Machine Learning for Infectious Disease Risk Prediction: A Survey

MUTONG LIU, Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong
YANG LIU, Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong
JIMING LIU, Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong

Infectious diseases place a heavy burden on public health worldwide. In this article, we systematically investigate how machine learning (ML) can play an essential role in quantitatively characterizing disease transmission patterns and accurately predicting infectious disease risks. First, we introduce the background and motivation for using ML for infectious disease risk prediction. Next, we describe the development and application of various ML models for infectious disease risk prediction, categorizing them according to the models' alignment with vital public health concerns specific to two distinct phases of infectious disease propagation: (1) the pandemic and epidemic phases (the P-E phases) and (2) the endemic and elimination phases (the E-E phases), with each presenting its own set of critical questions. Subsequently, we discuss challenges encountered when dealing with model inputs, designing task-oriented objectives, and conducting performance evaluations. We conclude with a discussion of open questions and future directions.

CCS Concepts: • **Computing methodologies** → **Machine learning approaches**; **Simulation types and techniques**; **Learning paradigms;**

Additional Key Words and Phrases: Machine learning, data-driven modeling, epidemiology-inspired learning, infectious disease risk prediction, transmission dynamics characterization

ACM Reference Format:
Mutong Liu, Yang Liu, and Jiming Liu. 2025. Machine Learning for Infectious Disease Risk Prediction: A Survey. *ACM Comput. Surv.* 57, 8, Article 212 (March 2025), 39 pages. https://doi.org/10.1145/3719663

1 Introduction

The propagation of infectious diseases, whether emergent (e.g., coronavirus disease 2019 (COVID-19), which has caused nearly 7 million deaths worldwide so far[1]) or long-standing (e.g., malaria,

[1]https://covid19.who.int/. Accessed September 02, 2024.

This work was supported in part by the National Science and Technology Major Project under Grant No. 2021ZD0112500, the General Research Fund from the Research Grant Council of Hong Kong SAR under Projects RGC/HKBU12203122 and RGC/HKBU12200124, the NSFC/RGC Joint Research Scheme under Project N_HKBU222/22, and the Guangdong Basic and Applied Basic Research Foundation under Project 2024A1515011837.
Authors' Contact Information: Mutong Liu, Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong; e-mail: csmtliu@comp.hkbu.edu.hk; Yang Liu, Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong; e-mail: csygliu@comp.hkbu.edu.hk; Jiming Liu, Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong; e-mail: jiming@comp.hkbu.edu.hk.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

© 2025 Copyright held by the owner/author(s).
ACM 0360-0300/2025/03-ART212
https://doi.org/10.1145/3719663

which has an ancient history and still causes more than 600 thousand deaths every year [119]), significantly affects human well-being and social development on a global scale [28, 119]. Thus, the battle against infectious disease is never-ending. Humanity's progress in developing countermeasures against various diseases has relied on conceptual innovation and scientific advancements across multiple disciplines. In recent decades, machine learning has proven particularly effective in infectious disease research due to its ability to handle vast and diverse datasets and uncover intrinsic, complex patterns. This proficiency has led to its wide and successful application in critical functions related to understanding and combating the spread of diseases [5, 149]. Among its applications, which range from propagation source identification [92, 145] and individual infection detection and inference [1, 111, 167] to intervention planning [196] and drug-virus and microbe–disease interaction predictions [101, 102], the modeling and prediction of transmission risks are of great importance as they inform public health decisions and shape intervention strategies [107]. Therefore, this survey is specifically dedicated to exploring the role of machine learning in modeling and predicting the transmission risks of infectious diseases.

The strategic emphasis in managing disease transmission risks shifts in accordance with the various phases of infectious disease spread, each phase demanding its tailored public health responses, such as prevention, mitigation, or containment. Consequently, the objectives of disease modeling and the prediction of disease risk are also dynamic, aligning with the specific public health goals related to each phase of the disease's propagation. With reference to [63], the initial or very early stage of infectious disease development is called the "watch phase". During this period, an infectious disease has not yet occurred in human populations but possibly exists in the environment, potentially in close proximity to human habitats. This phase is characterized by vigilant monitoring, wherein machine learning and data-driven models are employed to screen and identify potential hosts, such as wild animals, that could carry pathogens. The primary objective in the watch phase is to leverage these modeling techniques to prevent the transmission of pathogens from natural hosts to humans, ultimately aiming at preventing outbreaks among human communities.

In our article, we focus on the review and discussion of research efforts centered on modeling and predicting transmission risks during the critical phases where pathogens have engaged with human hosts, resulting in confirmed cases of infection. This focus is distinct from analyses of the aforementioned "watch phase". By concentrating on these later phases, our aim is at evaluating the methodologies and insights that directly address the challenges posed once an infectious disease begins to actively spread among humans. In such critical phases where human infection has been established, machine learning can significantly enhance our understanding of transmission dynamics. This understanding is vital for public health authorities to implement appropriate measures [18, 129]. By leveraging data-driven insights, we can inform decision-making processes, guiding the containment of disease spread and mitigating its impact. Predictions of epidemic trends, facilitated by mathematical modeling, data science, and machine learning, enable proactive actions, such as the strategic allocation of resources or the implementation of quarantine measures [33, 107]. Moreover, by retrospectively analyzing disease trends through machine learning models, we can uncover transmission patterns that will empower us to manage future outbreaks more effectively than current ones [156].

The field of infectious disease risk modeling and prediction has evolved significantly, beginning with the construction of epidemiological and statistical models. Now, the discipline has advanced further with the development of machine learning algorithms designed specifically for this purpose. Initially, the focus was on designing models to minimize predictive errors by capturing implicit data dependencies. However, as the practical application of these models has grown, so has the necessity for trustworthy, informative, and interpretable predictions. Recent advancements have seen the integration of epidemiological insights with data-driven techniques, giving rise to

epidemiology-inspired machine learning models. These models not only strive for accuracy but also provide meaningful information that is helpful in disease prevention and control efforts. Our survey will present a comprehensive examination of the aforementioned machine learning technologies, particularly in the context of modeling and predicting transmission risks during the most critical phases of an infectious disease's spread. We aim to showcase how machine learning can be leveraged to equip public health authorities with the insights needed to protect populations against the severe impacts of infectious diseases. In Appendix A, we present an overview of existing literature reviews related to modeling infectious disease risk, highlighting their focuses and taxonomies.

1.1 Contributions and Organization

In this article, we introduce a novel perspective for categorizing the literature on infectious disease risk modeling and prediction. Diverging from most of the existing reviews that sort studies by their computational model types, such as mathematical, statistical, or machine learning methodologies, our categorization is grounded in an analysis of how different methods address two pivotal questions in public health, tailored to the distinct phases of infectious disease propagation: (1) **the pandemic and epidemic phases** (**the P-E phases**) and (2) **the endemic and elimination phases** (**the E-E phases**). During the P-E phases, the priority is to model and extract the intrinsic dependencies from observational and transmission-related data over space and time to predict transmission dynamics. Meanwhile, in the E-E phases, the research focus shifts to leveraging disease/scenario-specific knowledge and heterogeneous risk-related factors for informing predictions of potential risks. Accordingly, we introduce a multi-level framework, as illustrated in Figure 1, for systematically categorizing the existing research:

- At the first level, we differentiate works based on the phases of infectious disease transmission–the P-E phases or E-E phases–acknowledging the distinct research emphases and challenges inherent to each phase.
- At the second level, we further group, analyze, and discuss the literature within each phase, organizing it according to various public health concerns related to risk prediction.
- Finally, for each specific public health concern, we summarize the computational solutions designed to address it, classifying them based on the nature of the methodological approaches.

By systematically categorizing methods according to the phases of infectious disease transmission and associated public health concerns, this survey clarifies the appropriate application of various machine learning techniques and, more crucially, offers strategic guidance for their design and deployment precisely customized to meet the needs of public health challenges. Specifically, our contributions to the machine learning community can be summarized as follows:

- By categorizing existing research according to public health phases of disease spread, we provide unique insights into the evolution of machine learning techniques in response to these challenges. This allows the machine learning community to trace algorithm and framework development beyond traditional reviews.
- We identify and discuss the computational challenges inherent to each phase of disease spread and the corresponding machine learning solutions. This serves as both a repository of problem-specific solutions and a foundation for developing novel methods adaptable to other domains.
- Our multi-level framework encourages machine learning researchers to think beyond traditional boundaries, inspiring the creation of novel machine learning models and techniques that not only address the current challenges in infectious disease prediction but are also adaptable to emerging and future public health crises.

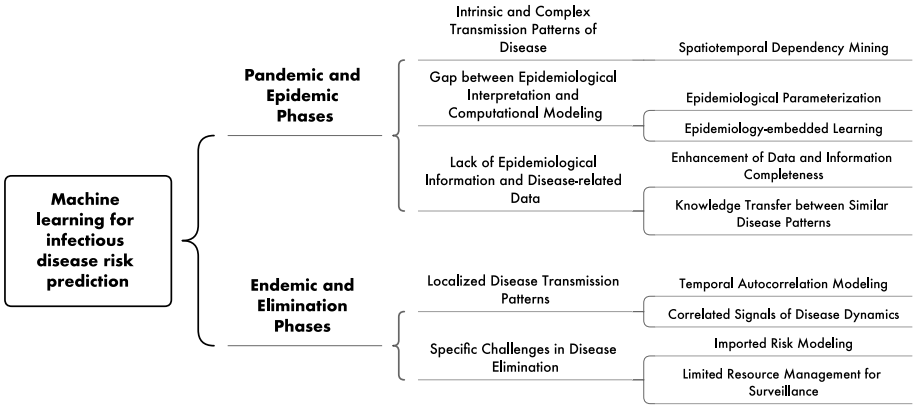

Fig. 1. The taxonomy of machine learning for infectious disease risk prediction. It categorizes existing machine learning research according to the distinct phases of disease propagation: (1) Pandemic and Epidemic Phases (P-E Phases) detailed in Section 2; (2) Endemic and Elimination Phases (E-E Phases) outlined in Section 3.

This article is organized as follows: Section 2 explores machine learning models for the P-E Phases, and Section 3 examines models for the E-E Phases. Both sections review models based on the specific public health concerns and computational strategies within each phase. Section 4 discusses technical challenges in predicting infectious disease transmission risks, including data-related, task-related, and evaluation-related issues, along with techniques to address them. Finally, Section 5 concludes the article and suggests future research directions. Appendix B details our literature search methodology and selection criteria.

2 Machine Learning in Pandemic and Epidemic Phases (P-E Phases)

One influential phase of infectious disease transmission is when it evolves into an epidemic or pandemic, as exemplified by COVID-19, which has had a profound global impact in recent years. During such a phase, the disease spreads rapidly across cities, countries, and continents and potentially affects the entire world. This widespread propagation of disease will result in a significant and dramatic increase in incidence and mortality, leading to substantial losses to society and livelihood, and bringing heavy burdens to public health systems. In this stage, a primary challenge for public health officials is to devise models that can capture and interpret the inherent dependencies within observational and transmission data, which are spatial and temporal in nature, to predict disease dynamics accurately. Addressing this challenge involves tackling multiple issues associated with risk prediction. In this section, we have identified and outlined key concerns for disease risk prediction from a public health perspective: (1) the intrinsic and complex spatiotemporal patterns of disease transmission; (2) the gap between the epidemiological interpretation of the disease and its computational modeling; and (3) the lack of comprehensive epidemiological data and related information on the disease. In the following, we will present various computational strategies tailored to address each of these public health concerns. These solutions will be organized and discussed according to their methodological characteristics. The problem statement of machine learning for infectious disease risk prediction is provided in Appendix E.1.

2.1 Modeling Intrinsic and Complex Transmission Patterns

2.1.1 Spatiotemporal Dependency Mining. Various types of machine learning methods have been developed to model the spatiotemporal patterns of disease transmission and mine complex

Table 1. Brief Summary of Spatiotemporal Dependency Learning Models

Categories		References
Traditional machine learning	Matrix factorization and nearest neighbor	[29]
	Generalized linear models	[194, 215]
	Gaussian process	[155]
Deep learning	RNN	[114, 132, 170, 176, 177]
	GNN	[25, 51, 75]
	Mixed deep modules	[17, 41, 72, 99, 108, 152, 175, 197, 199, 200, 221, 224]
	Encoder-decoder	[3, 74]
Other machine learning techniques	Unsupervised learning	[98]
	Self-supervised learning	[94]
	Multi-task learning	[198]
	Factorization machine	[135]
	Neural autoregressive model	[7]
	Intermediate fusion network	[123]

dependencies from data. Table 1 presents a brief summary of these machine learning methods. We have categorized the body of work into distinct subclasses based on the structural characteristics of the models, which include matrix factorization and nearest neighbor approaches, generalized linear models, Gaussian processes, a variety of deep learning architectures, and other machine learning techniques. We have expanded Table 1 to offer a more detailed summary of each work, which can be found in Table 3 in Appendix F. In the following, we will introduce the specifics of these methodologies, offering a detailed examination and discussion of their applications.

Matrix factorization and nearest neighbor. These models, which are popular in the field of recommender systems [84, 151], have also been utilized to predict disease risks. For instance, Chakraborty et al. proposed **matrix factorization with nearest-neighbor** (**MFN**) regression, which incorporates MF regression and nearest-neighbor-based regression, for **influenza-like illness** (**ILI**) count prediction [29]. 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.

Generalized linear models. Some studies used **generalized linear models** (**GLMs**) to predict disease risks in a single location or multiple locations. A general formulation of the GLMs is provided in Appendix E.2. Although GLMs share a similar regression framework for generating predictions, the specific structures of the developed GLMs vary across different studies. These variations are designed based on various assumptions to accurately reflect specific disease transmission processes. For example, Zhang et al. used the Poisson distribution to model case numbers as integer values and incorporated the effects of intra-regional, inter-regional, and external factors on disease transmission risk into a unified Poisson regression-based framework [215]. The **dynamic Poisson autoregressive model with exogenous** (**DPARX**) inputs variables proposed by Wang et al. [194] also modeled ILI case count as the Poisson distribution. Different from [215], a variant of the standard **autoregressive exogenous** (**ARX**) model with parameters dynamically changing over time was developed to consider the intra-location transmission as the regression of historical data and cross-location transmission as the regression of exogenous variables [194].

Gaussian processes. The **Gaussian process** (**GP**) model has also been utilized to predict infectious disease risks. GP models treat data points as random variables that follow a joint Gaussian distribution. With a designed kernel function, the GP model evaluates the similarity between data points by calculating their covariance matrix. The general formulation of the GP model is provided in Appendix E.3. 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 [155]. 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).

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. The general formulation of the loss function for deep learning models is provided in Appendix E.4. 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. RNNs are widely used to model the temporal dependency of time series data, such as voice or text data. 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 (I^{2}DRNN)** model [170] uses stacked RNN modules to capture spatiotemporal dependencies from heterogeneous and multiple-scale risk-related data. RNN architectures based on a gating mechanism, such as **long short-term memory** (**LSTM**) networks, have also been used in recent studies for disease risk prediction, due to their ability to preserve the long-term information of data sequences [114, 132, 176, 177]. A detailed discussion of these works can be found in Appendix D.1.1.

In contrast to RNN models, which capture the temporal dependency of sequential data, GNN models can deal with data with graphical structures [222]. Given their ability to capture characteristics within graph structures, GNN models are well-suited for representing spatial patterns of disease dynamics, which can be viewed as being driven by a disease transmission network [25, 51, 75]. For example, the **spatio-temporal graph neural network** (**STGNN**) proposed by Kapoor et al. utilizes daily mobility data from Google to construct the structure of time-varying disease transmission networks [75]. Based on the constructed networks, they designed two types of edges, i.e., edges between nodes within the network at the same time, and edges between nodes across the time, to characterize varying spatiotemporal dependencies driven by cross-regional human mobility and the effect of historical risk trends, respectively. Fritz et al. combined a distributional regression model and a GNN model to characterize the structured and unstructured data, respectively, so as to mine the spatiotemporal dependency [51]. 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, such as in the approach developed by Cao et al. [25]. A more detailed description can be found in Appendix D.1.2.

Mixture of deep learning modules. In addition to employing specific types of deep learning modules to characterize disease patterns, recent studies have increasingly applied combinations of multiple neural network structures to model the complex spatiotemporal patterns of disease transmission. 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. Some models use CNN modules to encode spatial dependencies and RNN modules to encode temporal dependencies [197, 200, 200]. For example, Wu et al. proposed a model named CNNRNN-Res, which incorporates CNN, RNN, and residual structures to capture spatiotemporal dependencies in historical disease dynamics [197]. Additionally, other works have constructed models that combine multi-scale convolution modules and the LSTM in parallel, as well as serial integrations of CNN and GRU. These approaches are discussed in more detail in Appendix D.1.3. In recent deep learning architectures with hybrid modules, there has been a notable shift from employing CNNs to capture spatial dependencies to utilizing GNNs, owing to their flexibility in encoding features [17, 41, 72, 99, 108, 152, 199, 221, 224]. For instance, the **cross-location attention-based GNN** (**ColaGNN**) is designed based on CNN, RNN, and GNN for long-term ILI prediction [41]. A more detailed description of several works that mix GNNs with other deep learning modules for infectious disease risk prediction can be found in Appendix D.1.4.

Encoder–decoder framework. In deep learning, the encoder–decoder is a popular architecture to process sequential data, particularly when dealing with inputs and outputs of variable lengths. It can incorporate a variety of deep learning modules, such as CNN, RNN, and attention mechanisms. Initially popularized by its application in machine translation [34], the encoder-decoder architecture has since expanded its applications to other domains, including public health. Adhikari et al. [3] introduced EpiDeep, an approach that integrates an encoder-decoder framework with deep clustering components to predict the **weighted ILI** (**wILI**). 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 [61], 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. Kao et al. proposed two **autoencoder** (**AE**) architectures: (1) **convolutional AE** (**CAE**) and (2) CAE with LSTM, aiming at predicting the spatiotemporal disease risk dynamics [74].

Other machine learning techniques. Combining various deep learning modules is a common strategy for capturing spatial and temporal dependencies in disease risk prediction. Beyond this integration, several machine learning techniques, such as unsupervised learning [98], self-supervised learning [94], online learning and multi-task learning [198], factorization machine [135], autoregressive models with exogenous inputs [7], and intermediate fusion networks [123], have been employed alongside deep learning components to enhance predictive performance. A more detailed description of those works can be found in Appendix D.1.5.

2.2 Bridging Epidemiology with Machine Learning

Although data-driven machine learning approaches capture the spatiotemporal transmission patterns of infectious diseases and improve the accuracy of disease risk prediction, they still struggle to provide insights to facilitate disease control. To overcome this drawback, epidemiological models, also referred to as mechanistic, compartmental, mathematical, or physics-based models in various references, have been revisited and integrated with machine learning methods. For epidemiological models, their parameters and the overall structures have clear epidemiological

Table 2. Brief Summary of Epidemiology-inspired Machine Learning Models

Categories			References
Epidemiological parameterization	Epidemiological parameter inference from data	Data assimilation	[47, 126, 154, 156, 157, 206, 209]
		Monte Carlo maximum likelihood analysis	[172, 212]
		MSE loss with traditional models	[24, 86, 93, 109, 168, 189, 191, 217, 226]
		MSE loss with deep learning	[44, 50, 69, 78, 89, 100, 113, 211, 220]
	Epidemiological parameters modeling	Generalized additive model	[12]
		Mixed effects model	[14, 23]
Epidemiology-embedded learning	Epidemiological mechanism-guided models	VCAP/EIR	[162, 163, 223]
		NGM	[97]
		Meta-population epidemiological models	[128, 182]
	Epidemiological regularization and constraints for optimization	Tensor factorization	[76]
		Bayesian inference	[66, 120, 195]
		RNNs	[20, 181, 184–186]
		Mixed deep modules	[26, 53, 96, 183]
		EINNs	[117, 141, 173]

meanings. However, such structures are typically based on relatively simplified assumptions, and the simulation/prediction of these models is sensitive to the setting of the initial values and epidemiological parameters, 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 article, we denote this type of model as "epidemiology-inspired machine learning" and divide it into two categories: (1) epidemiological parameterization and (2) epidemiology-embedded learning. In the following, we will introduce existing studies of the two aforementioned categories of epidemiology-inspired machine learning and describe how each category combines epidemiological prior knowledge with machine learning methods. A preliminary introduction to epidemiological models can be found in Appendix C.1. Table 2 provides a brief summary and classification of related works on epidemiology-inspired machine learning models. A more detailed summary of each model, including the targeted disease, the involved epidemiological components, and the machine learning components, is provided in the extended Tables 2 and 3 in Appendix F.

2.2.1 Epidemiological Parameterization. Epidemiological parameterization involves refining existing epidemiological models, such as **susceptible–infected–recovered** (**SIR**) and **susceptible–exposed–infected–recovered** (**SEIR**) models, to predict disease transmission more accurately. The initial conditions and parameters within these compartmental models, defined by **ordinary differential equations** (**ODEs**), are crucial for simulating and predicting disease dynamics. Due to the inherently simplified nature of ODEs, which may not capture the complexity of real-world scenarios, any inaccuracies in setting initial values and parameters can lead to skewed predictions that do not align with actual disease dynamics. This discrepancy highlights the need for

calibration techniques to adjust model parameters, ensuring that predictions accurately reflect observed data and correcting any inherent biases. In contrast to traditional simulation of mechanism-based models that rely on static or predetermined epidemiological parameters, epidemiological parameterization actively incorporates disease-related data, such as the number of infections, to estimate and adjust variables and parameters within the models. To achieve this, a variety of machine learning methods are employed, and these can be broadly categorized into two groups: methods that infer epidemiological parameters directly from data, and methods that represent epidemiological parameters with functional models. By utilizing these approaches, machine learning can enhance the predictive capabilities of epidemiological models, thereby providing a data-driven framework for understanding disease dynamics.

Epidemiological parameter inference from data. Data assimilation techniques, which are widely applied in atmospheric and oceanic sciences and in numerical weather forecasting [180], aim at utilizing observations to optimize mechanism-based models. Thus, they have also been applied in disease dynamic prediction to calibrate the epidemiological models [47, 126, 154, 156, 157, 206, 209]. 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 [156]. They proposed the SIRS–EAKF framework, which uses the **ensemble adjustment Kalman filter** (**EAKF**) and a **particle filter** (**PF**) to assimilate the observations of infections (i.e., estimates of influenza infections from Google Flu Trends) into the **susceptible–infectious–recovered–susceptible** (**SIRS**) model [159]. The SIRS–EAKF framework can estimate the posterior of probabilistic distributions of system state (e.g., susceptible populations and infected populations) and epidemiological parameters (e.g., the mean infectious period and the average duration of immunity) in the used SIRS model. The formulation of the posterior in this framework is detailed in Appendix E.5. Subsequently, Shaman et al. used similar data assimilation techniques to generate weekly influenza forecasts for the influenza season in 2012 and 2013 across 108 cities in the USA [157]. In addition, a series of similar studies have utilized KF/PF methods and epidemiological models at the metapopulation or population levels to forecast influenza [47, 126, 206], dengue [209], and COVID-19 [154]. Further details on these studies are provided in Appendix D.2.1.

The Monte Carlo maximum likelihood method is also used with the stochastic compartmental model, e.g., **global epidemic and mobility** (**GLEaM**) model, to calibrate the model parameters [172, 212]. GLEaM model is a computational framework designed to simulate the spread of infectious diseases across extensive geographical areas [15, 16]. Utilizing the GLEaM model, Tizzoni et al. conducted a study on the 2009 H1N1 influenza pandemic, applying the Monte Carlo maximum likelihood technique to infer unknown parameters of the disease spread [172]. In a related application of the GLEaM model, Zhang et al. developed an epidemic computational framework [212]. A detailed introduction to the steps of this method is given in Appendix D.2.2.

Aside from data-assimilation methods and GLEaM simulation-based methods, other 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. Its general formulation is detailed in Appendix E.6. There is a branch of work that focuses on formulating this type of loss function to estimate epidemiological parameters [24, 86, 93, 109, 168, 189, 191, 217, 226]. For example, Zou et al. formulated a loss function with a logarithmic-type **mean square error** (**MSE**) [226]. Based on this loss function, parameters can be optimized by the general gradient-based optimizer. They also developed a novel compartmental model, named the **SuEIR** model—an improved **SEIR** model that considers a scenario of untested or unreported cases of COVID-19—and trained it using

their machine learning approach. Various deep learning methods, such as the RNNs (e.g., LSTM and **gated recurrent units (GRU)**) [44, 50, 78, 89, 100, 113, 211, 220] and the **fully connected neural networks** (**FNN**) [69], have also been employed to estimate the time-varying parameters in epidemiological models. For example, Zheng et al. proposed an **improved susceptible–infected** (**ISI**) model and used an LSTM to estimate the infection rate from historical disease dynamics [220]. La Gatta et al. used the GCN and LSTM models to infer epidemiological parameters of the SIR and **susceptible–infected–recovered–deceased** (**SIRD**) models [89]. Appendix D.2.3 provides detailed descriptions of the literature related to the development of MSE-based models for inferring epidemiological parameters from data.

Epidemiological parameters modeling. In addition to models that infer values or probabilistic distributions of model parameters from observations, some methods aim at modeling and estimating the variation of epidemiological parameters and formulate them as functions of covariates. Arik et al. recently proposed the use of time-varying functions to model parameters [12]. This study is interesting for its 'explainability by design' approach, which achieves clarity without compromising accuracy. Specifically, instead of using the static epidemiological parameters in the traditional compartmental model, they used learnable functions to estimate parameter values from various covariates, which enables parameter values to vary over time. And they used the generalized additive model to encode the effects of covariates on epidemiological parameters. The interpretable encoders employed in the proposed methodology are not only effective for the current model but also hold the potential for designing other risk prediction models where both interpretability and accuracy are critical requirements. Baek et al. predicted the disease dynamics of multiple regions by using a stochastic SIR model [14]. 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. Buch et al. aimed at modeling and estimating the time-varying transmission rate of the SIR model, for the situations of single region and multiple regions, using a semiparametric log-linear mixed-effects model and a smooth GP model, which enable the description of the explained effects of covariates and the unexplained temporal heterogeneity [23].

2.2.2 Epidemiology-embedded Learning. Unlike the methods that infer parameters for epidemiological models (e.g., ODEs) to predict disease dynamics, many approaches use machine learning models to predict disease dynamics directly, while leveraging mechanism-based models to guide, regularize, or constrain these predictions.

Epidemiological mechanism-guided models. These models often incorporate domain-specific knowledge to shape their model structure and enhance their predictive accuracy. For example, Shi et al. proposed a spatial transmission model and an RNN-based model for predicting malaria transmissions [163]. The spatial model assesses the potential for disease transmission in various locations by calculating two key epidemiological indicators: **vectorial capacity** (**VCAP**) and **entomological inoculation rate** (**EIR**). These indicators are derived from ODEs that describe the transmission dynamics of vector-borne diseases and are influenced by environmental factors such as temperature and rainfall. Specifically, VCAP quantifies the daily potential for disease spread from a single infected human through mosquito bites, while EIR measures the daily average of infectious bites a person receives [166]. Additionally, the model incorporates the concept of a transmission network, which is based on the road transportation network, to understand the spread of disease across different regions. There are also some other works using the same epidemiological indicators, i.e., VCAP/EIR [162, 223] and similar epidemiological concepts, i.e., **next-generation matrix** (**NGM**) [97] to construct their non-linear models with epidemiological

parameters. Additionally, some works formulate the disease risk prediction for multiple regions as the network inference problem by directly using the formulation of meta-population models [128, 182]. A detailed discussion of these models can be found in Appendix D.3.

Epidemiological regularization and constraints for optimization. Some studies have added epidemiological constraints and regularizations, which are derived from epidemiological models, to standard objective functions of supervised machine learning models to aid model parameter optimization. The general formulation of the loss function for this type of approaches can be found in Appendix E.7. In classical machine learning models, techniques such as tensor factorization [76] and Bayesian inference [66, 120, 195] have been explored to incorporate epidemiological constraints and regularization. For instance, Kargas et al. applied epidemiological constraints in tensor factorization approaches to predict disease dynamics by devising **spatio-temporal tensor factorization with epidemiological regularization (STELAR)** [76]. 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., **canonical polyadic decomposition** (**CPD**). For the Bayesian models, Hua et al. proposed the **social media based simulation** (**SMS**) model for influenza dynamics prediction [66]. 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 discrepancy between the health status derived from the social media space and that from the simulation space at the population level.

Some deep learning models, such as RNNs [20, 181, 184–186] and mixed deep learning modules [26, 53, 96, 183], also incorporate epidemiological models to constrain the learning of model structures and parameters. This integration ensures that the models more accurately reflect the realistic dynamics of disease transmission. A representative example of such deep learning models is the **spatio-temporal attention network** (**STAN**) proposed by Gao et al., which is a **graph attention network (GAT)** model with epidemiological constraints designed for long-term prediction of pandemics [53]. Epidemiological constraints are incorporated into STAN learning and prediction. In addition to disease dynamics predictions, their model further generates the prediction of epidemiological parameters (i.e., the transmission rate and recovery rate). They also design the loss function for model optimization based on the above-mentioned two kinds of outputs: (1) the prediction loss which captures short-term trends by calculating errors between the dynamics predicted by the deep modules and real case numbers, and (2) 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. Another representative example is the **causal-based graph neural network** (**CausalGNN**) model proposed by Wang et al., which constrains the dynamic attention-based GNN module with an epidemiological model (i.e., the SIRD model) [183]. Similar to [53], the CausalGNN model also generates the prediction of case numbers and epidemiological parameters, and defines corresponding loss functions. Unlike the STAN model, the CausalGNN model feeds the simulations obtained from the SIRD model together with the input features into the data-driven model to generate the model outputs.

Epidemiology-informed neural networks (**EINNs**) [117, 141, 173] is a novel physics-informed deep learning approach designed specifically for forecasting the risk of infectious diseases. The idea of utilizing a **physics-informed neural networks** (**PINNs**) [134] to learn the latent epidemic dynamics provides a promising solution for integrating epidemiological insights with empirical data. As introduced by Rodriguez et al. in [141], there are two modules in EINNs. One module is the time module in which the neural network learning includes several types of loss

functions: (1) ODE loss that minimizes the difference between the automatic gradients of states from the neural network and the differential values calculated by ODEs; (2) data loss that reduces the inconsistency between model output of states and training data; and (3) monotonicity constraints that ensure the increase of the recovery population and the decrease of the susceptible population to overcome the difficulty of learning when assuming some states are unobservable. Unlike PINNs, which are usually used to solve ODEs, EINNs are able to make predictions of future disease risks by involving another module: a feature module in which several risk-related time-series features are inputted for disease risk prediction. The learning objective of the feature module is to maximize the consistency between gradients obtained from the feature module and those from the time module, while minimizing the difference between the ground truth and the output of the time module. Detailed introductions to the remaining related works on the aforementioned types can be found in Appendix D.4.

2.3 Overcoming Data and Information Scarcity in Epidemiology

In the initial stages of an outbreak within a population, particularly concerning newly emerging infectious diseases like COVID-19, there is often a lack of clear understanding regarding the epidemiological characteristics of the pathogen, including its pathogenicity, infectivity, and incubation period. Additionally, the modes of transmission, which can range from airborne to vector-borne to vertical, can be ambiguous. Moreover, the critical disease-related data, such as the number of confirmed cases and the mobility patterns of those infected individuals, could be insufficient or even scarce. This deficit in epidemiological insight and corresponding data tends to vary across different regions and diseases. In regions equipped with advanced and robust disease surveillance and reporting infrastructure, data may be plentiful, enabling more accurate risk assessments. Conversely, for newly emerging diseases or areas lacking comprehensive data collection, the challenge becomes leveraging high-quality, relevant data from analogous diseases or comparable regions to inform epidemic forecasting for the disease in question. Additionally, while infectious disease surveillance is essential for timely prevention and control efforts, offering real-time or near-real-time data, regions with sub-optimal surveillance systems often yield incomplete datasets. The data acquisition and consolidation process, being both time-intensive and costly, may further introduce delays in data availability. To enhance disease risk prediction under these circumstances, various approaches have been developed, which will be introduced in the subsequent subsections. These methodologies aim at bridging the data gaps and refine the predictive models, even when faced with limited or delayed information.

2.3.1 *Enhancing Data and Information Completeness.*

Lack of epidemiological information. Computational representation of epidemiological characteristics is essential for modeling infectious disease transmission and predicting disease risk. However, obtaining sufficient quantitative data on disease transmission can be challenging. For example, the specific timeframes associated with the exposure-infection process of COVID-19, such as the incubation period, could be unknown or untraceable. To tackle this issue, Cui et al. implemented an encoder-decoder framework designed to approximate the various exposure-infection intervals, aiming at enhancing the prediction of COVID-19 pandemic dynamics [37]. Within this framework, the encoder utilizes multiple multi-channel CNN modules with differing kernel sizes to perform temporal convolution to extract temporal patterns of multiple cross-ranges from case numbers and regional visitor counts. Moreover, they introduced the **graph-based within-range exposure-infection** (**GRE**) module, which characterizes both within-range temporal and spatial patterns from the temporal embedding by constructing a graph where timeslices and regions are represented as nodes, and dependencies between them are depicted as edges.

Data missing. Tan et al. addressed the challenge of spatiotemporal missing data in infectious disease risk prediction [171]. They developed a deep embedding technique for inferring missing reported cases by leveraging available data on reported cases and associated risk factors. In a separate study, Elimam et al. proposed a suite of three imputation methods to handle missing data: the **last observation carried forward** (**LOCF**) method, the **centered moving average imputation** (**CMA**) method, and the **time evidential k-nearest neighbors** (**TEKNN**) imputation method [48]. The LOCF method fills in missing values using the last available data point, assuming that subsequent missing values can be approximated by this last observation. The CMA method, in contrast, imputes missing values by calculating the average of the nearest known data points both before and after the gap, providing a temporally balanced estimate. The TEKNN method also utilizes known data points surrounding the missing values for imputation but differs by selecting the neighbors according to temporal proximity rather than assuming an equal distribution of value positions around the missing data points.

Data latency. Traditional surveillance data of infectious diseases, such as reports of case numbers from government institutions, are often subject to delays spanning several weeks due to the time required for data collection, organization, and verification. This leads to a significant issue known as data latency. To deal with this problem, Gao et al. developed a set of deep learning modules that incorporate attention mechanisms, specifically designed to be aware of spatial and temporal delays, which they termed **spatial latency-aware attention** (**S-LAtt**) and **temporal latency-aware attention** (**T-LAtt**). These mechanisms are used to integrate spatial and temporal embeddings derived from both real-time and latent data [54]. In their study, they assume the existence of an undirected network that underpins disease transmission dynamics and introduce a population-level disease prediction model called PopNet. PopNet begins by learning the network's structure, using population and geographical distance to calculate the similarity of each pair of locations. Utilizing this inferred network structure, PopNet then employs two GATs to generate node embeddings from both the immediate disease data and the subsequently revised data. These embeddings are then synthesized through the S-LAtt and T-LAtt mechanisms, sequentially. S-LAtt applies a feature similarity-based attention to adjust the node embeddings, taking into account the marginal effects of time latency on final predictions. T-LAtt employs GRU networks to capture temporal dependencies. Finally, PopNet concatenates the learned node embeddings to generate final predictions. The way of utilizing proxy data illustrated in this work is useful, not only in spatial network structure modeling, but also in more general scenarios where the targeted data are scarce or even unavailable.

2.3.2 Knowledge Transfer between Similar Disease Patterns. In scenarios where specific epidemiological information for an emerging infectious disease is sparse but abundant data exists for other diseases with analogous characteristics, one strategy is to utilize the existing datasets to inform model training. For instance, Yang et al. employed a deep learning architecture, i.e., LSTM, leveraging statistical data from the 2003 SARS outbreak alongside COVID-19 epidemiological parameters to forecast the incidence of new COVID-19 cases [207]. Another advanced technique applied under these conditions is **transfer learning** (**TL**). This machine learning framework involves a "source task" from which the model learns, and a "target task" that the model aims at executing. The TL architecture is specifically designed to apply the insights and knowledge gained from the source task to enhance the learning process for the target task [121]. A recent, representative work is the **COVID augmented ILI deep network** (**CALI-NET**), which is a **heterogeneous transfer learning** (**HTL**) framework for COVID-ILI forecasting [142]. It utilizes the EpiDeep model [3] to extract temporal patterns from historical wILI data, serving as the source model. The CALI-NET

framework includes the **COVID-augmented exogenous model** (**CAEM**), which captures spatiotemporal features from exogenous COVID-19 data using Laplacian regularization of a geographical adjacency matrix and a GRU module. These features are then integrated into the target model. Additionally, CALI-NET incorporates a **knowledge distillation** (**KD**) loss function, which is composed of a hint loss that aligns the representations from the source and target models, and an imitation loss that aligns the source model's predictions with the actual data, ensuring an effective transfer of knowledge. By introducing an interesting idea of "steering" historical disease forecasting models towards new scenarios, this work represents a fresh approach to knowledge and model transfer, ensuring that forecasting remains robust even when specific data are lacking. Roster et al. also investigated the effect of knowledge transfer between related diseases (e.g., dengue and Zike, influenza and COVID-19) on improving prediction accuracy by using the TL methods [144]. Different from the above studies, which focus on the transfer of knowledge between different but similar diseases, Ren et al. introduced a novel deep transfer learning framework. This model, referred to as TransCode, is designed to leverage fine-grained disease transmission patterns derived from the visiting records of COVID-19 confirmed cases. It enables the prediction of COVID-19 trends and the inference of transmission dynamics in regions lacking such detailed data [139].

3 Machine Learning in Endemic and Elimination Phases (E-E Phases)

In contrast to the widespread reach in the P-E phases, disease transmission during the E-E phases is typically confined to more specific and limited areas. In the endemic stage, an infectious disease is persistently present and maintains a relatively high incidence within a certain region. Such diseases often exhibit unique patterns influenced by local environmental factors. For instance, malaria and dengue fever are considered endemic in certain Southeast Asian countries, while seasonal influenza, which predominates in some regions during the winter months, can also be classified as an endemic. In the elimination stage, or as a region approaches elimination, cases of indigenous diseases tend to decrease. Nonetheless, the risk of disease persists due to the susceptibility of the local environment to disease transmission and the potential importation of the disease from other high-risk areas. In this stage, the disease risk may exist in specific foci and could emerge sporadically. Hence, a crucial issue in E-E phases is the application of disease- and scenario-specific knowledge, along with diverse risk factors, to predict potential risks. We have identified two primary public health concerns in these phases: (1) the localized patterns of disease transmission that are specific to certain areas, and (2) the specific challenges encountered during the elimination stage of a disease. In the following sections, we will explore the existing body of literature that addresses these concerns.

3.1 Modeling Localized Transmission Patterns

3.1.1 Temporal Autocorrelation Analysis. Infectious disease risk data, such as the number of infected cases and deaths, are typically represented in time-series format, positioning the task of predicting infectious disease risk as a time-series forecasting challenge. As a result, some representative statistical models have been employed to characterize the temporal dependencies and patterns inherent in this data. Despite their utility, the effectiveness of statistical models is often constrained by their intrinsic structural limitations. To overcome these constraints, a variety of machine learning models have been incorporated, enhancing the predictive capabilities of traditional statistical approaches. The superior ability of machine learning methods to capture complex and nonlinear dependencies makes them particularly well-suited for managing time-related information. Consequently, pure machine learning methods have also been widely used in addressing temporal dependencies and modeling time-series data. In this subsection, we will explore a range of hybrid models that integrate statistical methods with machine learning techniques.

Additionally, we will examine a selection of pure machine learning models that have been developed for the task of modeling time-series data related to disease spread. Appendix C.2 further provides a brief introduction to several classical statistical models that are fundamental to time-series data analysis.

Machine learning with time-series statistical models. Several studies have sought to enhance the performance of **autoregressive-based** (**AR**) and **moving average-based** (**MA**) models by integrating them with other machine-learning techniques to overcome their inherent limitations. This approach is utilized in various works, including those by Zhang et al. [214], K Abdul Hamid et al. [70], Chakraborty et al. [30], Swaraj et al. [169], and Wang et al. [192, 193]. Zhang et al. addressed the issue of non-stationary trends with zero counts, high proportional low counts, and wave patterns caused by the seasonal effects and human interventions by combining a segmented Poisson model with **autoregressive integrated moving average** (**ARIMA**) models [214]. They proposed a two-stage algorithm. Firstly, the segmented Poisson model identifies turning points in the time series, dividing it into distinct segments and modeling each with specific parameters. Subsequently, the ARIMA models analyze the residuals between the actual data and the segmented Poisson model's output to refine the prediction accuracy. K Abdul Hamid et al. presented the **ARIMA–least-squares support vector machine** (**ARIMA-LSSVM**) model, integrating the strengths of ARIMA in handling linear time series and the advantages of SVM in modeling the non-linear dependencies to improve the prediction performance [70]. Wang et al. used a hybrid approach by applying the **seasonal ARIMA** (**SARIMA**) and **nonlinear autoregressive network** (**NAR**) to analyze pertussis incidence data [192]. This data was initially decomposed into its linear and non-linear components using the **discrete wavelet transform** (**DWT**). The rationale behind this methodology is to make use of the strength of SARIMA in capturing and forecasting linear trends within the time series data while leveraging the flexibility of neural networks to model and interpret the non-linear patterns. Similarly, Chakraborty et al. combined ARIMA with **neural network autoregressive** (**NNAR**) models to simultaneously capture linear and non-linear components within time series data [30]. Swaraj et al. proposed a hybrid method, which combines the ARIMA and NAR to respectively model the linear and non-linear parts in time series data of COVID-19 cases [169]. In another work, Wang et al. integrated **ensemble empirical mode decomposition** (**EEMD**) with ARIMA and **nonlinear autoregressive artificial neural network** (**NARANN**) to conduct time series prediction [193]. This method aims at decomposing complex time series into simpler components using EEMD, apply NARANNs and ARIMA to model the dynamics of each of these components, and aggregate the outputs from all individual NARANNs and ARIMA to generate final predictions.

Pure machine learning models. In addition to autoregressive-based models that characterize temporal dependencies using intuitive and easily understandable structures, other classical machine learning models, such as Dirichlet process [115], regression-based models [77, 150, 225], logistic model and Prophet model [188], empirical Bayesian [22], and Kalman filter [165], are also designed for capturing intrinsic temporal patterns within time-series data. As discussed in Section 2.1, some deep learning modules, such as RNN and its variants LSTM and GRU, have been specifically designed for capturing sequential patterns. Therefore, it is not surprising that these modules have been widely adopted in various works of infectious disease modeling for characterizing the temporal dynamics of time series data that are related to disease transmission [62, 83, 187]. For instance, Wang et al. proposed a novel LSTM model, which designs a rolling update mechanism to train the model for long-term prediction of daily confirmed COVID-19 cases [187]. This mechanism functions by incorporating each day's predicted case numbers into the existing dataset after each training iteration. As a result, the model is continually trained on the most recent predictions,

which allows it to generate forecasts for several days ahead. In some other works, the CNN also has been integrated with the RNN for temporal modeling [112, 178, 201]. For example, Muhammad et al. proposed the CNN-LSTM algorithm [112]. In their algorithm, they adopted the CNN serving as the encoder, which consists of two layers of one-dimensional CNN convolution and pooling layers to embed the features, and an LSTM serving as the decoder to capture the short- and long-term temporal relationship. Similarly, Xu et al. [201] and Verma et al. [178] also designed a series of recurrent and convolutional neural network-based models to predict COVID-19 cases. Some other deep learning models, such as the NeuralProphet model [39] and dendritic neural regression [45], have also been developed for the temporal autocorrelation analysis. Introductions to other works on various types of pure machine learning models are included in Appendix D.5.

3.1.2 Correlated Signals of Disease Dynamics. In addition to exploring the temporal patterns of disease risk trends via a range of machine learning techniques, numerous studies have collected, processed, and utilized a wealth of diverse risk-related factors. These factors contribute to enhancing prediction accuracy by uncovering, analyzing, and modeling their correlations with disease dynamics. In this subsection, we will discuss the methodologies that incorporate these correlated risk factors into temporal models of disease transmission. This integration allows for a more comprehensive understanding of how these factors interact with the propagation of the disease over time, thereby improving the predictive capabilities of epidemiological models.

Web-based activity. One of the popular correlated indicators is web-based activities, which involve several different human behaviors that may reflect the disease dynamics implicitly. To be specific, lots of work shows that the information contained in search activity on the search engine [38, 58, 60, 73, 153, 205, 208], posted content in social media [2, 174, 179], the article of new release and press release [31, 81, 82], and Internet-based surveys/surveillance [127] can be extracted to improve the prediction of disease risks.

Considering the search engine data, 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 [58]. Although the Google Flu Trends service was discontinued in August 2015,[2] the **Google Extended Trends** (**GET**) application programming interface remains accessible and provides the statistics of online search trends at various temporal and geographical granularities, providing a valuable resource for researchers developing their own disease models [205]. Appendix D.6.1 provides a supplementary discussion on a series of publications that utilize search engine data from Google and Baidu.

Social media content, particularly from platforms like X (formerly known as Twitter), can also offer insights into the health trends of its active users. For instance, Achrekar et al. [2] discovered a linear relationship between the number of X users posting about influenza and the rate of physician visits for ILI. They calculated Pearson correlation coefficients and developed a regression model to support their findings. Building on this, they introduced the **social network enabled flu trends** (**SNEFT**) framework, which employs an ARX model to predict ILI cases using data from ILI physician visits and X posts. However, the study reported in [2] limited its analysis to the count of users posting influenza-related content and the number of tweets containing influenza-specific keywords. It did not leverage the full potential of the textual data available, omitting other textual features that could further refine the estimation of influenza infection rates. In contrast, Volkova et al. extracted detailed linguistic features and communication patterns as latent embeddings from posts on X, and fed these data, together with ILI data, into a joint neural network model based

[2] https://ai.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html. Accessed February 13, 2025.

on LSTM modules to predict ILI dynamics [179]. Tran et al. collected the multi-modal data from multiple different sources, such as X posts and government stringency features, and extracted the informative signals from human-generated text data in X by BertTweet, which is a pre-trained language model and constructed a graph structure to represent the correlation and interactions between the users [174].

Compared to the above-mentioned search activity statistics and social media posts, which may contain the subjective tendencies of users, the official sources of information, such as news articles, are likely to provide more accurate and rigorous information on infectious disease dynamics. For instance, Kim et al. collected 7,769 articles published by Centre for Health Protection in Hong Kong from 2004 to 2018 related to infectious diseases, aiming at exploring the usage of news data for influenza prediction [81]. After that, they extracted keywords that were the most relevant to influenza from the news article with Word2vec and consequently used the SVM to predict the increase or decrease of influenza patient numbers. Recently, they also used Word2Vec to find the words related to COVID-19 from news articles from the New York Times and adopted the Seq2Seq with the attention mechanism to predict the COVID-19 outbreak [82]. Different from extracting the words information relevant to the disease transmission by the **natural language processing** (**NLP**) technique, Chen et al. collected more than 60 official press releases about COVID-19 in Hubei, China, from January 2020 to May 2020, to identify risk dynamic data as time-series data with 10 features, such as numbers of in-hospital monitoring, and adopted a multivariate LSTM model to capture their temporal dynamics [31].

In addition to the above-mentioned web-based activities, which usually post content or publish information that can be accessed by anyone publicly, there are also some specific activities, such as online surveys, initiated by public health institutions, aiming at collecting instant information from the population and surveillance disease dynamics in real-time. For example, alongside the traditional surveillance data reported by the doctors, Perrotta et al. used web-based surveillance data collected by Internet-based surveys from a real-time participatory system, Influweb, to forecast the influenza activity in Italy by ARX model in a more time-immediate way [127].

Physical-world signals. Apart from the web-based information, which is virtual, some works considered the factors in the physical world that are highly related to disease dynamics. For example, Mcgough et al. collected the available weather information (i.e., air temperature and precipitation) and trained a series of SVM models to generate the ensemble prediction of dengue dynamics [110]. Kumar et al. used the symptoms data to optimize the prediction of COVID-19 cases and deaths by a **deep reinforcement learning (DRL)** model after they adopted the **modified LSTM** (**MLSTM**) model, which has a new activation function, to predict COVID-19 dynamics [87]. Zhao et al. utilized the mobility data collected by Google which can be divided into six types according to the locations reflecting the mobility behaviors and built a Poisson autoregression model that integrates the GLM model and the autoregressive count data model to predict the confirmed cases of COVID-19 in Sweden [216]. Zheng et al. also utilized the mobility data that includes the pre-illness activity tracking of COVID-19 patients in Macau to predict the spatial distribution of COVID-19 risk [219].

The intervention strategies applied by the government during the disease outbreak also play an important role in affecting the disease trends. Ali et al. collected the COVID-19 intervention data over time from the website of **Asia Regional Information Center** (**ARIC**) at **Seoul National University** (**SNU**).[3] They fed these features as well as the historical dynamic data into the proposed **stacked bi-directional LSTM** (**Stacked Bi-LSTM**) model to capture the temporal dynamics, thereby providing better predictions of cases, deaths, recovered patients, and

[3]https://sites.google.com/view/snuaric/home?authuser=0. Accessed February 13, 2025.

quarantined people number of COVID-19 [8]. Lai et al. proposed to utilize the **wastewater-based epidemiology** (**WBE**) information to predict the COVID-19 cases by gradient-boosting tree-based ML models [90, 91]. Price et al. collected the patient-level information of all SARS-CoV-2 tests and aggregated them to create the dataset with several public health metrics, such as number of cases, testing rates by county and day, and vaccination rate. Then they designed a multi-layer deep LSTM network combined with a sliding window approach to predict the COVID-19 cases [130].

Multi-type data. In addition to using correlated factors of similar types, many existing studies fuse data from multiple sources to enhance prediction performance. These approaches can be broadly categorized into two lines: (1) regression-based and classical machine learning models [35, 43, 59, 67, 190], and (2) deep learning models [9, 40, 49, 79, 95, 136, 148, 204]. In the first category, a representative work by Jain et al. adopted the **generalized additive models** (**GAMs**) to capture the correlation between the features from meteorological data, clinical data, disease surveillance data, socioeconomic data, and spatial encoding to predict the dengue transmission [67]. Another typical example is provided by Gong et al., who utilized 19 variables, including ten climate indicators (e.g., moisture), six geographical indicators (e.g., landform), and three social-economic features (e.g., gross domestic product) to predict the spatial distribution of schistosomiasis transmission risk with a statistical quantitative analysis method (i.e., information value) and a series of machine learning models, such as logistic regression, **random forest** (**RF**), **generalized boosted model** (**GBM**), and their combinations [59]. In addition to the aforementioned classical machine learning approaches, there has been a shift towards leveraging deep learning techniques to integrate features from diverse sources. For instance, MLPs have been utilized in works by Kiang et al. [79] and Liu et al. [95]. **Nonlinear autoregressive models with exogenous inputs (NARX)** have been adopted by Eltoukhy et al. [49], while LSTM networks have been applied in forecasting models by Said et al. [148], Rashed et al. [136], Amendolara et al. [9], and Yang et al. [204]. Additionally, auto-encoders have been explored by De et al. [40]. A detailed introduction to these models in both categories can be found in Appendix D.6.2.

3.2 Tackling Challenges in Disease Elimination

3.2.1 Modeling Imported Risks. As regions progress toward the elimination phase of infectious diseases, the incidence of indigenous cases is expected to decline and ultimately approach zero. The success of disease elimination in this phase is contingent upon controlling the influence of imported cases and the susceptibility of the local environment. A confluence of these factors can lead to a resurgence of the disease. In contrast to the widespread transmission seen in epidemic and pandemic phases, the elimination phase is characterized by a reduced and more localized disease presence. Here, the primary concern shifts to the risk posed by imported cases, which can significantly impact local disease dynamics. To effectively address these challenges, specialized methodologies have been developed to accurately monitor and predict disease transmission patterns during this phase. These include techniques for characterizing seasonal variations and identifying underlying trends, as proposed by Shi et al. [164] and Kamana et al. [71], as well as network-based modeling strategies, such as those introduced by Yang et al. [203]. These approaches are specifically designed to capture the complex and evolving transmission patterns unique to the elimination phase, enabling more precise and efficient public health interventions. To be specific, Shi et al. conducted a focused study on the risk of imported malaria in the border cities of Tengchong and Ruili in Yunnan, China, spanning the years 2006 to 2010 [164]. Utilizing data from the China CDC, they categorized the recorded malaria cases into two groups: imported and local. For the analysis of imported cases, the Loess method was employed to perform seasonal and trend decomposition, thereby uncovering temporal patterns. Regarding local malaria transmission, they incorporated climate factors such as temperature and rainfall, along with the VCAP index, into their predictive

models. They applied both **linear regression models** (**LRM**) and GAMs to estimate the local malaria risk, demonstrating the influence of environmental conditions on disease transmission within these communities. Yang et al. introduced the concept of the heterogeneous diffusion network and developed a network-based algorithm to model spatiotemporal disease transmission between different locations [203]. Their research focused on malaria transmission dynamics along the Yunnan-Myanmar border, with a particular emphasis on the risk associated with Yunnan residents who travel to Myanmar for work, contract malaria, and then return home. Kamana et al. investigated the problem of malaria resurgence in China [71]. They collected data on cases of *Plasmodium falciparum* across China's 31 provinces and recorded instances of malaria imported from 45 African countries to China. In order to understand and forecast how these imported cases might influence the risk of malaria domestically, they proposed a hybrid model called ARIMA-RNN. This model inherits the strengths of both statistical methods and deep learning by integrating the ARIMA model with an RNN that uses GRU. This combination is designed to effectively capture the relationships and dependencies between the time series data of local and imported malaria cases.

3.2.2 Optimizing Limited Resources for Disease Surveillance. In underdeveloped regions, the distribution of resources dedicated to disease prevention and control, including surveillance resources, frequently falls short of the demands across all impacted areas, especially in the elimination phase. To address this issue, some researchers have turned to machine learning for solutions. A typical example is the sentinel selection problem, which is an aspect of active surveillance in public health. This problem involves determining the most representative areas from a broader target region for conducting disease surveillance when resources are constrained. By applying machine learning techniques, researchers aim at optimizing the selection process to ensure effective monitoring with the available resources. Pei et al. introduced a multivariate regression approach called the **group sparse Bayesian learning** (**GSBL**) to enhance the allocation of limited surveillance resources in the field of disease prevention [124, 125]. The GSBL model strategically identifies key regions for surveillance within a disease transmission network, operating under the predictive framework of infectious disease dynamics. The model is designed to learn a row-sparse matrix representing the transmission network, where the sentinel locations are marked by non-zero rows. Utilizing the disease data from these sentinel nodes, the GSBL model is capable of reconstructing or forecasting the overall disease dynamics across all targeted areas. Notably, the model relies solely on historical case data to infer the transmission network, without the need for additional prior knowledge about the disease's spread. This, therefore, enables the algorithm to be broadly applied to a variety of diseases and potentially transferable to other domains. In the context of striving for malaria elimination, Zheng et al. adopted a distinct approach focused on the spatial clustering of malaria risk foci in Baoshan, Yunnan, aiming at enhancing active surveillance efforts during the elimination phase [218]. To achieve this, they proposed to use the **multivariate auto-regressive state-space** (**MARSS**) model. This model is designed to identify the optimal combination of multiple time series, which allows for a more comprehensive explanation of the variations observed in malaria incidence.

4 Technical Challenges in Infectious Disease Risk Prediction

In Sections 2 and 3, we presented a wide range of machine learning approaches for predicting infectious disease risks, organized according to the proposed three-tiered hierarchical taxonomy. Our review of the extensive literature reveals that a diverse array of machine learning architectures, as well as their integration with other epidemiological and statistical models, can be used to implicitly and/or explicitly model disease transmission and accurately predict disease dynamics, thus tackling various public health concerns during different phases of disease propagation, from P-E to E-E.

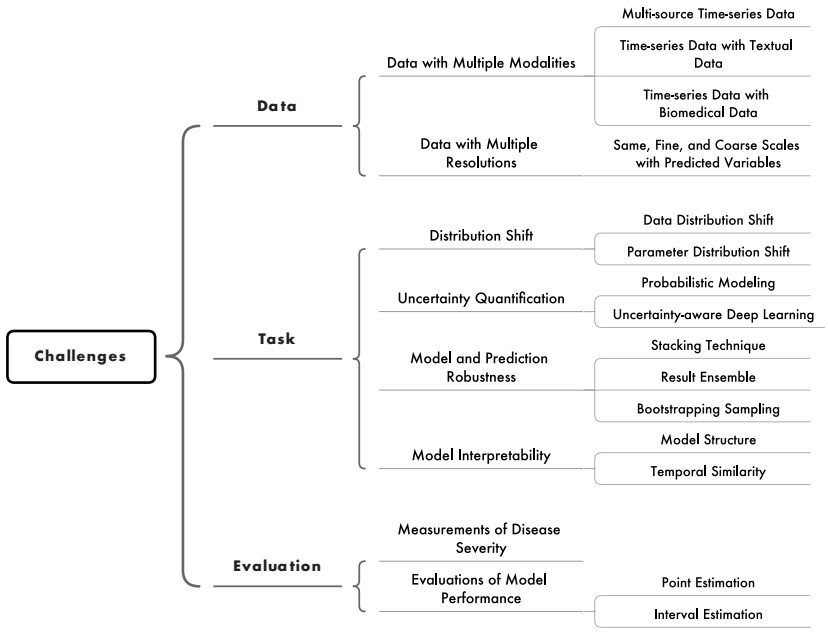

Fig. 2. Three main challenges in machine learning for infectious disease risk prediction: (1) data-related challenges (Section 4.1); (2) task-related challenges (Section 4.2); and (3) evaluation-related challenges (Section 4.3).

When developing methods for predicting disease risks, it is crucial to consider not only the public health concerns during different phases of disease propagation but also the technical challenges that arise. These include managing heterogeneous and multi-modal data, tackling the computational and design issues inherent in the prediction task, and establishing criteria for defining and evaluating model performance–areas not thoroughly explored in previous reviews. It is also important to recognize that no single methodology is adequate for categorizing all infectious disease risk prediction models. The unique strengths, limitations, interconnections, and practical applicability of various models become evident only when they are examined from multiple perspectives. In light of this, this section aims at categorizing and briefly review a subset of models, focusing on the specific challenges they address related to input data, the complexities of the prediction task, and the metrics used for output evaluation. To facilitate a clear understanding, we provide an illustrative taxonomy of these challenges in Figure 2.

4.1 Data Challenges

The predictive modeling of infectious diseases has seen some algorithms leverage historical disease dynamics to forecast future risks, relying on the temporal autocorrelation within a single region or correlations across multiple regions. However, the transmission of infectious diseases is inherently complex, and a growing body of research indicates that incorporating diverse, disease-related data can significantly improve prediction accuracy. To this end, a number of factors related to disease risk have been investigated and integrated into a variety of machine learning models. Such integration, however, comes with its own set of challenges, primarily due to the heterogeneity of the data involved. These challenges include managing data that vary in modality – that is, data presented in different forms or formats – as well as data that differ in resolution, such as information collected at varying spatial or temporal scales. In this section, we will discuss these

two challenges associated with data heterogeneity and summarize the methods that have been developed to address these issues.

4.1.1 Data with Multiple Modalities. The spread of disease is closely related to the interplay among humans, the environment, and pathogens. As a result, various types of data are continuously collected from sensors monitoring the physical world–including environmental factors–as well as from the public health sector, which tracks the dynamics of disease propagation. In the digital age, the popularity of the Internet and social media platforms has introduced new forms of data that can reflect human interactions and public perceptions, which are also indicative of disease transmission patterns and the perceived severity of outbreaks. Consequently, researchers have turned to these rich, multi-modal information sources to extract and analyze a diverse set of disease-related indicators, aiming at constructing a holistic view of how diseases disseminate through populations.

Multi-source time-series data. In infectious disease risk prediction, time-series data–such as historical case counts–serve as the primary tool for tracking and modeling the evolution of disease severity within and across regions over time. Additional time-series data, like climate records, further refine these models. Although these datasets share a common framework of time and space, they differ significantly in meaning and their influence on disease spread. To effectively integrate this heterogeneous data, specialized approaches have been developed to assimilate it in a manner that enhances predictive accuracy. Some studies incorporate this data indirectly by linking it to epidemiological parameters based on established relationships with disease risk [97, 156, 157, 163]. Other research directly inputs these variables into regression or deep learning models to autonomously discover their correlations with disease risk [79, 177, 215].

Take the climate data as an example. Considering the influence of climate on diseases like influenza and malaria, the correlation between climate trends and disease dynamics is a critical aspect of their behavior. Prior to leveraging climate variables in machine learning models, empirical studies have often sought to quantify the relationships between these variables and disease dynamics. However, the causal links and correlations are typically complex and non-obvious. As a result, many studies have either utilized established empirical formulas or employed statistical methods to uncover relationships between risk factors and disease incidence. In the context of influenza, a notable example is the work of Shaman and Kohn [158], who revisited laboratory data from guinea pig experiments [104] to show that **absolute humidity** (**AH**) is a significant factor in both the **influenza virus transmission** (**IVT**) and **influenza virus survival** (**IVS**) in temperate regions. Building on this, Shaman et al. developed a model that uses AH to predict influenza's seasonal patterns, by linking it to the virus's basic reproduction number [159]. This model was then applied to forecast influenza activity in various U.S. cities [156, 157]. Similarly, for vector-borne diseases like malaria, climate conditions play a significant role by influencing vectors and pathogens–for example, the survival rate of mosquitoes and the incubation period of *Plasmodium*. Research by Ceccato et al. quantified how temperature and rainfall correlate with vectorial capacity for malaria [27]. Following this approach, Shi et al. integrated temperature and rainfall data into their model, which accounts for VCAP/EIR, to enhance the prediction of malaria outbreaks that exhibit clear seasonal patterns [162]. In contrast to studies employing compartmental models and VCAP/EIR for indirect inclusion of risk factors, Zhang et al. collected data on several disease-related variables to construct feature vectors as the input of the model directly without the requirement for disease-specific knowledge in their analysis [215]. Additionally, Venna et al. adopted a symbolic time-series analysis to capture the nonlinear interactions between climate factors and influenza trends, offering a novel perspective on how environmental conditions impact influenza dynamics [177].

Time-series data with textual data. Textual data related to diseases, such as online search queries, social media posts, and news articles, provide valuable insights into disease transmission and severity. To leverage this information, machine learning methods have been developed to uncover hidden dependencies between text-based online information and time-series data on disease dynamics, thus enhancing predictive accuracy. As mentioned in Section 3.1.2, researchers have identified correlations between online behaviors and infectious disease levels–for instance, using search engine data, e.g., from Google [58, 60, 73, 153, 205] and Baidu [38, 208], to track ILI trends. This kind of work usually involves a two-stage modeling process: initially, key textual information is extracted and its relationship with disease patterns is determined through statistical or machine learning techniques, such as identifying and quantifying disease-related search terms [58]. Subsequently, these features are incorporated into predictive models, including regression analysis [58, 153, 208], ARIMA [73], ARX [205], and neural networks [60], to forecast disease dynamics. Another important online textual information is the posted content and virtual interactions on social media platforms of individual users. Unlike search engine queries, social media posts often contain more extensive content and personal details. While not explicitly quantifying disease spread, this data can be analyzed to extract valuable health-related insights. Prior research, as discussed in Section 3.1.2, has utilized X data to inform disease forecasting models [2, 174, 179]. Several studies have investigated the relationship between the volume of flu-related tweets and ILI rates, applying autoregressive models to capture this dynamic [2, 212]. Others have utilized neural networks, feeding them with features derived from post content alongside historical ILI data to predict disease trends [179]. Additional research has identified individual health states from X data (e.g., healthy, exposed, infectious) and integrated these at the population level into a simulation model to refine epidemiological predictions [66]. Online news articles serve as an additional textual data source reflecting disease severity. Typically, information is extracted from these articles by employing the Word2Vec model to convert text into vectorized form. These vectorized features are then inputted into regression or deep learning models to automatically identify correlations with disease data [81, 82].

Time-series data with biomedical data. Biomedical data, when combined with population-level disease dynamics such as incidence rates or case numbers, can enhance our understanding of disease transmission and healthcare resource usage. Typically, researchers extract statistical information from individual patient data and integrate it with disease risk data for predictive modeling. For instance, Gao et al. utilized daily hospitalization, ICU admissions, and diagnostic code frequencies from IQVIA's medical claim data[4] [53] to enrich the COVID-19 case data (including active, confirmed, and death cases) from Johns Hopkins University.[5] These combined datasets were then used as dynamic inputs in a graph neural network, aiding in the identification of spatiotemporal disease patterns. Similarly, Gao et al. leveraged disease-related statistics from the IQVIA dataset to inform a deep learning model for disease prediction [54].

4.1.2 Data with Multiple Resolutions. 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/granularity. 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 [170], 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

[4]https://www.iqvia.com/solutions/real-world-evidence/real-world-data-and-insights. Accessed February 13, 2025.
[5]https://github.com/CSSEGISandData/COVID-19. Accessed February 13, 2025.

(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.

4.2 Task Challenges

In addition to the challenges associated with the data, the task of infectious disease risk prediction faces other challenges when modeling disease transmission. These challenges include: (1) how to address the distribution shift of disease dynamics and transmission patterns; (2) How to take into account the uncertainty in modeling processes; (3) how to enhance the robustness of models and predictions; and (4) how to interpret the model and its outcome. In this section, we summarize several computational concerns about the above task-related challenges and introduce how they are addressed by various models.

4.2.1 Distribution Shift. Generally, machine learning methods that are trained based on empirical risk minimization face an inherent issue: the generalization ability. A model requires a good ability of generalization to make accurate predictions when receiving inputs that it has never seen. However, machine learning models for epidemic prediction also struggle to improve their 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 [10, 85]. A few studies have examined distribution shifts as part of the topic of epidemic prediction. Wang et al. [189] 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. They 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.

4.2.2 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 [65]. 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 [161]. 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. Among them, the

probabilistic modeling approaches, represented by Bayesian learning-based models and stochastic processes-based models, are popularly used to provide uncertainty measurements naturally.

Probabilistic modeling 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 [22]. Other Bayesian inference methods, including the KF and its variants [11], and PF [13], have also been used in disease dynamic prediction [126, 156, 157, 172, 212]. 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 [19]. Some representative studies have used the GP model to predict disease dynamics and provide the uncertainty of results [72, 155, 225]. Furthermore, some studies used deep learning models to perform stochastic processes; this type of model is called **neural processes** (**NPs**) [55]. Its extensions, such as the **functional NP** (**FNP**) [103] and the **recurrent NP** (**RNP**) [80, 131], have also been developed to capture complex dependency [72].

In addition to employing uncertainty quantification and estimation techniques for predicting disease risks at population and meta-population levels, some methods have been developed for the identification of individual health conditions. Specifically, uncertainty-aware deep learning models have been introduced to quantify diagnostic uncertainty for medical imaging, thereby improving the precision of disease diagnosis [57, 160]. A brief introduction of those two works can be found in Appendix D.7. While uncertainty-aware deep learning has shown promise in the analysis of medical image data and clinical time series for individual diagnoses, its application at the population or meta-population level for predicting infectious disease risks remains relatively unexplored. This presents an interesting direction for future research.

4.2.3 Model and Prediction Robustness. The robustness of a model and its predictions is essential and crucial for achieving reliable results when applying infectious disease risk models in complex real-world settings. Ensemble methods are a widely adopted approach to enhance the robustness of model predictions. These methods typically develop and train multiple models to generate predictions, which are then aggregated to produce either a weighted outcome or a probabilistic distribution.

To construct ensemble models, several studies have employed model stacking, which combines individual machine learning models into an integrated model in a hierarchical way to improve prediction performance [32, 138, 202]. For instance, the FluSight Network, a research consortium of four teams from the U.S. CDC-hosted 2017/2018 seasonal influenza forecasting challenge, utilized stacking to integrate the results from 21 different models and achieved second place in the challenge [138]. The 21 models involved in the ensemble strategy contain a variety of model types, including Bayesian hierarchical models, epidemiological models, and various statistical models. Each individual model contributes a predictive distribution and is assigned a learned weight reflecting its importance in the ensemble. The description of methods proposed in [32, 202] can be found in Appendix D.8.

Some models have adopted other machine learning-based approaches to learn the weights that contribute to the final prediction. Kuo et al. devised a GLM model that integrates predictions from eight machine learning models [88]. Adiga et al. combined statistical, machine learning, and mechanistic methods using Bayesian ensembling, which treats predictions and weights as distributions rather than fixed values [4]. Similarly, Olmo and Sanso-Navarro built an ensemble

predictor consisting of a set of Poisson regression models with different covariate combinations based on the Bayesian ensembling framework [116]. Jin et al. employed the RL method to learn the weights for a series of machine learning methods, including **temporal convolutional network** (**TCN**), GRU, and **deep belief networks** (**DBN**) [68].

In contrast to ensemble models that determine prediction weights through an additional machine learning method, as seen in stacking, some models adopt more straightforward ensemble strategies, such as simple averaging [137], median calculations [105], unweighted voting [213], or summation [122]. The COVID-19 Forecast Hub,[6] for example, aggregates forecasts from various models of different institutions using equal-weight averaging to generate U.S. COVID-19 death data [137]. Lucas et al. proposed the COVID-LSTM, which is an ensemble of ten LSTM networks, producing predictions by calculating the median value [105]. Zhang et al. developed an ensemble of five machine learning models, choosing the final prediction based on the most votes for four risk levels [213]. Panja et al. introduced the **ensemble wavelet neural network** (**EWNet**) model, combining the **maximal overlap discrete wavelet transform** (**MODWT**), which decomposes the time series, with the **autoregressive neural network** (**ARNN**), which learns decomposed time series, before summating their output of each ARNN to predict time series [122].

Bootstrapping has also been utilized to enhance individual model training by randomly sampling features [36, 52, 143, 210]. For instance, Rodriguez et al. used bootstrapping to resample a training dataset into multiple subsets for training diverse models, allowing for the calculation of prediction confidence intervals [143]. More details on this type of method are provided in Appendix D.8.

4.2.4 Model Interpretability. Deep learning models have broad applications in domains like healthcare, and their interpretability has been widely studied [6, 46, 147]. For epidemic prediction models, interpretability is crucial because misinterpretation can lead to poor decisions, negatively impacting human well-being and wasting resources. Therefore, researchers must be cautious when interpreting these models and their results.

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 [12], whereas other models use a linear model structure, such as AR and MA-based models [42, 106], 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, Adhikari et al. 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 [3]. 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 [72] is based on similar assumptions and uses a functional neural process module to learn the correlation between the predicted season and past seasons.

4.3 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.

[6] https://covid19forecasthub.org. Accessed February 13, 2025.

How to identify and use appropriate measurements is thus also a challenge to the modeling task. In this subsection, we summarize the common outcomes of prediction models and common measurements that have been widely used to evaluate these outcomes.

4.3.1 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 public health policy and on data availability. Various indicators of disease severity have been used. The most commonly used indicators include disease incidence [143, 184], case numbers [37, 41, 75, 155, 197, 221], death counts [37, 75, 221], patient visit counts related to the disease [41, 72], and disease activity levels [41, 197]. 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 [22, 212, 225].

4.3.2 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 [21]. 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**), the **mean absolute error** (**MAE**), the **mean absolute percentage error** (**MAPE**), and the **root mean squared percent error** (**RMSPE**). These indicators calculate the deviation of predicted values from ground truth. The Pearson **correlation coefficient** (**CORR**) is used to evaluate the correlation between the predicted trend and the real trend. The equations used to calculate the aforementioned indicators are presented in Equations (S13)–(S17) in Appendix E.8.

Interval estimation. Some indicators are widely used for interval estimation [21]. 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 [137]; **c**alibration **s**core is the integral of $\|k_M(c) - c\|$ over c from $0, \cdots, 1$, as shown in Equation (S18) in Appendix E.8 [72], and a calibration plot shows the relationship between c and $k_M(c)$ [72]. 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 take the natural logarithm of the calculated sum to obtain the final score [225].

5 Conclusions and Future Directions

5.1 Conclusive Remarks

In this survey, we explore the evolution of machine learning in the context of infectious disease risk prediction. Our approach diverges from most of the existing surveys on infectious disease risk modeling and prediction that typically organize the body of work according to the nature of computational models–such as mathematical, statistical, machine learning, and deep learning models. Instead, we offer a fresh perspective for categorizing the literature, one that hinges on the methodologies' alignment with two critical public health questions respectively associated with two phases of infectious disease propagation in accordance with distinct properties: (1) the P-E phases and (2) the E-E phases. Our multi-tiered framework commences with a categorization

of research according to the specific phase of disease transmission, recognizing that each phase presents unique research focuses and technical issues. In the second tier, we go deeper within each phase, arranging the literature around key public health considerations that inform risk prediction. The third tier summarizes the computational strategies devised to tackle each particular public health issue, arranging them by the foundational principles of the methodologies employed. Furthermore, we present the common challenges encountered across three main facets: data handling, prediction model design, and performance assessment. We discuss these challenges in detail, offering concrete examples to illustrate how prior research has coped with these challenges.

5.2 Future Research Directions

This subsection outlines several promising directions for advancing the research and application of machine learning in predicting infectious disease risks.

5.2.1 Integrating Multifaceted Data Streams for Enhanced Predictive Models.

- **Multi-modal data fusion.** Our survey has highlighted the potential of leveraging diverse data sources, such as travel patterns, social media trends, and climate and environmental conditions, to enrich infectious disease prediction models. A vital issue worthy of further investigation lies in preserving the inherent structure of these varied data types–ranging from simple vectors to higher-order tensors to complex networks–during integration. Future research should focus on developing fusion techniques that maintain the spatial and temporal integrity as well as other high-order dependencies of such data, thereby enabling a more comprehensive understanding of disease dynamics.
- **Interdisciplinary knowledge integration.** The synthesis of insights from epidemiology, climatology, sociology, and other relevant fields is crucial for a comprehensive approach to predicting disease risks. Mere data fusion is inadequate; it is of great importance to integrate domain-specific knowledge into the foundational structure of our predictive models. For instance, incorporating epidemiological factors and environmental variables in an appropriate manner has been shown to significantly enhance the predictive accuracy of even basic regression models, surpassing that of more sophisticated models that rely solely on data-driven methodologies [97]. Therefore, a multidisciplinary synthesis of science and knowledge, beyond simple data fusion, is essential for devising robust and accurate predictive models.

5.2.2 Advancing Machine Learning Methodologies.

- **Model transferability.** Research on the transferability of models between regions and diseases can expedite response strategies, especially where data is limited or in the face of newly emerging diseases. Future investigation along this direction could focus on analyzing the regional/disease-specific factors that influence disease spread and model performance, developing novel machine learning models for effective knowledge transfer across geographical regions or diseases; and theoretically and empirically evaluating the model transferability in various scenarios.
- **Model adaptability.** The adaptability of machine learning models to rapidly evolving transmission patterns is crucial for public health intervention. The concept of performative prediction [64], which posits that predictions can influence the outcomes they forecast, holds significant promise in infectious disease modeling. Future research should aim at understanding how predictive models alter public behaviors and intervention policies, and how these changes should be incorporated into model training.
- **Large foundation models (LFMs).** LFMs such as GPT-4, BERT, and other transformer-based architectures have demonstrated their capability in processing large datasets,

capturing complex patterns and interactions between multiple variables, and transferring learned knowledge to new tasks with minimal fine-tuning [133]. This adaptability is particularly beneficial in infectious disease risk prediction, where timely and accurate models can significantly impact public health [118]. Future work could focus on fine-tuning LFMs with disease-specific data, exploring their ability to incorporate real-time changes in disease dynamics, and evaluating their performance in different geographical and socio-economic settings.

5.2.3 Addressing Practical Issues in Machine Learning Applications.

- **Ethical considerations in machine learning deployment.** It is critical to address ethical issues, such as data biases that could lead to healthcare inequalities. This research should also consider the potential consequences of both false positives and false negatives in disease risk prediction.
- **Advancing explainable AI (XAI) for Public Health.** XAI can enhance transparency and trust in machine learning models. By providing understandable predictions, XAI enables health officials to make more informed decisions and potentially identify underlying factors driving disease outbreaks [146]. Research in this area should focus on developing models that can clearly interpret the significance and impact of various risk-related factors, such as social distancing measures or vaccination rates, on the predicted spread of an infectious disease.
- **Establishing benchmark standards.** In infectious disease modeling, researchers from epidemiology, public health, data science, and other areas bring diverse methodologies to data collection, preprocessing, and various approaches to model evaluation and interpretation. This diversity, while enriching, leads to a notable absence of standardized datasets and metrics specifically tailored for infectious disease risk prediction [56]. Consequently, the establishment of accessible, widely accepted benchmarks that adhere to data privacy standards is of great importance. Such benchmarks would enable consistent comparative evaluations and accelerate advancements in the field.
- **Privacy-preserving techniques.** With the sensitive nature of infectious disease related data, privacy-preserving methodologies like federated learning warrant further exploration [140]. This approach allows for collaborative model training across decentralized devices holding local data samples, ensuring privacy while still benefiting from diverse data sources.

References

[1] Asmaa Abbas, Mohammed M. Abdelsamea, and Mohamed Medhat Gaber. 2021. 4S-DT: Self-supervised super sample decomposition for transfer learning with application to COVID-19 detection. *IEEE Transactions on Neural Networks and Learning Systems* 32, 7 (2021), 2798–2808.

[2] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using Twitter data. In *Proceedings of the 2011 IEEE Conference on Computer Communications Workshops.* IEEE, Piscataway, NJ, USA, 702–707. DOI: https://doi.org/10.1109/INFCOMW.2011.5928903

[3] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B. Aditya Prakash. 2019. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM, New York, NY, USA, 577–586. DOI: https://doi.org/10.1145/3292500.3330917

[4] Aniruddha Adiga, Lijing Wang, Benjamin Hurt, Akhil Peddireddy, Przemyslaw Porebski, Srinivasan Venkatramanan, Bryan Leroy Lewis, and Madhav Marathe. 2021. All models are useful: Bayesian ensembling for robust high resolution covid-19 forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* . ACM, New York, NY, USA, 2505–2513. DOI: https://doi.org/10.1145/3447548.3467197

[5] Said Agrebi and Anis Larbi. 2020. Use of artificial intelligence in infectious diseases. In *Proceedings of the Artificial Intelligence in Precision Health.* Elsevier, Amsterdam, The Netherlands, 415–438.

[6] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.* ACM, New York, NY, USA, 559–560. DOI: https://doi.org/10.1145/3233547.3233667

[7] Mahmood Akhtar, Moritz U. G. Kraemer, and Lauren M. Gardner. 2019. A dynamic neural network model for predicting risk of Zika in real time. *BMC Medicine* 17 (2019), 1–16. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1389-3

[8] Furqan Ali, Farman Ullah, Junaid Iqbal Khan, Jebran Khan, Abdul Wasay Sardar, and Sungchang Lee. 2023. COVID-19 spread control policies based early dynamics forecasting using deep learning algorithm. *Chaos, Solitons & Fractals* 167 (2023), 112984. https://www.sciencedirect.com/science/article/pii/S0960077922011638

[9] Alfred B. Amendolara, David Sant, Horacio G. Rotstein, and Eric Fortune. 2023. LSTM-based recurrent neural network provides effective short term flu forecasting. *BMC Public Health* 23, 1 (2023), 1788.

[10] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. (2016). arXiv:1606.06565. Retrieved from https://arxiv.org/abs/1606.06565

[11] Jeffrey L. Anderson. 2001. An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review* 129, 12 (2001), 2884–2903.

[12] Sercan Arik, Chun-Liang Li, Jinsung Yoon, Rajarishi Sinha, Arkady Epshteyn, Long Le, Vikas Menon, Shashank Singh, Leyou Zhang, Martin Nikoltchev, et al. 2020. Interpretable sequence learning for COVID-19 forecasting. *Advances in Neural Information Processing Systems* 33 (2020), 18807–18818. https://proceedings.neurips.cc/paper/2020/hash/d9dbc51dc534921589adf460c85cd824-Abstract.html

[13] M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* 50, 2 (2002), 174–188.

[14] Jackie Baek, Vivek F. Farias, Andreea Georgescu, Retsef Levi, Tianyi Peng, Deeksha Sinha, Joshua Wilde, and Andrew Zheng. 2020. The limits to learning an SIR Process: Granular Forecasting for COVID-19. (2020). arXiv:2006.06373v1. Retrieved from https://arxiv.org/abs/2006.06373v1

[15] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21484–21489.

[16] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J. Ramasco, Vittoria Colizza, and Alessandro Vespignani. 2010. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of Computational Science* 1, 3 (2010), 132–145.

[17] Soumyanil Banerjee, Ming Dong, and Weisong Shi. 2022. Spatial–temporal synchronous graph transformer network (stsgt) for covid-19 forecasting. *Smart Health* 26 (2022), 100348. https://www.sciencedirect.com/science/article/pii/S2352648322000824

[18] 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. 2016. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases* 16, 1 (2016), 1–10.

[19] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY.

[20] Arthur Bousquet, William H. Conrad, Said Omer Sadat, Nelli Vardanyan, and Youngjoon Hong. 2022. Deep learning forecasting using time-varying parameters of the SIRD model for Covid-19. *Scientific Reports* 12, 1 (2022), 3030.

[21] Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. 2021. Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology* 17, 2 (2021), e1008618.

[22] Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. 2015. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Computational Biology* 11, 8 (2015), e1004382.

[23] David A. Buch, James E. Johndrow, and David B. Dunson. 2023. Explaining transmission rate variations and forecasting epidemic spread in multiple regions with a semiparametric mixed effects SIR model. *Biometrics* 79, 4 (2023), 2987–2997.

[24] Edgar Camargo, Jose Aguilar, Yullis Quintero, F. Rivas, and D. Ardila. 2022. An incremental learning approach to prediction models of SEIRD variables in the context of the COVID-19 pandemic. *Health and Technology* 12, 4 (2022), 867–877.

[25] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems* 33 (2020), 17766–17778. https://proceedings.neurips.cc/paper/2020/hash/cdf6581cb7aca4b7e19ef136c6e601a5-Abstract.html

[26] Qi Cao, Renhe Jiang, Chuang Yang, Zipei Fan, Xuan Song, and Ryosuke Shibasaki. 2022. MepoGNN: Metapopulation epidemic forecasting with graph neural networks. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer Nature Switzerland, Cham, 453–468.

[27] Pietro Ceccato, Christelle Vancutsem, Robert Klaver, James Rowland, and Stephen J. Connor. 2012. A vectorial capacity product to monitor changing malaria transmission potential in epidemic regions of Africa. *Journal of Tropical Medicine* 2012, 1 (2012), 595948. DOI: https://doi.org/10.1155/2012/595948 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1155/2012/595948

[28] Indranil Chakraborty and Prasenjit Maity. 2020. COVID-19 outbreak: Migration, effects on society, global environment and prevention. *Science of the Total Environment* 728 (2020), 138882. https://www.sciencedirect.com/science/article/pii/S0048969720323998

[29] 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. 2014. Forecasting a moving target: Ensemble models for ILI case count predictions. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, Philadelphia, USA, 262–270. DOI: https://doi.org/10.1137/1.9781611973440.30 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611973440.30

[30] Tanujit Chakraborty, Swarup Chattopadhyay, and Indrajit Ghosh. 2019. Forecasting dengue epidemics using a hybrid methodology. *Physica A: Statistical Mechanics and its Applications* 527 (2019), 121266. https://www.sciencedirect.com/science/article/abs/pii/S0378437119307320

[31] Shi Chen, Rajib Paul, Daniel Janies, Keith Murphy, Tinghao Feng, and Jean-Claude Thill. 2021. Exploring feasibility of multivariate deep learning models in predicting COVID-19 epidemic. *Frontiers in Public Health* 9 (2021), 661615. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.661615/full

[32] Hao-Yuan Cheng, Yu-Chun Wu, Min-Hau Lin, Yu-Lun Liu, Yue-Yang Tsai, Jo-Hua Wu, Ke-Han Pan, Chih-Jung Ke, Chiu-Mei Chen, Ding-Ping Liu, et al. 2020. Applying machine learning models with an ensemble approach for accurate real-time influenza forecasting in Taiwan: Development and validation study. *Journal of Medical Internet Research* 22, 8 (2020), e15394.

[33] Nakul Chitnis, Allan Schpira, David Smith, Simon I. Hay, Thomas Smith, Richard Steketee, et al. 2010. *Mathematical Modelling to Support Malaria Control and Elimination*. World Health Organization, Geneva.

[34] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. (2014). arXiv:1406.1078. Retrieved from https://arxiv.org/abs/1406.1078

[35] Mark Ciaccio, Chris Schneiderman, Abhishek Pandey, Robert Fowler, Kevin Chiou, Gage Koeller, David Hallett, Whitney Krueger, and Leon Raskin. 2023. A time-course prediction model of global COVID-19 mortality. *Frontiers in Public Health* 11 (2023), 1232531. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2023.1232531/full

[36] Shaoze Cui, Yanzhang Wang, Dujuan Wang, Qian Sai, Ziheng Huang, and TCE Cheng. 2021. A two-layer nested heterogeneous ensemble learning predictive method for COVID-19 mortality. *Applied Soft Computing* 113 (2021), 107946. https://www.sciencedirect.com/science/article/pii/S1568494621008681

[37] Yue Cui, Chen Zhu, Guanyu Ye, Ziwei Wang, and Kai Zheng. 2021. Into the unobservables: A multi-range encoder-decoder framework for COVID-19 prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* ACM, New York, NY, USA, 292–301. DOI: https://doi.org/10.1145/3459637.3482356

[38] Shangfang Dai and Litao Han. 2023. Influenza surveillance with Baidu index and attention-based long short-term memory model. *PLoS One* 18, 1 (2023), e0280834.

[39] Sujata Dash, Sourav Kumar Giri, Saurav Mallik, Subhendu Kumar Pani, Mohd Asif Shah, and Hong Qin. 2024. Predictive healthcare modeling for early pandemic assessment leveraging deep auto regressor neural prophet. *Scientific Reports* 14, 1 (2024), 5287.

[40] Emerson Vilar de Oliveira, Dunfrey Pires Aragão, and Luiz Marcos Garcia Gonçalves. 2024. A new auto-regressive multi-variable modified auto-encoder for multivariate time-series prediction: A case study with application to COVID-19 pandemics. *International Journal of Environmental Research and Public Health* 21, 4 (2024), 497.

[41] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. 2020. Cola-GNN: Cross-location attention based graph neural networks for long-term ILI prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* ACM, New York, NY, USA, 245–254. DOI: https://doi.org/10.1145/3340531.3411975

[42] Marcel Dettling. 2013. Applied Time Series Analysis. Lecture notes, CH-8401. (2013). Retrieved March 4, 2025 from https://stat.ethz.ch/education/semesters/ss2015/atsa/ATSA_Scriptum_v1_SS15.pdf

[43] Ousmane Diao, P-A Absil, and Mouhamadou Diallo. 2023. Generalized linear models to forecast malaria incidence in three endemic regions of senegal. *International Journal of Environmental Research and Public Health* 20, 13 (2023), 6303.

[44] Zhiwei Ding, Feng Sha, Yi Zhang, and Zhouwang Yang. 2023. Biology-informed recurrent neural network for pandemic prediction using multimodal data. *Biomimetics* 8, 2 (2023), 158.

[45] Minhui Dong, Cheng Tang, Junkai Ji, Qiuzhen Lin, and Ka-Chun Wong. 2021. Transmission trend of the COVID-19 pandemic predicted by dendritic neural regression. *Applied Soft Computing* 111 (2021), 107683. https://www.sciencedirect.com/science/article/pii/S1568494621006049

[46] Finale Doshi-Velez and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. (2017). arXiv:arXiv:1702.08608. Retrieved from https://arxiv.org/abs/1702.08608

[47] Vanja Dukic, Hedibert F. Lopes, and Nicholas G. Polson. 2012. Tracking epidemics with Google flu trends data and a state-space SEIR model. *J. Amer. Statist. Assoc.* 107, 500 (2012), 1410–1426.
[48] Rayane Elimam, Nicolas Sutton-Charani, Stéphane Perrey, and Jacky Montmain. 2022. Uncertain imputation for time-series forecasting: Application to COVID-19 daily mortality prediction. *PLOS Digital Health* 1, 10 (2022), e0000115.
[49] Abdelrahman E. E. Eltoukhy, Ibrahim Abdelfadeel Shaban, Felix T. S. Chan, and Mohammad A. M. Abdel-Aal. 2020. Data analytics for predicting COVID-19 cases in top affected countries: Observations and recommendations. *International Journal of Environmental Research and Public Health* 17, 19 (2020), 7080.
[50] Lin Feng, Ziren Chen, Harold A. Lay Jr, Khaled Furati, and Abdul Khaliq. 2022. Data driven time-varying SEIR-LSTM/GRU algorithms to track the spread of COVID-19. *Mathematical Biosciences and Engineering* 19, 9 (2022), 8935–8962.
[51] Cornelius Fritz, Emilio Dorigatti, and David Rügamer. 2022. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. *Scientific Reports* 12, 1 (2022), 3930.
[52] Joseph Galasso, Duy M. Cao, and Robert Hochberg. 2022. A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. *Chaos, Solitons & Fractals* 156 (2022), 111779. https://www.sciencedirect.com/science/article/pii/S0960077921011334
[53] Junyi Gao, Rakshith Sharma, Cheng Qian, Lucas M. Glass, Jeffrey Spaeder, Justin Romberg, Jimeng Sun, and Cao Xiao. 2021. STAN: Spatio-temporal attention network for pandemic prediction using real-world evidence. *Journal of the American Medical Informatics Association* 28, 4 (2021), 733–743.
[54] Junyi Gao, Cao Xiao, Lucas M. Glass, and Jimeng Sun. 2022. PopNet: Real-time population-level disease prediction with data latency. In *Proceedings of the ACM Web Conference.* ACM, New York, NY, USA, 2552–2562. DOI: https://doi.org/10.1145/3485447.3512127
[55] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Eslami, and Yee Whye Teh. 2018. Neural Processes. (2018). arXiv:1807.01622. Retrieved from https://arxiv.org/abs/1807.01622
[56] Soudeh Ghafouri-Fard, Hossein Mohammad-Rahimi, Parisa Motie, Mohammad A. S. Minabi, Mohammad Taheri, and Saeedeh Nateghinia. 2021. Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review. *Heliyon* 7, 10 (2021), e08143. DOI: https://doi.org/10.1016/j.heliyon.2021.e08143
[57] Biraja Ghoshal and Allan Tucker. 2020. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. (2020). arXiv:2003.10769. Retrieved from https://arxiv.org/abs/2003.10769
[58] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.
[59] Yan-Feng Gong, Ling-Qian Zhu, Yin-Long Li, Li-Juan Zhang, Jing-Bo Xue, Shang Xia, Shan Lv, Jing Xu, and Shi-Zhu Li. 2021. Identification of the high-risk area for schistosomiasis transmission in China based on information value and machine learning: A newly data-driven modeling attempt. *Infectious Diseases of Poverty* 10 (2021), 1–11. https://link.springer.com/article/10.1186/s40249-021-00874-9
[60] Daniel Alejandro Gónzalez-Bandala, Juan Carlos Cuevas-Tello, Daniel E. Noyola, Andreu Comas-García, and Christian A García-Sepúlveda. 2020. Computational forecasting methodology for acute respiratory infectious disease dynamics. *International Journal of Environmental Research and Public Health* 17, 12 (2020), 4540.
[61] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. 2017. Improved deep embedded clustering with local structure preservation. In *Proceedings of the 26th International Joint Conferences on Artificial Intelligence.* International Joint Conferences on Artificial Intelligence Organization, California, USA, 1753–1759. DOI: https://doi.org/10.24963/ijcai.2017/243
[62] Meenu Gupta, Rachna Jain, Soham Taneja, Gopal Chaudhary, Manju Khari, and Elena Verdú. 2021. Real-time measurement of the uncertain epidemiological appearances of COVID-19 infections. *Applied Soft Computing* 101 (2021), 107039. https://www.sciencedirect.com/science/article/pii/S1568494620309777
[63] Barbara A. Han and John M. Drake. 2016. Future directions in analytics for infectious disease intelligence: Toward an integrated warning system for emerging pathogens. *EMBO Reports* 17, 6 (2016), 785–789.
[64] Moritz Hardt and Celestine Mendler-Dünner. 2023. Performative prediction: Past and future. (2023). arXiv:2310.16608. Retrieved from https://arxiv.org/abs/2310.16608
[65] Inga Holmdahl and Caroline Buckee. 2020. Wrong but useful–what covid-19 epidemiologic models can and cannot tell us. *New England Journal of Medicine* 383, 4 (2020), 303–305.
[66] Ting Hua, Chandan K. Reddy, Lei Zhang, Lijing Wang, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. 2018. Social media based simulation models for understanding disease dynamics. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence.* International Joint Conferences on Artificial Intelligence Organization, California, USA, 3797–3804. DOI: https://doi.org/10.24963/ijcai.2018/528
[67] Raghvendra Jain, Sra Sontisirikit, Sopon Iamsirithaworn, and Helmut Prendinger. 2019. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. *BMC Infectious Diseases* 19 (2019), 1–16. https://link.springer.com/article/10.1186/s12879-019-3874-x

[68] Weiqiu Jin, Shuqing Dong, Chengqing Yu, and Qingquan Luo. 2022. A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning. *Computers in Biology and Medicine* 146 (2022), 105560. https://www.sciencedirect.com/science/article/pii/S0010482522003523

[69] Se Young Jung, Hyeontae Jo, Hwijae Son, and Hyung Ju Hwang. 2020. Real-world implications of a rapidly responsive COVID-19 spread model with time-dependent parameters via deep learning: Model development and validation. *Journal of Medical Internet Research* 22, 9 (2020), e19907.

[70] Abdul Aziz K Abdul Hamid, Wan Imanul Aisyah Wan Mohamad Nawi, Muhamad Safiih Lola, Wan Azani Mustafa, Siti Madhihah Abdul Malik, Syerrina Zakaria, Elayaraja Aruchunan, Nurul Hila Zainuddin, RU Gobithaasan, and Mohd Tajuddin Abdullah. 2023. Improvement of time forecasting models using machine learning for future pandemic applications based on COVID-19 data 2020–2022. *Diagnostics* 13, 6 (2023), 1121.

[71] Eric Kamana and Jijun Zhao. 2023. Deep learning hybrid model for analyzing and predicting the impact of imported malaria cases from Africa on the rise of Plasmodium falciparum in China before and during the COVID-19 pandemic. *Plos One* 18, 12 (2023), e0287702.

[72] Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. 2021. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. *Advances in Neural Information Processing Systems* 34 (2021), 19796–19807. https://proceedings.neurips.cc/paper/2021/hash/a4a1108bbcc329a70efa93d7bf060914-Abstract.html

[73] Sasikiran Kandula, Daniel Hsu, and Jeffrey Shaman. 2017. Subregional nowcasts of seasonal influenza using search trends. *Journal of Medical Internet Research* 19, 11 (2017), e370.

[74] I-Hsi Kao and Jau-Woei Perng. 2021. Early prediction of coronavirus disease epidemic severity in the contiguous United States based on deep learning. *Results in Physics* 25 (2021), 104287. https://www.sciencedirect.com/science/article/pii/S2211379721004216

[75] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O'Banion. 2020. Examining Covid-19 forecasting using spatio-temporal graph neural networks. (2020). arXiv:2007.03113. Retrieved from https://arxiv.org/abs/2007.03113

[76] Nikos Kargas, Cheng Qian, Nicholas D. Sidiropoulos, Cao Xiao, Lucas M. Glass, and Jimeng Sun. 2021. STELAR: Spatio-temporal tensor factorization with latent epidemiological regularization. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6 (2021), 4830–4837. DOI: https://doi.org/10.1609/aaai.v35i6.16615

[77] Shwet Ketu and Pramod Kumar Mishra. 2021. Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection. *Applied Intelligence* 51, 3 (2021), 1492–1512.

[78] Junaid Iqbal Khan, Farman Ullah, and Sungchang Lee. 2022. Attention based parameter estimation and states forecasting of COVID-19 pandemic using modified SIQRD Model. *Chaos, Solitons & Fractals* 165 (2022), 112818.

[79] Richard Kiang, Farida Adimi, Valerii Soika, Joseph Nigro, Pratap Singhasivanon, Jeeraphat Sirichaisinthop, Somjai Leemingsawat, Chamnarn Apiwathnasorn, and Sornchai Looareesuwan. 2006. Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. *Geospatial Health* 1, 1 (2006), 71–84.

[80] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. 2019. Attentive neural processes. (2019). arXiv:1901.05761. Retrieved from https://arxiv.org/abs/1901.05761

[81] Juhyeon Kim and Insung Ahn. 2019. Weekly ILI patient ratio change prediction using news articles with support vector machine. *BMC Bioinformatics* 20 (2019), 1–16. https://link.springer.com/article/10.1186/s12859-019-2894-2

[82] Yeongha Kim, Chang-Reung Park, Jae-Pyoung Ahn, and Beakcheol Jang. 2023. COVID-19 outbreak prediction using Seq2Seq+ Attention and Word2Vec keyword time series data. *Plos One* 18, 4 (2023), e0284298.

[83] László Róbert Kolozsvári, Tamás Bérczes, András Hajdu, Rudolf Gesztelyi, Attila Tiba, Imre Varga, B Ala'a, Gergő József Szőllősi, Szilvia Harsányi, Szabolcs Garbóczy, et al. 2021. Predicting the epidemic curve of the coronavirus (SARS-CoV-2) disease (COVID-19) using artificial intelligence: An application on the first and second waves. *Informatics in Medicine Unlocked* 25 (2021), 100691. https://www.sciencedirect.com/science/article/pii/S2352914821001751

[84] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[85] Wouter M. Kouw and Marco Loog. 2018. An introduction to domain adaptation and transfer learning. (2018). arXiv:1812.11806. Retrieved from https://arxiv.org/abs/1812.11806

[86] Olga Krivorotko, Mariia Sosnovskaia, Ivan Vashchenko, Cliff Kerr, and Daniel Lesnic. 2022. Agent-based modeling of COVID-19 outbreaks for New York state and UK: Parameter identification algorithm. *Infectious Disease Modelling* 7, 1 (2022), 30–44.

[87] R. Lakshmana Kumar, Firoz Khan, Sadia Din, Shahab S. Band, Amir Mosavi, and Ebuka Ibeke. 2021. Recurrent neural network and reinforcement learning model for COVID-19 prediction. *Frontiers in Public Health* 9 (2021), 744100. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.744100/full

[88] Cheng-Pin Kuo and Joshua S. Fu. 2021. Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions. *Science of The Total Environment* 758 (2021), 144151. https://www.sciencedirect.com/science/article/pii/S0048969720376828

[89] Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperli. 2020. An epidemiological neural network exploiting dynamic graph structured data applied to the COVID-19 outbreak. *IEEE Transactions on Big Data* 7, 1 (2020), 45–55.

[90] Mallory Lai, Yongtao Cao, Shaun S. Wulff, Timothy J. Robinson, Alexys McGuire, and Bledar Bisha. 2023. A time series based machine learning strategy for wastewater-based forecasting and nowcasting of COVID-19 dynamics. *Science of The Total Environment* 897 (2023), 165105. https://www.sciencedirect.com/science/article/pii/S0048969723037282

[91] Mallory Lai, Shaun S. Wulff, Yongtao Cao, Timothy J. Robinson, and Rasika Rajapaksha. 2023. An interpretable time series machine learning method for varying forecast and nowcast lengths in wastewater-based epidemiology. *MethodsX* 11 (2023), 102382. https://www.sciencedirect.com/science/article/pii/S2215016123003783

[92] Liang Li, Jianye Zhou, Yuewen Jiang, and Biqing Huang. 2021. Propagation source identification of infectious diseases with graph convolutional networks. *Journal of Biomedical Informatics* 116 (2021), 103720. https://www.sciencedirect.com/science/article/pii/S1532046421000496

[93] Zhifang Liao, Peng Lan, Zhining Liao, Yan Zhang, and Shengzong Liu. 2020. TW-SIR: Time-window based SIR for COVID-19 forecasts. *Scientific Reports* 10, 1 (2020), 22454.

[94] Chen Lin, Jianghong Zhou, Jing Zhang, Carl Yang, and Eugene Agichtein. 2023. Graph neural network modeling of web search activity for real-time pandemic forecasting. In *Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics*. IEEE, Piscataway, NJ, USA, 128–137. DOI: https://doi.org/10.1109/ICHI57859.2023.00027

[95] Kang Liu, Ling Yin, Meng Zhang, Min Kang, Ai-Ping Deng, Qing-Lan Li, and Tie Song. 2021. Facilitating fine-grained intra-urban dengue forecasting by integrating urban environments measured from street-view images. *Infectious Diseases of Poverty* 10 (2021), 1–16. https://link.springer.com/article/10.1186/s40249-021-00824-5

[96] Mutong Liu, Yang Liu, and Jiming Liu. 2023. Epidemiology-aware deep learning for infectious disease dynamics prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* . ACM, New York, NY, USA, 4084–4088. DOI: https://doi.org/10.1145/3583780.3615139

[97] Mutong Liu, Yang Liu, Ly Po, Shang Xia, Rekol Huy, Xiao-Nong Zhou, and Jiming Liu. 2023. Assessing the spatiotemporal malaria transmission intensity with heterogeneous risk factors: A modeling study in Cambodia. *Infectious Disease Modelling* 8, 1 (2023), 253–269.

[98] Qian Liu, Daryl L. X. Fung, Leann Lac, and Pingzhao Hu. 2021. A novel matrix profile-guided attention LSTM model for forecasting COVID-19 cases in USA. *Frontiers in Public Health* 9 (2021), 741030. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.741030/full

[99] Xu-dong Liu, Bo-han Hou, Zhong-jun Xie, Ning Feng, and Xiao-ping Dong. 2024. Integrating gated recurrent unit in graph neural network to improve infectious disease prediction: An attempt. *Frontiers in Public Health* 12 (2024), 1397260. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1397260/full

[100] Xu-Dong Liu, Wei Wang, Yi Yang, Bo-Han Hou, Toba Stephen Olasehinde, Ning Feng, and Xiao-Ping Dong. 2023. Nesting the SIRV model with NAR, LSTM and statistical methods to fit and predict COVID-19 epidemic trend in Africa. *BMC Public Health* 23, 1 (2023), 138.

[101] Yahui Long, Jiawei Luo, Yu Zhang, and Yan Xia. 2021. Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Briefings in Bioinformatics* 22, 3 (2021), bbaa146.

[102] Yahui Long, Yu Zhang, Min Wu, Shaoliang Peng, Chee Keong Kwoh, Jiawei Luo, and Xiaoli Li. 2022. Heterogeneous graph attention networks for drug virus association prediction. *Methods* 198 (2022), 11–18. https://www.sciencedirect.com/science/article/pii/S1046202321002012

[103] Christos Louizos, Xiahan Shi, Klamer Schutte, and Max Welling. 2019. *The Functional Neural Process*. Curran Associates Inc., Red Hook, NY, USA, 8746–8757.

[104] Anice C. Lowen, Samira Mubareka, John Steel, and Peter Palese. 2007. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathogens* 3, 10 (2007), e151.

[105] Benjamin Lucas, Behzad Vahedi, and Morteza Karimzadeh. 2023. A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA. *International Journal of Data Science and Analytics* 15, 3 (2023), 247–266.

[106] Helmut Lütkepohl and Markus Krätzig. 2004. *Applied Time Series Econometrics*. Cambridge University Press, Cambridge.

[107] 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. 2019. Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples. *BMC Public Health* 19, 1 (2019), 1–12.

[108] Zhuanghu Lv, Jing Li, Dafeng Liu, Yue Peng, and Benyun Shi. 2021. STANN: Spatio-temporal attention-based neural network for epidemic prediction. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. ACM, New York, NY, USA, 314–319. DOI: https://doi.org/10.1145/3498851.3498972

[109] Piklu Mallick, Sourav Bhowmick, and Surajit Panja. 2022. Prediction of COVID-19 infected population for Indian States through a State Interaction Network-based SEIR Epidemic Model. *Ifac-papersonline* 55, 1 (2022), 691–696.
[110] Sarah F. McGough, Leonardo Clemente, J. Nathan Kutz, and Mauricio Santillana. 2021. A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles. *Journal of The Royal Society Interface* 18, 179 (2021), 20201006.
[111] Mika A. Merrill and Tim Althoff. 2023. Self-supervised pretraining and transfer learning enable flu and COVID-19 predictions in small mobile sensing datasets. In *Proceedings of the Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research)*. Bobak J. Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C. Ho (Eds.), Vol. 209, PMLR, Cambridge, MA, USA, 191–206. Retrieved from https://proceedings.mlr.press/v209/merrill23a.html
[112] L. J. Muhammad, Ahmed Abba Haruna, Usman Sani Sharif, and Mohammed Bappah Mohammed. 2022. CNN-LSTM deep learning based forecasting model for COVID-19 infection cases in Nigeria, South Africa and Botswana. *Health and Technology* 12, 6 (2022), 1259–1276.
[113] Duc Q. Nguyen, Nghia Q. Vo, Thinh T. Nguyen, Khuong Nguyen-An, Quang H. Nguyen, Dang N. Tran, and Tho T. Quan. 2022. BeCaked: An explainable artificial intelligence model for COVID-19 forecasting. *Scientific Reports* 12, 1 (2022), 7969.
[114] Behnam Nikparvar, Md Mokhlesur Rahman, Faizeh Hatami, and Jean-Claude Thill. 2021. Spatio-temporal prediction of the COVID-19 pandemic in US counties: Modeling with a deep LSTM neural network. *Scientific Reports* 11, 1 (2021), 21715.
[115] Elaine O. Nsoesie, Scotland C. Leman, and Madhav V. Marathe. 2014. A Dirichlet process model for classifying and forecasting epidemic curves. *BMC Infectious Diseases* 14 (2014), 1–12. https://link.springer.com/article/10.1186/1471-2334-14-12
[116] Jose Olmo and Marcos Sanso-Navarro. 2021. Modeling the spread of COVID-19 in New York City. *Papers in Regional Science* 100, 5 (2021), 1209–1230.
[117] Ebenezer O. Oluwasakin and Abdul Q. M. Khaliq. 2023. Data-Driven deep learning neural networks for predicting the number of individuals infected by COVID-19 Omicron variant. *Epidemiologia* 4, 4 (2023), 420–453.
[118] Mahmud Omar, Dana Brin, Benjamin Glicksberg, and Eyal Klang. 2024. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A Systematic Review. *American Journal of Infection Control* 52, 9 (2024), 992–1001. DOI: https://doi.org/10.1016/j.ajic.2024.03.016
[119] World Health Organization. 2023. *World Malaria Report 2023*. World Health Organization, Geneva. Retrieved March 4, 2025 from https://books.google.com.hk/books?id=u6UOEQAAQBAJ
[120] Dave Osthus, James Gattiker, Reid Priedhorsky, and Sara Y. Del Valle. 2019. Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Analysis* 14, 1 (2019), 261–312.
[121] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2009), 1345–1359.
[122] Madhurima Panja, Tanujit Chakraborty, Uttam Kumar, and Nan Liu. 2023. Epicasting: An ensemble wavelet neural network for forecasting epidemics. *Neural Networks* 165 (2023), 185–212. https://www.sciencedirect.com/science/article/abs/pii/S0893608023002939
[123] Eirini Papagiannopoulou, Matías Nicolás Bossa, Nikos Deligiannis, and Hichem Sahli. 2024. Long-term regional influenza-like-illness forecasting using exogenous data. *IEEE Journal of Biomedical and Health Informatics* 28, 6 (2024), 3781–3792. DOI: https://doi.org/10.1109/JBHI.2024.3377529
[124] Hongbin Pei, Bo Yang, Jiming Liu, and Kevin Chen-Chuan Chang. 2022. Active surveillance via group sparse bayesian learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022), 1133–1148. DOI: https://doi.org/10.1109/TPAMI.2020.3023092
[125] Hongbin Pei, Bo Yang, Jiming Liu, and Lei Dong. 2018. Group sparse bayesian learning for active surveillance on epidemic dynamics. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018), 800–807. DOI: https://doi.org/10.1609/aaai.v32i1.11344
[126] Sen Pei, Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. 2018. Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2752–2757.
[127] Daniela Perrotta, Michele Tizzoni, and Daniela Paolotti. 2017. Using participatory Web-based surveillance data to improve seasonal influenza forecasting in Italy. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 303–310. DOI: https://doi.org/10.1145/3038912.3052670
[128] Bastian Prasse, Massimo A. Achterberg, Long Ma, and Piet Van Mieghem. 2020. Network-inference-based prediction of the COVID-19 epidemic outbreak in the Chinese province Hubei. *Applied Network Science* 5 (2020), 1–11. https://link.springer.com/article/10.1007/s41109-020-00274-2
[129] Kiesha Prem, Yang Liu, Timothy W. Russell, Adam J. Kucharski, Rosalind M. Eggo, Nicholas Davies, Stefan Flasche, Samuel Clifford, Carl A. B. Pearson, James D. Munday, et al. 2020. 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 (2020), e261–e270.

[130] Bradley S. Price, Maryam Khodaverdi, Brian Hendricks, Gordon S. Smith, Wes Kimble, Adam Halasz, Sara Guthrie, Julia D. Fraustino, and Sally L. Hodder. 2024. Enhanced SARS-CoV-2 case prediction using public health data and machine learning models. *JAMIA Open* 7, 1 (2024), ooae014.

[131] Shenghao Qin, Jiacheng Zhu, Jimmy Qin, Wenshuo Wang, and Ding Zhao. 2019. Recurrent attentive neural process for sequential data. (2019). arXiv:1910.09323. Retrieved from https://arxiv.org/abs/1910.09323

[132] Zongxi Qu, Beidou Zhang, and Hongpeng Wang. 2023. A multivariate deep learning model with coupled human intervention factors for COVID-19 forecasting. *Systems* 11, 4 (2023), 201.

[133] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[134] Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 378 (2019), 686–707. https://www.sciencedirect.com/science/article/abs/pii/S0021999118307125

[135] Ankit Ramchandani, Chao Fan, and Ali Mostafavi. 2020. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *IEEE Access* 8 (2020), 159915–159930. https://ieeexplore.ieee.org/abstract/document/9179729

[136] Essam A. Rashed and Akimasa Hirata. 2021. One-year lesson: Machine learning prediction of COVID-19 positive cases with meteorological data and mobility estimate in Japan. *International Journal of Environmental Research and Public Health* 18, 11 (2021), 5736.

[137] Evan L. Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y. Cramer, Johannes Bracher, Andrew Zheng, Teresa K. Yamana, Xinyue Xiong, et al. 2020. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the US. (2020). DOI: https://doi.org/10.1101/2020.08.19.20177493 medrxiv:2020.08.19.20177493

[138] Nicholas G. Reich, Craig J. McGowan, Teresa K. Yamana, Abhinav Tushar, Evan L. Ray, Dave Osthus, Sasikiran Kandula, Logan C. Brooks, Willow Crawford-Crudell, Graham Casey Gibson, et al. 2019. Accuracy of real-time multimodel ensemble forecasts for seasonal influenza in the US. *PLoS Computational Biology* 15, 11 (2019), e1007486.

[139] Jinfu Ren, Mutong Liu, Yang Liu, and Jiming Liu. 2023. TransCode: Uncovering COVID-19 transmission patterns via deep learning. *Infectious Diseases of Poverty* 12, 1 (2023), 14.

[140] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ Digital Medicine* 3, 1 (2020), 1–7.

[141] Alexander Rodríguez, Jiaming Cui, Naren Ramakrishnan, Bijaya Adhikari, and B. Aditya Prakash. 2023. EINNs: Epidemiologically-informed neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (2023), 14453–14460. DOI: https://doi.org/10.1609/aaai.v37i12.26690

[142] Alexander Rodríguez, Nikhil Muralidhar, Bijaya Adhikari, Anika Tabassum, Naren Ramakrishnan, and B. Aditya Prakash. 2021. Steering a historical disease forecasting model under a pandemic: Case of flu and COVID-19. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6 (2021), 4855–4863. DOI: https://doi.org/10.1609/aaai.v35i6.16618

[143] Alexander Rodriguez, Anika Tabassum, Jiaming Cui, Jiajia Xie, Javen Ho, Pulak Agarwal, Bijaya Adhikari, and B. Aditya Prakash. 2021. DeepCOVID: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (2021), 15393–15400. DOI: https://doi.org/10.1609/aaai.v35i17.17808

[144] Kirstin Roster, Colm Connaughton, and Francisco A. Rodrigues. 2022. Forecasting new diseases in low-data settings using transfer learning. *Chaos, Solitons & Fractals* 161 (2022), 112306. https://www.sciencedirect.com/science/article/abs/pii/S0960077922005161

[145] Xiaolei Ru, Jack Murdoch Moore, Xin-Ya Zhang, Yeting Zeng, and Gang Yan. 2023. Inferring patient zero on temporal networks via graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 8 (2023), 9632–9640. DOI: https://doi.org/10.1609/aaai.v37i8.26152

[146] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[147] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys* 16 (2022), 1–85. https://projecteuclid.org/journals/statistics-surveys/volume-16/issue-none/Interpretable-machine-learning-Fundamental-principles-and-10-grand-challenges/10.1214/21-SS133.full

[148] Ahmed Ben Said, Abdelkarim Erradi, Hussein Ahmed Aly, and Abdelmonem Mohamed. 2021. Predicting COVID-19 cases using bidirectional LSTM on multivariate time series. *Environmental Science and Pollution Research* 28, 40 (2021), 56043–56052.

[149] Omar Enzo Santangelo, Vito Gentile, Stefano Pizzo, Domiziana Giordano, and Fabrizio Cedrone. 2023. Machine learning and prediction of infectious diseases: A systematic review. *Machine Learning and Knowledge Extraction* 5, 1 (2023), 175–198.

[150] Mohd Saqib. 2021. Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. *Applied Intelligence* 51, 5 (2021), 2703–2713.

[151] Badrul M. Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2002. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the International Conference on Computer and Information Technology*. 291–324.

[152] Nevasini Sasikumar and Krishna Sri Ipsit Mantri. 2023. STAGCN: Spatial-temporal attention based graph convolutional networks for COVID-19 forecasting. In *Proceedings of the 2023 ICLR 1st Workshop on Machine Learning Global Health*. ICLR, Appleton, USA, 1–9. Retrieved from https://openreview.net/forum?id=k0E_VMXLXl

[153] Paul P. Schneider, Christel JAW van Gool, Peter Spreeuwenberg, Mariëtte Hooiveld, Gé A Donker, David J. Barnett, and John Paget. 2020. Using web search queries to monitor influenza-like illness: An exploratory retrospective analysis, Netherlands, 2017/18 influenza season. *Eurosurveillance* 25, 21 (2020), 1900221.

[154] Abdennour Sebbagh and Sihem Kechida. 2022. EKF-SIRD model algorithm for predicting the coronavirus (COVID-19) spreading dynamics. *Scientific Reports* 12, 1 (2022), 13415.

[155] Ransalu Senanayake, Simon O'Callaghan, and Fabio Ramos. 2016. Predicting spatio-temporal propagation of seasonal influenza using variational Gaussian process regression. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (2016), 3901–3907. DOI: https://doi.org/10.1609/aaai.v30i1.9899

[156] Jeffrey Shaman and Alicia Karspeck. 2012. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences* 109, 50 (2012), 20425–20430.

[157] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. 2013. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications* 4, 1 (2013), 1–10.

[158] Jeffrey Shaman and Melvin Kohn. 2009. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences* 106, 9 (2009), 3243–3248.

[159] Jeffrey Shaman, Virginia E. Pitzer, Cécile Viboud, Bryan T. Grenfell, and Marc Lipsitch. 2010. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biology* 8, 2 (2010), e1000316.

[160] Afshar Shamsi, Hamzeh Asgharnezhad, Shirin Shamsi Jokandan, Abbas Khosravi, Parham M. Kebria, Darius Nahavandi, Saeid Nahavandi, and Dipti Srinivasan. 2021. An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis. *IEEE Transactions on Neural Networks and Learning Systems* 32, 4 (2021), 1408–1417.

[161] Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. 2023. Post-hoc uncertainty learning using a dirichlet meta-model. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 8 (2023), 9772–9781. DOI: https://doi.org/10.1609/aaai.v37i8.26167

[162] Benyun Shi, Shan Lin, Qi Tan, Jie Cao, Xiaohong Zhou, Shang Xia, Xiao-Nong Zhou, and Jiming Liu. 2020. Inference and prediction of malaria transmission dynamics using time series data. *Infectious Diseases of Poverty* 9, 1 (2020), 1–13.

[163] Benyun Shi, Jiming Liu, Xiao-Nong Zhou, and Guo-Jing Yang. 2014. Inferring Plasmodium vivax transmission networks from tempo-spatial surveillance data. *PLoS Neglected Tropical Diseases* 8, 2 (2014), e2682.

[164] Benyun Shi, Jinxin Zheng, Hongjun Qiu, Guo-Jing Yang, Shang Xia, and Xiao-Nong Zhou. 2017. Risk assessment of malaria transmission at the border area of China and Myanmar. *Infectious Diseases of Poverty* 6, 04 (2017), 55–63.

[165] Koushlendra Kumar Singh, Suraj Kumar, Prachi Dixit, and Manish Kumar Bajpai. 2021. Kalman filter based short term prediction model for COVID-19 spread. *Applied Intelligence* 51, 5 (2021), 2714–2726.

[166] David L. Smith and F. Ellis McKenzie. 2004. Statics and dynamics of malaria infection in Anopheles mosquitoes. *Malaria Journal* 3, 1 (2004), 1–14.

[167] Kyungwoo Song, Hojun Park, Junggu Lee, Arim Kim, and Jaehun Jung. 2023. COVID-19 infection inference with graph neural networks. *Scientific Reports* 13, 1 (2023), 11469.

[168] Jichao Sun, Xi Chen, Ziheng Zhang, Shengzhang Lai, Bo Zhao, Hualuo Liu, Shuojia Wang, Wenjing Huan, Ruihui Zhao, Man Tat Alexander Ng, et al. 2020. Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. *Scientific Reports* 10, 1 (2020), 21122.

[169] Aman Swaraj, Karan Verma, Arshpreet Kaur, Ghanshyam Singh, Ashok Kumar, and Leandro Melo de Sales. 2021. Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India. *Journal of Biomedical Informatics* 121 (2021), 103887. https://www.sciencedirect.com/science/article/pii/S1532046421002161

[170] Qi Tan, Yang Liu, and Jiming Liu. 2021. Demystifying deep learning in predictive spatiotemporal analytics: An information-theoretic framework. *IEEE Transactions on Neural Networks and Learning Systems* 32, 8 (2021), 3538–3552.

[171] Qi Tan, Yang Liu, Jiming Liu, Benyun Shi, Shang Xia, and Xiao-Nong Zhou. 2021. Heterogeneous neural metric learning for spatio-temporal modeling of infectious diseases with incomplete data. *Neurocomputing* 458 (2021), 701–713. https://www.sciencedirect.com/science/article/abs/pii/S0925231220316842

[172] Michele Tizzoni, Paolo Bajardi, Chiara Poletto, José J. Ramasco, Duygu Balcan, Bruno Gonçalves, Nicola Perra, Vittoria Colizza, and Alessandro Vespignani. 2012. Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm. *BMC Medicine* 10, 1 (2012), 1–31.

[173] Thomas Torku, Abdul Khaliq, and Fathalla Rihan. 2023. SEINN: A deep learning algorithm for the stochastic epidemic model. *Mathematical Biosciences and Engineering* 20, 9 (2023), 16330–16361.

[174] Khanh-Tung Tran, Truong Son Hy, Lili Jiang, and Xuan-Son Vu. 2024. MGLEP: Multimodal graph learning for modeling emerging pandemics with big data. *Scientific Reports* 14, 1 (2024), 16377.

[175] Anil Utku. 2023. Deep learning based hybrid prediction model for predicting the spread of COVID-19 in the world's most populous countries. *Expert Systems with Applications* 231 (2023), 120769. https://www.sciencedirect.com/science/article/pii/S095741742301271X

[176] Shashank Reddy Vadyala, Sai Nethra Betgeri, Eric A. Sherer, and Amod Amritphale. 2021. Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM. *Array* 11 (2021), 100085. https://www.sciencedirect.com/science/article/pii/S2590005621000333

[177] Siva R. Venna, Amirhossein Tavanaei, Raju N. Gottumukkala, Vijay V. Raghavan, Anthony S. Maida, and Stephen Nichols. 2019. A novel data-driven model for real-time influenza forecasting. *IEEE Access* 7 (2019), 7691–7701. https://ieeexplore.ieee.org/abstract/document/8581423

[178] Hanuman Verma, Saurav Mandal, and Akshansh Gupta. 2022. Temporal deep learning architecture for prediction of COVID-19 cases in India. *Expert Systems with Applications* 195 (2022), 116611. https://www.sciencedirect.com/science/article/pii/S0957417422001038

[179] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D. Corley. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS One* 12, 12 (2017), e0188941.

[180] Bin Wang, Xiaolei Zou, and Jiang Zhu. 2000. Data assimilation and its applications. *Proceedings of the National Academy of Sciences* 97, 21 (2000), 11143–11144.

[181] Haoyu Wang, Xihe Qiu, Jinghan Yang, Qiong Li, Xiaoyu Tan, and Jingjing Huang. 2023. Neural-SEIR: A flexible data-driven framework for precise prediction of epidemic disease. *Mathematical Biosciences and Engineering* 20, 9 (2023), 16807–16823.

[182] Jingyuan Wang, Xiaojian Wang, and Junjie Wu. 2018. Inferring metapopulation propagation network for intra-city epidemic control and prevention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, NY, USA, 830–838. DOI: https://doi.org/10.1145/3219819.3219865

[183] Lijing Wang, Aniruddha Adiga, Jiangzhuo Chen, Adam Sadilek, Srinivasan Venkatramanan, and Madhav Marathe. 2022. CausalGnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (2022), 12191–12199. DOI: https://doi.org/10.1609/aaai.v36i11.21479

[184] Lijing Wang, Jiangzhuo Chen, and Madhav Marathe. 2019. DEFSI: Deep learning based epidemic forecasting with synthetic information. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (2019), 9607–9612. DOI: https://doi.org/10.1609/aaai.v33i01.33019607

[185] Lijing Wang, Jiangzhuo Chen, and Madhav Marathe. 2020. TDEFSI: Theory-guided deep learning-based epidemic forecasting with synthetic information. *ACM Transactions on Spatial Algorithms and Systems* 6, 3 (2020), 39 pages. https://doi.org/10.1145/3380971

[186] Peipei Wang, Haiyan Liu, Xinqi Zheng, and Ruifang Ma. 2023. A new method for spatio-temporal transmission prediction of COVID-19. *Chaos, Solitons & Fractals* 167 (2023), 112996. https://www.sciencedirect.com/science/article/pii/S0960077922011754

[187] Peipei Wang, Xinqi Zheng, Gang Ai, Dongya Liu, and Bangren Zhu. 2020. Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. *Chaos, Solitons & Fractals* 140 (2020), 110214. https://www.sciencedirect.com/science/article/pii/S096007792030610X

[188] Peipei Wang, Xinqi Zheng, Jiayang Li, and Bangren Zhu. 2020. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals* 139 (2020), 110058. https://www.sciencedirect.com/science/article/pii/S0960077920304550

[189] Rui Wang, Danielle Maddix, Christos Faloutsos, Yuyang Wang, and Rose Yu. 2021. Bridging Physics-based and Data-driven modeling for Learning Dynamical Systems. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control (Proceedings of Machine Learning Research)*, Vol. 144. PMLR, Cambridge, MA, USA, 385–398. Retrieved from https://proceedings.mlr.press/v144/wang21a.html

[190] Xin Wang, Yijia Dong, William David Thompson, Harish Nair, and You Li. 2022. Short-term local predictions of COVID-19 in the United Kingdom using dynamic supervised machine learning algorithms. *Communications Medicine* 2, 1 (2022), 119.

[191] Xiunan Wang, Hao Wang, Pouria Ramazi, Kyeongah Nah, and Mark Lewis. 2022. From policy to prediction: Forecasting COVID-19 dynamics under imperfect vaccination. *Bulletin of Mathematical Biology* 84, 9 (2022), 90.

[192] Yongbin Wang, Chunjie Xu, Zhende Wang, Shengkui Zhang, Ying Zhu, and Juxiang Yuan. 2018. Time series modeling of pertussis incidence in China from 2004 to 2018 with a novel wavelet based SARIMA-NAR hybrid model. *PloS One* 13, 12 (2018), e0208404.

[193] Yongbin Wang, Chunjie Xu, Sanqiao Yao, Lei Wang, Yingzheng Zhao, Jingchao Ren, and Yuchun Li. 2021. Estimating the COVID-19 prevalence and mortality using a novel data-driven hybrid model based on ensemble empirical mode decomposition. *Scientific Reports* 11, 1 (2021), 21413.

[194] Zheng Wang, Prithwish Chakraborty, Sumiko R. Mekaru, John S. Brownstein, Jieping Ye, and Naren Ramakrishnan. 2015. Dynamic Poisson autoregression for influenza-like-illness case count prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. S ACM, New York, NY, USA, 1285–1294. DOI: https://doi.org/10.1145/2783258.2783291

[195] Gregory L. Watson, Di Xiong, Lu Zhang, Joseph A. Zoller, John Shamshoian, Phillip Sundin, Teresa Bufford, Anne W. Rimoin, Marc A. Suchard, and Christina M. Ramirez. 2021. Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. *PLoS Computational Biology* 17, 3 (2021), e1008837.

[196] Lander Willem, Sean Stijven, Ekaterina Vladislavleva, Jan Broeckhove, Philippe Beutels, and Niel Hens. 2014. Active learning to understand infectious disease models and improve policy making. *PLoS Computational Biology* 10, 4 (2014), e1003563.

[197] Yuexin Wu, Yiming Yang, Hiroshi Nishiura, and Masaya Saitoh. 2018. Deep learning for epidemiological predictions. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval* . ACM, New York, NY, USA, 1085–1088. DOI: https://doi.org/10.1145/3209978.3210077

[198] Zipeng Wu, Chu Kiong Loo, Unaizah Obaidellah, and Kitsuchart Pasupa. 2023. A novel online multi-task learning for COVID-19 multi-output spatio-temporal prediction. *Heliyon* 9, 8 (2023), e18771. DOI: https://doi.org/10.1016/j.heliyon.2023.e18771

[199] Feng Xie, Zhong Zhang, Liang Li, Bin Zhou, and Yusong Tan. 2022. EpiGNN: Exploring spatial transmission with graph neural network for regional epidemic forecasting. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer Nature Switzerland, Cham, 469–485.

[200] Feng Xie, Zhong Zhang, Xuechen Zhao, Bin Zhou, and Yusong Tan. 2022. Inter- and Intra-series embeddings fusion network for epidemiological forecasting. (2022). arXiv:2208.11515. Retrieved from https://arxiv.org/abs/2208.11515

[201] Lu Xu, Rishikesh Magar, and Amir Barati Farimani. 2022. Forecasting COVID-19 new cases using deep learning methods. *Computers in Biology and Medicine* 144 (2022), 105342. https://www.sciencedirect.com/science/article/pii/S0010482522001342

[202] Vitaliy Yakovyna, Nataliya Shakhovska, and Aleksandra Szpakowska. 2024. A novel hybrid supervised and unsupervised hierarchical ensemble for COVID-19 cases and mortality prediction. *Scientific Reports* 14, 1 (2024), 9782.

[203] Bo Yang, Hua Guo, Yi Yang, Benyun Shi, Xiaonong Zhou, and Jiming Liu. 2014. Modeling and mining spatiotemporal patterns of infection risk from heterogeneous data for active surveillance planning. *Proceedings of the AAAI Conference on Artificial Intelligence* 28, 1 (2014), 493–499. DOI: https://doi.org/10.1609/aaai.v28i1.8762

[204] Liuyang Yang, Gang Li, Jin Yang, Ting Zhang, Jing Du, Tian Liu, Xingxing Zhang, Xuan Han, Wei Li, Libing Ma, et al. 2023. Deep-learning model for influenza prediction from multisource heterogeneous data in a megacity: Model development and evaluation. *Journal of Medical Internet Research* 25 (2023), e44238. https://www.jmir.org/2023/1/e44238/

[205] Shihao Yang, Mauricio Santillana, and Samuel C. Kou. 2015. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences* 112, 47 (2015), 14473–14478.

[206] Wan Yang, Benjamin J. Cowling, Eric HY Lau, and Jeffrey Shaman. 2015. Forecasting influenza epidemics in Hong Kong. *PLoS Computational Biology* 11, 7 (2015), e1004383.

[207] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease* 12, 3 (2020), 165.

[208] Qingyu Yuan, Elaine O. Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S. Brownstein. 2013. Monitoring influenza epidemics in china with search query from baidu. *PloS One* 8, 5 (2013), e64323.

[209] Qinghui Zeng, Xiaolin Yu, Haobo Ni, Lina Xiao, Ting Xu, Haisheng Wu, Yuliang Chen, Hui Deng, Yingtao Zhang, Sen Pei, et al. 2023. Dengue transmission dynamics prediction by combining metapopulation networks and Kalman filter algorithm. *PLOS Neglected Tropical Diseases* 17, 6 (2023), e0011418.

[210] Choujun Zhan, Yufan Zheng, Haijun Zhang, and Quansi Wen. 2021. Random-forest-bagging broad learning system with applications for COVID-19 pandemic. *IEEE Internet of Things Journal* 8, 21 (2021), 15906–15918.

[211] Gengpei Zhang and Xiongding Liu. 2021. Prediction and control of COVID-19 spreading based on a hybrid intelligent model. *Plos One* 16, 2 (2021), e0246360.

[212] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. 2017. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th*

International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 311–319. DOI: https://doi.org/10.1145/3038912.3052678

[213] Tianyu Zhang, Fethi Rabhi, Xin Chen, Hye-young Paik, and Chandini Raina MacIntyre. 2024. A machine learning-based universal outbreak risk prediction tool. *Computers in Biology and Medicine* 169 (2024), 107876. https://www.sciencedirect.com/science/article/pii/S0010482523013410

[214] Xiaolei Zhang and Renjun Ma. 2023. Forecasting waved daily COVID-19 death count series with a novel combination of segmented Poisson model and ARIMA models. *Journal of Applied Statistics* 50, 11-12 (2023), 2561–2574.

[215] Yu Zhang, William K. Cheung, and Jiming Liu. 2015. A unified framework for epidemic prediction based on poisson regression. *IEEE Transactions on Knowledge and Data Engineering* 27, 11 (2015), 2878–2892.

[216] Jing Zhao, Mengjie Han, Zhenwu Wang, and Benting Wan. 2022. Autoregressive count data modeling on mobility patterns to predict cases of COVID-19 infection. *Stochastic Environmental Research and Risk Assessment* 36, 12 (2022), 4185–4200.

[217] Weiping Zhao, Yunpeng Sun, Ying Li, and Weimin Guan. 2022. Prediction of COVID-19 data using hybrid modeling approaches. *Frontiers in Public Health* 10 (2022), 923978. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.923978/full

[218] Jinxin Zheng, Benyun Shi, Shang Xia, Guojing Yang, and Xiao-Nong Zhou. 2021. Spatial patterns of Plasmodium vivax transmission explored by multivariate auto-regressive state-space modelling-A case study in Baoshan Prefecture in southern China. *Geospatial Health* 16, 1 (2021), 205–212. DOI: https://doi.org/10.4081/gh.2021.879

[219] Liang Zheng, Yile Chen, Shan Jiang, Junxin Song, and Jianyi Zheng. 2023. Predicting the distribution of COVID-19 through CGAN–Taking Macau as an example. *Frontiers in Big Data* 6 (2023), 1008292. https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2023.1008292/full

[220] Nanning Zheng, Shaoyi Du, Jianji Wang, He Zhang, Wenting Cui, Zijian Kang, Tao Yang, Bin Lou, Yuting Chi, Hong Long, et al. 2020. Predicting COVID-19 in China using hybrid AI model. *IEEE Transactions on Cybernetics* 50, 7 (2020), 2891–2904.

[221] Shun Zheng, Zhifeng Gao, Wei Cao, Jiang Bian, and Tie-Yan Liu. 2021. HierST: A unified hierarchical spatial-temporal framework for COVID-19 trend forecasting. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. ACM, New York, NY, USA, 4383–4392. DOI: https://doi.org/10.1145/3459637.3481927

[222] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81. https://www.sciencedirect.com/science/article/pii/S2666651021000012

[223] Guanghu Zhu, Jiming Liu, Qi Tan, and Benyun Shi. 2016. Inferring the spatio-temporal patterns of dengue transmission from surveillance data in Guangzhou, China. *PLoS Neglected Tropical Diseases* 10, 4 (2016), e0004633.

[224] Xiaofeng Zhu, Yi Zhang, Haoru Ying, Huanning Chi, Guanqun Sun, and Lingxia Zeng. 2024. Modeling epidemic dynamics using Graph Attention based Spatial Temporal networks. *Plos One* 19, 7 (2024), e0307159.

[225] Christoph Zimmer and Reza Yaesoubi. 2020. Influenza forecasting framework based on Gaussian processes. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 119. PMLR, Cambridge, MA, USA, 11671–11679. Retrieved from https://proceedings.mlr.press/v119/zimmer20a.html

[226] Difan Zou, Lingxiao Wang, Pan Xu, Jinghui Chen, Weitong Zhang, and Quanquan Gu. 2020. Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States. (2020). https://www.medrxiv.org/content/10.1101/2020.05.24.20111989v1

Received 30 April 2023; revised 20 September 2024; accepted 4 February 2025