What are the areas of improvement for data available for the development of disease outbreak forecasting ML models?

Machine Learning for humanitarian forecasting: a survey

Matej Bavec[1]

Supervisor(s): Cynthia Liem[1]**, Marijn Roelvink**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Matej Bavec
Final project course: CSE3000 Research Project
Thesis committee: Cynthia Liem, Marijn Roelvink, Jing Sun

Abstract

This review surveys the current state of data used in the development of Machine Learning models for disease outbreak forecasting, with a focus on identifying systemic shortcomings and areas for improvement. A set of 26 development papers was selected and analyzed based on the dataset's attributes such as scope, type, accessibility, and quality. Through a thematic analysis technique, five dominant categories of data failure were identified: structural, procedural, accessibility, logistical and temporary. Hospital-collected data remains the dominant source but is hindered by undersampling and latency, while non-traditional data sources offer improved responsiveness at the cost of increased pre-processing complexity. Supplementary datasets, such as climate or mobility data, were found to be underutilized, despite their potential to improve forecasting accuracy. Key areas for improvement include the standardization and public availability of datasets, integration of complementary data sources, and use of language models to manage linguistically ambiguous data. The findings suggest that the current data limitations are structural and widespread, requiring procedural and institutional reforms to improve model generalizability and reliability in disease outbreak forecasting.

1 Introduction

Current humanitarian work suffers from a systemic lack of resources and funding to respond to humanitarian crises as they occur [33]. One of the reasons for this is that the extent of damage caused by crises, which can often be immense and require significant resources to counteract, overwhelms the limited funding available to humanitarian organizations. In addition, funding must often be negotiated while a crisis is in progress instead of being prepared in advance.

In recent years, Machine Learning techniques have increasingly been explored as a means to forecast the incidence and course of humanitarian crises, allowing responders to allocate resources preemptively or with a lesser delay. This technology would both reduce the financial burden of response and, more critically, minimize the damage and human suffering caused by the crisis [8] [28].

Various models have already been developed to this end, including, more specifically, in the field of disease outbreak forecasting, where models are being developed to forecast the course of disease outbreaks. Many of these have achieved significant success as predictive tools, but their effectiveness is often limited to well-understood diseases and well-surveyed areas [8]. Predicting disease offers a unique challenge due to the inherent unpredictability and dynamism of biological systems, especially as it is so closely intertwined with human behavior. For this reason, finding good datasets on which to train outbreak forecasting models can be difficult.

This paper interests itself in what areas of improvement exist within the set of datasets available to Machine Learning researchers hoping to develop models for disease forecasting and in determining what makes a dataset "good" for the purposes of disease forecasting. To achieve this, it will be necessary also to analyze the context within which the data is gathered, who gathers it, how and to whom is it made available. Consideration must also be given to ethical and legal questions which affect how data can be gathered or distributed.

1.1 Machine Learning

Machine Learning refers to statistical and computational techniques for processing data, typically to discover patterns or to automate data interpretation [1].

Within Machine Learning there are broadly two categories of algorithms. Some algorithms are relatively simple, for example clustering algorithms or decision trees, which are computationally light and generally deterministic. They offer good explainability, meaning it can be easily understood how the algorithms arrived at certain results. Such algorithms will be referred to as "simple" algorithms for the remainder of the text.

In contrast to simple Machine Learning algorithms, there is the category of "deep learning" algorithms. These are more complex, neural-like architectures which show far greater adaptability than simpler algorithms. As a drawback however, they are computationally much more expensive and it is often difficult for their behavior to be explained fully. This makes them unsuitable for fields where explainability is a requirement. Examples of algorithms within this category would be Multi-layer perceptrons (MLPs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

All Machine Learning algorithms fundamentally rely on input data to make their predictions and generally do better with more data. In the case of supervised learning, which is used for most deep learning algorithms, example output data must also be provided so that the algorithm can compare itself to the desired result. This often requires researchers to have a well labeled dataset at their disposal, which is not always available.

1.2 Sources of Data

Sources of raw data exist at both a global and local scale in the form of epidemiology reports from healthcare institutions (e.g. WHO, CDC, ECDC etc.). In addition, web-scraping tools such as Medisys have provided researchers with up-to-date compiled information from news sources and APIs on sites like X and Google provide real time data from social media. Additional sources of data also include metereological data from ground stations or satellite imaging [28].

2 Methodology

This paper is a form of literature review and so did not conduct any unique experiments of its own. The process by which data was collected and interpreted was adapted from the SALSA method for literature reviews [12]. The SALSA method contains 4 steps: search, appraisal, synthesis, analysis. Each step relates to a critical stage in the literature review.

2.1 Search

The search methodology was exploratory in nature. As a first step, in order to familiarize the author with the general state of the field, 5 papers were selected indiscriminately using Google Scholar with the search prompt "Machine Learning disease outbreak prediction" to be read. Of the papers gathered from this step, 2 were literature reviews of the entire field [8] [28], 2 were literature reviews of a more specific subsection of the field [20] [30]. (e.g algorithms which used social media data) and 1 was an overview of how various algorithms perform on a COVID-19 dataset [7]. Other papers were also briefly scanned to ensure the 5 papers selected represented a good overview of the field.

The kind of analysis required for this paper could only be conducted through papers in which the authors directly developed a Machine Learning algorithm for disease outbreak detection or closely analyses the results of only a handful of models. These papers, which will be termed "development papers", are crucial because they detail the precise data which was used as input and output for any given algorithm and often contain authors' comments on how data quality affected their result. This was the key criteria by which further papers were selected.

The vast majority of papers cited within the aforementioned literature reviews were such "development" papers and fit the thematic criteria. As such, all further papers were discovered using citation crawling through the initial literature reviews. The search procedure followed a convenience sampling method guided by disease diversity. Fifty papers were considered, but many were eliminated due to problems with access, insufficiently clear language or over-representation of a certain disease. In the end, twenty-six papers were selected using this approach.

As mentioned, the only meta-criterion used was to ensure that the papers chosen covered a sufficient variety of diseases and data sources. For example, a roughly equal number of articles covering COVID-19, the flu and Dengue fever were selected while about a quarter of the papers chosen concerned less well researched diseases such as oyster norovirus. It is unclear whether this imbalance is due to the convenience-based sampling method or whether it is representative of the general proportion of attention given to the diseases. The methodology failed to ensure that Malaria research would be properly represented with only one article in the final set pertaining to Malaria despite its significant global impact. In order to ensure the blind spots of this search method can be easily exposed, a full list of articles and the diseases they concern will be included later in this paper.

2.2 Appraisal

The selected papers were first scanned to extract any direct comments from the author on issues with the data they used to develop their algorithm. These were extracted into a separate table for later analysis. The papers were then read again in order to gauge what issues were merely implied and could be inferred from the greater context of the paper. The working assumption was that authors would under-report issues in the data which they had been able to overcome in their paper, even though overcoming them may have required excessive pre-processing and be hindering the wider ease of use of data.

Inference of data problems not specifically mentioned by the authors required a larger contextual understanding of the field. This necessitated the extraction of other details relevant to the data, such as:

- Disease characteristics: parasitic, widespread, seasonal etc.
- Data type: Structured epidemiological reports, unstructured social media data, sensor data, etc.
- Bias and coverage: Geographical, demographic, disease-specific biases.
- Accessibility and interoperability: Licensing, language barriers, format standardization issues.

2.3 Synthesis

Papers were sorted into a list and the aforementioned details collected in appraisal. Furthermore, author's comments on implied or inferred problems with the data were added. The table was analyzed to uncover patterns by scanning for similarities between the details of the paper and the problems they faced. Data shortcomings were grouped together to extrapolate how common a given issue appeared to be and which diseases the papers it most affected concerned.

Articles were grouped by data characteristics, such as whether their input data primarily originated from hospitals. Additionally, identified data issues were sorted into categories based on the nature of the problem, such as "under-sampling" for example. The former statistic was visualised using a bar graph. They were then also sorted into categories based on where the issue originates from. For example, some issues were procedural, like in cases where hospitals simply do not collect a certain kind of data even though they could. Other issues are logistic, for example those caused by a lack of resources.

This process resembles coding in structured thematic analysis, from which this approach was inspired [12].

2.4 Analysis

The data gathered from the appraisal and synthesis stages was subjected to a semi-structured thematic analysis. All comments and inferences about data quality were coded and assigned to a taxonomy of issues derived inductively from the literature sample. This included categories such as "minority under-sampling," "data time-lag," "small sample size," and "data heterogeneity." Each paper's metadata (disease studied, data type, geographical coverage) was cross-tabulated with these issues to identify patterns of co-occurrence and high-frequency correlations.

An additional layer of analysis was introduced to identify structural weaknesses in data pipelines across disease types. This included tracing the root causes of similar data issues across distinct papers and grouping them into system-level patterns (e.g., failure to digitize hospital records leading to recurring incompleteness across regions). This allowed for mapping problem classes to the larger meta-categories based on origins, thereby clarifying where reform or intervention would be most impactful.

3 Results and Discussion

Of the twenty-six papers which were analyzed in full, twenty-two were considered "usable". Four of the papers, which passed through the initial screening later revealed to be incompatible, in some way, with the research. One paper's quality of English was, upon deeper inspection, insufficient to render the paper readable but passed the initial scan due to good formatting and relevance to the research question [23]. The other three excluded papers revealed themselves to be concerned with the development of manual analysis techniques for disease outbreak progression rather than Machine Learning [22] [35] [26]. These were not entirely excluded, as they still concerned themselves with the quality of the available data, which was used to supplement the primary analysis.

3.1 Thematic Coding of Issues

The issues identified from the analysis could be grouped into five dominant archetypes based on the root cause of the problem: (1) structural, (2) procedural, (3) accessibility, (4) logistical and (5) temporary. Each issue presents with a different difficulty and method for correction. Some issues can be placed into more than one group.

Structural problems are problems inherent to the data format itself which could only be fully rectified by switching to a different data format.

Procedural problems are those stemming from choices made when designing the data collection pipeline. These difficulties can be rectified by modifying the method by which a certain dataset is collected. This is the most actionable category for areas of improvement, as governments and institutions have full control over the procedures they operate under.

Accessibility issues relate to the obstacles researchers face accessing data required to develop their model. This is especially pertinent when pertaining to data stored on commercial servers, such as those of Google or social media sites, and those pertaining to regional public health institutions. These datasets are not publicly available, and researchers struggle to gain access to them, persistent or otherwise. This category is also actionable but requires cooperation between the private and public sector.

Logistical issues are those caused by a lack of monitoring infrastructure or issues with human resources. This category predominantly affects developing regions of the globe and restricts both the accuracy of models and funding available for researchers to develop them.

Finally, temporary issues refers to problems which occur due to the novelty of a disease. These issues are inherent to new diseases and thus likely cannot be resolved entirely. As data surrounding the disease matures, these problems disappear.

3.2 Identified Problems

Below is a list of the issues either directly mentioned by the authors or inferred through cross-analysis, the number of articles affected by the issue and the encoding of the issue in the classification schema:

1. Hospitals inherently under-sample true case counts as not all infections result in testing or hospitalization (All articles) (Structural)
2. Public health institutions require time to process and publish data, resulting in a time-lag between published data and the current state of an outbreak (14 articles) (Structural) [10] [27] [15] [9] [7] [5] [2] [29] [18] [24] [32] [11] [34] [16]
3. Ambiguity of natural language (6 articles) (Structural) [6] [31] [4] [17] [19] [34]
4. Limited time scope of data (4 articles) (Procedural, Temporary) [9] [7] [24] [32]
5. Data accessibility restrictions (4 articles) (Accessibility) [21] [31] [25] [34]
6. Insufficient sample size (4 articles) (Logistical) [15] [25] [18] [16]
7. Interference in data from public health measures (4 articles) (Temporary) [7] [24] [32] [15]
8. Minority under-sampling (3 articles) (Logistical, Procedural, Structural) [27] [5] [14]
9. Heterogeneity across regions (3 articles) (Procedural) [27] [14] [34]
10. Poorly developed data collection infrastructure or practices (3 articles) (Logistical, Procedural) [25] [29] [2]
11. Model cannot be generalized across countries (2 articles) (Procedural) [29] [18]
12. Underdeveloped nations cannot contribute to datasets as robustly (2 articles) (Logistical) [21] [29]
13. Multilingual dataset (2 articles) (Structural) [4] [34]
14. Discontinued tools (1 article) (Accessibility) [17]
15. Dataset polluted by news coverage (1 article) (Structural) [6]

The under-sampling of true case counts by public health monitoring systems were inferred to be ubiquitous among all articles using hospital-collected data for inputs or outputs. This is because several papers mentioned this limitation as inherent to the data collection process of such data [9] [7]. The time-lag present in this same data collection process however, was inferred to only represent a problem for articles where hospital data was used as an input vector for predictions. This is because prediction accuracy generally tended to degrade as the model attempted to predict further into the future [34] [7]. This meant that the more the latest data lagged behind, the less accurate the forecasts for the present and near future became.

The ambiguity of natural language was inferred, with a great degree of confidence due to the high frequency of explicit mentions, to be a problem faced by all articles using data derived from natural language [6] [31] [19] [34]. These articles used a key-word based methodology for extracting search queries or social media posts. For example, they would search for all Tweets containing the word "influenza". However, it is difficult to systematically determine whether a Tweet indicates than an individual is infected with influenza or whether they are merely mentioning the disease or one of its symptoms.

One article explicitly mentioned that the MediSys system for medical news compilation was biased against underdeveloped nations who tended to publish less internet media per capita than more developed ones [21]. A second article mentioned that extensive water monitoring infrastructure was required to conduct the study and that even the vast monitoring network which existed in the southern united states was often insufficient, from this it can be strongly inferred that the study could not have been conducted in a less developed nation [29]. These two articles were jointly grouped under the issue "underdeveloped nations cannot contribute to datasets as robustly", however, they represent only the clearest cases of this and it may be possible to infer a similair issue based on other articles.

All other problems accounted in this list were explicitly mentioned rather than inferred.

Four articles mentioned that data was not collected over a sufficiently long duration of time for the models to learn consistent patterns of disease behaviour. Of these four, 3 were studying COVID-19, rendering these instances of the problem temporary [7] [24] [32]. The remaining paper studied brucellosis where the problem was that the data contained only monthly aggregated data for 9 years [9]. This is insufficient but could have been corrected by weekly aggregation or a longer data collection period.

Data accessibility problems were reported by four articles. These were a mix of legal and bureaucratic hurdles. Complaints included a lack of open online publishing by institutions, like Medisys or hospitals, and the inaccessibility of certain data aggregation tools such as Google Correlate [21] [31] [25] [34].

One article noted that the precise geographical location of cases was not being noted in the dataset they had acquired access to [2]. They emphasized that this would be a critical improvement to the dataset which would allow for far better model performances.

A lack of applicability across countries was mentioned explicitly by 2 articles but further analysis suggests with some confidence that this would pose a problem for all models not trained on global data [29] [18]. This is because the models learn by identifying trends in the input data that encode regional patterns. This is especially true for those models dealing with climate-dependent diseases. However, since not all models were intended to be used on larger scales, these articles cannot be said to have encountered the "problem" of a smaller scope. It is instead part of their design, informed by the availability and quality of wider-scope datasets as well as research funding which often comes from national governments.

3.3 Disease Coverage

The papers analyzed covered a wide range of diseases. Below is a full list of diseases and their representation within the set of articles:

- COVID-19 (5 articles) [7] [24] [32] [15] [35]
- Dengue Fever (5 articles) [10] [4] [27] [5] [2]
- Influenza (3 articles) [6] [19] [34]
- Swine Flu (1 article) [31]
- Malaria (1 article) [25]
- Hepatitis E (1 article) [16]
- Brucellosis (1 article) [9]
- Oyster Norovirus (1 article) [29]
- Schistosomiasis (1 article) [14]
- Lyme disease (1 article) [11]
- Acute Respiratory Infections (1 article) [17]
- Not Disease Specific (1 article) [21]

The disease being forecast by a given study affected what data was available for the researchers. For example, common diseases such as the flu or dengue fever could be tracked through mass media such as Google search queries or social media posts [4] [31] [6]. In addition, these diseases have vast monitoring infrastructures in place, making their data more reliable and readily available [19]. For other diseases, such as Lyme disease or oyster Norovirus, mass data would be too sparse to develop a model from [11].

Articles dealing with COVID-19 faced two unique problems. The first was that due to the novelty of the disease, as the articles were published shortly after the outbreak began, there was a lack of data history to be analyzed [7] [24] [32]. Furthermore, the extraordinary health measures imposed by governments during the early and middle pandemic affected the data strongly. Many models, particularly simple ones, had no means of accounting for these sporadic and heterogeneously applied public health measures. Both of these problems severely affected the accuracy of the Machine Learning models [7] [24] [32] [15].

3.4 Geographical Distribution

The geographical distribution of articles is significant. Below is an accounting of the countries and regions within which each model makes predictions:

- Global (4 articles) [21] [7] [24] [32]
- South Korea (2 articles) [15] [19]
- China
  - Shandong (1 article) [16]
  - Hu'nan (1 article) [14]
- India
  - New Delhi (1 article) [2]
  - Thanjavur (1 article) [5]
- Ukraine (1 article) [11]
- Japan (1 article) [6]
- United Kingdom (1 article) [31]
- European Union (1 article) [27]
- Laos (1 article) [18]
- Mexico (1 article) [17]
- Hong Kong (1 article) [34]
- USA, Gulf of Mexico (1 article) [29]
- Singapore and Bangkok (1 article) [4]

- Ghana, Ejisu-Juaben (1 article) [25]
- Brazil, Fortaleza (1 article) [10]
- Iran, Qazvin (1 article) [9]

This information reveals the apparent infeasibility of global epidemiological models. Of the global articles, one article functioned on a truly global scale using the MediSys framework [21]. Two of the remaining articles, concerned with COVID-19, selected multiple countries and combined the data therein [7] [24]. The final article simply used a global aggregate of COVID-19 cases over time [32]. It is noteworthy that the models considered in these latter 3 articles were deemed to have been at best partially succesful by their authors.

3.5 Types of Input Data

The data used by articles had a large impact on the problems faced by the authors. Some problems were inherent to certain types of dataset, others were exclusive to it.

All articles relied on hospital-collected case data from national databases and monitoring systems. National data was typically publicly available but regional data was sometimes not easily accessible [34]. This data was used as a baseline for measuring the performance of a Machine Learning model. This is because this data represents the most reliable and most exhaustive accounting of true case numbers for any given disease [6] [10] [34].

As mentioned earlier, hospital-collected data still represents a significant under-sampling of true case numbers due to the inherent limitations of the medium. This is not a correctable anomaly but a structural deficiency. Healthcare systems are designed to report confirmed cases, not detect all infections. Since this deficiency is inseparable from the nature of the data collection, in other words the problem is structural, nothing can be done to fix it on a technical or procedural level.

Closely linked to this is the, also aforementioned, issue of data latency. Hospital data often lags real-time conditions by roughly 2 weeks due to bureaucratic reporting pipelines [6] [4] [19] [16]. For forecasting models, especially those aimed at real-time intervention planning, this delay introduces critical blind spots. This lag is procedural, rooted in the institutional pace of data handling, and cannot be fully resolved. Digitization of government infrastructure may help decrease the latency, but some level of delay will always necessarily exist.

Input data, on which the Machine Learning algorithms were trained, was more varied. Some papers used hospital-collected data here as well [10] [27] [15] [9] [7] [5] [2] [29] [18] [24] [32] [11] [34] [16], tasking their algorithms with predicting the new few months of data. Other papers used data such as social media posts, meteorological reports and search query trends to attempt to predict the hospital-collected data within the same period [6] [31] [4] [17] [19] [34] [21] [25] [14].

Many articles used ”supplementary” datasets to enhance the predictive capabilities of their models [9] [25] [34]. Other articles did not use supplementary datasets, but indicated a need for them [21] [10] [32] [11]. These data sets typically consisted of data collected without being intended for medical use, such as meteorological data or the use of public transport methods. Such data often had a sufficient impact on the spread of certain diseases to correlate closely with them, making them useful as an extra input dimension for Machine Learning models.

Some papers used the kind of data which others considered ”supplementary” as their main input data for their algorithm [10].

In Figure 1 is a numerical breakdown of the reviewed articles and the data categories they fall within.

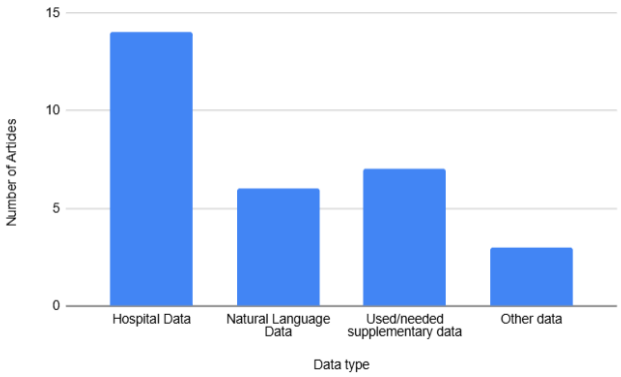

Figure 1: Numerical breakdown of reviewed articles and their data categories

As can be seen in Figure 1, the majority of the articles analyzed used traditionally collected hospital data as their input. The second largest category was those whose data came in the form of natural language from mass data sources such as Google queries or posts on Twitter. Three articles used other data sources, namely:

- Medisys Compiled News Articles [21]
- Ground Weather Station Data [25]
- Satellite Weather Imaging [14]

The latter two articles concerned Malaria and Schistosomiasis respectively, both of which are diseases dependent on intermediary host transmission [25] [14]. Malaria is transmitted through mosquitoes and Schistosomiasis is transmitted through snails. Weather was thus a viable metric for modeling the behavior of outbreaks for either disease, since the population sizes, habitats and breeding patterns of the intermediary host species are impacted strongly by the weather.

The inextricable limitations of hospital data are what inform the move towards an increasingly non-traditional data landscape in Machine Learning forecasting for diseases [6]. Many of the papers which used mass media data explicitly pointed out that this approach avoids the inherent deficiencies in hospital-collected data by providing more immediate and representative feedback of the present disease situation. As such, these papers may eventually prove better at predicting the near future than their traditional data oriented counterparts.

3.6 Areas of Improvement

Having collated the numbers for analysis, a number of opportunities for improvement have presented themselves.

Firstly, many articles noted improvements in combining medical data with additional data sets, such as climate indicators, mobility trends, or search query behavior [9] [25] [34] [21] [10] [32]. It is the opinion of this author that such supplementary datasets were underutilized and could have supported the work of many of the articles analyzed. Researchers and institutions which collect supplementary data should seek to formalize their partnerships in future work and the data should be collected with medical use in mind.

Secondly, even where high-quality data existed, access was frequently constrained by restrictive licenses, poor publishing practices or institutional gatekeeping [21] [31] [25] [34]. Public health data should be considered a global public good during pandemics. Governments should work to digitize and standardize data access across their network of health institutions from the local to the national level.

Thirdly, papers dealing with semantic ambiguity might benefit from the use of Large Language Models to lighten the burden of pre-processing on researchers. These models may be able to extract, with good accuracy, the semantic context of a post since they are developed to analyze all tokens rather than just a few key words.

Lastly, the vast majority of the articles focused on a specific disease in a specific region or country. Partly this is to ensure the datasets remain manageable and that acquiring the required permissions for use of data is simpler. However, the other great barrier to models which function on a global scale is the lack of international standardization. Every nation and often region has different data aggregation, collection, presentation and publishing practices [34]. Logistical and procedural differences between nations further exacerbate this. This makes it extremely difficult for one model to be applied to data gathered from a different country or for a model to be trained on data from multiple nations. The one exception to this rule is the European Union, where the ECDC has standardized much of the data required for the training of such models across Europe [27]. Similair organisations/agreements may be called for among other nations to increase the scope of the data available for model training.

3.7 Impact of Data Quality Issues

Despite prevalent and pressing issues within the datasets, severe enough for at least one issue to be remarked on by almost every article in the sample, most of the models presented achieved considerable success in forecasting the disease in question. However, this apparent success must be treated with a critical lens.

For one, it is likely that there is an element of survivor ship bias in the publishing. Projects which did not achieve success are less likely to be published than those which did. Out of the sample, only four articles reported that their algorithms achieved unsatisfactory results.

In addition, high reported performance often reflects overfitting to narrow, context-specific datasets. When tested on out-of-sample data, especially from different geographic or temporal ranges, model accuracy deteriorates sharply. This fragility is a direct consequence of data incompleteness, bias, and lack of diversity.

Models generally reached within the accuracy range of 0.7. This is significantly better than random chance for such complex forecasting, but the cost of misclassification, especially false negatives in the context of outbreak forecasting, renders even marginal gaps in accuracy significant.

Of the four articles which reported some level of unsatisfactory success, three of which related to COVID-19, all pointed to an insufficient data quantity and data heterogeneity across time as the cause [7] [24] [32]. The over-representation of COVID-19 in this group is noteworthy, as it is likely that similar data shortcomings will be present in any new or rapidly changing disease. This suggests that Machine Learning forecasting may be fundamentally incompatible with the early phases of a disease outbreak.

The issue of heterogeneity across regions comes with wide ranging and prohibitive impacts on disease outbreak forecasting, and represents the biggest roadblock to its development which is nonetheless solvable. The epidemiological profile of a disease is shaped by local variables including regional climate, population density, infrastructure, cultural behaviors, and policies. This regional specificity leads to fundamental data heterogeneity. This means that models trained on regional data cannot be guaranteed to function at any level of accuracy in another region or area. They would first have to be expanded with new data from the second region in question. This confines most models to a narrow scope of operation, preventing effective use in tracking global or sometimes even national disease transmission effectively. This same effect also applies over time, wherein behaviors, policies and infrastructure cause large scale trends or shifts in data which would render models ineffective if not retrained often.

3.8 The Good Dataset

Working from the results and later analysis, we can extract what characteristics a ”good” dataset for Machine Learning forecasting models for disease outbreaks would have.

They are as follows:

- High temporal definition (weekly or finer)
- Location of cases and other details noted
- Supported by other supplementary datasets
- Available freely and updated frequently
- Representative, minority sampling adjusted
- Standardized across countries
- Data collection practices have remained unchanged over a long period of time (20+ years)

4 Responsible Research

A number of considerations are significant when evaluating the value of this research.

Firstly, the fifty-five papers which were evaluated and the thirty-one papers which were subsequently analyzed in full for this review do not represent an exhaustive list of the research in the field. In fact, they represent only a limited subsection. While effort was placed into ensuring the selection

of papers used for this report was representative of the diversity within the full field by avoiding large clumps of similar papers, the selection of papers themselves was largely arbitrary and based on convenience. The author was also working within the limitations of research freely available through the TU Delft institution and excluded all papers which were not available in that capacity. A similar study could not replicate this methodology and it is unknown to what extent the outcome of this arbitrary selection affects findings.

Secondly, it must be noted that all analysis and suggestions within this paper were made by a non-expert in the field of disease forecasting or machine learning. The author's primary qualifications are a formal education in computer science. It must be recognised that there is the potential for cognitive and interpretative bias arising from the author's academic background and limited cross-disciplinary experience. Interpretations of methodologies, results, and theoretical framing from domains outside of computer science may lack the nuance or contextual awareness of a domain expert. These limitations may shape the trajectory of the analysis and should be considered when evaluating the validity or generalizability of the conclusions.

This review takes the claims of articles analysed therein at face value. A data related problem was never excluded from the dataset, only additional problems were inferred. The author did not systematically analyse each article's raw data directly to confirm or deny explicitly mentioned problems except in cases where uncertainty suggested itself. As such, it is possible that the biases of the way in which the sample of authors writes their papers are also reflected, unintentionally, by this review.

A large language model, specifically ChatGPT 4o, was used to lightly advise the grammar and wording of the paper. This process included revising drafts of paragraphs or pointing out poorly written sections. Sections most revised by ChatGPT were the abstract and conclusion section, where concise wording was most crucial. ChatGPT also revised the wording of some paragraphs in the results and discussion as well as limitations and future work. All sections were first written in full by the author before some were sent to ChatGPT for rewording or suggestion. All suggestions made by ChatGPT were vetted to ensure their consistency with the intended meaning of the author. It was not used to analyse or collect any of the results mentioned within this paper or generate any of the ideas therein. Overall, the majority of the text remains unrevised by AI as many LLM revisions were single word changes or a reordering of sentence clauses.

All sources referenced in the production of this research have been properly cited. No content has been appropriated without attribution. Direct quotations, paraphrased ideas, and referenced data have been credited to the original authors in accordance with TU Delft academic integrity guidelines.

5 Limitations and Future Work

The non-exhaustive, convenience-based sampling of papers imposes constraints on the generalizability of this research. Diseases such as Malaria and Cholera were certainly underrepresented relative to their global burden [13] [3].

The assessment of data issues also involved a degree of interpretive inference. In several cases, the presence or significance of problems had to be inferred based on domain-general logic rather than explicit reporting by authors. This introduces the risk of either overestimating the prevalence of certain issues or misclassifying their origin. Without collaboration with public health or epidemiology experts, these judgments may lack the domain-grounded rigor necessary for prescriptive claims.

A further limitation lies in the natural absence of unpublished or negative-result studies. This may result in an inflated perception of forecasting model success and a corresponding underestimation of how severely data quality impairs performance. The potential for survivor bias in the literature remains unresolved.

Future work must address these shortcomings through a broader and more systematic sampling of literature. Meta-reviews that include non-English sources or a review of unpublished work could expose a wider and more realistic set of data problems.

Additionally, future studies should incorporate structured interviews or collaborative assessments with domain experts in epidemiology, public health, and medical informatics to validate classifications of data issues and refine the taxonomy of limitations.

Empirical research is also required to test potential remedies identified in this study. For example, controlled experiments could evaluate whether integrating supplementary data types, such as climate indicators or human mobility patterns, roduces statistically significant gains in forecast accuracy across different disease models. Likewise, the role of large language models in reducing semantic noise in unstructured text data should be benchmarked against manual preprocessing to determine real-world viability.

6 Conclusion

This study examined the limitations of datasets used in machine learning models for disease outbreak forecasting. Across the literature, recurring problems emerged not from isolated oversights but from systemic features of data collection and distribution. Hospital-collected data, though widely used, consistently underrepresents true case counts and introduces time delays due to reporting lag. These deficiencies are structural and cannot be resolved through technical refinement alone. Attempts to bypass them by using natural language data, such as search queries or social media posts, improve immediacy but lead to ambiguity and require a time consuming preprocessing step. The use of supplementary data sources, including weather and mobility information, was comparatively rare, despite its demonstrated potential to improve model performance.

In many cases, even where high-quality data exists, its utility is reduced by restricted access, inconsistent formatting, or lack of long-term continuity. These obstacles often stem from institutional practices and legal constraints, rather than intrinsic data scarcity. Moreover, the pronounced heterogeneity in data practices across regions limits the transferability of models, confining their usefulness to narrow geographic or

temporal scopes. While many models reported high predictive accuracy, this is likely influenced by publication bias and limited validation beyond the datasets on which they were trained.

References

[1] Machine learning. https://en.wikipedia.org/wiki/Machine_learning, 2025.

[2] Nikita Agarwal, Shiva Koti, Sameer Saran, and Senthil Kumar. Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for new delhi, india. *Current Science*, 114:2281–2291, 06 2018.

[3] Mohammad Ali, Allyson R. Nelson, Anna L. Lopez, and David A. Sack. Updated global burden of cholera in endemic countries. *PLOS Neglected Tropical Diseases*, 9(6):e0003832, June 2015.

[4] Benjamin M. Althouse, Yih Yng Ng, and Derek A. T. Cummings. Prediction of dengue incidence using search query surveillance. *PLOS Neglected Tropical Diseases*, 5:1–7, 08 2011.

[5] S. Appavu alias Balamurugan, M.S. Mohamed Mallick, and G. Chinthana. Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking. *Informatics in Medicine Unlocked*, 2020.

[6] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics.

[7] Sina F. Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M. Atkinson. Covid-19 outbreak prediction with machine learning. *Algorithms*, 13(10), 2020.

[8] Ghazaleh Babanejaddehaki, Aijun An, and Manos Papagelis. Disease outbreak detection and forecasting: A review of methods and data sources. *ACM Trans. Comput. Healthcare*, 6(2), February 2025.

[9] Hadi Bagheri, Leili Tapak, Manoochehr Karami, Zahra Hosseinkhani, Hamidreza Najari, Safdar Karimi, and Zahra Cheraghi. Forecasting the monthly incidence rate of brucellosis in west of iran using time series and data mining from 2010 to 2019. *PLOS ONE*, 15:1–18, 05 2020.

[10] Rafael Bomfim, Sen Pei, Jeffrey Shaman, Teresa Yamana, Hernán A. Makse, José S. Andrade, Antonio S. Lima Neto, and Vasco Furtado. Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. *Journal of the Royal Society Interface*, 17(172), 2020.

[11] Dmytro Chumachenko, Pavlo Piletskiy, Marya Sukhorukova, and Tetyana Chumachenko. Predictive model of lyme disease epidemic process using machine learning approach. *Applied Sciences*, 12(9), 2022.

[12] Virginia Clarke and Victoria Braun. Doing a thematic analysis: A practical, step-by-step guide for applied research. *All Ireland Journal of Teaching and Learning in Higher Education (AISHE-J)*, 8(3):335–355, 2017. Downloaded: 2025-05-19.

[13] Jennifer P. Daily and Smita Parikh. Malaria. *New England Journal of Medicine*, 392(13):1320–1333, April 2025.

[14] T. Fusco, Y. Bi, H. Wang, et al. Data mining and machine learning approaches for prediction modelling of schistosomiasis disease vectors. *International Journal of Machine Learning and Cybernetics*, 11:1159–1178, 2020.

[15] Taewan Goo, Catherine Apio, Gyujin Heo, Doeun Lee, Jong Hyeok Lee, Jisun Lim, Kyulhee Han, and Taesung Park. Forecasting of the covid-19 pandemic situation of korea. *Genomics & Informatics*, 19(1), 2021.

[16] Yanhui Guo, Yi Feng, Fuli Qu, Li Zhang, Bingyu Yan, and Jingjing Lv. Prediction of hepatitis e using machine learning models. *PLOS ONE*, 15:1–12, 09 2020.

[17] Daniel Alejandro Gónzalez-Bandala, Juan Carlos Cuevas-Tello, Daniel E. Noyola, Andreu Comas-García, and Christian A García-Sepúlveda. Computational forecasting methodology for acute respiratory infectious disease dynamics. *International Journal of Environmental Research and Public Health*, 17(12), 2020.

[18] A.A. Hemedan, M. Abd Elaziz, P. Jiao, et al. Prediction of the vaccine-derived poliovirus outbreak incidence: A hybrid machine learning approach. *Scientific Reports*, 10, 2020.

[19] Beakcheol Jang, Inhwan Kim, and Jong Wook Kim. Effective training data extraction method to improve influenza outbreak prediction from online news articles: Deep learning model study. *JMIR Med Inform*, 9(5):e23305, May 2021.

[20] Ravikiran Keshavamurthy, Samuel Dixon, Karl T. Pazdernik, and Lauren E. Charles. Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches. *One Health*, 15:100439, 2022.

[21] J. Kim and I. Ahn. Infectious disease outbreak prediction using media articles with machine learning models. *Scientific Reports*, 11:4413, 2021.

[22] T. Liu, Y. Zhang, H. Lin, et al. A large temperature fluctuation may trigger an epidemic erythromelalgia outbreak in china. *Scientific Reports*, 5, 2015.

[23] Yasas Mahima and Thepul Ginige. Covid-19 spread prediction based on food categories using data science. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Nagarjuna College of Engineering and Technology, Bangalore, India, 2020.

[24] Zohair Malki, El-Sayed Atlam, Aboul Ella Hassanien, Guesh Dagnew, Mostafa A. Elhosseini, and Ibrahim

Gad. Association between weather data and covid-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons Fractals*, 138:110137, 2020.

[25] Babagana Modu, Nereida Polovina, Yang Lan, Savas Konur, A. Taufiq Asyhari, and Yonghong Peng. Towards a predictive analytics-based intelligent malaria outbreak warning system. *Applied Sciences*, 7(8), 2017.

[26] Fotios Petropoulos and Spyros Makridakis. Forecasting the novel coronavirus covid-19. *PLOS ONE*, 15:1–8, 03 2020.

[27] Donald Salami, Carla Alexandra Sousa, Maria do Rosário Oliveira Martins, and César Capinha. Predicting dengue importation into europe, using machine learning and model-agnostic methods. *Scientific Reports*, 10, 2020.

[28] Omar Enzo Santangelo, Vito Gentile, Stefano Pizzo, Domiziana Giordano, and Fabrizio Cedrone. Machine learning and prediction of infectious diseases: A systematic review. *Machine Learning and Knowledge Extraction*, 5(1):175–198, 2023.

[29] Shima Shamkhali Chenar and Zhiqiang Deng. Development of artificial intelligence approach to forecasting oyster norovirus outbreaks along gulf of mexico coast. *Environment international*, 111:212–223, 12 2017.

[30] Rameshwer Singh and Rajeshwar Singh. Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – a review. *Materials Today: Proceedings*, 81:1006–1011, 2023. International Virtual Conference on Sustainable Materials (IVCSM-2k20).

[31] Martin Szomszor, Patty Kostkova, and Ed de Quincey. #swineflu: Twitter predicts swine flu outbreak in 2009. In Martin Szomszor and Patty Kostkova, editors, *Electronic Healthcare*, pages 18–26, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[32] Shreshth Tuli, Shikhar Tuli, Rakesh Tuli, and Sukhpal Singh Gill. Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing. *Internet of Things*, 11:100222, 2020.

[33] United Nations Office for the Coordination of Humanitarian Affairs. Global humanitarian overview 2025: February update. https://humanitarianaction.info/document/global-humanitarian-overview-2025-monthly-updates/article/february-update-0, 2025. Accessed: 2025-05-11.

[34] Qinneng Xu, Yulia R. Gel, L. Leticia Ramirez Ramirez, Kusha Nezafati, Qingpeng Zhang, and Kwok-Leung Tsui. Forecasting influenza in hong kong with google search queries and statistical model fusion. *PLOS ONE*, 12:1–17, 05 2017.

[35] Xiaolei Zhang, Renjun Ma, and Lin Wang. Predicting turning point, duration and attack rate of covid-19 outbreaks in major western countries. *Chaos, Solitons Fractals*, 135:109829, 2020.