Machine Learning Models for Early Detection and Monitoring of Infectious Diseases: AI Approaches for Enhancing Surveillance and Response Strategies

By Dr. Jean-Pierre Berger

Associate Professor of Artificial Intelligence, Université Claude Bernard Lyon 1, France

1. Introduction to Infectious Diseases and Surveillance

Infectious diseases have been at the forefront of major global pandemics, episodic regional outbreaks that cripple public health, and ongoing threats recurrently experienced at national and subnational levels. Emerging and reemerging pathogens have the potential to cause large-scale social, economic, and political disruptions. Highly interconnected societies, combined with unprecedented ease and speed of global travel, make the rapid surveillance, detection, and response to especially infectious diseases a real necessity. Surveillance systems are an important part of the public health response to infectious diseases and can be used to monitor changing trends, to identify possible outbreaks, to indicate areas where additional data collections are needed, and to evaluate the impact of control and prevention programs. Certain infectious diseases evolve rapidly and necessitate robust surveillance methods for early detection and monitoring activities to mitigate the impact of a disease.

Because of the complexity and globalization of infectious disease, impairments in local, regional, or national surveillance systems can hinder international efforts to detect and respond to human health threats. A clear need exists to integrate technology into existing surveillance systems with additional mechanisms necessary to be incorporated into stand-alone platforms. Some of these platforms are aimed at enhancing surveillance capabilities in countries and regions heavily impacted by global health security challenges. One such strategy involves the adaptation of advancements in machine learning applied in other sectors. In this context, this text aims to present a survey of both framed and grey literature to determine what current machine learning models can offer in the way of early detection and monitoring of infectious diseases.

2. Role of Machine Learning in Early Detection

Early detection is a cornerstone of a successful strategy to contain and limit the impact of an outbreak of infectious diseases. Today, a small number of interventions have been designed based on machine learning algorithms: techniques that are able to process, automatically classify, and learn from datasets characterized by high volume, high dimensionality, and heterogeneity. Different ML methodologies can be used to model data and predict outcomes regarding outbreaks. On one hand, unsupervised learning algorithms can automatically identify and provide new insights when analyzing unique datasets, where the absence of clear examples does not allow a training phase; on the other hand, supervised learning algorithms can tackle predictive modeling through training with historical datasets.

The use of electronic health records and other everyday datasets could transform not only the way health is being monitored but may, with the application of novel machine learning methodologies to detect signals of infectious diseases emerging from such data, save individual lives and prevent weekend food traders from the double disaster of illness and loss of customers. Health monitoring systems for epidemics from infectious agents have traditionally collected data from only a small number of syndromes and system performance parameters that are perceived to be particularly relevant. Such data are recorded and transmitted remotely. The range of syndromes and time delays between the onset of symptoms and consultation time are smaller when such data are recorded online compared to data for hospital inpatients for whom there may be significant delays between the onset of symptoms and laboratory investigation. Real-time processing of large data streams for monitoring infectious diseases provides the opportunity to quickly identify atypical health patterns.

An increased investment of support is required to enhance the development of machine learning algorithms that have the capability to detect the signals of infectious diseases from a range of data flows. It is advocated that government departments responsible for trade, travel, and domestic affairs should examine whether they can and should alter response strategies that should encourage the incorporation of these new technologies into generic response strategies in line with the updated regulations concerning systems for global alerts and response.

3. Key Machine Learning Models for Infectious Disease Surveillance

In this review, we examine the role of machine learning models in the detection of infectious diseases using digital data sources. We chose to focus on three main types of models that are increasingly employed in surveillance practices, including supervised, unsupervised, and deep learning models. In recent years, models included in these categories have been particularly gaining prominence in digital disease detection literature. By utilizing a range of data sources, they promise higher predictive accuracy, notably in the context of public health surveillance scenarios, and enhance the decision-making process to support timely responses to infectious disease threats.

Supervised learning approaches come up most frequently in the studies included in our review. These models are trained on labeled data to make predictions and build new models for unseen data. They have shown great potential in the detection and monitoring of infectious diseases, for example, in tracking seasonal and/or pandemic influenza. Additional examples exist in drafts on COVID-19 modeling that are included in this special issue. Unsupervised learning models come next and allow for discovering novel patterns from data without pre-existing labels. They have been applied to biosurveillance aims to identify anomalies in data that might relate to infectious disease threats. For instance, the deployment of these models may lead to the early detection of unnatural disease outbreaks, such as bioterrorist attacks. Lastly, deep learning architectures have, in the last few years, started to make an appearance in the surveillance of infectious disease literature. Deep learning models are used in tasks that require extracting high-level information, processing large volumes of data, or in studies that use unstructured data. For example, these models have been used to detect diseases in radiological images. In infectious disease surveillance settings, the use of deep learning models has focused on processing and analyzing unstructured textual data, such as news, social media, or online reports about infectious diseases. These models, especially, hold great potential for early detection studies that employ digital media data.

3.1. Supervised Learning Algorithms

The objective of the supervised learning task is to predict specific outputs based on available input data. In this approach, a model is trained on a labeled dataset that involves input data with their corresponding outputs. Labeled data are crucial because they play an important role in the process of training the model to correctly capture the underlying patterns in the data. Therefore, a training dataset should be chosen carefully to be as representative as

possible of the whole population, which can guarantee that the model will also perform well on unseen instances. In the context of infectious disease surveillance, such algorithms can be trained on labeled datasets including disease cases that have been confirmed or reported and utilized to make accurate predictions between the input datasets and the number of cases occurring in specific regions, or the likelihood that an outbreak will occur.

Various studies have demonstrated the effectiveness of employing supervised algorithms in accurately identifying the occurrence of infectious disease outbreaks and their development trends. A time series regression analysis was proven to effectively correlate the weekly patient trend with the cases of influenza-like illness. In a Bayesian Gaussian Process methodology, a set of input data divided into two models, which are the cumulative and the previous cases of zoonotic TB, has shown the capability to effectively reproduce the number of new TB cases in wildlife. Another field where these algorithms have been widely applied is in syndromic surveillance. Using the time course of laboratory results, a change in the proportion of incident HIV cases was illustrated to be accurately identified using decision tree learning. Ensemble models and support vector machines were also proven to perform better in recognizing HIV patients compared to individual learners. The choice of methods between these algorithms will ultimately be determined by the applications and the complexity of the characteristics that are planned to be trained on the system. However, biases may occur when using training datasets contaminated with false positives or false negatives, which will compromise the learning and inference ability of these algorithms. Furthermore, overfitting refers to a learning model that cannot be generalized to predict new cases beyond the training dataset. Hence, solving these limitations and developing newer advanced algorithms is a prospective suggestion.

3.2. Unsupervised Learning Algorithms

In contrast to supervised learning, unsupervised learning algorithms do not utilize labeled data for training and rely on discovering patterns directly from input data. This makes these methods particularly useful for dealing with outbreak situations of infectious disease. Traditional methods of surveillance using supervised learning would be impossible because we need at least a reasonable amount of labeled data to have a potential predictive susceptibility model. In many scenarios, labeled data are rare or non-existent. There are various unsupervised learning algorithms. Clustering techniques are beneficial to identify

hidden subpopulations and underlying structures in epidemiological data, while anomaly detection identifies any outliers from the majority. Unsupervised learning algorithms are also used to detect atypical clusters or extreme values in space and/or time; hence, they can be used for the early detection of epidemic events and the determination of the type of outbreak transmission dynamics. In causal inference studies, this approach can also be used to identify different patterns of individuals.

A major advantage of unsupervised machine learning algorithms is the flexibility and adaptability to the rapidly changing epidemiological context. The model does not require careful calibration between parameters, nor does it require updating to adjust to new emerging infectious diseases. However, limitations and challenges arise when deploying unsupervised learning algorithms in practice. These algorithms are data-driven, and researchers may face a trade-off between data quality, dataset size, algorithm complexity, model parameterization, and model interpretability. Complex machine learning models and algorithm results are often difficult to explain. Researchers have tried to validate methodologies utilizing either simulated scenarios or real data focused on local outbreaks. In such cases, these early warning systems were largely retrospective, and only data available at the start of suspect outbreaks were considered, contrary to real-time surveillance scenarios where streaming data is continuously fed into predictive models. These approaches do not supply real-time insights to inform early warning responses. Therefore, unsupervised learning algorithms are relevant to improve health surveillance. They help to create links between upstream and downstream health data, as in the case of syndromic surveillance from medical diagnoses coded in patients' healthcare records. This type of approach is a reminder of how such a system can readily adapt to a range of animal diseases as well.

3.3. Deep Learning Architectures

Deep learning is a subset of machine learning that is based on the use of multi-layered neural networks. When models are given a large set of data to learn from, the neurons within the network adjust the parameters to discover underlying patterns, commonalities, and relationships that exist in the data. The feature hierarchies that the models can build are its chief advantage, making them adept at analyzing high-dimensional, complex, and 'messy' data, such as images, sound, and text. Deep learning has revolutionized image processing and has achieved the state-of-the-art in computer vision and natural language processing tasks.

Few-shot and one-shot learning techniques, generative adversarial networks, and representation learning have accelerated deep learning in solving a plethora of computer vision and natural language modeling tasks, applications that have significance to the field of epidemiology.

Deep learning-based models have been used in some applications relevant to infectious disease surveillance, such as the creation of an autoencoder-based anomaly detection system that used images to estimate poverty in urban areas. The model detected a poverty outbreak caused by a natural disaster up to 16 days before official estimates were released; a side-by-side visual comparison with the purchase activity of a large retailer similarly indicated that the retailer's data were three weeks behind the satellite data. Another convolutional neural network compared synoptic surveillance images before and after releases in a hospital's endoscopy unit and flagged 89% of the cases associated with infection outcomes, while a count-based method that monitored reports and observations of personnel indicated the same outcome in only 29% of cases. A long short-term memory recurrent neural network that made use of just-in-time learning data achieved model training, inference, and validation results better than baseline models, indicating its strength in analyzing temporally varying electronic health records. A body of research also exists demonstrating the growing value of deep learning to infectious disease outbreak prediction and monitoring. For example, machine and deep learning algorithms were applied to social media data to improve the accuracy, efficiency, and timeliness of cholera outbreak detection. Research documented that social media data improved the specificity of an outbreak-predicting model significantly and made it possible to issue an alert earlier than traditional surveillance systems. Overall, these developments provide strong examples of the potentially transformative impact that deep learning could have across the fields of infectious disease forecasting, detection, and monitoring.

A final point of consideration is that deep learning requires an extensive amount of computational resources, in particular to understand the 'deep' and complex relationships in the data. Moreover, as the learning process in deep learning is reliant upon the characteristics of the data, such as the density of the data points and the decision boundaries between different classes, biases that exist within data samples can be learned by the model. Another challenge in deep learning models is the lack of interpretability; as the models are complex and non-linear, it is difficult to understand how the model comes to its conclusion. It can thus

be difficult to interpret the processes that underpin the models' outputs, as well as the rationale for an operating threshold or values. Since clear and interpretable indicators and decision thresholds are key to informing a public health action, model interpretability and/or explainability is paramount when operationalizing deep learning models for health security.

4. Challenges and Limitations in Implementing ML Models

The implementation of ML models for infectious disease surveillance is associated with several challenges. Firstly, the quality of used data is an issue. Syndromic data sources may present poor data quality, including incomplete records, noise, or potential bias. Moreover, data are sometimes pre-processed, which can impact the observed time series. There may also be delays between the time when an event occurs, when it is reported, and when it is available for analysis, which can impact the system's timeliness and resolution. Model training requires a significant amount of annotated data. Given the complexity of health data annotation, access to expert-annotated data from the field is limited. Even when large quantities of data are available, large variations in health status, ethnicity, cultural, and social factors as well as access to health care across geographic regions will diminish model performance. Thus, models must be retrained using local data. The use of global models trained using data from different geographic regions can penalize model performance. Consequently, before they can be operationalized, models need to overcome the systematization of inequalities in different populations.

In order to promote the use of these models in public health, it is important to make the model transparent and interpretable. For public trust, it is important to ensure that a decision is not driven by an invisible characteristic. Therefore, models used in the context of health surveillance must systematically account for the elements of explainable AI, not only to predict potential health outcomes but also to ensure that surveillance activities will not be biased. Moreover, data privacy may be a concern if patient records are used to train models. Concerns over access and ownership of personal data may create issues for modeling approaches. Ethical concerns related to potential harm during a public health emergency can emerge. Collaborative studies can help to address those challenges. Interdisciplinary research activities can lead to improving access to experts and can open new possibilities. Clearly, more research and development in AI-based health surveillance is needed. Model limitations—linked to data input and algorithmic constraints—are highlighted and will continue to impact

the operationalization of any model due to insufficient data collection and common regulatory restrictions.

5. Case Studies and Applications in Infectious Disease Surveillance

5. Case Studies and Applications in Infectious Disease Surveillance

5.1 Prediction of the 2015 Ebola Outbreak Machine learning analyses leveraging public internet-based surveillance data have been used to predict infectious disease outbreaks and assist in the monitoring and characterization of ongoing outbreaks. More granular forecasts of infectious disease spread have also been generated using an asymptotic growth model and a generative linkage model trained on subnational infectious disease case count data. Machine learning can help identify appropriate and interpretable features within non-traditional data streams and provide alternative methods to inform public health responses.

5.4.1 Dengue Fever in Brazil In addition to malaria, the GFDD has been used for subnational infectious disease surveillance. The histogram intersection-based SVM algorithm was applied to the GFDD to monitor dengue fever in Brazil. The algorithm was used to classify the spatial patterns of dengue outbreaks, and the output was validated using the Pearson correlation coefficient between the modeled and recorded suspect cases of dengue in the State of São Paulo, Brazil. The study showed that the algorithm is able to capture actual dengue outbreaks from the start, but the intensity of the outbreaks is far from the recorded ones. Additionally, false positive detections of dengue fever patterns were concentrated in the same area, indicating the presence of similar conditions to areas actually affected by dengue fever. The researchers addressed the limitations in human case data and the need for timely and reliable data to validate the algorithm.

6. Future Direction

With the pilot work conducted so far and while acknowledging mainly the limitations related to available infectious disease datasets, data characteristics, and operational surveillance systems, in this final section we identify a few areas for future work and discussion in this field. Given the increasing availability of digital data and the increasing interest in modeling the spread of infectious diseases at sub-national scales, such as small-scale spatial and temporal phenomena, it will be necessary to develop surveillance methods and modeling results that are directly applicable to small-scale responding and incremental.

In the development of models that integrate machine learning and other big data methodologies, we suggest that the importance of building a solid base of quality data be emphasized in future work. This goes beyond ensuring internal model validity but also reliable interpretation of model results and outputs. A multi-stakeholder approach developed between academia, governments, public health authorities, and industries is likely to provide the best way forward for integrating non-traditional data sources to develop disease dynamics surveillance applications. Open communication and transparent information sharing within these collaborations may also help to improve the integration of new methods and technologies and foster dialogue relating to the ethical, equity, and human rights implications that could arise. Policymakers and ethicists should prepare for the use of AI models in this area and ensure that robust regulatory frameworks are in place to protect vulnerable and minority populations. This holds particularly true if decentralized data systems become a future source of disease dynamics models, as the possibilities to scrutinize and ensure data integrity will increase. We recommend that new models should start to move away from batch learning methods toward those that are capable of incremental learning and can deal with streaming, continuous data and knowledge integration.

7. Conclusion

Infectious diseases are among the main causes of death worldwide. Disruptive events such as travel, urbanization, vaccine policies, international conflicts, among others, can have a substantial effect on the global burden of infectious diseases. As such, the accumulation of high-quality data and comprehensive understanding of the mechanisms that underpin the surveillance and response strategies have a critical effect on preventing, controlling and mitigating global outbreaks. Machine learning offers the potential to support and enhance infectious disease surveillance through increasing early detection capability, reducing the time to establish the cause, route of transmission, and the best prevention and mitigation responses to outbreaks, and/or reducing the costs, number of deaths, and burden to society caused by epidemic episodes mostly affecting diseases of neglected importance in Low and Middle Income Countries. AI is an integral point of interest across infectious diseases and wider global health efforts and initiatives. Within 'infectious disease AI' there is a proliferation of possibilities as well as responsibilities. The broad and constantly evolving disciplinary areas of AI and bioinformatics, computational social sciences and the range of theoretical algorithms and models will require further elucidation and, in some cases, cultural

modification, so that innovation and intervention go hand in hand. Without undermining the exploitation of vulnerabilities and responses, AIs of learning and decision can challenge our traditional models of public health. It is clear that early and lightweight public health surveillance is increasingly at the forefront. This pragmatic and technical review has presented the 'what' and the 'how' of deploying machine learning strategies applicable to a range of infectious disease monitoring including zoonoses, emerging and re-emerging infections, and 'endemic' pathogens. There remains a need for a consensus approach in terms of nomenclature, practical applications, dissonances, and data analysis. The strengths and limitations of these new strategies must be characterized and regularly updated to ensure that stakeholders, governments, and tech companies advocate for the public benefit.

Reference:

1. Pushadapu, Navajeevan. "AI-Driven Solutions for Enhancing Data Flow to Common Platforms in Healthcare: Techniques, Standards, and Best Practices." Journal of Computational Intelligence and Robotics 2.1 (2022): 122-172.
2. Bao, Y.; Qiao, Y.; Choi, J.E.; Zhang, Y.; Mannan, R.; Cheng, C.; He, T.; Zheng, Y.; Yu, J.; Gondal, M.; et al. Targeting the lipid kinase PIKfyve upregulates surface expression of MHC class I to augment cancer immunotherapy. Proc. Natl. Acad. Sci. USA 2023, 120, e2314416120.
3. Gayam, Swaroop Reddy. "AI for Supply Chain Visibility in E-Commerce: Techniques for Real-Time Tracking, Inventory Management, and Demand Forecasting." Distributed Learning and Broad Applications in Scientific Research 5 (2019): 218-251.
4. Nimmagadda, Venkata Siva Prakash. "AI-Powered Risk Management and Mitigation Strategies in Finance: Advanced Models, Techniques, and Real-World Applications." Journal of Science & Technology 1.1 (2020): 338-383.
5. Putha, Sudharshan. "AI-Driven Metabolomics: Uncovering Metabolic Pathways and Biomarkers for Disease Diagnosis and Treatment." Distributed Learning and Broad Applications in Scientific Research 6 (2020): 354-391.

6. Sahu, Mohit Kumar. "Machine Learning Algorithms for Enhancing Supplier Relationship Management in Retail: Techniques, Tools, and Real-World Case Studies." Distributed Learning and Broad Applications in Scientific Research 6 (2020): 227-271.

7. Kasaraneni, Bhavani Prasad. "Advanced Machine Learning Algorithms for Loss Prediction in Property Insurance: Techniques and Real-World Applications." Journal of Science & Technology 1.1 (2020): 553-597.

8. Kondapaka, Krishna Kanth. "Advanced AI Techniques for Optimizing Claims Management in Insurance: Models, Applications, and Real-World Case Studies." Distributed Learning and Broad Applications in Scientific Research 5 (2019): 637-668.

9. Kasaraneni, Ramana Kumar. "AI-Enhanced Cybersecurity in Smart Manufacturing: Protecting Industrial Control Systems from Cyber Threats." Distributed Learning and Broad Applications in Scientific Research 5 (2019): 747-784.

10. Pattyam, Sandeep Pushyamitra. "AI in Data Science for Healthcare: Advanced Techniques for Disease Prediction, Treatment Optimization, and Patient Management." Distributed Learning and Broad Applications in Scientific Research 5 (2019): 417-455.

11. Kuna, Siva Sarana. "AI-Powered Techniques for Claims Triage in Property Insurance: Models, Tools, and Real-World Applications." Australian Journal of Machine Learning Research & Applications 1.1 (2021): 208-245.

12. Nimmagadda, Venkata Siva Prakash. "Artificial Intelligence for Automated Loan Underwriting in Banking: Advanced Models, Techniques, and Real-World Applications." Journal of Artificial Intelligence Research and Applications 2.1 (2022): 174-218.

13. Pushadapu, Navajeevan. "Advanced AI Algorithms for Analyzing Radiology Imaging Data: Techniques, Tools, and Real-World Applications." Journal of Machine Learning for Healthcare Decision Support 2.1 (2022): 10-51.

14. Gayam, Swaroop Reddy. "AI-Driven Customer Support in E-Commerce: Advanced Techniques for Chatbots, Virtual Assistants, and Sentiment Analysis." Distributed Learning and Broad Applications in Scientific Research 6 (2020): 92-123.

15. Nimmagadda, Venkata Siva Prakash. "Artificial Intelligence and Blockchain Integration for Enhanced Security in Insurance: Techniques, Models, and Real-World Applications." African Journal of Artificial Intelligence and Sustainable Development 1.2 (2021): 187-224.

16. Putha, Sudharshan. "AI-Driven Molecular Docking Simulations: Enhancing the Precision of Drug-Target Interactions in Computational Chemistry." African Journal of Artificial Intelligence and Sustainable Development 1.2 (2021): 260-300.

17. Sahu, Mohit Kumar. "Machine Learning for Anti-Money Laundering (AML) in Banking: Advanced Techniques, Models, and Real-World Case Studies." Journal of Science & Technology 1.1 (2020): 384-424.

18. Kasaraneni, Bhavani Prasad. "Advanced Artificial Intelligence Techniques for Predictive Analytics in Life Insurance: Enhancing Risk Assessment and Pricing Accuracy." Distributed Learning and Broad Applications in Scientific Research 5 (2019): 547-588.

19. Kondapaka, Krishna Kanth. "Advanced AI Techniques for Retail Supply Chain Sustainability: Models, Applications, and Real-World Case Studies." Journal of Science & Technology 1.1 (2020): 636-669.

20. Kasaraneni, Ramana Kumar. "AI-Enhanced Energy Management Systems for Electric Vehicles: Optimizing Battery Performance and Longevity." Journal of Science & Technology 1.1 (2020): 670-708.

21. Pattyam, Sandeep Pushyamitra. "AI in Data Science for Predictive Analytics: Techniques for Model Development, Validation, and Deployment." Journal of Science & Technology 1.1 (2020): 511-552.

22. Kuna, Siva Sarana. "AI-Powered Solutions for Automated Underwriting in Auto Insurance: Techniques, Tools, and Best Practices." Journal of Science & Technology 1.1 (2020): 597-636.