

**PREDICTING DISEASE OUTBREAKS USING
HEALTHCARE DATA TRENDS**

PRESENTED BY: SNEHA BABARIYA

STUDENT NUMBER: 2409818

**MODULE: RESEARCH METHODS AND PROFESSIONAL
DEVELOPMENT (ACFA7004)**

TABLE OF CONTENTS

1. Introduction	3
1.1. Context and Background	3
1.2. Research Problem	4
1.3. Research Aim and Objectives	5
1.4. Significance of the Research	5
2. Literature Review	6
2.1. Introduction and Scope of Literature	6
2.2. Literature Review	6
2.3. Frameworks and Relation to Current Study	14
2.4. Summary of Gaps Identified	15
3. Research Design and Methodology	15
3.1. Research Design	15
3.2. Research Methods	16
3.3. Research Materials	16
4. Expected Outcomes:	17
5. Timeline	18
6. Reference List:	19

1. Introduction

1.1. Context and Background

The number and effect of infectious disease outbreaks have increased substantially in the past two decades, highlighting the importance of rigorous, data-informed predictive systems in public health contexts. Classical epidemiological methods work well for retrospective analysis but are often too cumbersome and not sensitive enough to pick up early signals of emerging diseases. Therefore, public health organisations are more and more leveraging data analytics and artificial intelligence (AI) tools for predicting and controlling infectious diseases [1]. The application of AI and ML in the prediction of pandemics represents a black-box escape from the paradigm of passive surveillance to active intervention-based solutions that can effectively guide decision-makers in resource allocation and control implementation.

Data that informs these predictive models is drawn from a variety of platforms, including electronic health records (EHRs), social media feeds, IOT-derived health monitoring systems, environmental sensors, and digital behaviour footprints [19]. Furthermore, with the rapid advances in big data analytics and cloud computing, processing of "big and heterogeneous" data sets could easily be achieved in near real-time, allowing for insights into transmission dynamics, vulnerable populations and likely hotspots. This change in approach enables immediate action and the agility of healthcare systems to new threats.

In the last few years, ML predictions for disease outbreaks have performed unprecedentedly. Supervised learning algorithms, such as Random Forest, Support Vector Machines (SVM) and deep learning models, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have shown high predictive precision in the pandemic modelling [3]. These technologies were critical in estimating case numbers, mortality, and policy effects during the COVID-19 pandemic.

However, there are still major challenges in integrating and implementing predictive models in the infrastructure of public health. Obstacles consist of the absence of universal data sources, ethical issues around privacy, and the lack of real-world deployment of AI-based systems. However, the capacity for AI and ML to be applied significantly to predict the outcome of diseases is undeniable and is generating attention at the academic and institutional levels [8]. With the threat of rapidly spreading diseases via globalisation, this capability to

forecast outbreaks with mathematical models has become a vital tool in the modern public health practitioner's arsenal.

1.2. Research Problem

While AI and ML have great potential to improve disease outbreak prediction, current applications are scattered and generally limited to experimental frameworks. At the root of the problem is the dependence on manually reported numbers that cause time lags and limit the agility of public health systems. The majority of predictive models continue to rely on historical and static data sources that do not fully capture the dynamics of disease spread or contemporaneous human activity [16]. There is thus an urgent imperative to implement models that can consume and process streaming data of various natures (e.g., EHR, social media, environmental signals) for time-sensitive prediction.

Additionally, current predictive models tend to suffer from a lack of uniform standards in data quality and completeness. The datasets at the heart of ML model training often are biased, incomplete and suffer from poor quality labelling, which affects the accuracy and generalisation of the model. Moreover, the parameters and feature selection are commonly made and implemented without formal validation or domain relevance-oriented tailoring, leading to (near-) suboptimal experimental results when applied in the field [17]. These technical limitations are compounded by operational challenges, including a lack of computational resources in resource-poor environments and insufficient partnerships between data scientists and public health practitioners.

Additionally, there is a clear mismatch in the incorporation of AI systems into official disease surveillance systems. The use of such models in routine practice within health systems or international health agencies is, nevertheless, limited, although they can perform well within simulations or retrospectively. The disconnect underscores the motivation for mechanisms that can bridge between model development and field implementation, allowing AI-based predictions to translate into practical applications [11]. With the increasing frequency and complexity of outbreaks, tackling these issues is of paramount importance to improve global health preparedness and for the development of robust epidemiological surveillance systems.

1.3. Research Aim and Objectives

Aim:

The objective of this study is to build and test ML models for the prediction of disease outbreaks in secondary health data, with the specific aim of comparing model performance across different methods of ML. The study aims to investigate whether AI could work as a tool that enables public health authorities to detect and act early.

Objectives:

- To search and gather relevant secondary datasets regarding disease surveillance from public domains like Kaggle that include factors contributing to the transmission of disease.
- To prepare the fetched data, using data mining techniques, such as data cleaning and feature selection, transformation, and normalisation, to enable its use in ML model training.
- To apply supervised machine learning techniques, e.g., RF, SVM, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks, as well as deep learning models, to predict disease outbreak events.

1.4. Significance of the Research

Implications: This study might help improve public health systems' readiness and responsiveness by using artificial intelligence. Through the application of machine learning techniques, which can handle big data in a real-time manner, this study provides a direction for proactive outbreak prevention and decision support [7]. Additionally, the paper makes contributions to the literature by comparing classical algorithms and deep learning techniques over a multitude of datasets, helping to inform the comparison by providing alternative views into the relative applicability and strength of the methods. Such a multi-dimensional approach is indispensable for finding the most appropriate model for a given epidemic context, e.g. that of respiratory diseases, vector-transmitted diseases or zoonotic emergencies [12]. Beyond the technical advance, including its implications for system integration, the present research provides solutions to practical issues in terms of ethical data use and scalability in resource-limited settings. These are issues that must be addressed to move AI models from research-

industry environments to working health infrastructure. The findings of this study have potential implications for how colleges plan, identify, and react to future outbreaks of illness.

2. Literature Review

2.1. Introduction and Scope of Literature

The literature review intends to describe the existing studies that have used artificial intelligence (AI) and machine learning (ML) algorithms to forecast outbreaks of diseases. As an increasing amount of medical data is accessible from different sources, including EHRs, social media, IoT devices and environmental sensors, researchers have begun to emphasise the use of computational models for early outbreak detection. Here, an overview of studies from 2018 to 2024 has been discussed, which offer the recent progress on predictive modelling, and especially the application of supervised learning models (e.g., Random Forest, SVM) and deep learning approaches (e.g., CNN, LSTM). The coverage is technical methods, integration issues, performance measurements and ethical aspects. Through uncovering major trends, theoretical frameworks and current research to this end, this section underpins the development of an AI-based solution to advance disease surveillance and response in real-world applications within healthcare.

2.2. Literature Review

Data Analytics for Predicting Disease Outbreaks

review in detail and depth how data analytics platforms are transforming disease outbreak prediction [1]. The paper classifies available models into statistical, machine learning (ML) and hybrid models and suggests the inadequacies of traditional surveillance systems. The authors highlight the transition from traditional retrospective approaches to predictive models that use structured and unstructured data sourced from EHRs, mobile health apps, social media, and environmental data. An important contribution of the paper is the classification of the tools and platforms used to base outbreak modelling including Python libraries, cloud platforms or visual analytics dashboards. It focuses on outline studies of cholera and influenza outbreaks, at which early detection rates have been enhanced by ML models, such as Decision Trees and Support Vector Machines (SVM). The paper contains a figure that represents well a multi-source data pipeline: it is a sketch of how to receive different types of inputs (e.g., hospital records, tweets, IOT sensors) and process them into structured data sets for modelling. This

value highlights the challenge of incorporating online streams and the importance of having strong preprocessing pipelines which incorporate quality filtering and imputation techniques.

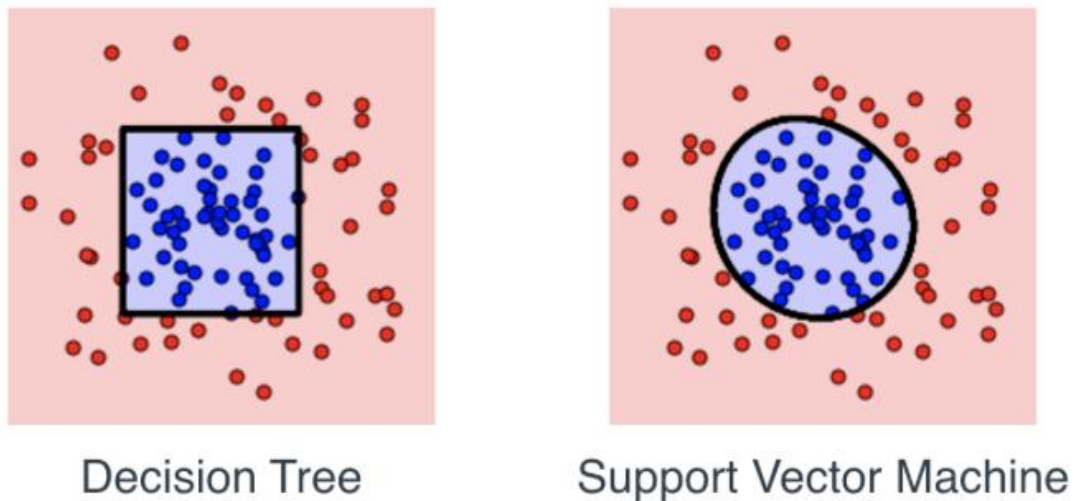


Figure 1: Decision Tree and Support Vector Machine

(Source: [linkedin.com/pulse](https://www.linkedin.com/pulse/))

This article provides a basis for the choice of data processing pipelines and model architectures in the intended study. It supports the conclusion of using multi-source data for better prediction performance. The number of multi-source integrations is given for comparison and the advocated approach in this paper. Its conclusion, on the success of hybrid tools, promotes the use of both ML (Random Forest, SVM) as well as DL (CNN, LSTM) models, in the research designs.

How AI/ML Is Transforming Disease Forecasting

focus on the revolutionary potential of AI and ML in disease prediction in the post-COVID period [16]. The research divides AI methods according to how well they handle structured versus unstructured data and shows how approaches like natural language processing (NLP) and anomaly detection are disrupting the field of outbreak prediction. The authors illustrate through pandemic timelines how algorithms like SVM, Random Forest, and LSTM can model temporal correlations in spreading. The paper acknowledges obstacles such as data scarcity, overfitting, and non-inferentiality and suggests ensemble learning and attention mechanisms as possible remedies. An especially critical contribution to this article is the "AI Outbreak

Forecasting Loop," a feedback loop (from data collection → feature extraction → model training → prediction → evaluation → decision support → informed data collection). This loop demonstrates the iterative process and continuous learning of predictive modelling.

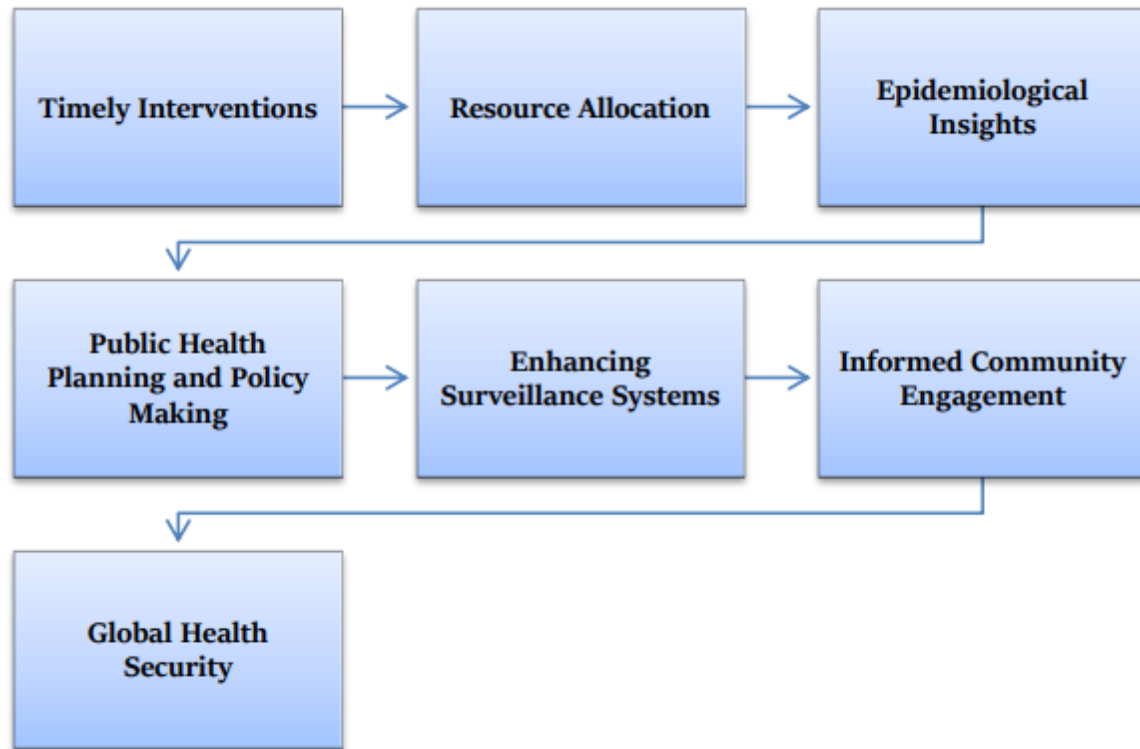


Figure 2: The Importance of Accurate Disease Forecasting

(Source: Reddy *et al.*, 2021)

prediction loop has significant similarity in structure to the proposed system, in particular with respect to using real-time feedback for improving the model [16]. It also accommodates the common strategy of combining static datasets for building baseline models and streaming live data for updates. The paper justifies the use of temporal models, such as LSTM and their capacity to adapt to developing health trends. Furthermore, the elicited challenges will inform risk management during model evaluation in this project.

Media-Based Disease Outbreak Prediction Using ML

concentrate on a fairly new way of disease prediction: utilising media news and online articles as a data source [11]. Their research constructs ML models with text-mining and NLP over thousands of media articles to identify early signals of infectious outbreaks. The authors use logistic regression, naïve Bayes, and SVM classifiers on the extracted keywords, location

references, and time markers. In some cases, their models have accurately predicted the occurrence of outbreaks two weeks before official announcements by public health agencies. The article also considers keyword drift and media bias as sources of data-poor quality. The "Media Text Processing Framework" is mapped by a central character in the article. This process includes steps such as media scraping, topic modelling (eg LDA), keyword scoring, geographic tagging, and model training. This representation emphasises the role of pre-processing and makes it possible for raw narrative data to become predictive signals.

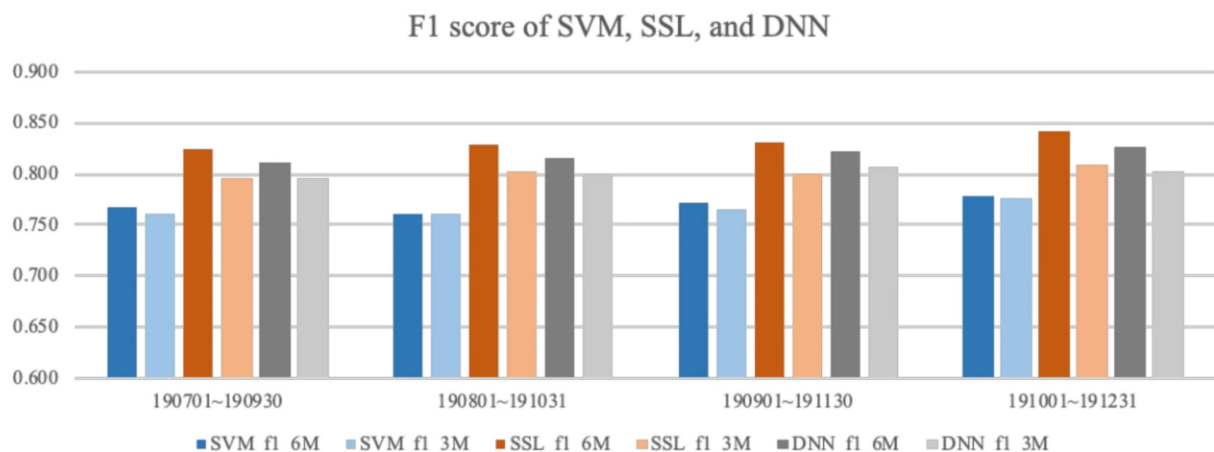
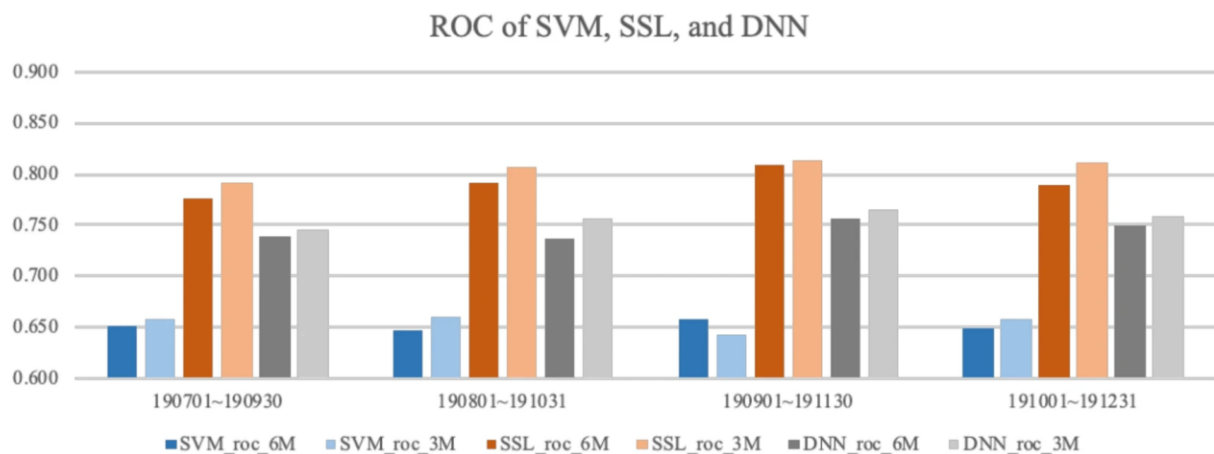
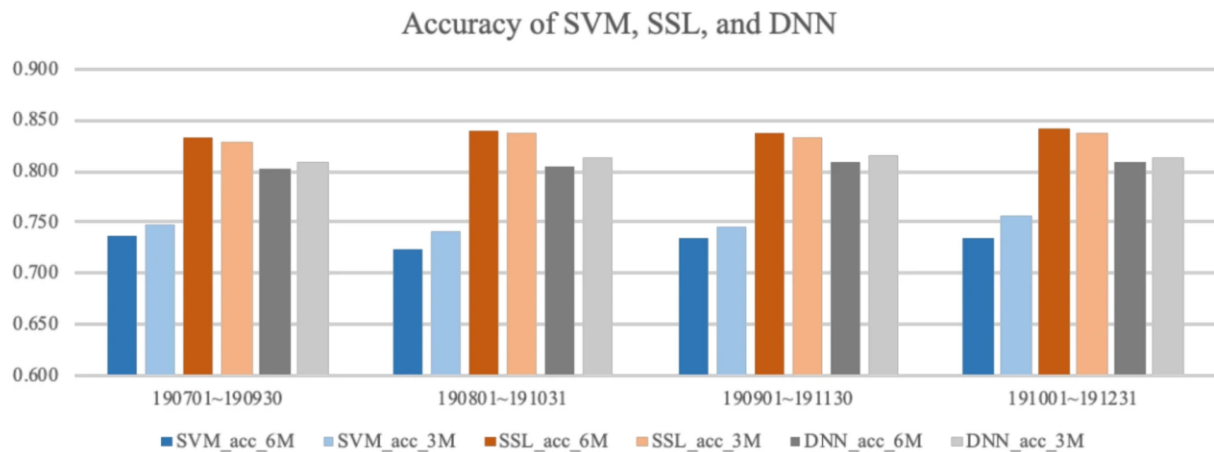


Figure 3: Accuracy, ROC, and F1 score of each validation data set period by each model, respectively

(Source: Kim and Ahn, 2021)

This paper paves the way for the literature by including unstructured narrative information from news or social media. While the parent project relies on structured healthcare data sources, this approach may be examined as a second model input. In the figure, the complex NLP pipeline that the figure represents could be replaced in the process of preprocessing, since such data has been included. It also raises potential bias concerns and recommends wider model checks for other types of data.

Predictive Modelling for Zoonotic Diseases

treats the escalating importance of zoonotic infectious diseases, with a focus on predictive modelling for early detection and monitoring [8]. It provides an overview of several models that have been used for monitoring and predicting forms of zoonosis other than the novel coronavirus, such as West Nile virus and Ebola, with a heavy emphasis on public health surveillance. It assesses the performance of supervised ML algorithms and emphasises the significance of ECMS that use environmental, behavioural and healthcare data. The importance of model sensitivity and specificity in discriminating real outbreaks and noise is emphasised, and models that are opaque in real-time are criticised. It also discusses challenges with model deployment, such as poor digital infrastructure in developing countries and a lack of synchrony between health practitioners and data scientists. One of the feature images of the article shows a surveillance network that brings in mobile data collection, regional health databases, and AI detection hubs. This illustration demonstrates how cascade-based data aggregation can support near real-time monitoring of disease risk zones.

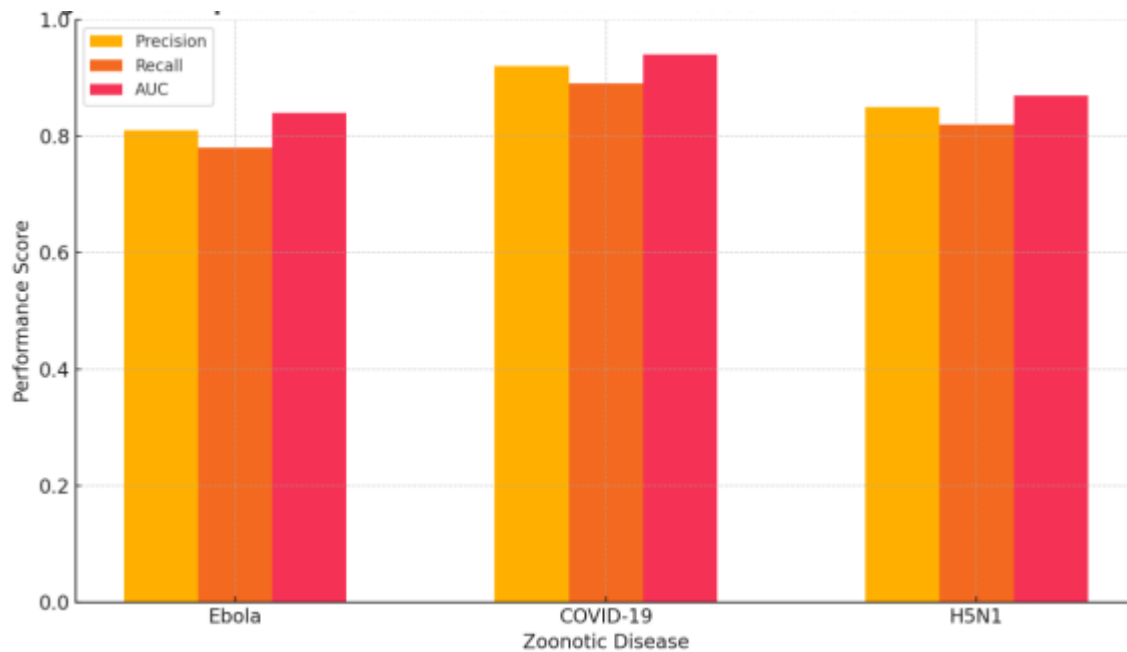


Figure 4: Comparative Performance of Prediction Models in Historical Zoonotic Outbreaks[8]

The work of Folasole is in line with efforts to incorporate environmental and behavioural factors into disease forecasting. The surveillance framework figure, therefore, provides a theoretical underpinning for the organisation of the PIPPC’s multi-faceted approaches to the collection of data and evaluation. It also highlights the need for public health integration and the potential scalability of the outputs of the project to low-resource settings, if infrastructure (e.g. cloud cloud-based models) is provided.

AI for Pandemic Preparedness and Disease Surveillance

offers an extensive analysis of how AI has generated changes to pandemic readiness, especially for what concerns COVID-19 [3]. This study assesses the role of AI in outbreak detection, transmission modelling, and policy optimisation. Through case studies of South Korea, the U.S. and India, the piece also contrasts various AI applications and the contribution of these systems to real-time decision-making. It focuses on deep learning models (including CNNs and LSTMS) that are well-known in the literature for their efficient pattern recognition in time-series epidemiological data. The paper also delves into how AI can inform government-level decision-making by incorporating mobility data, clinical symptoms, and social behaviour in outbreak maps. The accompanying visual, shown below, is a “Pandemic AI Dashboard” mock-

up of how various ML model outputs (case predictions, R-values, resource needs) flow into an analysed, consolidated view for public health officials. This kind of real-time visualisation shows the power of AI at work, steering interventions.

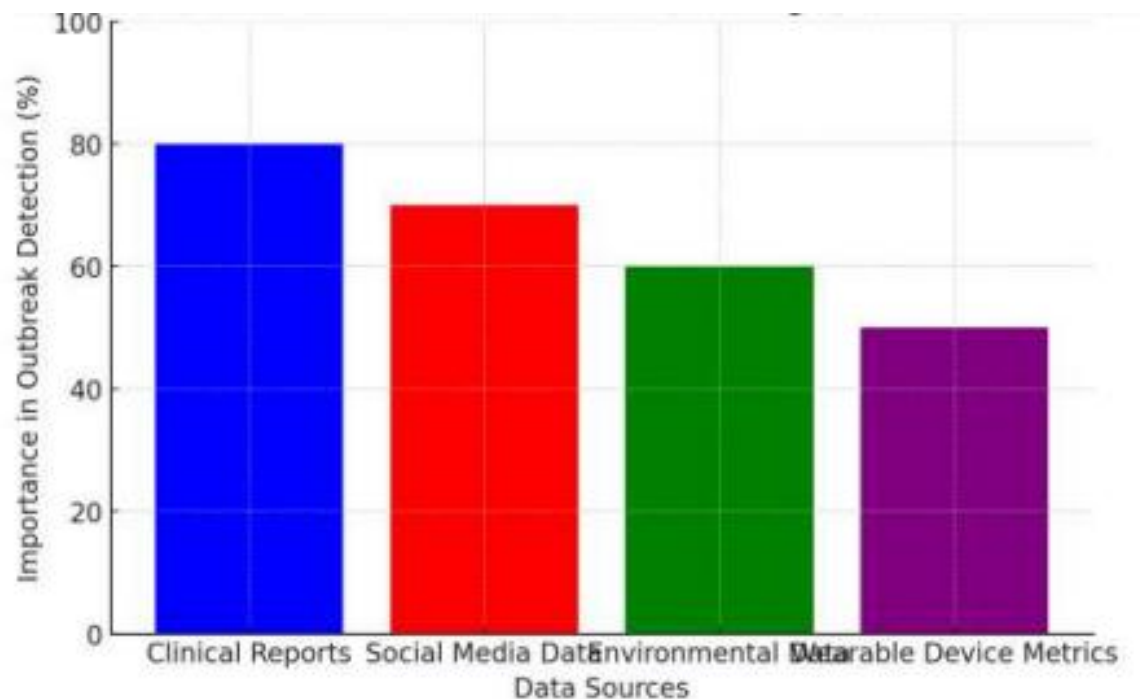


Figure 5: Framework for AI-driven outbreak detection using real-time data sources[3]

This focus of research is strongly justified by Ali's work. The application of CNN and LSTM models, as it has been highlighted in this work, is compatible with the technique's mindset of this project. The AI dashboard also presents an intriguing visualisation model that could be used to seed the prototype phase of our study. The policy integration dimensions also provide additional avenues for which to 'translate' this research to more practical purposes outside of academia.

A Framework for Pandemic Prediction Using Big Data Analytics

present a structured framework for pandemic forecasting, which is based on big data analytics, specifically focusing on COVID-19 [2]. Using data from various sources, such as Google Trends, social media posts and national health records, they aim to predict waves of outbreaks. The paradigm considers stages like data fusion, correlation analysis, and the use of ensemble machine learning models, among others. The authors show that hybrid predictive models that employ both good medical data and bad public opinion metrics significantly bolster the early

warning powers of such predictive models. They also face the difficulty of coping with heterogeneous data and the need for real-time processing. The central node represents the “Big Data Predictive Cycle,” composed of data ingestion, pre-processing (including NLP and image recognition), model training (Random Forest and Deep Neural Networks), and public health production. It draws attention to how information streams from raw sources into usable knowledge.

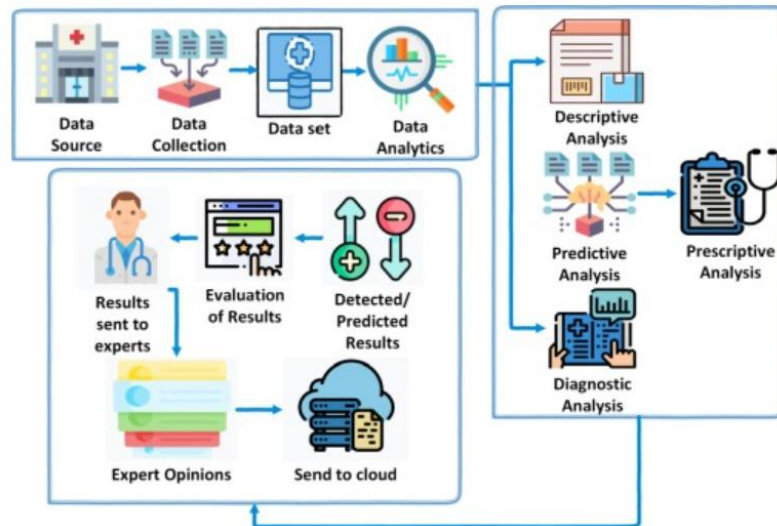


Figure 6: Proposed IOT-based data analytics framework for pandemic analysis and prediction[2]

This paper provides further evidence supporting the utility of a multistream and non-traditional data approach to outbreak prediction. Its big data cycle number is consistent with the number of the CRISP-DM framework in this project. It also confirms the research design as pertains to the models (traditional and deep learning) used, and for testing the cross-data correlation strategies that provide better predictive outputs.

Machine Learning and Prediction of Infectious Diseases

have systematically reviewed machine learning based applications in global studies that focus on predicting infectious diseases [17]. In a paper, the models are divided into supervised, unsupervised, and combined methods and popular algorithms such as Random Forest, SVM, k-NN, and LSTM are emphasised. The diseases include COVID-19, dengue, malaria and influenza, and the review compares model performances for the sensitivity, accuracy and data constraints. It reflects a common problem: differences in data and the absence of a standard entry format. The writers urge serious pre-processing standards and model interpretability for

public health acceptance. A taxonomy map of ML methods with respect to concerns (e.g., vector borne diseases vs. airborne spreading) is pictured in the paper. The minimum penalty as a function of β for selected outbreaks and different cases. The following figure shows that the choice of the algorithm depends on the specific type of outbreak, as well as the structure of the data and the temporal complexity of the outbreak.

The model selection reasoning is supported by this systematic review. This figure provides further evidence that CNN and LSTM models are suitable in time-sensitive situations for airborne disease prediction. The article also guides the choice of evaluation metrics and reveals the importance of comparative performance assessment. It confirms the importance of studying model generalisation across diseases and geographies.

Forecasting COVID-19 Using Google Trends and Deep Learning

present a hybrid method for disease prediction based on the combination of Google search trends and deep learning, which is achieved using a Grey Wolf-based LSTM [15]. The paper shows how online behaviour can be used as a proxy indicator for epidemiological trends when the formal reporting lags. Our self-attention permutation problems include the prediction of future keyword occurrences from a keyword frequency time series, using data across several countries during the COVID-19 pandemic, and location-tagged tweet classification. A key piece in the paper compares the accuracy of the GWO-optimised LSTM model with LSTM and ARIMA. The visualisation illustrates that the RMSE and MAE performance gain is significant because of optimisation, especially in the prediction of peaks and troughs in pandemic waves.

This paper provides a new approach to integrating behavioural data sources within the project. This is intended as a principle that can be established and applied in the course of research; an architecture that is suitable for a dataset in which the core data is associated with healthcare can be enriched or partially validated with search trends. The performance comparison figure also gives useful information for when and how optimisation can be performed during the training of the model.

2.3. Frameworks and Relation to Current Study

In the present work, the CRISP-DM (Cross-Industry Standard Process for Data Mining) model paradigm is used as a methodological basis to provide a systematic process for the design,

development, and testing of disease outbreak predictive methods. This process supports following a clear process, starting from the domain understanding and the data inspection, through preprocessing, model construction, evaluation, and in the end, the deployment. The CRISP-DM model has been slightly useful in the application of healthcare analytics (due to the large scale and heterogeneity of data) through a process of preparing data and refining iteratively to actionable insights [2]. This framework guarantees a transparent and reproducible research process. The work is also built on deep learning models, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), as they have shown attractive results for time-series prediction and pattern recognition in the epidemic modelling context.

2.4. Summary of Gaps Identified

At present, there are critical unmet needs in the real-time application of predictive models to public health infrastructures that relate to the heterogeneity of the data, transparency of the models, and pragmatic operations. Moreover, the lack of multi-source data fusion and poor validation procedures impact the credibility and adaptability of current machine learning-based outbreak forecasting schemes.

3. Research Design and Methodology

3.1. Research Design

This study uses a quantitative research design to develop and evaluate machine learning models for predicting infectious disease outbreaks. The research follows a descriptive and analytical approach, leveraging secondary data sourced from publicly available healthcare datasets. These datasets may include outbreak timelines, case counts, hospital records, and relevant environmental indicators. The quantitative design allows the application of statistical techniques and machine learning algorithms to identify temporal patterns, correlations, and trends in outbreak behavior.

A cross-sectional structure is used to capture a snapshot of historical outbreak conditions, which will be analyzed using models such as Random Forest, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. These models will help determine how different input variables contribute to outbreak prediction accuracy. This design supports

the identification of early-warning signals and the evaluation of how data-driven techniques can enhance preparedness for future public health threats.

3.2. Research Methods

This study adopts a quantitative research approach supported by secondary data analysis. No primary human participants are involved. Instead, the research relies on publicly available, anonymized datasets containing healthcare records, time-series outbreak data, and environmental or demographic variables. These datasets will be sourced from reputable platforms such as the World Health Organization (WHO), Kaggle, and government health data portals. The quantitative data will be used to train and evaluate machine learning models aimed at predicting potential disease outbreaks.

To ensure a comprehensive evaluation of model effectiveness, the study will use classification metrics including accuracy, precision, recall, and F1-score. Comparative analysis will be performed between Random Forest, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) models. These methods allow for both broad statistical analysis of outbreak patterns and nuanced learning from historical healthcare trends. This approach enables the development of an early warning framework to support public health decision-making.

3.3. Research Materials

This study relies entirely on secondary data sources to develop and evaluate predictive models for disease outbreaks. No survey or interview data will be collected, and no human participants are involved. The data used will consist of anonymized, publicly available health-related datasets sourced from platforms such as the World Health Organization (WHO), Kaggle, and other reputable open-access repositories. These datasets may include time-stamped disease case counts, symptom records, hospitalization data, and associated environmental variables.

The study will use Python for data preprocessing, feature engineering, and machine learning model development. Quantitative analysis will be conducted using descriptive and inferential statistics to assess model performance. Models including Random Forest, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks will be trained and evaluated using standard metrics such as accuracy, precision, recall, and F1-score. No qualitative data or coding is required, and no primary data collection or participant recruitment

will be conducted. This approach ensures both ethical compliance and methodological focus on real-time predictive modeling.

4. Expected Outcomes:

The research is expected to identify the most suitable machine learning model for early detection of infectious disease outbreaks based on healthcare data trends. It will provide comparative performance results for models such as Random Forest, CNN, and LSTM. Additionally, the project aims to demonstrate how secondary data can support real-time predictions and contribute to more effective public health planning. The study may also uncover limitations in current datasets and propose recommendations for improving data readiness and AI integration in health surveillance systems.

Limitations of the Study:

Despite the potential of machine learning models in disease outbreak prediction, this study faces several limitations. First, the reliance on publicly available secondary datasets may introduce inconsistencies in data quality, structure, or completeness. Many datasets may not include real-time data, making the predictive models less effective in dynamic outbreak conditions. Second, the models are trained on historical data, which may not fully capture emerging patterns or novel diseases. Third, limitations in computational resources may restrict the complexity or scalability of deep learning models like LSTM and CNN. Additionally, the absence of access to domain experts or healthcare professionals may affect the clinical interpretation of results. Finally, since the project is a simulation rather than a deployed real-time system, the practical impact of predictions could not be validated in live settings.

5. Timeline

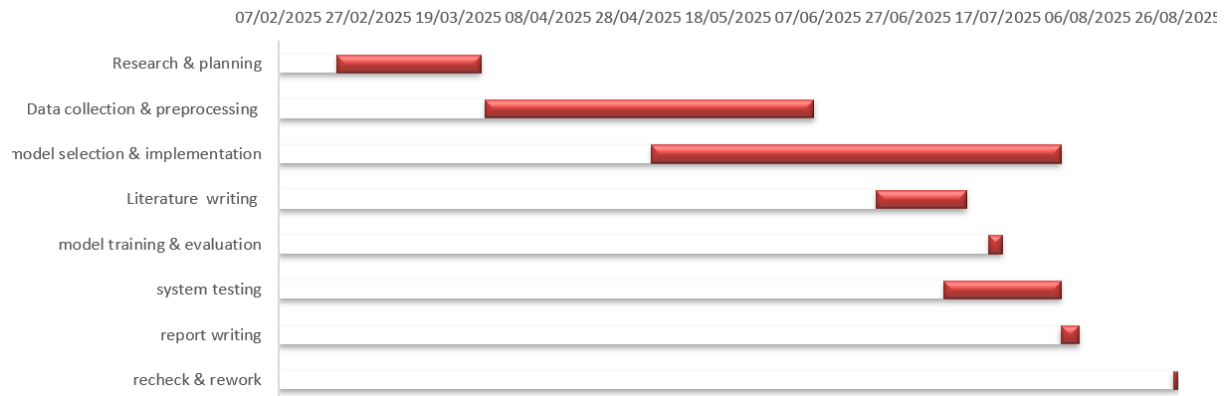


Figure 7: Gantt chart

6. Reference List:

- [1] B. O. Adegoke, T. Odugbose, and C. Adeyemi, "Data analytics for predicting disease outbreaks: A review of models and tools," **Int. J. Life Sci. Res. Updates**, vol. 2, no. 2, pp. 1–9, 2024. [Online]. Available: <https://www.researchgate.net/publication/380207270>
- [2] I. Ahmed, M. Ahmad, G. Jeon, and F. Piccialli, "A framework for pandemic prediction using big data analytics," **Big Data Res.**, vol. 25, p. 100190, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2214579621000071>
- [3] H. Ali, "AI for pandemic preparedness and infectious disease surveillance: predicting outbreaks, modelling transmission, and optimizing public health interventions," **Int. J. Res. Publ. Rev.**, vol. 5, no. 8, pp. 4605–4619, 2024. [Online]. Available: <https://www.researchgate.net/publication/389138161>
- [4] K. E. ArunKumar et al., "Forecasting the dynamics of cumulative COVID-19 cases using ARIMA and SARIMA," **Appl. Soft Comput.**, vol. 103, p. 107161, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1568494621000843>
- [5] S. Dash, C. Chakraborty, S. K. Giri, and S. K. Pani, "Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics," **Pattern Recognit. Lett.**, vol. 151, pp. 69–75, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865521002762>
- [6] S. Dash et al., "BIFM: Big-data driven intelligent forecasting model for COVID-19," **IEEE Access**, vol. 9, pp. 97505–97517, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9474452/>
- [7] K. J. Devi et al., "Artificial intelligence in healthcare: diagnosis, treatment, and prediction," in **E3S Web Conf.**, vol. 399, p. 04043, 2023. [Online]. Available: https://www.e3s-conferences.org/articles/e3sconf/abs/2023/36/e3sconf_iconnect2023_04043/e3sconf_iconnect2023_04043.html
- [8] A. Folasole, "Data analytics and predictive modelling approaches for identifying emerging zoonotic infectious diseases," **Int. J. Eng. Technol. Res. Manag.**, vol. 7, no. 12, p. 292, 2023. [Online]. Available: <https://www.researchgate.net/publication/390768835>

- [9] W. Huang, B. Cao, G. Yang, N. Luo, and N. Chao, “Turn to the internet first? Using online medical behavioural data to forecast COVID-19 epidemic trend,” **Inf. Process. Manag.**, vol. 58, no. 3, p. 102486, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306457320309754>
- [10] R. P. Jones, “Multidisciplinary insights into health care financial risk and hospital surge capacity,” **J. Health Care Finance**, 2021. [Online]. Available: https://www.academia.edu/download/65590468/242_492_1_SM.pdf
- [11] J. Kim and I. Ahn, “Infectious disease outbreak prediction using media articles with machine learning models,” **Sci. Rep.**, vol. 11, p. 4413, 2021. [Online]. Available: <https://www.nature.com/articles/s41598-021-83926-2>
- [12] M. U. Kraemer et al., “Artificial intelligence for modelling infectious disease epidemics,” **Nature**, vol. 638, no. 8051, pp. 623–635, 2025. [Online]. Available: <https://www.nature.com/articles/s41586-024-08564-w>
- [13] D. B. Olawade et al., “Using artificial intelligence to improve public health: a narrative review,” **Front. Public Health**, vol. 11, p. 1196397, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2023.1196397/full>
- [14] O. Panahi, “The Future of Healthcare: AI. Public Health and the Digital Revolution,” **Medi Clin Case Rep J.**, vol. 3, no. 1, pp. 763–766, 2025. [Online]. Available: <https://www.researchgate.net/publication/389673845>
- [15] S. Prasanth et al., “Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-deep learning approach,” **Chaos Solitons Fractals**, vol. 142, p. 110336, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960077920307311>
- [16] M. S. Reddy et al., “Predicting Tomorrow's Ailments: How AI/ML Is Transforming Disease Forecasting,” **ESP J. Eng. Technol. Adv.**, vol. 1, no. 2, pp. 188–200, 2021. [Online]. Available: <https://www.academia.edu/download/119235855>
- [17] O. E. Santangelo et al., “Machine learning and prediction of infectious diseases: a systematic review,” **Mach. Learn. Knowl. Extr.**, vol. 5, no. 1, pp. 175–198, 2023. [Online]. Available: <https://www.mdpi.com/2504-4990/5/1/13>

- [18] N. Sharma et al., “A heterogeneous ensemble forecasting model for disease prediction,” *New Gener. Comput.**, 2021. [Online]. Available: <https://link.springer.com/article/10.1007/s00354-020-00119-7>
- [19] S. Yousefinaghani et al., “Prediction of COVID-19 waves using social media and Google search: a case study of the US and Canada,” *Front. Public Health**, vol. 9, p. 656635, 2021. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2021.656635/full>
- [20] D. Zeng, Z. Cao, and D. B. Neill, “Artificial intelligence-enabled public health surveillance—from local detection to global epidemic monitoring and control,” in *Artificial Intelligence in Medicine**, Academic Press, pp. 437–453, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780128212592000223>