizos, Xiahan Shi, Klamer Schutte, and Max Welling. The functional neural process. *Advances in Neural Information Processing Systems*, 32, 2019.
- [120] Shenghao Qin, Jiacheng Zhu, Jimmy Qin, Wenshuo Wang, and Ding Zhao. Recurrent attentive neural process for sequential data. *arXiv preprint arXiv:1910.09323*, 2019.
- [121] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.
- [122] Evan L Ray, Nutch Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRxiv*, 2020.

- [123] Alexander Rodriguez, Anika Tabassum, Jiaming Cui, Jiajia Xie, Javen Ho, Pulak Agarwal, Bijaya Adhikari, and B Aditya Prakash. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15393–15400, 2021.
- [124] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 559–560, 2018.
- [125] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [126] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.
- [127] Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, 17(2):e1008618, 2021.
- [128] Christoph Zimmer, Sequoia I Leuba, Reza Yaesoubi, and Ted Cohen. Use of daily internet search query data improves real-time projections of influenza epidemics. *Journal of The Royal Society Interface*, 15(147):20180220, 2018.

