

A comparative analysis of supervised vs. unsupervised machine learning in predicting disease outbreaks

Avantika Guha

Student

Abstract- Disease outbreaks continue to strain major public healthcare systems, and early detection of disease outbreaks is essential for timely interventions, efficient resource allocation, and improved disease surveillance, particularly for respiratory infections such as influenza and COVID-19. Recent advances in machine learning (ML) have provided effective tools for analysing large-scale surveillance data, which includes case reports, laboratory-confirmed diagnoses, hospital admissions, and environmental variables, to support outbreak prediction and monitoring. This study presents a comparative analysis of the efficiency and effectiveness of supervised and unsupervised machine learning approaches to predict respiratory disease outbreaks. Moreover, a conceptual research design was adopted to eliminate the use of primary surveillance data or model implementation, and findings were synthesized from the existing studies. Supervised learning techniques, including Logistic Regression, Decision Trees, and many more, were compared with unsupervised approaches such as K-clustering based on their data requirements, predictive performance, interpretability, and practical applications. The study showcases the areas where both learning approaches shine distinctly, and also brings forth where they start to fall short. Overall, the study concludes that the choice between the two learning models depends heavily on the availability and quality of the surveillance data, and unifying both learning approaches into one single framework can offer the highest potential for epidemic prediction and surveillance quality.

Keywords: Machine Learning, Supervised Learning, Unsupervised Learning, Disease Outbreak Prediction, Respiratory Disease Surveillance, Public Health Surveillance

1. INTRODUCTION

1.1 Disease Outbreaks and Public Health Importance

Disease outbreaks are the sudden escalation of an infectious illness that soon contributes to greater public health concerns, public safety, and negative impacts on healthcare infrastructure. Ultimately, the issue significantly concerns social and economic disruptions. Generally, it occurs when the number of disease cases exceeds the expected level within a particular population or geographical area during a specific period. According to WHO reports, these types of disease outbreaks can be caused by vector-borne, food and waterborne, airborne, and unknown etiology (WHO, 2023). However, it can be limited if effective outbreak predictions and efficient medical resources can be provided on time. Early warning systems are particularly important for reducing transmission, minimizing the burden on hospitals, and protecting vulnerable populations (McGaughy et al., 2021).

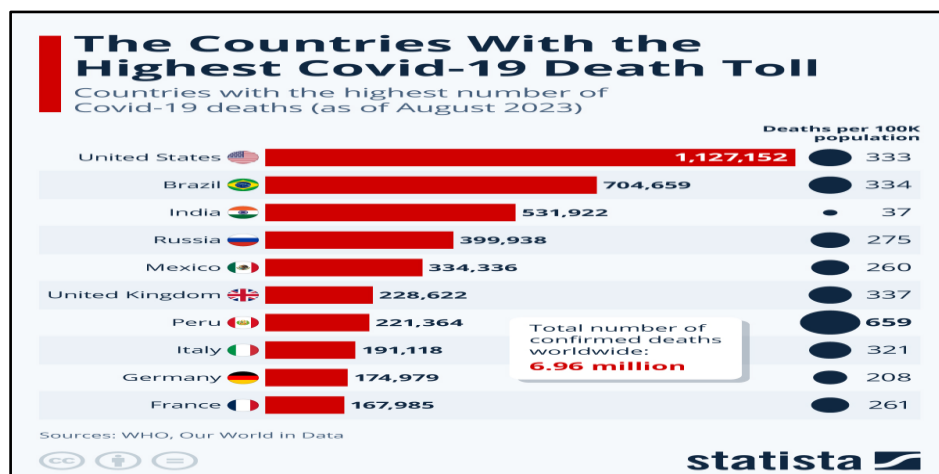


Figure 1.1.1: Deaths Caused by COVID-19

Source: (Fourreau, 2021)

However, in reality, severe infectious diseases continue to pose challenges to global health systems, and still, the outbreak prediction system and early warning systems are not effectively implemented. For instance, Influenza outbreaks occur seasonally and contribute to substantial morbidity and 2,90,000 to 6,50,000 respiratory deaths worldwide (World Health Organization, 2025). Another example of the lack of outbreak prediction and early warning systems can be demonstrated with the recent COVID-19 pandemic. COVID-19 highlights how quickly respiratory pathogens can spread across regions and countries, resulting in increased hospitalization rates and pressure on healthcare infrastructure, and leading to several deaths (Figure 1.1.1). Similarly, cholera outbreaks also have been another “significant global health threat for over two centuries”, leading to over “40 million deaths” and continue to take over lakhs of lives each year currently (Mohamed et al., 2025). While predictive surveillance can mitigate the spread of many infectious diseases, respiratory infections remain a major area of concern because of their high transmissibility and frequent occurrence.

Recent cases of respiratory diseases, including influenza, coronavirus infections, pneumonia, and other respiratory tract infections, account for a considerable proportion of global healthcare utilization. As their transmission can be influenced by multiple factors, including environmental conditions, population density, mobility patterns, and seasonal variations, their outbreaks have become an important objective for public health organizations to seek strategies and build a better response system as soon as possible.

1.2 Surveillance Data in Disease Prediction

Controlling disease outbreaks requires efficient surveillance systems that can continuously collect and analyse data. These systems can help by providing insights into disease occurrence, trends, and patterns over time. Routine surveillance data are commonly obtained through case notification systems, hospital admissions, laboratory-confirmed diagnoses, emergency department visits, and public health reporting mechanisms.

Surveillance data are deemed highly valuable as they enable health officials to identify and analyse unusual increases in disease incidence. In many cases, it also helps in the early detection of an outbreak in its early stage. In the context of respiratory infections, surveillance data includes laboratory test results, hospitalization records, demographic data, and environmental variables such as temperature and humidity (Al-Tawfiq et al., 2014). Therefore, the availability of such information supports evidence-based decision-making and strengthens the disease monitoring efforts.

1.3 Machine Learning in Public Health

Machine Learning (ML), in general, is a branch of artificial intelligence that enables computer systems to perform various complex calculations. In the health sector, ML is implemented to identify patterns and make predictions based on previous observational data provided by hospitals, research labs, and health programs. Unlike traditional analytical approaches, which rely on predefined assumptions, ML algorithms can learn directly from data. This makes them suitable for analysing large and complex datasets. Moreover, the increase in volume of these datasets has created new opportunities for ML applications in epidemiology and the health sector. Surveillance datasets often contain information collected from multiple sources over long periods, which makes the overall analysis difficult and time-consuming. Implementation of ML allows these datasets to be processed much faster and more efficiently while also identifying old and new patterns, trends, and associations.

ML techniques that are used for outbreak prediction are commonly divided into supervised and unsupervised learning approaches. Between them, supervised learning involves training algorithms using labelled datasets, where the outcomes are previously known regarding outbreak and non-outbreak periods (Kanevsky et al., 2016). On the other hand, unsupervised learning focuses on identifying patterns and groupings within data without relying on predefined labels. Both of these approaches are equally important in modern healthcare management and have gained significant attention in disease surveillance research.

1.4 Research Question and Objectives

This study is based on the following question:

1. How do supervised and unsupervised machine learning methods compare in predicting respiratory disease outbreaks using surveillance data?

To address this question, the following objectives are followed as mentioned below:

- To explain the principles and characteristics of supervised machine learning approaches used in disease outbreak prediction.
- To explain the principles and characteristics of unsupervised machine learning approaches used in disease outbreak prediction.
- To compare the strengths, limitations, and data requirements associated with supervised and unsupervised methods.
- To propose a conceptual framework for selecting appropriate machine learning techniques for outbreak prediction.

1.5 Scope of the Study

This paper adopts a conceptual and comparative approach, drawing on existing literature to derive results. It does not involve direct data collection of original data, like hospital records or patient-level analysis. Instead, it gathers the findings of existing studies that have applied supervised and unsupervised machine learning methods to disease surveillance and outbreak prediction. Primarily, this paper addresses respiratory infectious diseases because they represent a significant public health burden and are frequently examined within machine learning research. By drawing a comparison between both supervised and unsupervised approaches, this study aims to provide insights into their practical applications, advantages, limitations, and suitability under different surveillance conditions.

2. LITERATURE REVIEW

2.1 Epidemiological Foundations of Disease Outbreak Prediction

2.1.1 Disease Outbreaks and Epidemiological Terminology

Epidemiology acts as the cornerstone of public health, providing the theoretical understanding for understanding disease occurrence, transmission, and control within populations. Straif-Bourgeois et al. (2023) state that epidemiology examines the distribution and determinants of health-related events to support disease prevention and public health interventions. Incidence and prevalence are the two most important epidemiological measures, where incidence represents newly occurring cases within a specified period, and prevalence indicates the total disease burden in a population, as seen in Figure 2.1.1.1.

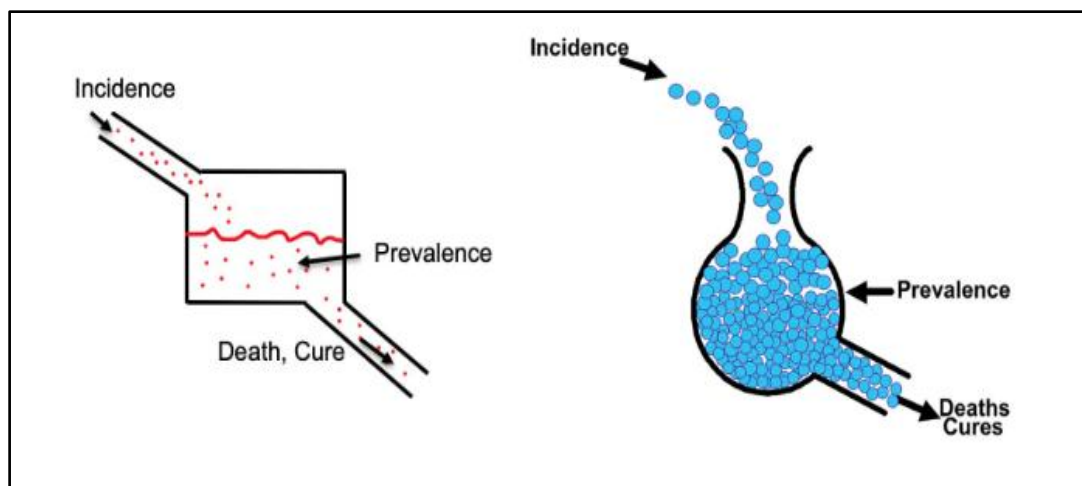


Figure 2.1.1.1: Conceptual Illustration of Disease Frequency Measures

Source: (Wainger, 2026)

Identifying and distinguishing outbreaks from endemic or epidemic conditions is essential for disease surveillance. As Riley (2019) suggests, endemic diseases occur at expected levels within populations, while epidemics involve a substantial

increase beyond normal expectations. To successfully transition from epidemic to endemicity, boosting immunity levels, increasing public health concerns and responses, and pathogen evolution are necessary, as Hadfield (2022) suggests. Hence, in surveillance practice, the outbreak threshold is considered a medium to identify abnormal increases in disease activity, marked as “early warning” systems to alert epidemiologists, as reported by WHO (2008). Additionally, Thayer et al. (2025) found that threshold-based approaches can facilitate earlier detection of disease transmission and improve public health decision-making. Similar findings are also presented by Thomson et al. (2023), as they highlight the importance of epidemiological indicators to mitigate disease burden and pandemic-level disease spread, while also helping with healthcare planning.

2.1.2 Infectious Disease Surveillance Systems

Public health surveillance systems play a critical role in monitoring infectious diseases and supporting outbreak preparedness by allowing authorities to map disease trends, detect outbreaks, and guide timely intervention. Buehler et al. (2019) define surveillance as the continuous collection, analysis, interpretation, and dissemination of health data for public health action. Figure 2.1.2.1 maps out these specific core functions, demonstrating how data collection flows systematically into core public health response mechanisms. Murray & Cohen (2017) also found that surveillance systems typically integrate information from case notifications, hospital admissions, laboratory reports, and mortality records to identify unusual disease patterns.

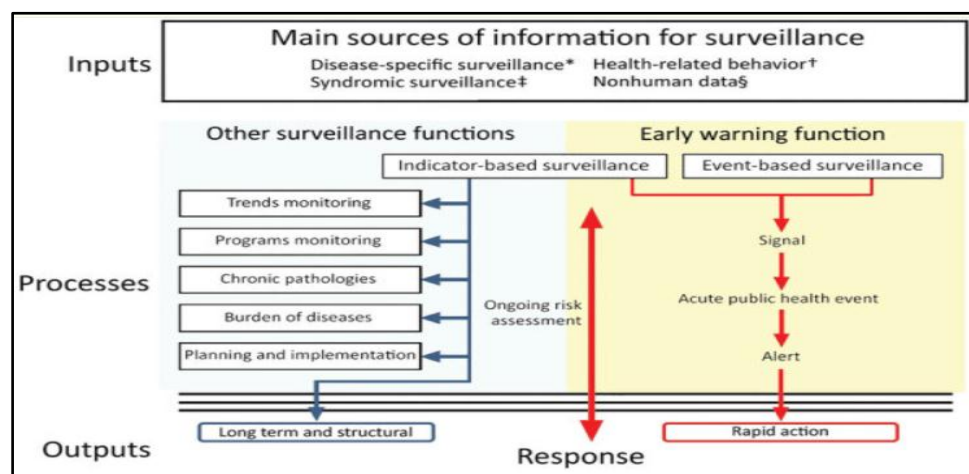


Figure 2.1.2.1: Public Health Surveillance and Response Functions Overview

Source: (Ricks et al., 2022)

Among these, respiratory disease surveillance has gained increasing importance because respiratory pathogens can spread rapidly across communities. In this matter, Al-Tawfiq et al. (2014) suggest that surveillance for emerging respiratory viruses is necessary for timely intervention and effective outbreak management. Similarly, Maleki et al. (2023) reported that continuous monitoring of respiratory viral infections provides valuable insights into transmission dynamics and changing disease patterns. The World Health Organization (WHO, 2024) reported that the Global Influenza Surveillance and Response System (GISRS) facilitates influenza monitoring, virus characterization, and pandemic preparedness, allowing for fewer life risks.

2.1.3 Surveillance Data for Respiratory Disease Prediction

Surveillance datasets contain real-life primary sources of information regarding outbreak prediction models. These sources can include weekly case counts, laboratory-confirmed infections, hospital admissions, demographic variables, and environmental indicators such as temperature and humidity, which could be used for disease outbreak prediction. Gao et al. (2025) further found that incidence-based surveillance data contain meaningful temporal patterns that can be exploited using machine learning approaches for early outbreak detection.

INFLUENZA	SARS-CoV-2 (COVID-19)	RSV
Epidemiology and Global Burden  3–5 million severe cases/year	 700+ million confirmed cases; >6.3 million deaths (as of 2024)	 Millions of hospitalizations in children <5 and elderly annually
Seasonality  Strong winter seasonality	 Strong winter seasonality	 Emerging winter pattern post-2021
 Acute illness, severe in elderly/comorbid	 Acute illness, severe in elderly/comorbid	 Severe in infants, elderly, and immunocompromised
 Annual vaccines; antigenic drift/shift requires updates	 Multiple vaccines; efficacy wanes; boosters needed	 No universal vaccine yet; monoclonal antibodies and candidate vaccines in nat
 Constant antigenic changes; annual reformulation	 Rapid variant evolution; immune escape; vaccine fatigue	 Limited prophylactic options; underdiagnosed burden

Figure 2.1.3.1: Comparative Seasonal Transmission Patterns and Global Impact of Dominant Respiratory Viruses

Source: (Shuaibu et al., 2025)

Acute identification and distinction between baseline disease activity and actual outbreak is particularly important because respiratory infections, like influenza or RSV, often overlap seasonal fluctuations and could also resemble outbreak conditions (Figure 2.1.3.1). AbdelGawad et al. (2020) evaluated several methods, including CSEM, the Proposed WHO method, and the Moving Epidemic Method (MEM), for determining influenza baselines and thresholds and found that surveillance-derived thresholds can improve the detection of unusual disease activity. From these discussions, it is certain that reliable surveillance systems and robust epidemiological indicators provide the foundation for developing predictive frameworks and provide ground-level data to help with identifying respiratory disease outbreaks.

2.2 Supervised Machine Learning in Outbreak Predictions

2.2.1 Principles of Supervised Learning

In modern and advanced healthcare, supervised machine learning is one of the most widely adopted AI approaches for disease prediction because it has repeatedly been trained using historical data with known outcomes. Al Meslamani et al. (2024) state that supervised learning algorithms establish relationships between input variables and predefined labels, allowing models to predict future outbreak events based on previously observed patterns. This enables distinguishing between common outbreak and non-outbreak periods, allowing models to classify future observations accordingly. As Kanevsky et al. (2016) illustrated in Figure 2.2.1.1, supervised machine learning consists of a multi-step pipeline spanning from initial data training and validation to the classification or prediction of novel outcomes, through which healthcare authorities can make new predictions related to outbreaks.

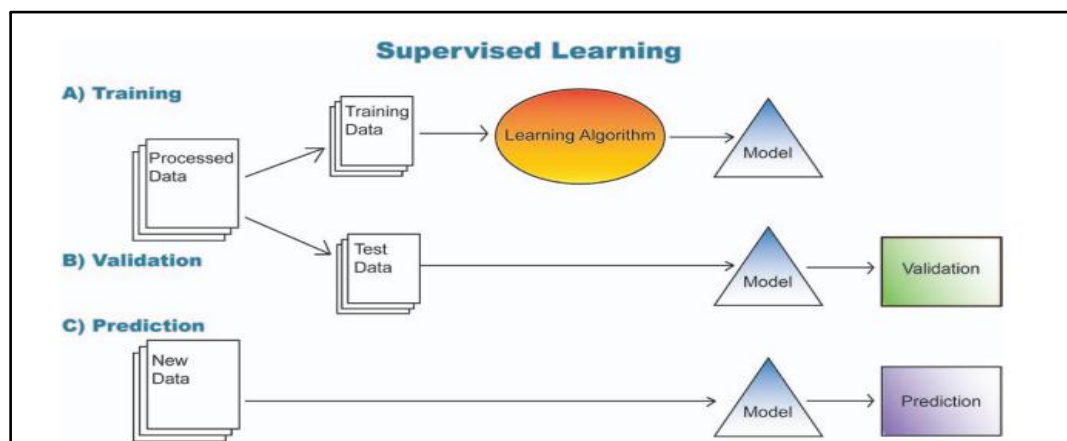


Figure 2.2.1.1: Multi-Step Pipeline of Supervised Machine Learning

Source: (Kanevsky et al., 2016)

Regarding the importance of this method, Zhang et al. (2024) suggest that supervised learning has become increasingly valuable for outbreak risk prediction due to its ability to integrate diverse epidemiological, environmental, and demographic variables into a single predictive framework. Similar views are also presented by Potnis & Tiple (2023), as they discussed how supervised learning techniques perform particularly well when trained using high-quality, well-labelled datasets, making them a cornerstone in modern healthcare applications and better outcomes. However, the quality of data used for the supervised training is addressed by Al Meslamani et al. (2024), as they highlight the interconnected process of data collection, preprocessing, feature selection, model training, validation, and performance evaluation. Hence, the role of data quality plays a significant role in determining the predictive capability of supervised models.

2.2.2 Common Supervised Algorithms in Epidemiology

There are several supervised learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and hidden Markov models, applied to infectious disease surveillance, and each of them offers prominent advantages based on the prediction tasks (Figure 2.2.2.1).

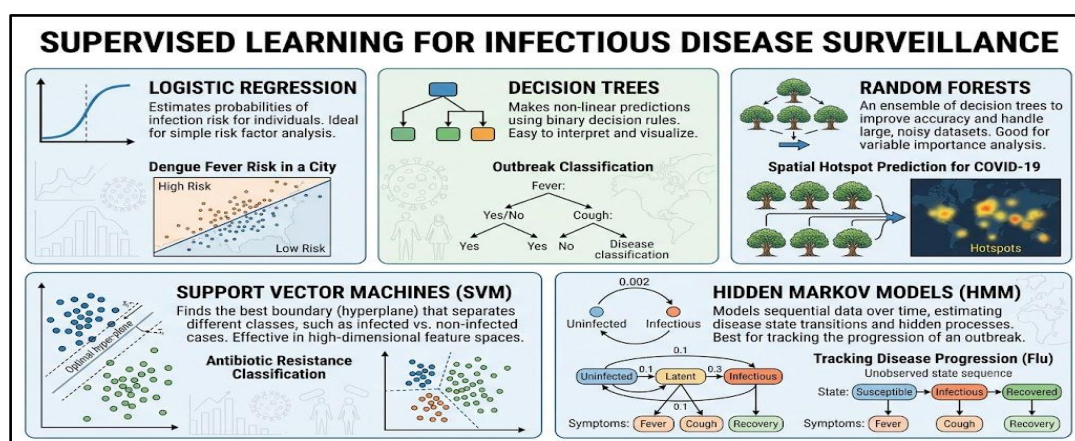


Figure 2.2.2.1: Common Supervised Algorithms

Source: (Self-Created)

Logistic regression is one of the simplest and most interpretable approaches for disease prediction. Nopour et al. (2022) found that logistic regression effectively estimates disease probabilities by modelling relationships between multiple clinical and epidemiological variables, making it suitable for binary classification problems such as outbreak and non-outbreak prediction. A decision tree classifier uses a recursive binary splitting approach to break down complex datasets into smaller and more uniform subsets. Regarding this, Battista et al. (2023) suggest that decision trees are particularly useful in public health surveillance because their graphical structure allows health professionals to interpret prediction pathways with more ease. Random forests further combat the high variance of individual models and create multiple subsets of the training data that learn from a slightly different set of observations. Aawar & Srivastava (2025) used Random Forest as an ensemble technique to combine predictions from multiple mechanistic models to maintain robustness and automated accuracy by bridging interpretable domain knowledge with data-driven machine learning, rather than relying on the forest to parse raw epidemiological interactions from scratch.

Support Vector Machines (SVMs) have also demonstrated strong predictive performance in healthcare applications. As Guido et al. (2024) state, SVMs are effective in analysing high-dimensional health datasets, and G et al. (2025) suggest that they perform particularly well in large-scale public health data where complex decision boundaries are required. Additionally, Hidden Markov Models (HMMs) are also used in epidemiology and bioinformatics to map the hidden states of disease on observable clinical symptoms or temporal disease patterns. Van Kessel et al. (2025) found that Poisson Hidden Markov Models successfully detected hospital outbreaks using routine surveillance data, whereas Zhou et al. (2021) demonstrated that HMMs can also analyse and visualize spatio-temporal COVID-19 transmission patterns.

2.2.3 Applications in Respiratory Disease Prediction

As healthcare and disease outbreak management integrate modern technologies, especially AI, the application of supervised learning to respiratory disease surveillance has expanded considerably in recent years. Cheng et al. (2025) reported that supervised machine learning algorithms consistently achieve high predictive performance for COVID-19 forecasting, especially after being trained using comprehensive epidemiological datasets. Likewise, G et al. (2023) reported that machine learning and deep learning models improve epidemic prediction by learning complex temporal relationships from COVID-19 surveillance data.

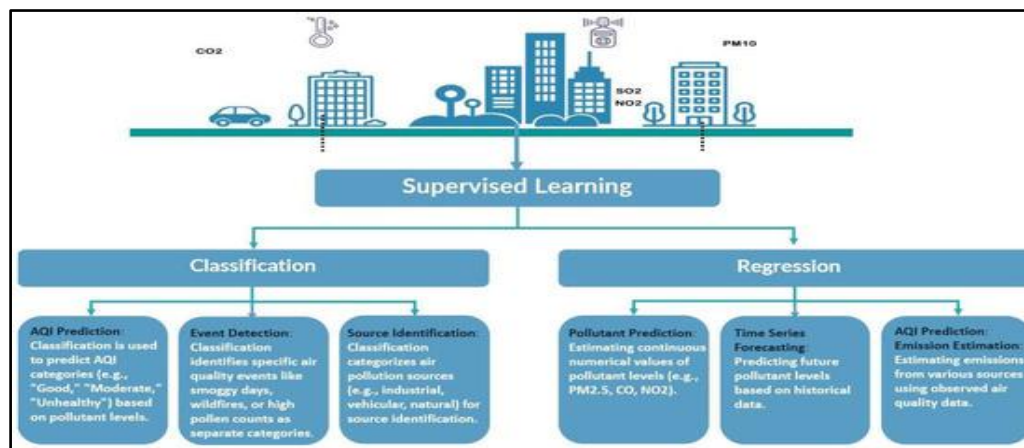


Figure 2.2.3.1: Supervised Learning Approaches in Air Quality Monitoring

Source: (Essamlali et al., 2024)

It is especially useful in respiratory disease prediction because different environmental factors are also being incorporated into supervised prediction models (Figure 2.2.3.1). For example, as Ku et al. (2022) found, climatic conditions and air pollution variables significantly improve the prediction of respiratory disease occurrence when combined with machine learning algorithms. Algarni (2024) further suggests that supervised learning techniques can support the early detection of respiratory diseases by analysing multiple patient and environmental characteristics simultaneously. While all these indicate that supervised learning offers reliable outbreak prediction when labelled surveillance datasets are available, their effectiveness depends heavily on the availability of accurate historical labels and high-quality surveillance data, without which predicted outcomes might not be achieved.

2.3 Unsupervised Machine Learning in Disease Surveillance

2.3.1 Principles of Unsupervised Learning

Unsupervised machine learning analyses datasets without predefined labels, enabling the discovery of hidden structures and relationships by grouping data points through their resemblance in characteristics, which is also known as a "clustering algorithm" (Figure 2.3.1.1). Several researchers, such as Wang et al. (2020) and Lan & Delmelle (2023), state that unsupervised ML systems have an effective impact in identifying latent disease clusters and patient subgroups from the given healthcare data. Moreover, it is highly valuable in disease surveillance when outbreak labels are unavailable or incomplete, making it suitable for exploratory epidemiological analyses.

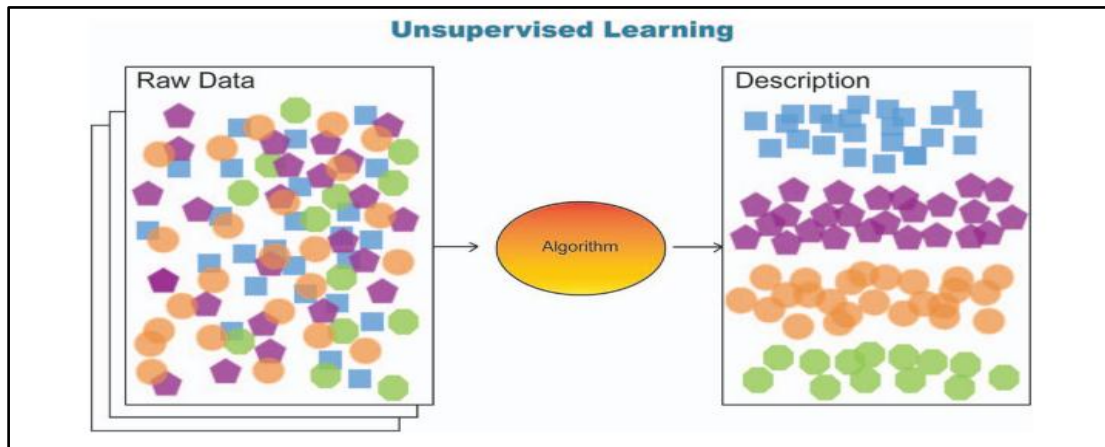


Figure 2.3.1.1: Unsupervised Machine Learning

Source: (Kanevsky et al., 2016)

2.3.2 Clustering Techniques Used in Epidemiology

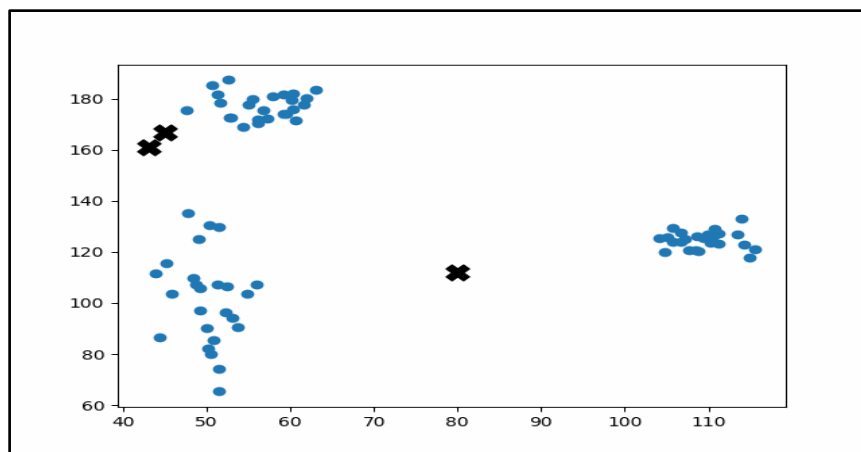


Figure 2.3.2.1: K-Means Clustering Partitioning a Dataset into Three Distinct, Globular Patient Cohorts

Source: (Schwartz, 2023)

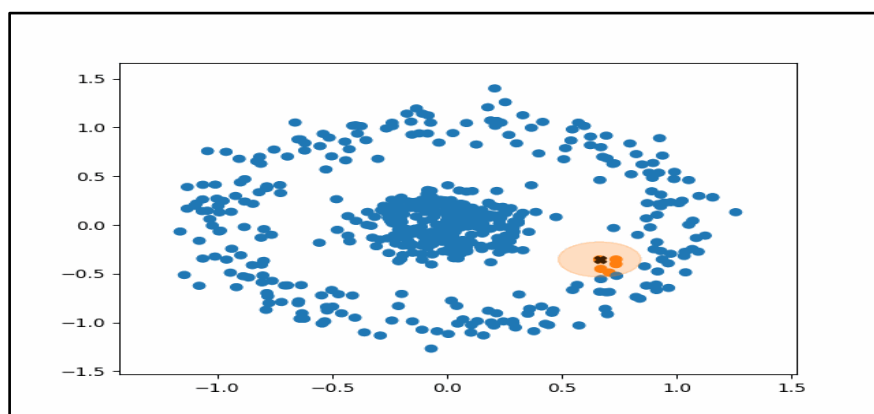


Figure 2.3.2.2: DBSCAN Algorithm Identifying an Irregularly Shaped Spatial Cluster while Filtering out Noise Points

Source: (Schwartz, 2023)

Clustering algorithms are the most commonly utilized method in unsupervised ML systems in epidemiology. According to Garcia-Vidal et al. (2024), “k-means clustering” can identify clinically distinct COVID-19 patient groups, as illustrated in Figure 2.3.2.1, whereas Kuo et al. (2018) demonstrated that DBSCAN can also effectively detect spatial disease clusters. Unlike distance-based methods, DBSCAN is uniquely suited for geographic or non-globular tracking because it maps dense, arbitrary shapes while also categorizing background anomalies as unclustered noise points (Figure 2.3.2.1). Moreover, further studies conducted by Que & Tsui (2008), Jung et al. (2020), Zhang et al. (2022), and Li et al. (2021) suggest that spatial, temporal, and hierarchical clustering methods support the identification of outbreak patterns and transmission trends.

2.3.3 Unsupervised Discovery of Disease Patterns

Apart from clustering datasets, unsupervised ML systems also assist in the identification of disease subgroups and outbreak-prone regions. Bej et al. (2022) found that clustering reveals meaningful disease sub-populations, while Kumar & Gidisu (2025) also reported on the effectiveness in identifying epidemiological hotspots for targeted interventions. Moreover, Amin & Vellinga (2026) further suggested that unsupervised clustering can detect respiratory tract infection patterns. Therefore, these findings demonstrate the overall usefulness of unsupervised ML systems when labelled surveillance data is highly scarce.

2.4 Evidence from Supervised Learning Studies

2.4.1 Salmonella and Campylobacter Outbreak Detection

Zacher & Czogiel (2022) appraised supervised ML models that utilise surveillance datasets to detect disease patterns, such as Salmonella and Campylobacter outbreaks in Germany. They found that these models have an improved outbreak detection system in comparison to traditional methods, with higher accuracy and lower false alarm rates. This study suggests that labelled surveillance data can enable algorithms to identify outbreak signals earlier. Additionally, it also supports timely public health interventions.

2.4.2 Influenza Forecasting Model

Several researchers, such as Tapak et al. (2019), compared different forecasting models that utilise surveillance data of influenza-like illness from healthcare centres across Iran. These authors found that machine learning-based models provide reliable predictions of influenza activity and support the detection of seasonal outbreaks at an early stage. This further solidifies that the effectiveness of these models depends largely on the quality and continuity of surveillance data.

2.4.3 Nosocomial Infection Risk Prediction

Researchers such as Chen et al. (2024) further examined the application of supervised machine learning methods to predict hospital-acquired infections through surveillance data. The authors’ findings suggest that algorithms like Random Forest and Support Vector Machine can accurately identify patients who are at high risk of infection by analysing demographic and clinical variables. This further suggests that supervised learning can implement strategies to detect and prevent infections. This also allows healthcare professionals to implement timely and targeted interventions. Therefore, these findings suggest that through supervised ML systems, it is possible to achieve accurate and reliable outbreak predictions. However, it is highly dependent on high-quality labelled surveillance data, which is also directly proportional to the value in infectious disease surveillance and public health decision-making.

2.5 Evidence from Unsupervised Learning Studies

2.5.1 Respiratory Tract Infection Clustering

Amin (2024) conducted an investigation on the application of unsupervised machine learning for predicting respiratory tract infection (RTI) outbreaks in acute Irish hospitals using surveillance data collected between 2016 and 2021. The study findings suggest that clustering techniques were successful in grouping hospital admission patterns into distinct risk categories without requiring predefined outbreak labels. The identified clusters highlighted periods with increased RTI

activity, which demonstrates that unsupervised learning can support early outbreak recognition. Additionally, they provide further assistance to healthcare institutions in prioritizing surveillance and resource allocation.

2.5.2 COVID-19 Country Clustering

Rizvi et al. (2021) applied clustering techniques to classify countries according to COVID-19 prevalence, healthcare capacity, and environmental indicators. Their study found that countries with similar epidemiological and socioeconomic characteristics formed distinct clusters, which enabled comparisons of disease transmission patterns across vast regions. The authors further suggest that country-level clustering can further help public health authorities to identify the shared risk factors and develop region-specific intervention strategies during large-scale infectious disease outbreaks.

2.5.3 Geospatial Hotspot Detection

Researchers such as Fatima et al. (2025) explored the spatial distribution of five major respiratory diseases in South Punjab, Pakistan, using geospatial clustering techniques. The study established significant disease hotspots that are consistent with higher infection rates and further distinguished high-risk areas from regions with lower burden rates. The findings of the study suggest that geospatial hotspot detection supports targeted surveillance, with efficient allocation of healthcare resources, and evidence-based public health interventions where detailed, labelled outbreak data are limited. Unsupervised ML systems are effective in identifying hidden disease patterns, spatial hotspots, and high-risk clusters without relying on curated datasets. The lack of human interpretation may lead to false categorization of data into clusters, which will require valuable human labour to make the datasets usable.

2.6 Literature Gap

Despite the increase in use of ML systems for the detection of infectious disease outbreaks, important research gaps remain. Both Bojja (2025) and Sogandi (2024) reported that most studies examine supervised and unsupervised learning separately, with few providing direct comparisons under similar settings, and there is limited evidence on how both approaches can be integrated successfully for outbreak prediction. Furthermore, Hamilton & Kelley (2021) and Wiemken & Kelly (2020) noted that supervised models dominate the literature despite their reliance on labelled data. Other researchers, such as Rudin et al. (2022) and Allen et al. (2023), emphasized the need for more interpretable models for public health use. Therefore, this study provides a conceptual comparison of both approaches for respiratory disease outbreak prediction using surveillance data.

3. METHODOLOGY

3.1 Research Design

This study adopts a qualitative, conceptual research design that is based on a comparative review of published literature. This research synthesizes existing evidence to evaluate the effectiveness of supervised and unsupervised ML systems to predict respiratory disease outbreaks, and is not based on empirical studies, which require primary data collection and model development. The implementation of a comparative review allows the study to identify the differences in data requirements, predictive capabilities, interpretability, and practical applicability across both learning approaches. Furthermore, this study also relies heavily on peer-reviewed journal articles, conference papers, and reports from recognized public health organisations such as the WHO and AIMS. Hence, no original surveillance datasets, patient records, or experimental ML models were used. Alternatively, findings from previous studies are integrated to develop a conceptual framework for comparing supervised and unsupervised approaches in disease outbreak prediction.

3.2 Surveillance Data and Prediction Variables

ML models for disease outbreak predictions are heavily trained using routinely collected surveillance data from both healthcare and public health systems, with the addition of heterogeneous datasets, including environmental, social, and genomic sequence patterns in modern times, to improve forecasting accuracy (Villanueva-Miranda et al., 2025). Generally, the training data involves weekly respiratory case counts, hospital admissions, laboratory-confirmed infections, demographic information such as age and sex, and environmental variables including temperature, humidity, and seasonal trends. All of these factors are interconnected, creating a pattern of disease occurrence and transmission.

Especially in supervised learning, such surveillance records are labelled according to predefined outbreak thresholds to help with the pattern creation and enable distinguishing between outbreak and non-outbreak periods. However, it is not required in unsupervised learning, as it allows algorithms to discover naturally occurring clusters and hidden patterns within the data, without labelling them.

3.3 Conceptual Framework for Supervised Learning and Unsupervised Learning

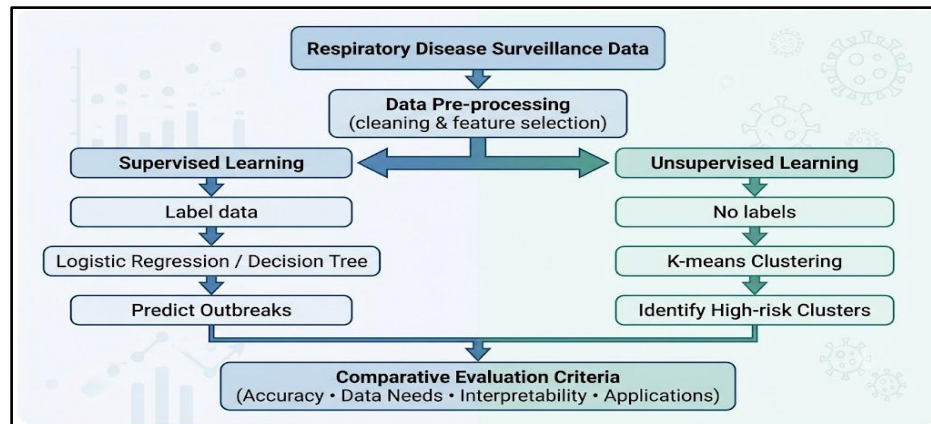


Figure 3.3.1: Conceptual Framework for Comparing Supervised and Unsupervised Machine Learning Approaches in Respiratory Disease Outbreak Prediction

Source: (Self-Created)

The above Figure 3.3.1 highlights the conceptual framework used in this paper to compare supervised and unsupervised ML approaches. It begins with the root, i.e., routine respiratory disease surveillance data, including case reports, hospital admissions, laboratory-confirmed cases, and many more, which are processed through data cleaning to remove inconsistencies and select only the relevant ones for analysis. Then, the workflow diverges into two ML approaches, including supervised and unsupervised. In the supervised learning pathway, labelled surveillance data are used to train algorithms such as Logistic Regression and Decision Trees to predict future outbreak occurrences. Comparatively, the unsupervised learning pathway analyses unlabelled data using K-means clustering to identify groups of observations with similar characteristics, enabling high-risk cluster detection and potential outbreak-prone patterns. In the following chapter, both approaches will be evaluated based on their common ground, i.e., prediction accuracy, data requirements, interpretability, and practical applications, so a thorough comparative discussion can be presented.

4. COMPARITIVE ANALYSIS AND DISCUSSION

4.1 Comparative Evaluation Criteria

To compare both machine learning approaches, this study utilises four evaluation criteria, which are derived from the reviewed literature: data requirements, predictive performance, interpretability, and practical applications, as presented in Table 4.1. These criteria were particularly selected as they reflect the key characteristics that influence the effectiveness of both ML models during outbreak forecasting. The following sections discuss each criterion to evaluate the strengths, limitations, and practical applicability of both approaches in respiratory disease surveillance.

Table 4.1: Comparison of Supervised and Unsupervised Machine Learning Approaches

Criterion	Supervised Learning	Unsupervised Learning
Data requirement	Labelled surveillance data required	No labelled data required
Predictive performance	Predict disease outbreaks when reliable labels are available	Identify hidden patterns and clusters rather than making

		direct predictions
Interpretability	Outputs are generally easier to interpret, but complexities remain	Cluster results are intuitive but require further interpretation
Practical applications	Best suited for outbreak prediction, forecasting, and risk classification	Best suited for hotspot detection, pattern discovery, and exploratory surveillance

4.2 Data Requirements

The overall effectiveness of ML models depends largely on the quality and the availability of surveillance data. In supervised learning, it is required to categorise datasets in which historical records are classified as outbreak or non-outbreak events. Zacher & Czogiel (2022) found that the accuracy in labelling can significantly improve outbreak detection performance, while Al Meslamani et al. (2024) further suggested that the absence of reliable labels can reduce the model's overall effectiveness. In contrast, unsupervised ML systems do not require predefined labels and can naturally identify groupings within the data. Supporting this, Wang et al. (2020) further stated that this characteristic makes unsupervised models particularly useful in surveillance systems where outbreak labels are incomplete or unavailable. Therefore, supervised learning is more suitable for mature surveillance systems, whereas unsupervised learning is advantageous in resource-constrained settings.

4.3 Prediction Accuracy and Performance

Many researchers have admitted that supervised learning approaches demonstrate superior predictive performance due to being trained using historical outbreak information on labels. For example, Cheng et al. (2025) found that supervised algorithms achieve high prediction accuracy in COVID-19 forecasting, while Aawar & Srivastava (2025) reported that "Random Forest models" effectively forecast influenza outbreaks. Similarly, Chen et al. (2024) demonstrated that supervised models accurately predict nosocomial infection risks using surveillance data. However, unsupervised learning methods are not suitable for direct outbreak prediction, as researchers like Garcia-Vidal et al. (2024) and Amin (2024) addressed; while clustering techniques successfully identify high-risk groups and outbreak-prone periods, their outputs are exploratory rather than predictive. Hence, according to the previous information, supervised learning offers better forecasting capabilities, whereas unsupervised ones provide complementary insights into the disease patterns.

4.4 Interpretability and Practical Application

Interpretability of data is also an important consideration in public health decision-making because health professionals also rely on model outputs before implementing any intervention. According to Rudin et al. (2022), interpretable ML systems can improve transparency and trust in healthcare applications. For instance, models such as Logistic Regression and Decision Trees provide clear explanations of prediction outcomes, whereas more complex algorithms, including Random Forests and Support Vector Machines, are often less transparent despite their higher predictive performance (Guido et al., 2024). Moreover, unsupervised ML systems produce clusters that are visually intuitive and also useful for identifying diseased hotspots. However, these clusters often require additional epidemiological interpretation before any policy decision can be made (Lan & Delmelle, 2023).

4.5 Best Practical Application of Both Approaches

Neither supervised nor unsupervised learning is universally superior, as their suitability is conditional and is highly dependent on surveillance objectives and the availability of data. Supervised learning is most appropriate when labelled datasets are available and accurate outbreak forecasting is required. For instance, influenza prediction or hospital infection risk assessment. Conversely, unsupervised learning is better suited for exploratory analyses, hotspot detection, and identifying hidden disease patterns in datasets with limited labels (Amin, 2024). Bojja (2025) argues that a combination of both approaches can improve overall surveillance by implementing unsupervised learning to discover meaningful patterns and supervised learning to generate accurate outbreak predictions. Consequently, such hybrid strategies may provide a more comprehensive framework for supporting public health surveillance and disease outbreak management.

5. LIMITATIONS AND FUTURE WORK

5.1 Limitations

This study is subject to several limitations. First, it uses a conceptual and literature-based approach. The findings are based on evidence reported in previous literature rather than any empirical validation. Therefore, it does not include any primary data collection or implementation of machine learning models that use real-world surveillance datasets. Secondly, the review focuses primarily on respiratory disease outbreaks, which limits the general usability of the findings to other existing diseases with different transmission patterns. Additionally, surveillance data are often affected by under-reporting, missing values, reporting delays, and inconsistencies, which also influence the overall performance of both machine learning models.

5.2 Future Work

Before conducting further studies, future research should validate the proposed conceptual framework using openly available respiratory disease surveillance datasets, such as influenza or COVID-19. Implementing comparative experimental studies will allow future researchers to evaluate both learning algorithms under identical conditions. It would provide a stronger understanding of their performance. Furthermore, implementing explainable artificial intelligence (XAI) techniques may improve model transparency and also increase the confidence of health professionals in ML-based decision-making. And finally, combining both learning approaches within a hybrid surveillance framework will also provide a promising direction in the enhancement of early outbreak detection and public health preparedness.

6. CONCLUSION

Machine learning in modern healthcare and management has emerged as a valuable asset by strengthening disease outbreak prediction and saving millions of lives using large and complete surveillance datasets. This study draws on the comparison between supervised and unsupervised machine learning approaches for predicting respiratory disease outbreaks using evidence from published research of the last 20 years. The thorough review highlights that supervised learning generally provides higher predictive accuracy with the help of reliable labelled surveillance data, which makes it suitable for outbreak forecasting and risk assessment. Unsupervised learning is equally effective, even if it is not suitable for direct outbreak prediction. This learning approach is used to identify hidden patterns, disease clusters, and geographical hotspots, even without labelled data. However, both approaches pose limitations, as their effectiveness depends on surveillance data quality, public health objectives, and available resources. Hence, it is better to use a combination of both of these methods to gain a more accurate prediction. Nonetheless, ML has achieved great success in improving disease surveillance and supporting timely public health decision-making, with continuous training to expect far greater results in the future.

REFERENCES

1. Aawar, M. A., & Srivastava, A. (2025). Random Forest of epidemiological models for Influenza forecasting. *Epidemics*, 53, 100862. <https://doi.org/10.1016/j.epidem.2025.100862>
2. AbdelGawad, B., Vega, T., El Houssinie, M., Mohsen, A., Fahim, M., Abu ElSood, H., Jabbour, J., Eid, A., & Refaey, S. (2020). Evaluating tools to define influenza baseline and threshold values using surveillance data, Egypt, season 2016/17. *Journal of Infection and Public Health*, 13(3), 430-437. <https://doi.org/10.1016/j.jiph.2019.05.005>
3. Al Meslamani, A. Z., Sobrino, I., & de la Fuente, J. (2024). Machine learning in infectious diseases: potential applications and limitations. *Annals of medicine*, 56(1), 2362869. <https://doi.org/10.1080/07853890.2024.2362869>
4. Algarni, A. (2024). Smart detection: Using supervised machine learning for respiratory diseases. *Advances and Applications in Statistics*, 91(12), 1607-1625. <https://doi.org/10.17654/0972361724082>
5. Allen, G. I., Gan, L., & Zheng, L. (2023). Interpretable machine learning for discovery: Statistical challenges and opportunities. *Annual Review of Statistics and Its Application*, 11. <https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-040120-030919>

6. Al-Tawfiq, J. A., Zumla, A., Gautret, P., Gray, G. C., Hui, D. S., Al-Rabeeah, A. A., & Memish, Z. A. (2014). Surveillance for emerging respiratory viruses. *The Lancet. Infectious diseases*, 14(10), 992–1000. [https://doi.org/10.1016/S1473-3099\(14\)70840-0](https://doi.org/10.1016/S1473-3099(14)70840-0)
7. Amin, D. (2024). Utilising unsupervised machine learning to predict outbreaks of respiratory tract infections in acute Irish hospitals (2016-2021). *Journal of Public Health and Epidemiology*, 16(2), 112–124. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12925592/>
8. Amin, D., & Vellinga, A. (2026). Utilising Unsupervised Machine Learning to Predict Outbreaks of Respiratory Tract Infections in Acute Irish Hospitals (2016-2021). *Public Health in Practice*, 100748. <https://www.sciencedirect.com/science/article/pii/S2666535226000273>
9. Battista, K., Diao, L., Patte, K. A., Dubin, J. A., & Leatherdale, S. T. (2023). Examining the use of decision trees in population health surveillance research: an application to youth mental health survey data in the COMPASS study. *Utilisation des arbres décisionnels dans la recherche en surveillance de la santé de la population : application aux données d'enquête sur la santé mentale des jeunes de l'étude COMPASS. Health promotion and chronic disease prevention in Canada : research, policy and practice*, 43(2), 73–86. <https://doi.org/10.24095/hpcdp.43.2.03>
10. Bej, S., Sarkar, J., Biswas, S., Mitra, P., Chakrabarti, P., & Wolkenhauer, O. (2022). Identification and epidemiological characterization of type-2 diabetes sub-population using an unsupervised machine learning approach. *Nutrition & Diabetes*, 12(1), 27. <https://www.nature.com/articles/s41387-022-00206-2>
11. Bojja, R. R. (2025). A comparative study of supervised and unsupervised learning approaches. *International Multidisciplinary Research Journal Reviews (IMRJR)*, 2, 80-88. https://www.researchgate.net/profile/Revanth-Bojja-2/publication/390798656_A_Comparative_Study_of_Supervised_and_Unsupervised_Learning_Approaches/links/68012e0d60241d51400d85e8/A-Comparative-Study-of-Supervised-and-Unsupervised-Learning-Approaches.pdf
12. Chen, Y., Zhang, Y., Nie, S., Ning, J., Wang, Q., Yuan, H., Wu, H., Li, B., Hu, W., & Wu, C. (2024). Risk assessment and prediction of nosocomial infections based on surveillance data using machine learning methods. *BMC public health*, 24(1), 1780. <https://doi.org/10.1186/s12889-024-19096-3>
13. Cheng, Y., Cheng, R., Xu, T., Tan, X., & Bai, Y. (2025). Machine Learning Techniques Applied to COVID-19 Prediction: A Systematic Literature Review. *Bioengineering (Basel, Switzerland)*, 12(5), 514. <https://doi.org/10.3390/bioengineering12050514>
14. Dairi, A., Harrou, F., Zeroual, A., Hittawe, M. M., & Sun, Y. (2021). Comparative study of machine learning methods for COVID-19 transmission forecasting. *Journal of biomedical informatics*, 118, 103791. <https://www.sciencedirect.com/science/article/pii/S1532046421001209>
15. Essamlali, I., Nhaila, H., & El Khaili, M. (2024). Supervised Machine Learning Approaches for Predicting Key Pollutants and for the Sustainable Enhancement of Urban Air Quality: A Systematic Review. *Sustainability*, 16(3), 976. <https://doi.org/10.3390/su16030976>
16. Fatima, M., Butt, I., Firouraghi, N., & Kiani, B. (2025). Exploring geospatial clusters of five primary respiratory diseases in South Punjab, Pakistan: An epidemiological examination. *Geography, Environment, Sustainability*, 18(1), 14-22. <https://ges.rgo.ru/jour/article/view/3986>
17. Fourreau, V. (2021, February 23). The Countries With the Highest COVID-19 Death Toll. Retrieved from Statista Daily Data website: https://www.statista.com/chart/24258/countries-with-the-highest-number-of-covid-19-deaths/?srsltid=afmboop7xawocaxhso0iuk55mgbqxx-hke_ihnzholztecyrlyza135
18. G, K., K P, I., Hasin A, J., M, L. F. J., Siluvai, S., & G, K. (2025). Support Vector Machines: A Literature Review on Their Application in Analyzing Mass Data for Public Health. *Cureus*, 17(1), e77169. <https://doi.org/10.7759/cureus.77169>
19. G., M., V, M., M, M., V, S., Luhach, A. Kr., & Kumar, S. (2023). Epidemic Prediction using Machine Learning and Deep Learning Models on COVID-19 Data. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(3), 377–393. <https://doi.org/10.1080/0952813X.2022.2058618>
20. Gao, S., Chakraborty, A. K., Greiner, R., Lewis, M. A., & Wang, H. (2025). Early detection of disease outbreaks and non-outbreaks using incidence data: A framework using feature-based time series classification and machine learning. *PLoS computational biology*, 21(2), e1012782. <https://doi.org/10.1371/journal.pcbi.1012782>
21. Garcia-Vidal, C., Teijón-Lumbreras, C., Aiello, T. F., Chumbita, M., Menendez, R., Mateu-Subirà, A., Peyrony, O., Monzó, P., Lopera, C., Gallardo-Pizarro, A., Méndez, R., Calbo, E., Xercavins, M., Cuesta-Chasco, G., Martínez, J. A.,

- Marcos, M. A., Mensa, J., & Soriano, A. (2024). K-Means Clustering Identifies Diverse Clinical Phenotypes in COVID-19 Patients: Implications for Mortality Risks and Remdesivir Impact. *Infectious diseases and therapy*, 13(4), 715–726. <https://doi.org/10.1007/s40121-024-00938-xn>
22. Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*, 15(4), 235. <https://doi.org/10.3390/info15040235>
23. Hadfield, A. (2022). From pandemic to endemic? Learning lessons from a global contagion. *Journal of Chinese Economic and Business Studies*, 20(3), 225-233. <https://www.tandfonline.com/doi/abs/10.1080/14765284.2021.2024486>
24. Hamilton, A. J., Strauss, A. T., Martinez, D. A., Hinson, J. S., Levin, S., Lin, G., & Klein, E. Y. (2021). Machine learning and artificial intelligence: applications in healthcare epidemiology. *Antimicrobial Stewardship & Healthcare Epidemiology*, 1(1), e28. <https://www.cambridge.org/core/journals/antimicrobial-stewardship-and-healthcare-epidemiology/article/machine-learning-and-artificial-intelligence-applications-in-healthcare-epidemiology/60FFBF2D360CF70446AC5B4A972B299A>
25. Jalali, N., Sahu, K. S., Oetomo, A., & Morita, P. P. (2020). Understanding user behavior through the use of unsupervised anomaly detection: proof of concept using internet of things smart home thermostat data for improving public health surveillance. *JMIR mHealth and uHealth*, 8(11), e21209. <https://mhealth.jmir.org/2020/11/e21209/>
26. James W. Buehler, M.D.,¹ Richard S. Hopkins, M.D.,² J. Marc Overhage, M.D.,³ Daniel M. Sosin, M.D.,² Van Tong, M.P.H.2. (2019). Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks. Retrieved from <https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5305a1.htm>
27. Jung, S., Moon, J., & Hwang, E. (2020). Cluster-Based Analysis of Infectious Disease Occurrences Using Tensor Decomposition: A Case Study of South Korea. *International journal of environmental research and public health*, 17(13), 4872. <https://doi.org/10.3390/ijerph17134872n>
28. Kanevsky, J., Corban, J., Gaster, R., Kanevsky, A., Lin, S., & Gilardino, M. (2016). Big data and machine learning in plastic surgery: a new frontier in surgical innovation. *Plastic and reconstructive surgery*, 137(5), 890e-897e. https://journals.lww.com/plasreconsurg/FullText/2016/05000/Big_Data_and_Machine_Learning_in_Plastic_Surgery_45.aspx
29. Ku, Y., Kwon, S. B., Yoon, J. H., Mun, S. K., & Chang, M. (2022). Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors. *Clinical and experimental otorhinolaryngology*, 15(2), 168–176. <https://doi.org/10.21053/ceo.2021.01536>
30. Kumar, A. D., & Gidisu, J. A. (2025). Unsupervised Machine Learning and Cluster Analysis of Epidemiological Outbreaks in Ghana: Identifying Spatial Hotspots for Targeted Interventions. *Brainae Journal of Business, Sciences and Technology*, 9(3), 529-539. <https://brainajournal.com/manuscripts/10.%20DK%202025.pdf>
31. Kuo, F. Y., Wen, T. H., & Sabel, C. E. (2018). Characterizing diffusion dynamics of disease clustering: a modified space–time DBSCAN (MST-DBSCAN) algorithm. *Annals of the American Association of Geographers*, 108(4), 1168–1186. <https://www.tandfonline.com/doi/abs/10.1080/24694452.2017.1407630>
32. Lan, Y., & Delmelle, E. (2023). Space-time cluster detection techniques for infectious diseases: A systematic review. *Spatial and Spatio-Temporal Epidemiology*, 44, 100563. <https://doi.org/10.1016/j.sste.2022.100563n>
33. Li, X., Chen, D., Zhang, Y., Xue, X., Zhang, S., Chen, M., ... & Ding, G. (2021). Analysis of spatial-temporal distribution of notifiable respiratory infectious diseases in Shandong Province, China during 2005–2014. *BMC Public Health*, 21(1), 1597. <https://link.springer.com/article/10.1186/s12889-021-11627-6>
34. Maleki, A., Mehrbod, P., Bokharaei-Salim, F., Eybpoosh, S., Tavakoli, M., Mohammadnejad, A. E., Hosseini, Z., Kashanian, S., Asadi, L. F., Salehi-Vaziri, M., & Fotouhi, F. (2023). Epidemiological surveillance of respiratory viral infections in SARS-CoV-2-negative samples during COVID-19 pandemic in Iran. *Virology journal*, 20(1), 296. <https://doi.org/10.1186/s12985-023-02226-5>
35. McGaughey, J., Fergusson, D. A., Van Bogaert, P., & Rose, L. (2021). Early warning systems and rapid response systems for the prevention of patient deterioration on acute adult hospital wards. *The Cochrane database of systematic reviews*, 11(11), CD005529. <https://doi.org/10.1002/14651858.CD005529.pub3>

36. Mohamed, M. G., Dabou, E. A. A., Abdelsamad, S., & Elsalous, S. H. (2025). Cholera outbreaks: Public health implications, economic burden, and preventive strategies. *AIMS public health*, 12(3), 767–795. <https://doi.org/10.3934/publichealth.2025039>
37. Murray, J., & Cohen, A. L. (2017). Infectious Disease Surveillance. *International Encyclopedia of Public Health*, 222–229. <https://doi.org/10.1016/B978-0-12-803678-5.00517-8>
38. Nopour, R., Shanbehzadeh, M., & Kazemi-Arpanahi, H. (2022). Using logistic regression to develop a diagnostic model for COVID-19: A single-center study. *Journal of education and health promotion*, 11, 153. https://doi.org/10.4103/jehp.jehp_1017_21
39. Potnis, N., & Tiple, B. (2023). Machine learning techniques for disease prediction. *ITM Web of Conferences*, 57, 01004. <https://doi.org/10.1051/itmconf/20235701004>
40. Que, J., & Tsui, F. C. (2008). A multi-level spatial clustering algorithm for detection of disease outbreaks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2008, 611–615. <https://pmc.ncbi.nlm.nih.gov/articles/PMC2655962/>
41. Ramudu, K., Mohan, V. M., Jyothirmai, D., Prasad, D. V. S. S. V., Agrawal, R., & Boopathi, S. (2023). Machine learning and artificial intelligence in disease prediction: Applications, challenges, limitations, case studies, and future directions. In *Contemporary Applications of Data Fusion for Advanced Healthcare Informatics* (pp. 297–318). IGI Global Scientific Publishing. <https://www.igi-global.com/chapter/machine-learning-and-artificial-intelligence-in-disease-prediction/327726>
42. Ricks, P. M., Njie, G. J., Dawood, F. S., Blain, A. E., Winstead, A., Popoola, A., Jones, C., Li, C., Fuller, J., Anantharam, P., Olson, N., Walker, A. T., Biggerstaff, M., Marston, B. J., Arthur, R. R., Bennett, S. D., & Moolenaar, R. L. (2022). Lessons Learned from CDC's Global COVID-19 Early Warning and Response Surveillance System. *Emerging infectious diseases*, 28(13), S8–S16. <https://doi.org/10.3201/eid2813.212544>
43. Riley, L. W. (2019). Differentiating epidemic from endemic or sporadic infectious disease occurrence. *Microbiology spectrum*, 7(4), 10–1128. <https://journals.asm.org/doi/abs/10.1128/microbiolspec.ame-0007-2019>
44. Rizvi, S. A., Umair, M., & Cheema, M. A. (2021). Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, solitons, and fractals*, 151, 111240. <https://doi.org/10.1016/j.chaos.2021.111240>
45. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16, 1–85. <https://projecteuclid.org/journals/statistics-surveys/volume-16/issue-none/Interpretable-machine-learning-Fundamental-principles-and-10-grand-challenges/10.1214/21-SS133.short>
46. Sagar, A., Kolluru, V., Jaiswal, U., Kumavat, G., Hole, S. R., & Kumar, A. (2025, March). Machine Learning and Artificial Intelligence for Predictive Modeling in Antimicrobial Resistance Data Sets, Challenges, and Future Directions. In *2025 3rd International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)* (pp. 1–6). IEEE. <https://ieeexplore.ieee.org/abstract/document/11009316/>
47. Schwartz, A. (2023, March 23). Visualizing Clustering Algorithms: K-Means and DBSCAN. Retrieved from Medium website: <https://levelup.gitconnected.com/visualizing-clustering-algorithms-k-means-and-dbscan-c4ce62de23c1>
48. Shuaibu, B. I., Momodu, A., Muhibi, M. A., Dutsinma, U. A., & Yelwa, I. S. (2025). A comparative review on the epidemiology of influenza, COVID-19 and respiratory syncytial viruses. *Discover Medicine*, 2(1), 274. <https://link.springer.com/article/10.1007/s44337-025-00518-z>
49. Sogandi, F. (2024). Identifying diseases symptoms and general rules using supervised and unsupervised machine learning. *Scientific Reports*, 14(1), 17956. <https://www.nature.com/articles/s41598-024-69029-8>
50. Straif-Bourgeois, S., Tonzel, J. L., Kretzschmar, M., & Ratard, R. (2023). Infectious disease epidemiology. In *Handbook of epidemiology* (pp. 1–79). New York, NY: Springer New York. https://link.springer.com/content/pdf/10.1007/978-1-4614-6625-3_34-1.pdf
51. Tapak, L., Hamidi, O., Fathian, M., & Karami, M. (2019). Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare centers in Iran. *BMC Research Notes*, 12, 353. <https://doi.org/10.1186/s13104-019-4393-y>
52. Thayer, M. B., Marzan-Rodriguez, M., Torres Aponte, J., Rivera, A., Rodriguez, D. M., Madewell, Z. J., Rysava, K., Paz-Bailey, G., Adams, L. E., & Johansson, M. A. (2025). Dengue epidemic alert thresholds for surveillance and decision-

- making in Puerto Rico: development and prospective application of an early warning system using routine surveillance data. *BMJ open*, 15(9), e106182. <https://doi.org/10.1136/bmjopen-2025-106182>
53. Thomson, A. M., McHugh, T. A., Oron, A. P., Teply, C., Lonberg, N., Tella, V. V., ... & Hay, S. I. (2023). Global, regional, and national prevalence and mortality burden of sickle cell disease, 2000–2021: a systematic analysis from the Global Burden of Disease Study 2021. *The Lancet Haematology*, 10(8), e585-e599. [https://www.thelancet.com/journals/lanhae/article/PIIS2352-3026\(23\)00118-7/fulltext?linkId=100000205683156](https://www.thelancet.com/journals/lanhae/article/PIIS2352-3026(23)00118-7/fulltext?linkId=100000205683156)
 54. Van Kessel, S., Wielders, C., Van de Kassteele, J., Verbon, A., & Schoffelen, A. (2025). The use of a Poisson hidden Markov model for automated detection of hospital outbreaks with vancomycin-resistant enterococci in routine surveillance data. *Journal of Hospital Infection*, 162, 62-67. <https://doi.org/10.1016/j.jhin.2025.04.025>
 55. Villanueva-Miranda, I., Xiao, G., & Xie, Y. (2025). Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review. *Frontiers in public health*, 13, 1609615. <https://doi.org/10.3389/fpubh.2025.1609615>
 56. Wainger, J. (2026). [Review of Measures of Disease Frequency - Stats - Medbullets Step 1]. Retrieved from Step1.medbullets.com website: <https://step1.medbullets.com/stats/101004/measures-of-disease-frequency>
 57. Wang, Y., Zhao, Y., Therneau, T. M., Atkinson, E. J., Tafti, A. P., Zhang, N., Amin, S., Limper, A. H., Khosla, S., & Liu, H. (2020). Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. *Journal of biomedical informatics*, 102, 103364. <https://doi.org/10.1016/j.jbi.2019.103364>
 58. WHO. (2008). Early Warning and Response to Outbreaks and other Public Health Events: A Guide. Retrieved from <https://iris.who.int/server/api/core/bitstreams/739d9953-31c0-4fe6-ba19-103592727d73/content>
 59. WHO. (2023). Disease Outbreaks. Retrieved from [www.who.int](https://www.who.int/teams/environment-climate-change-and-health/emergencies/disease-outbreaks) website: <https://www.who.int/teams/environment-climate-change-and-health/emergencies/disease-outbreaks>
 60. Wiemken, T. L., & Kelley, R. R. (2020). Machine learning in epidemiology and health outcomes research. *Annual review of public health*, 41, 21-36. <https://www.annualreviews.org/content/journals/10.1146/annurev-publhealth-040119-094437>
 61. World Health Organization. (2024). Global Influenza Surveillance and Response System (GISRS). Retrieved from [Www.who.int](https://www.who.int/initiatives/global-influenza-surveillance-and-response-system) website: <https://www.who.int/initiatives/global-influenza-surveillance-and-response-system>
 62. World Health Organization. (2025). Influenza (seasonal). Retrieved from [Who.int](https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)) website: [https://www.who.int/news-room/fact-sheets/detail/influenza-\(seasonal\)](https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal))
 63. Zacher, B., & Czogiel, I. (2022). Supervised learning using routine surveillance data improves outbreak detection of Salmonella and Campylobacter infections in Germany. *PLOS ONE*, 17(4), e0267510. <https://doi.org/10.1371/journal.pone.0267510>
 64. Zacher, B., Ullrich, A., & Ghazzi, S. (2019). Supervised Learning for Automated Infectious-Disease-Outbreak Detection. *Online Journal of Public Health Informatics*, 11(1), e261. <https://doi.org/10.5210/ojphi.v11i1.9770>
 65. Zhang, A., Surette, M. D., Schwartz, K. L., Brooks, J. I., Bowdish, D. M., Mahdavi, R., ... & MacFadden, D. (2022, July). The collapse of infectious disease diagnoses commonly due to communicable respiratory pathogens during the coronavirus disease 2019 pandemic: a time series and hierarchical clustering analysis. In *Open Forum Infectious Diseases* (Vol. 9, No. 7, p. ofac205). Oxford University Press. <https://academic.oup.com/ofid/article-abstract/9/7/ofac205/6570601>
 66. Zhang, T., Rabhi, F., Chen, X., Paik, H., & MacIntyre, C. R. (2024). A machine learning-based universal outbreak risk prediction tool. *Computers in Biology and Medicine*, 169, 107876. <https://doi.org/10.1016/j.compbimed.2023.107876>
 67. Zhou, S., Braca, P., Marano, S., Willett, P., Millefiori, L. M., Gaglione, D., & Pattipati, K. R. (2021). Application of Hidden Markov Models to analyze, group and visualize spatio-temporal COVID-19 data. *IEEE Access*, 9, 134384-134399. doi:10.1109/ACCESS.2021.3114384/<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9543670>