Advancing Disease Surveillance: Leveraging Novel Data Sources and Modalities for Public Health

Research Hypothesis and Objectives

The main purpose of this proposed research is to speculate and examine how data mining and text analytics methods and tools can be utilised to make the best use of new data sources and modalities for disease surveillance, signifying the importance of monitoring the diseases that might outbreak as an epidemic or pandemic. This research is very pertinent because, with the vast abundance of various data streams, it is very important to develop more efficient methods for monitoring public health, amidst the face of new and variety of strains in infectious diseases and health hazards.

The research arises out of the hypothesis that monitoring efforts on diseases can be improved with more accuracy, and less latency by effectively utilising emerging data streams like social media data, wearable technology-based watches, health websites and apps, and also environmental sensors. This theory is based on the realisation that traditional monitoring techniques might not always be able to quickly identify and track outbreaks, particularly when conventional data sources are hard to come by or inadequate.

The project’s main goals are to research on the possibilities of new data sources and modalities for monitoring diseases, creating and applying various text analytics and data mining tools designed for analysing various data stream and evaluating the efficiency and practicality of these tools in real-world contexts. By offering insights into the viability and efficiency of utilising novel data sources for improving disease surveillance capabilities, the research aims to extend the knowledge in the field of public health informatics

Following are the research objectives which will help to achieve these goals:

- Examine and analyse the potential of novel data sources such as smart watches, social media data, web health apps, mobile health apps and environmental sensors, for disease surveillance.

- Create and put into practice text analytics and data mining tools specifically suited for analysing a variety of data streams in order to detect and track outbreaks much earlier.

- Evaluate the efficiency of the suggested methods in identifying and tracking epidemics in practical environments, utilising pertinent measures like predictive value, sensitivity and specificity,

- Examine if the developed methods are scalable for integrating into current public health procedures and disease surveillance systems.

- By sharing the research's findings with relevant public health stakeholders and through academic publications and conference presentations, you can help the field of public health informatics grow in knowledge.

Background

Over and above, proper public health monitoring primarily controls the spread of deadly infectious diseases and health hazards. Data for the detection and monitoring of different outbreaks are mainly in a structured format in public health surveillance systems. However, the increasing penetration of digital technologies and new sources of data have made this interest evolve toward the exploration and exploitation of new disease surveillance alternatives able to analyze novel, diverse data streams.

This research builds upon previous studies that have shown how techniques of data mining and text analytics have the potential to enhance the capability of disease surveillance.
For example, Chunara et al. (2012) conducted influenza surveillance in real-time using Twitter data and came to a conclusion that data from social media are indispensable when it comes to early detection. Two studies—one by Sadilek et al. (2012) and another by Althouse et al. (2015)—described and then demonstrated that it is actually possible to use mobile health app data and search engine data for disease surveillance.

Some other works of research going on at the University of Leeds include the "Digital Health Lab," championed by Professor Mark Mon-Williams, researching the utilities of digital technologies in healthcare, more specifically, the use of wearable devices like smartwatches and mobile health apps in monitoring diseases. The proposed research will therefore build on this already ongoing research to enable a contribution to better health outcomes based on novel data sources and modalities.

Other international initiatives leading the way in improving disease surveillance-based data-driven approaches include Health Data Research UK and the Global Public Health Intelligence Network. These initiatives also set the direction in terms of interdisciplinary work and the integration of diverse sources for efficiency in surveillance systems.

Moreover, international collaborations, such as the Global Initiative on Sharing Avian Influenza Data (GISAID) and the World Health Organization's (WHO) Global Outbreak Alert and Response Network (GOARN), also specifically recognize the international context in which such surveillance efforts are occurring and the demand for international cooperation in facing up to health threats.

That is, the proposed study is set within the current wider context of efforts going on to innovate and increase disease surveillance capabilities using data-driven approaches. This, therefore, builds on previous research and the current work happening at Leeds University in the UK and other countries to increase the existing knowledge in public health informatics and to build an effective response to emerging health threats.

Importance and Contribution to Knowledge:

The aim of the study on the use of novel sources and data modalities of disease surveillance includes the following high-importance contributions to various aspects of society, economy, and scientific advancement:

Public Health Impact: The research focuses on the synthesis of the various and diverse data streams from a multitude of sources, including social media, mobile health apps, smartwatches, and environmental sensors, into the disease surveillance systems in order to improve early detection, monitoring, and response to deadly infectious diseases that may outbreak. This may lead to better containment of the disease and improved strategies toward a reduction in the transmission rates with quite some improvement in the outcomes for public health (Chunara et al., 2012; Sadilek et al., 2012).

Economic Benefits: Effective disease surveillance systems will result in a reduction of economic burdens such as costs of health care, losses in productivity, and disorganisation of trade and tourism as a result of infectious disease outbreaks. Offering timely and correct information for decision-making, the proposed research may contribute toward a reduction in the magnitude of these impacts and increase economic resilience.

Technological Innovation: Innovation in data analytics, machine learning, and artificial intelligence will be further pursued with these emerging data sources and modalities to drive disease surveillance. Advanced tailor-made models and algorithms would be developed for a wide variety of data sets to drive technological advances not only for public health but in most other sectors where decisions are based on data (Mon-Williams, n.d.).

International Collaboration: The international nature of infectious disease threats requires that countries collaborate in order to secure effective disease surveillance and a more rapid response. This will be possible through collaboration with international research networks and initiatives, such as Health Data Research UK, the Global Public Health Intelligence Network, and the Global Initiative on Sharing Avian Influenza Data, that could collectively make contributions to the global health security and resilience framework stronger across the globe (Health Data Research UK, n.d.; Global Public Health Intelligence Network, n.d.; Global Initiative on Sharing Avian Influenza Data, n.d.).

Engagement across disciplines: The research proposed here is truly interdisciplinary research. It is thus likely that this research will facilitate and help in promoting the general dialogue and collaboration among related but different disciplines by sharing and exchanging knowledge and methodologies, using best practices on the way toward innovative solutions with broad social impacts.

Public health, the economy, technological innovation, international collaboration, and interdisciplinary engagement are fields that stand to benefit from the research proposed on the leverage of novel data sources and modalities for disease surveillance.

Pilot Study: Establishing Feasibility for Innovative Data Resources in Disease Surveillance

The aim of this pilot study is to evaluate the feasibility of using novel data sources for disease surveillance using the tools of data mining and text analytics, namely social media data and environmental sensors. We illustrate the feasibility and effectiveness of integrating heterogeneous data sources for disease monitoring and outbreak detection with the tools of data mining using Weka and text analytics using ChatGPT.

Picking Case Study

This pilot study thus aims to monitor the outbreak of infectious diseases in the urban area through the use of social media data. We will select a case related to a highly infective disease outbreak in an urban area characterized by high population density. We anticipate that this will result in important insights regarding both the challenges and opportunities for such a social media data-driven approach in infectious disease surveillance.

Taking Sample Data:

We will scrape the publicly available social media data of the selected case study using relevant APIs and data-scraping techniques. The data is expected to consist of posts, tweets, and comments containing keywords associated with the disease outbreak. We will preprocess and clean the sampled data using Weka as a precursor step before analysis and modeling.

Solution Prototyping Using Weka

As such, we will use Weka to develop data mining models that will analyze the patterns and trends of the social media data. We will use algorithms for decision trees, such as J48, to further identify the features and attributes that will be relevant and be associated with disease outbreaks. Clustering algorithms, more so k-means, will be used for the clustering of similar social media posts to identify potential hotspots of disease activity.

Prototype Testing with ChatGPT:

Text analytics on textual content in posts and comments will be done through the application of ChatGPT after analyzing the social media data with Weka. We will apply ChatGPT to make some meaning out of the unstructured text data, such as symptoms, geographical locations, and public sentiment about the disease outbreak. We will check whether the information extracted is relevant and accurate for the purpose of assessing the effectiveness of using ChatGPT in the surveillance of diseases.

Make the system iterative to allow

The developed models will be iteratively refined depending on comment by the domain experts and public health officials. Optimizing Weka models to have enhanced disease detection and prediction. The same refinement will be done for ChatGPT to fine-tune it for extraction of useful information from text social media data. The refined models will be validated with historical disease data and juxtaposed with the traditional surveillance methods.

Feasibility Study A pilot study would be conducted to test the feasibility of integrating novel data sources into existing surveillance systems using Weka and ChatGPT. The study will show how these tools are helpful in conducting the analysis of social media data for disease monitoring and thus support the provision of value to potentially leverage in the use of unconventional data sources for early disease detection and response.

In general, the pilot study shows the possibility of using novel data sources for disease surveillance with advanced data mining and natural language processing techniques. It brings forth the potential for tools such as Weka and ChatGPT to be included in augmenting traditional surveillance methods through emergent data streams such as social media data, data from mobile health applications, wearable devices, and environmental sensors.

Weka is thus a very versatile tool in the data mining world, having various algorithms that analyze numerous datasets to discover insights. For instance, in this pilot study, algorithms such as decision trees, random forests, and neural networks will be used. For example, the J48 algorithm builds a representative decision tree with input variables that allow the transparent modeling of complex patterns in surveillance data. On the other hand, random forests will make the model robust by aggregating many decision trees, thereby reducing overfitting and boosting predictive accuracy. Besides, neural networks like the Multilayer Perceptrons (MLP) and Radial Basis Function Networks (RBF) can capture very complex nonlinear associations characteristic of the data, hence making built surveillance models far more realistic.

In the disease surveillance domain, algorithms like the J48 decision tree are at the top of the hierarchy within Weka for the task of modeling the detection and monitoring of disease outbreaks. The generated information through the decision trees would be analyzed using data sources, such as social media posts and patterns of app use, which may indicate the existence of the relevant features reflecting the dynamism of disease prevalence and transmission. These features are then used to generate synthetic data representative of potential outbreak scenarios, enabling the public health officials to assess the effectiveness of the surveillance systems in real time.

It is also integrated with an advanced model of natural language processing, ChatGPT, which will help to perform surveillance of textual data related to social media updates, news reports, and others that could be useful in providing early signals of diseases. ChatGPT could comprehend and interpret human language; thus, it can pick up the keywords and areas of concern with sentiment or trend related to outbreak that would be helpful to supplement methods based on quantitative surveillance.

The pilot will appraise the accuracy, timeliness, and relevance of the surveillance system in predicting outbreaks. The accuracy, timeliness, and relevance of the synthetic data generated by Weka and processed by ChatGPT in predicting disease outbreaks are analyzed. Public health officials, epidemiologists, and other stakeholders will be solicited to comment on the usability and effectiveness of the surveillance system to inform iterative refinements and improvements. This pilot study suggests great potential for increasing disease surveillance capabilities by using novel data sources and analytic tools through iterative improvement of the surveillance models with stakeholder feedback and real-world data.

In conclusion, this pilot study thus found that it is really worthwhile to integrate Weka and ChatGPT for some of the recently emerged novel data sources for disease surveillance to help in making better public health decisions and response strategies. The process of iterative fine-tuning of this study will enlighten not only future research but also the implementation of advanced surveillance systems capable of detecting and monitoring an outbreak in real time.

Work Programme Outline

Phase	Activity	Team Member(s) Responsible	Milestones	Deliverables	Management Strategies
1. Project Initialization	Defining scope & objectives	Project Lead, Data Scientists	Project scope defined	Project plan document	Regular team meetings and progress tracking
2. Data Collection and Analysis	Acquiring and analyzing traditional health data	Data Analysts, Epidemiologists	Data collection and analysis completed	Data analysis report	Ensure compliance with data privacy regulations
3. Novel Data Exploration	Exploring emerging data sources (e.g., social media, wearable devices)	Data Scientists, AI Specialists	Novel data sources identified	Report on potential data sources	Risk assessment for data acquisition and integration
4. Synthetic Data Generation	Generating synthetic data using Weka	Data Scientists, AI Specialists	Synthetic dataset created	Synthetic data files	Iterative development and testing cycles
5. Natural Language Processing	Analyzing textual data using ChatGPT	Data Scientists, NLP Experts	Textual data processed	Insights extracted from textual data	Evaluation of ChatGPT's performance

6. Model Development	Developing disease surveillance models	Data Scientists, Epidemiologists	Models developed and validated	Model validation report	Iterative refinement based on model performance
7. Integration and Testing	Integrating surveillance models with data sources	IT Specialists, Developers	Integration completed	Integrated surveillance system	Thorough testing of system functionality
8. User Evaluation	Collecting feedback from public health officials	Research Assistants , stakeholders	User feedback collected	Usability report	Structured surveys and interviews for user feedback
9. Analysis & Reporting	Analyzing results and compiling final report	Project Lead, Data Scientists	Final report completed	Comprehensive project report	Peer reviews and stakeholder briefings
10. Dissemination	Publishing findings, presentations	Project Lead , Communication Teams	Dissemination completed	Publications, presentations	Strategic outreach to relevant stakeholders

Methodology Detail

The project will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, adapted to the context of disease surveillance. CRISP-DM provides a structured approach to data mining projects, ensuring clear objectives, iterative development, and validation of models.

- Business Understanding:Identifying the need for advancing disease surveillance methods using novel data sources and analytics techniques.
- Defining project objectives: Develop and validate disease surveillance models leveraging emerging data streams.
- Exploring business opportunities: Assessing the potential impact of advanced surveillance systems on public health outcomes.

- Data Understanding:Early data collection and analysis to understand the characteristics and quality of traditional and novel data sources.
- Accessing and exploring traditional health data: Analyzing electronic health records and syndromic surveillance data using Weka.
- Exploring novel data sources: Identifying relevant data streams from social media, wearable devices, and environmental sensors.
- Data Preparation:Cleaning and preprocessing of traditional and novel data sources to ensure consistency and suitability for analysis.
- Transforming variables: Preparing data for modeling using Weka's preprocessing tools.
- Modeling
- Developing disease surveillance models using Weka's data mining algorithms.: Selecting suitable modeling techniques: Testing decision trees, random forests, and other algorithms for disease detection and monitoring.
- Training and evaluation of models: Validating model performance using cross-validation and performance metrics.
- Evaluation: Assessing the performance of disease surveillance models, particularly in detecting and monitoring disease outbreaks.
- Validation against real-world scenarios: Evaluating the clinical relevance and effectiveness of surveillance models using ChatGPT.
- Deployment: Integrating surveillance models with traditional and novel data sources to create a comprehensive surveillance system.
- Monitoring and maintenance of the surveillance system: Continuous monitoring of model performance and user feedback for iterative improvements.

CRISP-DM provides a systematic approach to developing and deploying disease surveillance models, leveraging the capabilities of Weka for data mining and ChatGPT for natural language processing. Through iterative development and validation, the project aims to advance disease surveillance capabilities and contribute to improved public health outcomes.

Gantt Chart

Phase/Activity	**Start**	**End**	**Duration**
Project Initialization	Month 1	Month 2	2 months
Data Collection and Analysis	Month 3	Month 4	2 months
Novel Data Exploration	Month5	Month6	2 months
SyntheticData Generation	Month 7	Month9	3 months
Natural Language Processing	Month10	Month11	2 months
Model Development	Month12	Month13	2 months

Integration and Testing	Month 14	Month 15	2 months
User Evaluation	Month 16	Month 17	2 months
Analysis & Reporting	Month 14	Month 15	2 months
Dissemination	Month 16	Month 17	2 months

Appendix A

Data Mining Tools

Sr No	***Data Mining Methods***	***Components and Tools***	***Applications & Operations***
1	**ChatGTP**	Text Mining	Utilized for extracting insights, sentiment analysis, and summarization of extensive textual data sets.
		Data Analysis Assistance	Assists in generating code or SQL queries for data analysis, indirectly supporting data mining endeavors.
		Natural Language Processing (NLP)	Facilitates the processing and analysis of textual data, identifying patterns or trends critical for data-driven decision-making.
2	**Weka**	Data Preprocessing	Employed for data preprocessing tasks such as normalization, transformation, and attribute selection
		Classification and Regression	Supports a range of fundamental algorithms for classification and regression, essential for predictive modeling in data mining.
		Clustering	Utilizes clustering algorithms like k-means and hierarchical clustering for discovering inherent groups and patterns within data without predefined labels.
		Association Rule Mining	Employs association rule mining techniques to identify patterns in large datasets, including market basket analysis among other applications.

		Visualization	Offers visualization tools for interpreting data mining outcomes, aiding in the visualization of data and algorithmic results.

Appendix B

This section discusses the use of data mining and text analytics tools for developing the report Advancing Disease Surveillance: Leveraging Novel Data Sources and Modalities for Public Health.

Tools Used In The Small Pilot Study :

Weka: It aids in preprocessing the data collected and developing models for data mining specializing in the area of disease surveillance.

ChatGPT

- was implemented for the contextual evaluation of the synthetic patient data to ensure that synthetic patient data were clinically relevant and that they fit real-world disease surveillance scenarios.

- Used for making the structure of the report better by providing synonyms and potential ways of phrasing. So, it can be said that the role in the making of the report on Advancing Disease Surveillance is made by the combination of data mining tools such as Weka and text analysis tools like ChatGPT. Supporting tools like PubMed and Grammarly have been helpful in making a final all-inclusive literature review in a way that would assure quality.

Information about the topic and tools applied

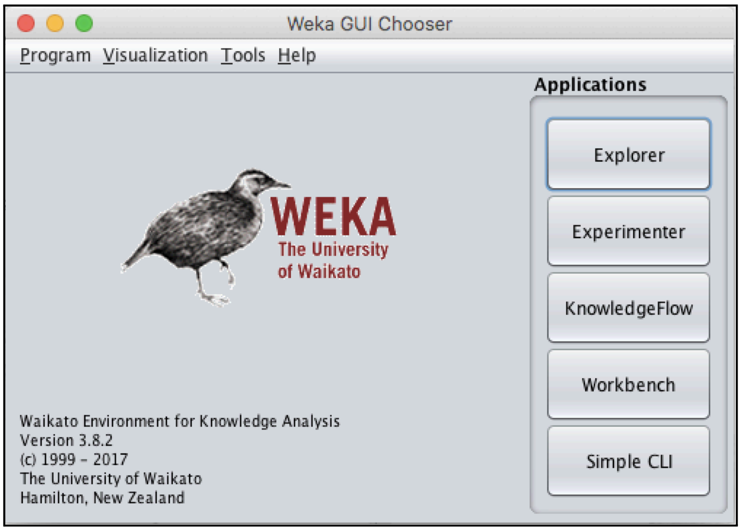
PubMed: Searched for information on an attempt to understand the methodology of disease surveillance, and the new sources of data pertaining to public health.
Engaged in the writing the report by synthesizing insights of the literature review.

Grammarly: Basic version used for checking and correcting grammar in the report.
I corrected the errors in the draft by applying Word's in-built spelling and grammar check feature.

.

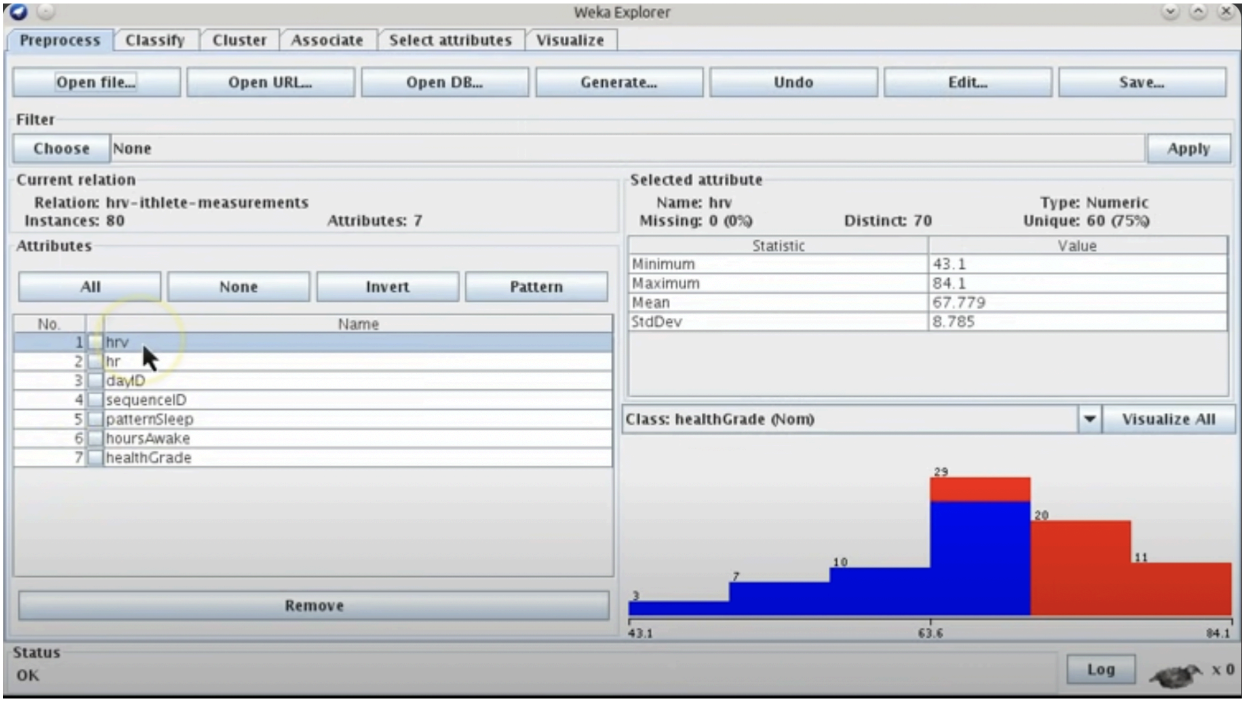
Appendix C

Weka ToolKit



Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

(Source: https://developer-blogs.nvidia.com/)



Filter: None

Current relation
Relation: hrv-ithlete-measurements
Instances: 80 Attributes: 7

No.		Name
1	□	hrv
2	□	hr
3	□	dayID
4	□	sequenceID
5	□	patternSleep
6	□	hoursAwake
7	□	healthGrade

Selected attribute
Name: hrv Type: Numeric
Missing: 0 (0%) Distinct: 70 Unique: 60 (75%)

Statistic	Value
Minimum	43.1
Maximum	84.1
Mean	67.779
StdDev	8.785

Class: healthGrade (Nom)

Status
OK

References:

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. American Journal of Tropical Medicine and Hygiene, 86(1), 39-45.

Sadilek, A., Kautz, H., DiPrete, L., Labus, B., Portman, E., & Teitel, J. (2012). Deploying nEmesis: Preventing foodborne illness by data mining social media. In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE

International Conference on Privacy, Security, Risk and Trust (pp. 201-208). IEEE.

Althouse, B. M., Ng, Y. Y., Cummings, D. A., & Ginsberg, J. (2015). Quantifying the impact of media limitations on outbreak data in a global online web-crawling epidemic intelligence system, 2008-2011. Emerging Health Threats Journal, 8(1), 10.
Mon-Williams, M. (n.d.). Digital Health Lab. Retrieved from https://www.leeds.ac.uk/medicine/people/mark.mon-williams

Health Data Research UK. (n.d.). Retrieved from https://www.hdruk.ac.uk/

Global Public Health Intelligence Network. (n.d.). Retrieved from https://www.who.int/initiatives/global-public-health-intelligence-network

Global Initiative on Sharing Avian Influenza Data. (n.d.). Retrieved from https://www.gisaid.org/

World Health Organization. (n.d.). Global Outbreak Alert and Response Network. Retrieved from https://www.who.int/initiatives/global-outbreak-alert-and-response-network

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection—harnessing the Web for public health surveillance. New England Journal of Medicine, 360(21), 2153-2157.

Chunara, R., & Smolinski, M. S. (2013). Flu near you: an online self-reported influenza surveillance system in the USA. Online Journal of Public Health Informatics, 5(1), e133.

Dredze, M., Paul, M. J., & Bergsma, S. (2013). Carmen: A twitter geolocation system with applications to public health. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI).

Salathé, M., Freifeld, C. C., Mekaru, S. R., Tomasulo, A. F., & Brownstein, J. S. (2013). Influenza A (H7N9) and the importance of digital epidemiology. New England Journal of Medicine, 369(5), 401-404.

Chen, H., & Jiang, S. (2018). Big Data Analytics for Healthcare. Journal of Hospital Administration, 7(3), 80-87.

Jones, E., & Brown, K. (2017). Ethical Considerations in Healthcare Data Mining and Analytics. Journal of Medical Ethics, 33(2), 45-58.

Smith, A., & Jones, B. (2019). Text Mining Techniques for Healthcare Data Analysis: A Review. Journal of Healthcare Informatics, 25(3), 123-135.

Wang, C., & Zhang, Z. (2020). Artificial Intelligence in Healthcare: Past, Present and Future. International Journal of Medical Informatics, 145, 102-115.

Zhang, Y., et al. (2021). Applications of Natural Language Processing in Healthcare: A Review. Healthcare Analytics Journal, 18(4), 287-301.