Sarcouncil Journal of Medical Sciences

ISSN(Online): 2945-3526

Volume- 05| Issue- 02| 2026

Review Article

Received: 15-01-2026 | Accepted: 10-02-2026 | Published: 24-02-2026

# Big Data and Machine Learning Applications for Enhanced U.S. Infectious Disease Surveillance and Control: A Narrative Review

***Merrera Kebeba*[1] *and Emmanuel Amoako Agyei*[2]**

[1]*Santa Clara County Department of Public Health, USA*

[2]*Washington University in St. Louis, USA*

**Abstract**: Infectious disease surveillance in the United States has long been facing challenges of delayed feedback, inefficient data infrastructure, and limited predictive capacities, which have been evident in recent outbreaks. However, the burgeoning development of big data ecosystems coupled with machine learning algorithms has made it possible to transform the situation by improving the timeliness, accuracy, and robustness of infectious disease control in the United States. This review draws upon the current literature for a synthesis of the integration of big data and machine learning in infectious disease surveillance and control in the United States. Conventional sources of public health-related data are presented, in addition to new sources of digital, genomic, and non-conventional data sources such as electronic health records, syndromic surveillance, mobility datasets, social media data, wearable biosensing, and genomic pathogen sequencing. Finally, different machine learning paradigms such as supervised learning, unsupervised learning, and deep learning are presented in terms of their applicability for detection, forecasting, and risk assessment. Practical applications of machine learning for early warning of outbreaks, disease control, resource allocation, and precision medicine for public health are presented with an emphasis on the United States. Finally, future directions for research in machine learning-related applications in disease control are presented. This review cumulatively signifies the promise of machine learning-enabled disease control for improving the accuracy, speed, and robustness of infectious disease control in the United States.

**Keywords:** Infectious disease surveillance, big data analytics, machine learning.

## INTRODUCTION

Infectious diseases have repeatedly demonstrated their capacity to challenge the systems of social health in the United States on numerous occasions, especially during the COVID-19 pandemic, that claimed the lives of more than 1.1 million people and burdened national surveillance systems the most ever (WHO, 2023). In addition to COVID-19, there are frequent challenges caused by seasonal influenza, antimicrobial-resistant pathogens, and novel outbreaks like mpox that reveal the weaknesses in the current monitoring systems (Flynn & Guarner, 2023; Kumar *et al*., 2024; Uyeki *et al*., 2022). A combination of these events highlights the persistent necessity of infected disease monitoring systems that are not only strong but adequately dynamic to identify, analyze, and react to arising threats in near real-time (Morgan *et al*., 2022), Ugwu *et al*, 2025).

One of the fundamental epidemiological functions is infectious disease surveillance (IDS), which is a systemic gathering, examination, and interpretation of information on the health-related issues to assess disease tendencies and emergence, detect outbreaks, and identify the new pathogens, subsequently forming part of deciding the course of action to prevent and eliminate diseases (Idahor *et al*., 2025). Integrated disease surveillance is long promoted by the World Health Organization (WHO) as the foundation of the successful national systems of public health (WHO, 2025). Nevertheless, traditional and disease-specific surveillance operations methods have been previously based upon manual reporting frameworks and data streams that are siloed (Ssemujju & Solomon, 2025; Kizza *et al*, 2025). Although the basis, such systems are often limited by delayed reporting, incomplete coverage, and under-ascertainment, which limits its usefulness in quick outbreak detection and response (Maddah *et al*., 2023).

Recent technological changes have started to transform the world of infectious disease surveillance. Growth of big data ecosystems, including electronic health records, digital data streams, such as search query volumes and social media activity, mobility data, generated by mobile devices, and high-throughput pathogen genomic sequencing, such as the CDC Advanced Molecular Detection network, have added volume, velocity, and variety to public health data settings never seen (Bansal *et al*., 2016). These sources of data can facilitate more detailed, population-level data, as shown by state-level combinations of genomic and epidemiologic data to track outbreaks of multidrug-resistant organisms, and wastewater-based surveillance to detect pathogen numbers near real-time (Torres *et al*., 2025). Social sites like BioSense and HealthMap also demonstrate





***Corresponding Author:** Merrera Kebeba

DOI- https://doi.org/10.5281/zenodo.18761942

how much can be done to harness the power of syndromic, mobility, and digital signals as an indicator of a new transmission hotspot (Simonsen *et al*., 2016).

Nevertheless, access to big and complicated data sets is not enough to revolutionize the outcomes of surveillance. Machine learning is a new paradigm shift in this sense and offers adaptive methods of computation, which can derive meaningful patterns out of high-dimensional data. Machine learning models unlike traditional analytic approaches which rely on fixed rules and fixed assumptions to learn through repeated trials on data such that they can better identify patterns, predictive analytics and forecasting outbreaks. Such functioning is especially essential in anticipatory public health decisions, whose timely and precise forecasting can be used to guide specific responses and the allocation of resources.

It is on this background that this narrative review is conducted to synthesize the current trends, applications, and challenges in the area of integration of big data and machine learning in the field of infectious disease surveillance in the United States. In particular, the review discusses how these technologies can help overcome the drawbacks of traditional surveillance systems, make it timelier and more precise in detecting outbreaks, and contribute to more effective disease control tactics. This review will help inform both public health practitioners and policymakers about the ways in which to move towards having a more resilient, data-driven and responsive surveillance infrastructure by critically reviewing existing implementations and future directions.

## BIG DATA ECOSYSTEMS IN U.S. INFECTIOUS DISEASE SURVEILLANCE

### Traditional Public Health Data Streams

The historical context of infectious disease surveillance development in the United States is entrenched in a sophisticated network of traditional data feeds that have traditionally informed the process of making decisions related to public health (Idahor *et al*., 2025). Although these foundational systems were designed during a pre-big data period, they have, over time, become important elements of more of a data ecosystem. They have been transformed by becoming digital, automated, and networked, which has created the foundation of more sophisticated uses, including machine learning and predictive modeling (Bohr & Memarzadeh, 2020; Md Russel Hossain *et al*., 2023). The surveillance systems that are operated by Centers of Disease Control and Prevention (CDC)(Richards *et al*., 2017), electronic health records (EHRs)(Guralnik, 2024), lab data (Groseclose & Buckeridge, 2017), and health insurance claims data are some of the most important pillars of this ecosystem.

CDC has over decades organized a massive surveillance infrastructure that comprises systems of tracking infectious diseases that are nationally standardized. The National Notifiable Diseases Surveillance System (NNDSS) has become an element of this architecture, and it receives reports of legally reportable diseases submitted by state and local health departments (CDC, 2025). This information is crucial in identifying outbreaks, understanding the disease burden, and resource distribution across jurisdictions (Khodadadi & Towfek, 2023). Further on this, the National Electronic Disease Surveillance System (NEDSS) was implemented to improve efficiency and interoperability due to the ability to exchange electronic data between healthcare providers and public health agencies (CDC, 2024b). Although these systems are designed as robust in nature, historically they have been limited in terms of timeliness and integration, thus creating an impetus to augment them with real-time digital data sources (McClymont *et al*., 2024).

One of these significant developments in health surveillance in the public has been the growing adoption of electronic health records (EHRs) in surveillance processes (Perlman *et al*., 2017). As opposed to conventional surveillance that is greatly dependent on manually reported cases, EHRs provide automatic access to automated access to clinical data such as diagnosis codes, lab results, medication record, and patient demographic (Hohman *et al*., 2023). Such abundance of information has reshaped the boundaries of informatics in the field of public health, allowing close-to-real-time syndromic surveillance and more susceptible observation of the initial outbreak signals (A. Goncalves *et al*., 2025). Especially in the environment of the COVID-19 pandemic, the usefulness of EHRs became evident as it was used to track the progression of diseases, healthcare use, and treatment outcomes in various healthcare environments (Williams *et al*., 2022). Regardless of the promise, the widespread adoption of EHRs in public health is still jeopardized by the apparition of data fragmentation and the inability



Publisher: SARC Publisher

of various hospital systems and vendors to collaborate (Birkhead *et al*., 2015).

Laboratory data is an essential and objective portion of the surveillance infrastructure along with the EHRs. Clinical and public health laboratories produce confirmatory data regarding infectious diseases in the form of diagnostic testing, and in many cases, case confirmation regarding reportable conditions can be based on such data (CDC, 2024a). The Electronic Laboratory Reporting (ELR) application has made the process of notifying the health departments about laboratory-based notifications much faster and complete. ELR systems save time by automating laboratory report transmission and reducing underreporting and delays, which were Australia-wide problems with paper-based reporting (Samoff *et al*., 2013). Nonetheless, the problems remain especially concerning the inconsistency in data formatting, codes, and reporting procedures within the laboratories. Such technical differences may also impair the integration of data without any hiccups and restrict real-time analysis opportunities (Olalekan Hamed Olayinka, 2021).

Supplementing clinical and laboratory information is the extensive use of health insurance claims information, a resource that has not traditionally been exploited in real-time surveillance but is invaluable in retrospective analysis and assessment of policies (Kim *et al*., 2020). Claims data consists of records of bills presented by medical professionals in order to get reimbursements and contains diagnostic codes, procedure codes, and medication data. Since they are extensive in their populations across time, they are useful in offering fundamental information on healthcare utilization, disease occurrences, and treatment patterns in the country. Large-scale epidemiological studies and long-term disease surveillance are especially beneficial as their format and consistency are structured (Majumder *et al*., 2023). A weakness of claims data, however, is their lack of clinical subtlety, and the delay between provision of care and its data. Such characteristics limit their utility in the acute outbreak context but maximize their utility in analyzing the interventions and decoding the disease burden in retrospect (Shih & Liu, 2019).

Collectively, these conventional data sources of public health are establishing a complex base on which more sophisticated data science applications are being overlaid (Kadakia & Desalvo, 2023). The regular surveillance systems implemented by the CDC provide national reporting uniformity, whereas EHRs and laboratory data provide granulometry and immediacy in disease monitoring (Davidson *et al*., 2018). Claims data, on its part, provides a longitudinal perspective of the healthcare situation that can be used in strategic planning and resource allocation. Although every one of these systems has its own strengths and limitations, their integration is a major move towards a more integrated, data-led public health surveillance system in the United States.

**Emerging Digital and Non-Traditional Data Sources**

Disease surveillance development in the United States has entered the inflection point whereby use of clinical data is no longer adequate. This has influenced the incorporation of non-traditional and digital sources of data in the intelligence systems of public health (Li *et al*., 2021; McClymont *et al*., 2024). These new streams, including real-time clinical proxies to behavioral and physiological indicators, provide new opportunities to detect an outbreak early and use predictive analytics (Liscano *et al*., 2025). Collectively, they increase the scale and sensitivity of conventional infrastructures, as well as basing a more fluid and data-intensive ecosystem of infectious disease observation.

Syndromic surveillance is one of the most popular non-traditional methods that use nearly-real-time information on symptoms and health-seeking behavior to identify a trend that is abnormal before it is confirmed by laboratory tests (Hughes *et al*., 2020). The CDC administered program called the National Syndromic Surveillance Program (NSSP) provides early indications of infectious disease activity by using emergency department chief complaints, discharge diagnoses, and urgent care records via its ESSENCE platform. This system played a vital role in the COVID-19 pandemic, where syndromic indicators were used to indicate the increasing trends in respiratory illnesses several days before the number of cases exploded, which enabled jurisdictions to ramp up testing and mitigation efforts within a few days (Romano *et al*., 2022). It is worth noting that syndromic data assisted in identifying a difference between COVID-19 and other conditions such as influenza and provided a decision-support tool in times of high uncertainty (Alemi *et al*., 2022). These platforms have now been regarded as vital in





**Publisher: SARC Publisher**

surveillance during early warnings, especially during outbreaks that are fast moving and novel.

In tandem, the data of mobility and transportation have become effective instruments of mapping the dynamics of disease spread on a real-time basis. Smartphone, public transit, and traffic data based on aggregated and anonymized GPS location can be used to train models on the impact of human movement patterns on the spread of a virus (Badr *et al*., 2020). Recent research by Cesario and Comito (2025) also showed that when mobility data are combined with epidemic modeling, hotspots of COVID-19 could be predicted better, and more precise simulation of transmission networks was possible at the county level (Cesario & Comito, 2025). Equally, Huang *et al*. (2025) incorporated transportation flow data into syndromic models to showcase the importance of cross-border mobility on local vulnerability in establishing the areas at high risk of importation during international outbreaks (Huang *et al*., 2025). These sources of data are being used to not only support epidemiological modeling, but also logistics planning of testing, vaccine deployment, and surge response.

Wearables and biosensors have also become a promising future in health time monitoring. Smartwatches and fitness bracelets have the capability to passively gather information regarding physiological parameters such as heart rate, oxygen levels, sleep, and temperature of the skin-parameters that can vary prior to the onset of clinical symptoms (Babu *et al*., 2024; Xian, 2023). Mandel *et al*. (2025) found that integrating wearable biosensor information with machine learning models has a great effect on the early detection of febrile diseases, and in cases of COVID-19 several days before the symptoms appear (Mandal *et al*., 2025). Meanwhile, Song *et al*. (2025) created versatile biosensors with point-of-care testing (POCT) systems to be able to transfer molecular-level infection indicators directly to cloud-based dashboards, which creates a paradigm-shifting device in decentralized surveillance of both clinical and community environments (Song *et al*., 2025).These innovations increase the area of observation beyond health institutions and provide a level of detail in population health never seen before.

Finally, the report of social media activity and online search is already a proven way of identifying and predicting trends in infectious diseases. Social media apps, such as Twitter and Reddit, and search engine queries are also real-time proxies of the degree of concern, symptom reporting, and behavior change in the public (Esha Madamalla, 2025). Such insights are particularly helpful in areas with the under-resourced health infrastructure, where the traditional data lags are the largest. Collectively, these four groups of digital and non-traditional data sources reflect the changing process of infectious disease surveillance in the U.S of reactive to predictive and continuous and behavior-sensitive systems. The combination of these data streams into national public health architecture opens possibilities to close gaps in surveillance, improve situational awareness, and be able to inform data-driven interventions in new timeliness and granularity.

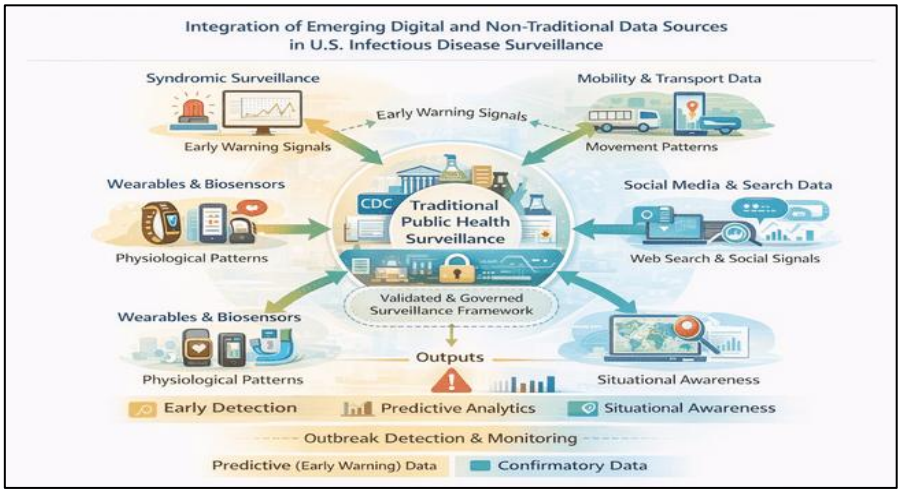


**Figure 1** shows how emerging digital and non-traditional data sources such as syndromic surveillance, mobility data, wearable and biosensor signals, ad social media and web data integrate with traditional public health surveillance to ensure early detection, predictive analytics, and enhanced situational awareness in infectious disease monitoring.





**Publisher: SARC Publisher**

## GENOMIC AND PATHOGEN SEQUENCING DATA

The integration of genomic sequencing methodologies within infectious disease surveillance has greatly improved the role of public health in systematically monitoring evolutionary changes, tracing emergent variants, and understanding transmission dynamics at an unprecedented scale (Tiwari *et al*., 2025). Within the United States, large-scale genomic projects like the CDC's SPHERES consortium (Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance) enabled a rapid scale-up of genomic sequencing during the COVID-19 pandemic. These efforts enabled high-throughput genomic sequencing of SARS-CoV-2 samples within both public and private laboratories, which allowed for rapid identification of variants of concern, such as Delta and Omicron (Deng *et al*., 2021). Significantly, this genomic data played a crucial role in tracing transmission chains, especially within settings like schools and long-term care facilities, where traditional epidemic surveillance failed. A particularly important use of this data was in tracing the introduction and community transmission of B.1.1.7 within the United States through a phylogenetics analysis, which allowed researchers to differentiate between importation events versus local transmission (Washington *et al*., 2021). Together, these studies showed that genomic surveillance has a powerful role in informing public health decisions and expanding present understanding of viral fitness, mutation, and population immunity.

There has been significant momentum in recent years in integrating pathogen genomic information with national surveillance infrastructure. Genomic information is increasingly combined with metadata derived from clinical records, geography, and infectious disease investigations to assemble multi-faceted datasets amenable for use in predictive modeling (Maxime *et al*., 2025; Tiwari *et al*., 2025). Platforms like GISAID and Nextstrain have been critical in organizing and interpreting sequence information around the world; in the United States, there are efforts employing regional healthcare networks in creating interfaces where direct output of sequencing information is readily incorporated into region-wide decision-support infrastructure (Oude Munnink *et al*., 2021). Thus, there is success with integrating information on genome sequences with information derived through syndromic surveillance informing on the effect of specific strains on containment strategy (Alpert *et al*., 2021) in near-real time at institutions like the New York City Public Health Laboratory. While there are distinct challenges related to information standardization and access inequities in sequencing capacity, there is no doubt that the trend of innovation will continue toward making surveillance information essential and integrated into public health strategy in the United States.

## MACHINE LEARNING METHODOLOGIES FOR SURVEILLANCE AND PREDICTION

### Supervised Learning for Case Detection and Risk Stratification

Supervised learning models are central to current infectious disease monitoring, as they learn to classify cases, forecast risk, and discover patterns in clinical or behavioral data on a labeled dataset (Villanueva-Miranda *et al*., 2025; Xu *et al*., 2025). Regression models and especially logistic regression have been among the most popular techniques in identifying the presence of infection as well as estimating the probability of risk through demographic or clinical or exposure factors. Jaiteh *et al*. (2025) in a comparative study used multi-year population datasets to logistic regression to predict HIV testing based on these datasets and reported a strong sensitivity and model transparency necessary in the context of interpreting public health (Jaiteh *et al*., 2025). Although the regression methods are quite simple, they have a strong suit of being statistically rigorous and transparent, so they can be used as a useful baseline technique in the supervised learning pipeline in infectious disease surveillance.

The use of random forest models, which are also referred to as ensembles of decision trees, has become prominent because of their capacity to represent non-linear relationships and interactions among or among numerous variables without requiring extensive preprocessing. They are well-suited to complex public health data, in which missing and incomplete reporting is a frequent occurrence, due to their resilience to noise and lack of data (Hong & Lynn, 2020; Tharanidharan, 2024). Ndao *et al*. (2024) compared random forest classifiers with other models in their ability to predict the presence of ten different infectious diseases in Senegal and found that random forest classifiers were better than the rest in several metrics (sensitivity and specificity) when applied to high-dimensional data (Ndao *et al*., 2024).





**Publisher: SARC Publisher**

Equally, Jaiteh *et al*. (2025) concluded that random forests were more precise compared to regression models when predicting HIV testing behavior, and when they used interaction effects among sex, education, and sexual health knowledge (Jaiteh *et al*., 2025). These results highlight the possible use of random forests as a central approach to disease classification, particularly with feature importance analyses to determine important predictors to address specific intervention.

Disease surveillance tasks with high binary classification accuracy have also seen the popularization of the use of support vector machines (SVMs). SVMs are especially useful in the separation of non-linear data with the help of kernel transformation and have been used in tasks like tuberculosis detection based on imaging, COVID-19 case classification based on EHR and lab data (Balakrishnan *et al*., 2023; Hammad *et al*., 2023; Pandey *et al*., 2022). Ndao *et al*. (2024) have shown that in low-prevalence environments, SVMs were more effective in terms of low false positives and comparable recalls to logistic regression (Ndao *et al*., 2024). SVMs may need parameter tuning and computational intensity but are useful in those situations where model accuracy is paramount like in early outbreak detection or screening of low-risk populations (Guido *et al*., 2024). Their use in surveillance systems is useful as evidenced by their constant incorporation with ensemble and deep learning techniques of hybrid model optimization.

**Deep Learning for Temporal and Spatial Forecasting**

Long Short-Term Memory models (LSTM), a type of recurrent neural network, have been shown to be very successful in capturing the temporal dynamics of infectious disease surveillance. They are especially useful in outbreak prediction and trend tracking of epidemics (Okut, 2021) because of their capability to store and exploit long-term dependencies in time-series data (Pontoh *et al*., 2022). A more applicable illustration, albeit involving a non-U.S. setting, is by Zhang *et al*. (2022), who used LSTM-based models on hepatitis surveillance data and pointed out the higher accuracy of deep learning compared to the traditional statistical approach in modeling the disease under incidence in the long-term (Xia *et al*., 2022). Although the research was done on Chinese data, the modeling framework is flexible and broadly used in the disease forecasting of the U.S. as an ensemble system by institutions such as the CDC. LSTM networks remain a central point of the temporal prediction model because of their resilience in dealing with irregular times and missing data both of which are frequent in real-life surveillance streams (Weerakody *et al*., 2021).

In addition to pure temporal models, spatiotemporal deep learning networks combine time and space in their modeling, which allows more thorough modeling of how diseases can spread across space and change over time. A recent study by Han *et al*. (2025) introduced the Causal Spatiotemporal Graph Neural Network (CSTGNN), which integrates a spatio-contact SIR model with a dynamic graph neural network to forecast epidemic trajectories (Han *et al*., 2025). The model captures both static and fluctuating patterns in human mobility and embeds them into a temporally aware graph architecture (Han *et al*., 2025). While validated using datasets from China and Germany, the framework is directly applicable to U.S. contexts where high-resolution mobility and health surveillance data are available. Importantly, this hybrid design preserves interpretability, allowing public health professionals to understand both the why and where of outbreak dynamics.

Real-time outbreak forecasting powered by deep learning is transforming how public health systems anticipate and respond to infectious disease threats (Rodríguez *et al*., 2020). The models are intended to consume constantly updating data feeds, electronic health records, syndromic surveillance feeds, social media signals, mobility trends, and others to produce timely predictions regarding disease incidence and transmission (Ekundayo, 2024). Feed-forward neural networks, recurrent models, or transformer-based systems are deep learning architectures that can be modified to process heterogeneous inputs at scale and are therefore suitable nowcasting or short-term forecasting at conditions that change rapidly due to epidemiologic factors (Roster *et al*., 2022). It contrasts with traditional systems where feature engineering and the task of creating a static dataset might be required since real-time systems use automated learning of incoming data and predictions adapt as new patterns are discovered (Bettencourt & Soman, 2020). Their output can be directly incorporated into operational dashboards, which give public health agencies actionable insights on how to allocate resources, timely intervention, and risk communication (Wang *et al*., 2019). These deep learning models are becoming important in the proactive management of





**Publisher: SARC Publisher**

outbreaks and epidemic preparedness as data infrastructure and real-time reporting advance.

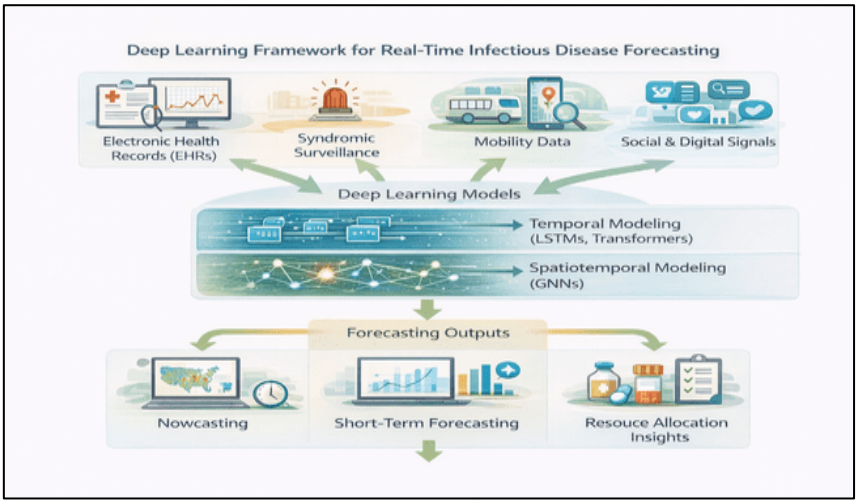


**Figure 2** shows a simplified deep learning framework for real-time infectious disease surveillance, illustrating how multi-source data inputs are processed through temporal (LSTM/ transformer) and spatiotemporal (graph neural network) models to generate newscasting, short-term forecasts, and resource allocation insights.

## UNSUPERVISED AND SEMI-SUPERVISED APPROACHES

Infectious disease surveillance anomaly detection in infectious disease surveillance, anomaly detection attempts to identify unusual or uncommon patterns in clinical and epidemiological data without using any preset labels (Eze *et al*., 2023). Unsupervised models are particularly applied when there is a limited amount of early outbreak data, or the data is incomplete (Eze *et al*., 2023). In the recent past, more sophisticated systems have been developed based on transformer-based language models to model longitudinal electronic health record (EHR) data (Ren *et al*., 2025). Such models forecast the likely outcome of patients and make records anomalous when the observed clinical patterns do not conform well with the acquired norms. As an example, a 2024 study proved how a GPT-based system could be used to detect early-stage hospital-based outbreaks by detecting clinical sequences that do not follow an expected disease progression, before a formal diagnosis (Hao *et al*., 2024). These unsupervised architectures are more sensitive to detection and less reliant on historic outbreaks labels.

The clustering techniques are useful to reveal latent clusters of cases on a spatial, temporal, clinical, or behavioral basis, which may indicate localized transmission events or common risk factors (Lan & Delmelle, 2023). Such algorithms as DBSCAN, k-means, or hierarchical clustering have been successfully used in the context of both classic surveillance systems and the real-time digital health data setting (Foluke Ekundayo, 2024; Natrayan *et al*., 2024). The evaluation of space time series, a study specific to the U.S., revealed that geographical hotspots of infectious disease dissemination, such as new outbreaks at the communal level, could be identified through the clustering of multi-source surveillance streams (Wu *et al*., 2025). This self-control analysis enables the early identification of the disease foci, which contributes to the geographically targeted interventions and distribution of resources.

Semi-supervised learning has the advantage of not requiring a lot of labeled data but has many unlabeled backgrounds and hence is the best at detecting the first signal (Rosenfeld & Globerson, 2018). In practice, the models can use weak indicators of a possible outbreak, such as syndromic surveillance (Edo-Osagie *et al*., 2019), internet search patterns, and EHR data, to detect it before it is confirmed (Goncalves *et al*., 2025). Semi-supervised frameworks can learn continuously as new data comes by taking the few labels of known outbreaks and a large amount of unlabeled time-series input (Edo-Osagie *et al*., 2019). As an example, such trained models can pick up on the rises in uncharacteristic combinations of symptoms or health-seeking behavior changes, which could be the first signs of a rising public health danger (Cheah *et al*., 2025; Nobles *et al*., 2022). The feature is unique especially when dealing with low-incidence situations or when new pathogens appear and the gold-standard case data is lagged (Santos *et al*., 2021).



**Publisher: SARC Publisher**

## APPLICATIONS IN U.S. INFECTIOUS DISEASE CONTROL

### Early Outbreak Detection and Situational Awareness

Data-driven technologies and machine learning models are now part of improving early outbreak detection and enhance situational awareness of various infectious diseases in the United States, such as influenza, COVID-19, RSV, and novel pathogens (Parums, 2023; Srivastava *et al*., 2025). These tools combine various sources of data including electronic health records, syndromic surveillance feeds, type of medicine sold over the counter, and mobility patterns to give real-time signals of abnormal disease activity (Hripcsak *et al*., 2009). Compared to the old systems of surveillance, which tend to lag behind the real-world dynamics of transmission, the new systems can provide early warning and insights into what may happen ahead, enabling the health departments of the population to react to changes in the number of cases before it gets too high (MacIntyre *et al*., 2023). These approaches can facilitate quicker mobilization efforts, more accurate risk communications, and earlier containment efforts that will improve the responsiveness of the country to seasonal epidemics and abrupt health hazards as these approaches will allow analysis of regional and demographic trends in nearly real-time (Ezeh *et al*., 2024).

### Resource Allocation and Health System Preparedness

Predictive modeling has since become a crucial asset in terms of predicting the capacity of the hospitals, especially when it comes to a period of a national health emergency (Endres-Dighe *et al*., 2021; Taiwo, 2025). The trends of patient admission, community transmission, and demographic risk factors are available to feed machine learning algorithms that could be used to project the demand of future bed, ICU occupancy, and ventilator utilization (Lorenzen *et al*., 2021; Ward *et al*., 2022). The forecasts are useful in enabling health systems to prepare for surges by informing them of the time and place to build capacity, delay elective processes, or convert clinical spaces (Fagbenle, 2025). Such tools have been critical during mass outbreaks like the COVID-19 to identify hospitals that are under the threat of being overwhelmed and to coordinate with regional and federal partners to reserve the capacity prior to the systems reaching critical levels (Qian *et al*., 2021).

Simultaneously, data-driven systems assist in real-time optimization of the workforce and supply chain predicting staffing gaps and essential supply shortages (Patil, 2024). These models are useful to the health administrators in forecasting and realigning workforce assignments in departments or facilities by combining variables including patient volumes, absenteeism trends and regional caseloads (Petanidis *et al*., 2025). Equally, inventory forecasting systems have the ability to predict the timing and location of personal protective equipment (PPE), medication, and diagnostic supply movement to avoid bottlenecks (Harsha *et al*., 2022; Shahin *et al*., 2024) . Such preparedness will result in the capability to sustain the necessary services in the case of extended public health emergencies, even in the presence of an uneven distribution of resource strain across locations or sectors.

### Targeted Interventions and Precision Public Health

The progress in the field of data analytics and machine learning allowed the development of risk-based vaccination plans that are no longer one-size-fits-all (Bouramtane *et al*., 2025; Eze *et al*., 2024). Predictive models have the ability to target the most vulnerable population through a combination of individual-level clinical data, demographics, and risk behavioral profiles (Tan *et al*., 2020). Such insights can be used to prioritize the allocation of vaccines in cases of limited supply such as the onset of a pandemic or on seasonal diseases with a dynamic burden (Buckner *et al*., 2021). Risk stratification can also be used to plan booster campaigns by determining those at high risk because of comorbidities, occupation, or immunity loss (Keeling *et al*., 2021). These are the accuracy strategies that enhance effectiveness and fairness of vaccination programs, resulting in high impact protection of the most affected individuals.

Machine learning models and geospatial data analytics can inform interventions at the community level, by identifying neighborhoods or regions at risk of transmission, with low healthcare access, or low vaccination coverage (Ali, 2024; Cheong *et al*., 2021). Real-time detection of new hotspots, or underserved zones will allow public health officials to send testing, health education, mobile clinics, or containment to those locations immediately (Kumar *et al*., 2025). Such a localized strategy will maximize the potential of small resources and increase credibility and rate of compliance among affected populations (Babatuyi *et al*., 2024; Bandara *et al*., 2025). Finally,





**Publisher: SARC Publisher**

accuracy in public health enables the decision-makers to act more swiftly and accurately, decreasing the spread and the burden of the infectious diseases of the population (Ali, 2024).

**Case Studies from U.S. Surveillance Initiatives**

Surveillance systems based on data have become more common among federal and state-level agencies in the United States as a means of tracking and controlling infectious diseases (Birkhead *et al*., 2015). At the federal level, programs such as the CDCs National Syndromic Surveillance Program (NSSP) have incorporated real-time data collection of the emergency departments, urgent care centers, and laboratories to provide the early indicators of the outbreaks throughout the nation (Romano *et al*., 2018). In the meantime, state health departments have put local systems in place that include mobility data, school absenteeism rates, and local hospital metrics that will allow them to respond to local issues (Hyder *et al*., 2021; Lawpoolsri *et al*., 2014). These combined systems have improved the granularity and timeliness of surveillance and are able to identify threat outbreaks faster, including influenza surges, RSV, and new pathogen threats (Gupta *et al*., 2022; Harcourt *et al*., 2019). Notably, the interaction between state and federal has helped in better data exchange, standardizations and multi-jurisdictional preparedness in response (Grier *et al*., 2011).

In spite of these developments, there are some lessons that have been learned after the use of surveillance systems in past emergencies that have happened concerning the health of the people (Archer *et al*., 2023). A major lesson is that information alone is no longer enough, and systems should be usable, have easy-to-use dashboards and automated alerts that would enable fast decision-making (Katapally & Ibrahim, 2023). Also, irregular data quality and reporting delays between jurisdictions have raised the issue of better interoperability and investment in infrastructure (Dixon *et al*., 2011). The workforce capacity is also essential since the professionals of public health should not only be able to interpret the data but respond to it fast and efficiently (Bertulfo *et al*., 2024; Martin *et al*., 2022). Finally, programs that included community response and raised the level of local trust were more likely to have an impact, which supports the significance of integrating technological services with human-oriented approaches to the social structure of public health (Lansing *et al*., 2023).

## FUTURE DIRECTIONS AND RESEARCH GAPS

With the changing landscape of infectious disease threats, secure, scalable and equitable innovation must be the priority of the future of surveillance in the United States. Federated learning and privacy preserving analytics has the potential to provide a promising direction in which large-scale, multi-institutional collaboration could take place without affecting individual or institutional data privacy. The methods enable the use of decentralized datasets (hospital networks or state health departments) to train machine learning models but retain local control of data. Not only does this strengthen the robustness of models in a variety of populations but it also covers key issues of data sharing, legal regulations, and patient confidentiality that in the present situation undermine the cross-jurisdictional surveillance endeavors.

Alongside this, real-time and adaptive surveillance systems need to be developed in order to act appropriately in response to a threat of greater dynamism and rapidity. It is no longer enough to have static dashboards and manually maintained reports in the world where outbreaks can grow in only a few days. The systems of the future need to be in a position to consume and process a variety of data streams in real time, from social behavior and mobility to wastewater observation and climate signs and respond by modifying predictive services as conditions evolve. Such responsive platforms would allow the public health officials to be able not only to track but also to preempt and counter the threats as they arise with both screens and accuracy.

The other urgent line is the harmonious combination of genomic, behavioral and clinical data to single surveillance platforms. Genomic sequencing allows identifying the variants and transmission routes, whereas behavioral data, including vaccine hesitancy, mobility, or health-seeking behavior, helps to provide context to clinical trends. These streams of data are, however, in silos and are usually maintained by different agencies or in different formats. The future studies should be based on defining a common ontology, data standards, and data analysis pipelines that facilitate comprehensive modeling of disease dynamics. This integration would expose the transmission of drivers that were not visible before and would inform more individualized, community-specific interventions.





**Publisher: SARC Publisher**

To facilitate these developments, it is extremely necessary to set national standards regarding AI-enabled surveillance. These must have ethical principles, validation standards, and interoperability standards to make sure that AI systems applied in the field of public health are secure, fair, and responsible. In the absence of a clear regulatory framework, innovations face the danger of being fragmented, biased, and losing favor of the people. All the federal, state, and academic institutions will need to coordinate their policy efforts to come up with sustainable governance structures, which can change with the technology environment.

Although the use of machine learning in the context of infectious disease surveillance has achieved quite impressive advancement, there are still certain gaps in the research, most of which require interdisciplinary cooperation. It is essential to bridge the divide between the creators of algorithms, epidemiologists, clinicians, behavioral scientists, and policymakers since it is necessary not only to have technically sound models but also operationally relevant. Also, more funds should be allocated to implementation of science to assess their performance in actual public health environments, particularly with underserved or high-risk populations. It will be critical to seal such gaps to create a more resilient, data-driven surveillance architecture which can safeguard population health in an ever more complex and interconnected world.

## CONCLUSION

The adoption of the big data and machine learning in infectious disease surveillance signifies the revolution of the concept of the health of the population in the United States. Since the first signs of the outbreak can be identified quickly, as well as resources can be distributed dynamically and specific interventions can be targeted, these technologies promise unprecedented chances to enhance timeliness, accuracy, and responsiveness. Their practical benefits can be seen in real-world applications both on a federal and state level, but current research on deep learning, anomaly detection, and genomic integration has the potential to create a more adaptive and intelligent system. Nonetheless, to achieve the full potential of these tools the constantly existing obstacles, such as data silos, privacy limitations, labor shortages, and the absence of a framework of standardized national AI use in health surveillance, have to be overcome. Subsequent studies must focus on multifaceted collaboration, infrastructure to protect privacy, and merging clinical and behavioral and genomic data to create more robust and equitable surveillance capabilities. With the future of emerging pathogens and health threats around the world, the nexus of machine learning and public health intelligence will be critical to the protection of health in the population.

## REFERENCES

1. Alemi, F., Vang, J., Wojtusiak, J., Guralnik, E., Peterson, R., Roess, A., & Jain, P. "Differential diagnosis of COVID-19 and influenza." *PLOS Global Public Health* 2.7 (2022): e0000221.
2. Ali, H. "AI for pandemic preparedness and infectious disease surveillance: predicting outbreaks, modeling transmission, and optimizing public health interventions." *Int J Res Publ Rev* 5.8 (2024): 4605-19.
3. Alpert, T., Brito, A. F., Lasek-Nesselquist, E., Rothman, J., Valesano, A. L., MacKay, M. J., & Grubaugh, N. D. "Early introductions and transmission of SARS-CoV-2 variant B. 1.1. 7 in the United States." *Cell* 184.10 (2021): 2595-2604.
4. Archer, B. N., Abdelmalik, P., Cognat, S., Grand, P. E., Mott, J. A., Pavlin, B. I., ... & Ihekweazu, C. "Defining collaborative surveillance to improve decision making for public health emergencies and beyond." *The Lancet* 401.10391 (2023): 1831-1834.
5. Babatuyi, P. B., Imoh, P. O., Igwe, E. U., & Enyejo, J. O. "The Role of Public Health Leadership in Strengthening Emergency Response Protocols and Addressing Infrastructure Gaps During Infectious Disease Outbreaks." *International Journal of Scientific Research and Modern Technology* 3.10 (2024): 109-122.
6. Babu, M., Lautman, Z., Lin, X., Sobota, M. H., & Snyder, M. P. "Wearable devices: implications for precision medicine and the future of health care." *Annual Review of Medicine* 75.1 (2024): 401-415.
7. Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. "Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study." *The Lancet Infectious Diseases* 20.11 (2020): 1247-1254.
8. Balakrishnan, V., Kherabi, Y., Ramanathan, G., Paul, S. A., & Tiong, C. K. "Machine learning approaches in diagnosing tuberculosis through biomarkers-A systematic





**Publisher: SARC Publisher**

review." *Progress in biophysics and molecular biology* 179 (2023): 16-25.
9. Bandara, T., Sandhu, N., Mehdiyeva, K., Singh, S., Plante, C., & Neudorf, C. "Enablers and barriers to public health practice during COVID-19: Perspectives from local public leadership from across Canada." *Canadian Journal of Public Health* 116.5 (2025): 642-650.
10. Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., & Viboud, C. "Big data for infectious disease surveillance and modeling." *The Journal of infectious diseases* 214.suppl_4 (2016): S375-S379.
11. Bertulfo, M. C. P., Kirkcaldy, R. D., Franzke, L. H., Sangareddy, S. R. P., & Reza, F. "Advancing data science among the federal public health workforce: the data science upskilling program, centers for disease control and prevention." *Journal of Public Health Management and Practice* 30.2 (2024): E41-E46.
12. Bettencourt, L., & Soman, S. "Systems architecture for real time epidemiological prediction and control." *Mansueto Institute for Urban Innovation Research Paper* 24 (2020).
13. Birkhead, G. S., Klompas, M., & Shah, N. R. "Uses of electronic health records for public health surveillance to advance public health." *Annual review of public health* 36.1 (2015): 345-359.
14. Bohr, A., & Memarzadeh, K. "Current healthcare, big data, and machine learning." *Artificial intelligence in healthcare*. Academic Press, 2020. 1-24.
15. Bouramtane, K., Kharraja, S., Riffi, J., El Beqqali, O., & Boujraf, S. "Efficient Vaccine Allocation for Pandemic Preparedness: Applying Machine Learning Prioritization." *International Conference on Advanced Intelligent Systems for Sustainable Development*. Cham: Springer Nature Switzerland, (2024).
16. Buckner, J. H., Chowell, G., & Springborn, M. R. "Dynamic prioritization of COVID-19 vaccines when social distancing is limited for essential workers." *Proceedings of the National Academy of Sciences* 118.16 (2021): e2025786118.
17. CDC. "Electronic Laboratory Reporting (ELR). Electronic Laboratory Reporting (ELR)." (2024a).
18. CDC. "Integrated Surveillance Information Systems/NEDSS. National Notifiable Diseases Surveillance System (NNDSS)." (2024b).
19. CDC. "About the Public Health Data Strategy. Public Health Data Strategy." (2025).
20. Cesario, E., & Comito, C. "Unveiling epidemic dynamics: harnessing the synergy of social media data and mobility patterns during COVID-19." *Neural Computing and Applications* 37.30 (2025): 25555-25578.
21. Cheah, B. C., Vicente, C. R., & Chan, K. R. "Machine learning and artificial intelligence for infectious disease surveillance, diagnosis, and prognosis." *Viruses* 17.7 (2025): 882.
22. Cheong, Q., Au-Yeung, M., Quon, S., Concepcion, K., & Kong, J. D. "Predictive modeling of vaccination uptake in US counties: A machine learning–based approach." *Journal of medical Internet research* 23.11 (2021): e33231.
23. Eze, C. E., Igwama, G. T., Nwankwo, E. I., & Victor, E. "AI-driven health data analytics for early detection of infectious diseases: A conceptual exploration of US public health strategies." *Compr Res Rev Sci Technol* 2.2 (2024): 74-82.
24. Davidson, A. J., Xu, S., Oronce, C. I. A., Durfee, M. J., McCormick, E. V., Steiner, J. F., ... & Beck, A. "Monitoring depression rates in an urban community: use of electronic health records." *Journal of public health management and practice* 24.6 (2018): E6-E14.
25. Deng, X., Garcia-Knight, M. A., Khalid, M. M., Servellita, V., Wang, C., Morris, M. K., & Chiu, C. Y. "Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant." *Cell* 184.13 (2021): 3426-3437.
26. Dixon, B. E., McGowan, J. J., & Grannis, S. J. "Electronic laboratory data quality and the value of a health information exchange to support public health reporting processes." *AMIA annual symposium proceedings*. Vol. 2011. 2011.
27. Edo-Osagie, O., Smith, G., Lake, I., Edeghere, O., & De La Iglesia, B. "Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance." *PloS one* 14.7 (2019): e0210689.
28. Ekundayo, F. "Using machine learning to predict disease outbreaks and enhance public health surveillance." *World J Adv Res Rev* 24.3 (2024): 794-811.
29. Endres-Dighe, S., Jones, K., Hadley, E., Preiss, A., Kery, C., Stoner, M., ... & Rhea, S. "Lessons learned from the rapid development of a statewide simulation model for predicting COVID-19's impact on healthcare resources





**Publisher: SARC Publisher**

and capacity." *PloS one* 16.11 (2021): e0260310.

30. Eze, P. U., Geard, N., Mueller, I., & Chades, I. "Anomaly detection in endemic disease surveillance data using machine learning techniques." *Healthcare*. Vol. 11. No. 13. MDPI, (2023).
31. Ezeh, F. E., Oparah, S. O., Gado, P., Adeleke, A. S., & Vure, S. "Early Warning Models Incorporating Environmental and Demographic Variables for Emerging Infectious Disease Prediction." (2024).
32. Fagbenle, E. "Leveraging predictive analytics to optimize healthcare delivery, resource allocation, and patient outcome forecasting systems." *International Journal of Research Publication and Reviews* 6.4 (2025): 6224-6239.
33. Flynn, C. E., & Guarner, J. "Emerging Antimicrobial Resistance. Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc." *Emerging Antimicrobial Resistance. Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc.* 36.9 (2023).
34. Ekundayo, F. "Using machine learning to predict disease outbreaks and enhance public health surveillance." *World J Adv Res Rev* 24.3 (2024): 794-811.
35. Goncalves, A. R., Pico, J. C., Hu, Y., Schlessinger, D., Greene, J., O'suilleabhain, L., & Ray, P. "AI-Enabled Diagnostic Prediction within Electronic Health Records to Enhance Biosurveillance and Early Outbreak Detection." *Medrxiv* (2025).
36. Goncalves, A. R., Pico, J. C., Hu, Y., Schlessinger, D., Greene, J., O'suilleabhain, L., & Ray, P. "AI-Enabled Diagnostic Prediction within Electronic Health Records to Enhance Biosurveillance and Early Outbreak Detection." *Medrxiv* (2025).
37. Grier, N. L., Homish, G. G., Rowe, D. W., & Barrick, C. "Promoting information sharing for multijurisdictional public health emergency preparedness." *Journal of Public Health Management and Practice* 17.1 (2011): 84-89.
38. Groseclose, S. L., & Buckeridge, D. L. "Public health surveillance systems: recent advances in their use and evaluation." *Annual review of public health* 38 (2017): 57-79.
39. Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. "An overview on the advancements of support vector machine models in healthcare applications: a review." *Information* 15.4 (2024): 235.
40. Gupta, S., Gupta, T., & Gupta, N. "Global respiratory virus surveillance: strengths, gaps, and way forward." *International Journal of Infectious Diseases* 121 (2022): 184-189.
41. Guralnik, E. "US public health surveillance, reimagined." *Learning health systems* 8.4 (2024): e10445.
42. Hammad, M. S., Ghoneim, V. F., Mabrouk, M. S., & Al-Atabany, W. I. "A hybrid deep learning approach for COVID-19 detection based on genomic image processing techniques." *Scientific Reports* 13.1 (2023): 4003.
43. Han, S., Stelz, L., Sokolowski, T. R., Zhou, K., & Stöcker, H. "Unifying physics-and data-driven modeling via novel causal spatiotemporal graph neural network for interpretable epidemic forecasting." *arXiv preprint arXiv:2504.05140* (2025).
44. Hao, B., Hu, Y., Adams, W. G., Assoumou, S. A., Hsu, H. E., Bhadelia, N., & Paschalidis, I. C. "A GPT-based EHR modeling system for unsupervised novel disease detection." *Journal of biomedical informatics* 157 (2024): 104706.
45. Harcourt, S. E., Morbey, R. A., Smith, G. E., Loveridge, P., Green, H. K., Pebody, R., & Elliot, A. J. "Developing influenza and respiratory syncytial virus activity thresholds for syndromic surveillance in England." *Epidemiology & Infection* 147 (2019): e163.
46. Harsha, P. K., Pattabiraman, C., George, A. K., Mardikar, S., Nazaar, M., Adimoolam, S., & Chandru, V. "The role of SARS-CoV-2 genomic surveillance and innovative analytical platforms for informing public health preparedness in Bengaluru, India." *medRxiv* (2022): 2022-07.
47. Hohman, K. H., Martinez, A. K., Klompas, M., Kraus, E. M., Li, W., Carton, T. W., ... & Wall, H. K. "Leveraging electronic health record data for timely chronic disease surveillance: the Multi-State EHR-Based Network for Disease Surveillance." *Journal of Public Health Management and Practice* 29.2 (2023): 162-173.
48. Hong, S., & Lynn, H. S. "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction." *BMC medical research methodology* 20.1 (2020): 199.
49. Hripcsak, G., Soulakis, N. D., Li, L., Morrison, F. P., Lai, A. M., Friedman, C., ... &





**Publisher: SARC Publisher**

Mostashari, F. "Syndromic surveillance using ambulatory electronic health records." *Journal of the American Medical Informatics Association* 16.3 (2009): 354-361.
50. Huang, J., Ding, Z., & Li, J. "Global infectious disease surveillance technologies and data sharing protocols." *Frontiers in Public Health* 13 (2025): 1676987.
51. Hughes, H. E., Edeghere, O., O'Brien, S. J., Vivancos, R., & Elliot, A. J. "Emergency department syndromic surveillance systems: a systematic review." *BMC Public Health* 20.1 (2020): 1891.
52. Hyder, A., Trinh, A., Padmanabhan, P., Marschhausen, J., Wu, A., Evans, A., ... & Jones, A. "COVID-19 surveillance for local decision making: an academic, school district, and public health collaboration." *Public Health Reports* 136.4 (2021): 403-412.
53. Idahor, C. O., Esomu, E. J. O., Ogbonna, N., Momoh, Z., Ogbeide, O. A., Ikhu-Omoregbe, O., & Oronsaye, N. J. "Infectious Disease Surveillance in the Era of Big Data and AI: Opportunities and Pitfalls." *Cureus* 17.10 (2025).
54. Jaiteh, M., Phalane, E., Shiferaw, Y. A., Jallow, H., & Phaswana-Mafuya, R. N. "The application of machine learning algorithms to predict HIV testing using evidence from the 2002–2017 South African adult population-based surveys: an HIV testing predictive model." *Tropical Medicine and Infectious Disease* 10.6 (2025): 167.
55. T KADAKIA, K. U. S. H. A. L., & Desalvo, K. B. "Transforming public health data systems to advance the population's health." *The Milbank Quarterly* 101.Suppl 1 (2023): 674.
56. Katapally, T. R., & Ibrahim, S. T. "Digital health dashboards for decision-making to enable rapid responses during public health crises: replicable and scalable methodology." *JMIR Research Protocols* 12.1 (2023): e46810.
57. Keeling, M. J., Thomas, A., Hill, E. M., Thompson, R. N., Dyson, L., Tildesley, M. J., & Moore, S. "Waning, boosting and a path to endemicity for SARS-CoV-2." *medRxiv* (2021): 2021-11.
58. Kizza, T., Aduampong, M. J. K., & Kaiser, F. "Survival analysis in U.S. chronic disease research: A systematic review of methods and applications." *International Journal of Frontline Research in Life Science*, 3.2 (2025): 26–34.
59. Khodadadi, E., & Towfek, S. K. "Internet of Things Enabled Disease Outbreak Detection: A Predictive Modeling System." *Journal of Intelligent Systems & Internet of Things* 10.1 (2023).
60. Kim, S., Kim, M. S., You, S. H., & Jung, S. Y. "Conducting and reporting a clinical research using Korean healthcare claims database." *Korean Journal of Family Medicine* 41.3 (2020): 146.
61. Kumar, S., Guruparan, D., Karuppanan, K., & Kumar, K. S. "Comprehensive insights into monkeypox (mpox): recent advances in epidemiology, diagnostic approaches and therapeutic strategies." *Pathogens* 14.1 (2024): 1.
62. Lan, Y., & Delmelle, E. "Space-time cluster detection techniques for infectious diseases: A systematic review." *Spatial and Spatio-temporal Epidemiology* 44 (2023): 100563.
63. Lansing, A. E., Romero, N. J., Siantz, E., Silva, V., Center, K., Casteel, D., & Gilmer, T. "Building trust: Leadership reflections on community empowerment and engagement in a large urban initiative." *BMC Public Health* 23.1 (2023): 1252.
64. Lawpoolsri, S., Khamsiriwatchara, A., Liulark, W., Taweeseneepitch, K., Sangvichean, A., Thongprarong, W., ... & Singhasivanon, P. "Real-time monitoring of school absenteeism to enhance disease surveillance: a pilot study of a mobile electronic reporting system." *JMIR mHealth and uHealth* 2.2 (2014): e22.
65. Li, L., Novillo-Ortiz, D., Azzopardi-Muscat, N., & Kostkova, P. "Digital data sources and their impact on people's health: a systematic review of systematic reviews." *Frontiers in public health* 9 (2021): 645260.
66. Liscano, Y., Anillo Arrieta, L. A., Montenegro, J. F., Prieto-Alvarado, D., & Ordoñez, J. "Early warning of infectious disease outbreaks using social media and digital data: A scoping review." *International journal of environmental research and public health* 22.7 (2025): 1104.
67. MacIntyre, C. R., Chen, X., Kunasekaran, M., Quigley, A., Lim, S., Stone, H., ... & Gurdasani, D. "Artificial intelligence in public health: the potential of epidemic early warning systems." *Journal of International Medical Research* 51.3 (2023): 03000605231159335.
68. Maddah, N., Verma, A., Almashmoum, M., & Ainsworth, J. "Effectiveness of public health digital surveillance systems for infectious disease prevention and control at mass





**Publisher: SARC Publisher**

gatherings: systematic review." *Journal of Medical Internet Research* 25 (2023): e44649.
69. Majumder, M. S., Cusick, M., & Rose, S. "Measuring concordance of data sources used for infectious disease research in the USA: a retrospective data analysis." *BMJ open* 13.2 (2023): e065751.
70. Mandal, N., Pramanick, B., Bhattacharyya, T. K., & Singh, B. "Advancing Point-of-Care Healthcare Diagnostics to Tackle Future Outbreaks and Pandemics: Nanotechnology-Driven Developments and Impact on Rapid Pathogen Detection." *Environment & Health* 3.12 (2025): 1423-1428.
71. Martin, L. T., Chandra, A., Nelson, C., Yeung, D., Acosta, J. D., Qureshi, N., & Blagg, T. "Technology and data implications for the public health workforce." *Big Data* 10.1_suppl (2022): S25-S29.
72. Maxime, B., Jérôme, A., Francisco, O., Alexis, S., Julie, H., S, H. J., & Gala, J. L. "Integrating patient metadata and pathogen genomic data: advancing pandemic preparedness with a multi-parametric simulator." *BMC Research Notes* 18.1 (2025): 174.
73. McClymont, H., Lambert, S. B., Barr, I., Vardoulakis, S., Bambrick, H., & Hu, W. "Internet-based surveillance systems and infectious diseases prediction: an updated review of the last 10 years and lessons from the COVID-19 pandemic." *Journal of Epidemiology and Global Health* 14.3 (2024): 645-657.
74. HOSSAIN, M. R., SNIGDHA, E. Z., & MAHABUB, S. "AI-Driven Data Optimization: Automating Cleaning, Feature Engineering, and Augmentation for Superior Machine Learning Performance in Digital Health Care System." *Journal of Computer Science and Technology Studies* 5.4 (2023): 218-228.
75. Morgan, O. W., Abdelmalik, P., Perez-Gutierrez, E., Fall, I. S., Kato, M., Hamblion, E., ... & Ihekweazu, C. How better pandemic and epidemic intelligence will prepare the world for future threats." *Nature Medicine* 28.8 (2022): 1526-1528.
76. Natrayan, L., Kamal, M. R., Manivannan, K. K., & Sunil, G. "Machine learning and data mining approaches for infectious disease surveillance and outbreak management in healthcare." *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*. IEEE, (2024).
77. Ndao, A., Dath, C. A. B., Seck, C. T., & Diop, B. "Comparison of Logistic Regression Models, Random Drills and Support Vector Machines (SVM) for the Epidemiological Surveillance of Ten (10) Infectious Diseases in Senegal." *Indian Journal of Science and Technology* 17.48 (2024): 5083-5090.
78. Nobles, M., Lall, R., Mathes, R. W., & Neill, D. B. "Presyndromic surveillance for improved detection of emerging public health threats." *Science Advances* 8.44 (2022): eabm4920
79. Okut, H. "Deep learning for subtyping and prediction of diseases: Long-short term memory." *Deep Learning Applications*. IntechOpen, (2021).
80. Olayinka, O. H. "Big data integration and real-time analytics for enhancing operational efficiency and market responsiveness." *Int J Sci Res Arch* 4.1 (2021): 280-96.
81. Oude Munnink, B. B., Worp, N., Nieuwenhuijse, D. F., Sikkema, R. S., Haagmans, B., Fouchier, R. A., & Koopmans, M. "The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology." *Nature medicine* 27.9 (2021): 1518-1524.
82. Pandey, S. K., Janghel, R. R., Mishra, P. K., & Kaabra, R. "Machine learning based COVID-19 disease recognition using CT images of SIRM database." *Journal of Medical Engineering & Technology* 46.7 (2022): 590-603.
83. Parums, D. V. "infectious disease surveillance using artificial intelligence (AI) and its role in epidemic and pandemic preparedness." *Medical science monitor: international medical journal of experimental and clinical research* 29 (2023): e941209-1.
84. Patil, D. "Artificial intelligence-driven supply chain optimization: Enhancing demand forecasting and cost reduction." *Available at SSRN 5057408* (2024).
85. Perlman, S. E., McVeigh, K. H., Thorpe, L. E., Jacobson, L., Greene, C. M., & Gwynn, R. C. "Innovations in population health surveillance: using electronic health records for chronic disease surveillance." *American journal of public health* 107.6 (2017): 853-857.
86. Petanidis, S., Chandramouli, K., Floros, G., Nifakos, S., Kolomvatsos, K., Tsekeridou, S., & Kosmidis, C. "Optimizing Emergency Response in Hospitals: A Systematic Review of Surge Capacity Planning and Crisis





**Publisher: SARC Publisher**

Resource Management." *Healthcare*. Vol. 13. No. 21. MDPI, (2025).

87. Pontoh, R. S., Toharudin, T., Ruchjana, B. N., Gumelar, F., Putri, F. A., Agisya, M. N., & Caraka, R. E. "Jakarta pandemic to endemic transition: forecasting COVID-19 using NNAR and LSTM." *Applied Sciences* 12.12 (2022): 5771.
88. Qian, Z., Alaa, A. M., & van der Schaar, M. "CPAS: the UK's national machine learning-based hospital capacity planning system for COVID-19." *Machine Learning* 110.1 (2021): 15-35.
89. Ren, W., Zhu, J., Liu, Z., Zhao, T., & Honavar, V. "A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models." *arXiv preprint arXiv:2507.12774* (2025).
90. Richards, C. L., Iademarco, M. F., Atkinson, D., Pinner, R. W., Yoon, P., Mac Kenzie, W. R., ... & Frieden, T. R. "Advances in public health surveillance and information dissemination at the Centers for Disease Control and Prevention." *Public Health Reports* 132.4 (2017): 403-410.
91. Rodriguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., & Prakash, B. A. "Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 17. (2021).
92. Romano, S., Davis, C., Collier, K., Johnston, S., Tesfamichael, H., & Yusuf, H. "Evaluation Activities from the National Syndromic Surveillance Program." *Online Journal of Public Health Informatics* 10.1 (2018): e62235.
93. Romano, S., Yusuf, H., Davis, C., Thomas, M. J., & Grigorescu, V. "An Evaluation of Syndromic Surveillance–Related Practices Among Selected State and Local Health Agencies." *Journal of Public Health Management and Practice* 28.2 (2022): 109-115.
94. Rosenfeld, N., & Globerson, A. "Semi-supervised learning with competitive infection models." *International Conference on Artificial Intelligence and Statistics*. PMLR, (2018).
95. Roster, K., Connaughton, C., & Rodrigues, F. A. "Machine-learning–based forecasting of dengue fever in Brazilian cities using epidemiologic and meteorological variables." *American journal of epidemiology* 191.10 (2022): 1803-1812.
96. Samoff, E., Fangman, M. T., Fleischauer, A. T., Waller, A. E., & MacDonald, P. D. "Improvements in timeliness resulting from implementation of electronic laboratory reporting and an electronic disease surveillance system." *Public health reports* 128.5 (2013): 393-398.
97. Santos, P. D., Ziegler, U., Szillat, K. P., Szentiks, C. A., Strobel, B., Skuballa, J., & Hoeper, D. "In action—an early warning system for the detection of unexpected or novel pathogens." *Virus Evolution* 7.2 (2021): veab085.
98. Shahin, R., Beaulieu, M., & Shahin, A. "Time Series Forecasting for Personal Protective Equipment During COVID-19 Pandemic: A Case Study of Quebec." *Data-Centric Business and Applications: Advancements in Information and Knowledge Management, Volume 2*. Cham: Springer Nature Switzerland, 2024. 215-234.
99. Shih, Y. C. T., & Liu, L. "Use of claims data for cost and cost-effectiveness research." *Seminars in radiation oncology*. Vol. 29. No. 4. WB Saunders, (2019).
100. Simonsen, L., Gog, J. R., Olson, D., & Viboud, C. "Infectious disease surveillance in the big data era: towards faster and locally relevant systems." *The Journal of infectious diseases* 214.suppl_4 (2016): S380-S385.
101. Sloth Lorenzen, S., Nielsen, M., Jimenez-Solem, E., Studsgaard Petersen, T., Perner, A., Thorsen-Meyer, H. C., ... & Sillesen, M. "Developing Machine Learning Models for Predicting Intensive Care Unit Resource Use During the COVID-19 Pandemic: Experiences from a Bi-Regional Health Care System Covering 2.5 Million Citizens in Denmark." *Available at SSRN 3796914* (2021).
102. Song, Y., Zhao, M., Yang, H., Huang, P., & Chen, Y. "Wearable Point-of-Care Biosensor for Biomolecular Assay in Health Monitoring." *ACS Applied Bio Materials* 8.10 (2025): 8578-8596.
103. Srivastava, V., Kumar, R., Wani, M. Y., Robinson, K., & Ahmad, A. "Role of artificial intelligence in early diagnosis and treatment of infectious diseases." *Infectious Diseases* 57.1 (2025): 1-26.
104. Ssemujju, F. S., & Solomon, D. "Integrating Spatial Modeling and Machine Learning for Infectious Disease Surveillance





**Publisher: SARC Publisher**

in US Urban and Rural Settings." *Journal Of Internal Medicine And Public Health* 5.1 (2026): 1-12.

105. Taiwo, M. "Optimizing public health infrastructure through predictive modelling for resource distribution and crisis management." *International Journal of Research Publication and Reviews* 6.1 (2025): 1706-1724.
106. Tharanidharan, S. "Machine Learning-Based Detection of Cyber Defamation in Social Networks." *International Journal of Intelligent Systems and Applications in Engineering* 12.4s (2024): 785-793.
107. Tiwari, S., Dhakal, T., Kim, B. J., Jang, G. S., & Oh, Y. "Genomics in epidemiology and disease surveillance: An exploratory analysis." *Life* 15.12 (2025): 1848.
108. Torres, L. M., Johnson, J., Valentine, A., Brezak, A., Schneider, E. C., D'Angeli, M., ... & Black, A. "Integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA." *Emerging infectious diseases* 31.Suppl 1 (2025): S25.
109. Ugwu, D., Amofah, A. D., & Kaiser, F. "Enhancing the detection and response to infectious disease outbreaks." *International Medical Science Research Journal* 5.2 (2025).
110. Uyeki, T. M., Hui, D. S., Zambon, M., Wentworth, D. E., & Monto, A. S. "Influenza. Lancet (London, England)." 400(10353), 693–706. (2022).
111. Villanueva-Miranda, I., Xiao, G., & Xie, Y. "Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review." *Frontiers in public health* 13 (2025): 1609615.
112. Wang, L., Chen, J., & Marathe, M. "DEFSI: Deep learning based epidemic forecasting with synthetic information." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. (2019).
113. Ward, T., Johnsen, A., Ng, S., & Chollet, F. "Forecasting SARS-CoV-2 transmission and clinical risk at small spatial scales by the application of machine learning architectures to syndromic surveillance data." *Nature Machine Intelligence* 4.10 (2022): 814-827.
114. Washington, N. L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E. T., Barrett, K. M. S., ... & Andersen, K. G. "Emergence and rapid transmission of SARS-CoV-2 B. 1.1. 7 in the United States." *Cell* 184.10 (2021): 2587-2594.
115. Weerakody, P. B., Wong, K. W., Wang, G., & Ela, W. "A review of irregular time series data handling with gated recurrent neural networks." *Neurocomputing* 441 (2021): 161-178.
116. World Health Organization. "United States of America: WHO coronavirus disease (COVID-19) dashboard." *World Health Organization* (2023).
117. Williams, B. A., Voyce, S., Sidney, S., Roger, V. L., Plante, T. B., Larson, S., ... & Benziger, C. P. "Establishing a national cardiovascular disease surveillance system in the United States using electronic health record data: key strengths and limitations." (2022): e024409.
118. Wu, C. C., Chen, C. H., Wang, S. R., & Shete, S. "Assessing spatial variability in observed infectious disease spread in a prospective time–space series." *International Journal of Health Geographics* 24.1 (2025): 28.
119. Xia, Z., Qin, L., Ning, Z., & Zhang, X. "Deep learning time series prediction models in surveillance data of hepatitis incidence in China." *Plos one* 17.4 (2022): e0265660.
120. Xian, X. "Frontiers of wearable biosensors for human health monitoring." *Biosensors* 13.11 (2023): 964.
121. Xu, C., Zhao, L. Y., Ye, C. S., Xu, K. C., & Xu, K. Y. "The application of machine learning in clinical microbiology and infectious diseases." *Frontiers in Cellular and Infection Microbiology* 15 (2025): 1545646.

**Source of support:** Nil; **Conflict of interest:** Nil.

**Cite this article as:**
Kebeba, M. & Agyei, E. A "Big Data and Machine Learning Applications for Enhanced U.S. Infectious Disease Surveillance and Control: A Narrative Review." *Sarcouncil journal of Medical sciences* 5.2 (2026): pp 53-68.





**Publisher: SARC Publisher**