International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predictive Algorithms Based on Machine Learning for Infectious Disease Outbreaks and Spread

Prof. Mrs Tejaswini Abhishek Puranik

Assistant Professor, Computer Science and Engineering, Shri Sant Gajanan Maharaj College of Engineering, Shegaon

Abstract

The study aims to demonstrate whether machine learning can be used to predict infectious disease outbreaks early. The Cochrane Collaboration's guidelines, the meta-analysis of observational studies in epidemiology, and the preferred reporting items for systematic reviews and meta-analyses were followed in this study. The suitable bibliography on PubMed/Medline and Scopus was searched by combining text, words, and titles on medical topics. At the end of the search, this systematic review contained 75 records. The studies analyzed in this systematic review demonstrate that it is possible to predict the incidence and trends of some infectious diseases; by combining several techniques and types of machine learning, it is possible to obtain accurate and plausible results.

Keywords: Machine Learning, **epidemiology**,PubMed, infectious diseases

1. Introduction

Millions of people die each year from infectious diseases, making them a growing part of the global health burden [1]. It is essential to identify the factors of disease diffusion in order to apply control and prevention measures in order to develop protective measures against infectious diseases [2]. Predictions that can assist policymakers in making appropriate decisions regarding, for instance, the purchase of vaccines, public awareness campaigns, and training programs for health professionals, could be made by determining the factors of disease diffusion [3,4].

Overview of Infectious Diseases and Machine Learning

Machine-learning (ML) algorithms have emerged as valuable tools in the control of infectious diseases, offering the capability to spatially and temporally predict their evolution and spread [5]. By analyzing large, complex, and often multidimensional datasets, ML techniques can detect patterns and trends that are not easily discernible through traditional analytical methods. This makes them particularly suitable for modeling the dynamics of infectious diseases, which are influenced by diverse factors such as population demographics, environmental conditions, and individual behaviors.

In recent years, an increasing number of studies have applied ML techniques to predict infectious disease outbreaks, with promising outcomes. However, the successful deployment of these techniques hinges on the availability and quality of data. Infectious disease surveillance systems typically collect a range of information, including reported cases, geographic spread, and patient demographics. Unfortunately, these datasets are often incomplete, noisy, or biased, which can significantly hinder

model performance. Furthermore, long incubation periods associated with many diseases can create a lag between infection and data reporting, complicating timely and accurate predictions.

2. Materials and Methods

2.1. Search Strategy and Data Sources

The guidelines for the meta-analysis of observational studies in epidemiology (MOOSE) [12] and the Cochrane Collaboration [11] were followed when carrying out the current systematic review. The procedure and the outcomes were reported using the preferred reporting items for systematic reviews and meta-analyses [13] and guidelines [14]. A bibliographic search was conducted on 9 November 2022, on the Scopus and PubMed/MEDLINE databases, combining keywords by using the Boolean operators "AND", "OR" and "NOT". Supplementary Table S1 details the search strategy. No time filter was used. It was not always possible to apply all of the items on the PRISMA checklist because of the innovative nature of the study and its recent field of application halfway between epidemiology, information technology, and medicine (more information is provided in the study limitations section). 2.2. Criteria for Inclusion and Exclusion Studies had to meet the following criteria to be considered eligible: (i) language: written in English; (ii) population: human; (iii) interventions: machine learning; (iv) comparators/control: infectious diseases; (v) outcomes: prediction/forecasting outbreaks infectious diseases; (vi) type of study: epidemiologic studies (case-control, cross-sectional, or cohort studies).

Exclusion criteria were as follows: (i) articles not published in English; (ii) not human; (iii) full text not available; (iv) interventions: not about machine learning; (v) comparators/control: not about infectious diseases; (vi) outcomes: not about prediction/forecasting outbreaks infectious diseases; (vii) type of study: review article, meta-analysis, trial, expert opinion, commentary, editorial, case report, letter to the editor, or book chapters. See Supplementary Table S2 where the detailed description of the inclusion/exclusion criteria is reported.

2.3. Selection Process and Data Extraction

Two reviewers (D.G. and C.F.) independently evaluated the titles and abstracts of the manuscripts found using the search strategy and those retrieved from additional sources. After that, the same authors independently reviewed the downloaded text and assessed the articles' eligibility. The case was discussed with a senior reviewer (O.E.S.) when there was an unresolved disagreement between the two evaluators. Full texts were downloaded only for potentially eligible studies.

Data extraction was conducted only for those articles that met all the inclusion criteria, and it was performed using a predefined and prepiloted spreadsheet elaborated in Microsoft Excel for Windows. The extracted data included the author, publication year, study period, country where the study was conducted, disease, data source, model and/or techniques, aim, main results, accuracy/best model, space/time resolution, order of magnitude modeled populations, funds, and conflicts of interest.

2.4. Method for the Synthesis of Data By following the PRISMA 2020 guidelines, a flowchart (Figure S1) was created showing the number of references at each stage of the review process [15]. Summary tables were created showing the qualitative results of the literature. A full report was produced; in this, there is a general overview of the main findings of the review.

2.5. Critical Evaluation A critical evaluation of the articles using the Newcastle–Ottawa scale (NOS) was independently carried out by two authors (O.E.S. and D.G.) [16]; this was a bias-risk assessment tool for observational studies that can assign up to nine points for the lowest risk of bias in three

domains: (i) study group selection; (ii) comparability; and (iii) assessment of exposure and outcomes for case-control and cohort studies, respectively.

An adapted version of the NOS was used to assess cross-sectional studies [17]. According to these criteria and on the standard cutoff used in the previous literature [18,19], studies were classified as being of high, moderate, or low quality when their NOS score was ≥7, 4–6, and ≤3, respectively.

3. Results

An extensive search of the literature was carried out using the **Scopus** and **PubMed/MEDLINE** databases. This initial search retrieved **375 records from Scopus** and **333 from PubMed/MEDLINE**, totaling **708 articles**.

After eliminating **89 duplicate entries**, **619 unique studies** remained for screening. Titles and abstracts were assessed first, leading to the exclusion of **537 records** based on the following criteria: the topic was unrelated (**n = 530**), the article was not original research (**n = 3**), the language was not English (**n = 3**), or the study did not involve human subjects (**n = 1**).

A total of **82 full-text articles** were then reviewed in detail. Of these, **7 were excluded** following full-text evaluation for not meeting the inclusion criteria.

This process resulted in the final inclusion of **75 studies** in the review [20–94]. A flowchart illustrating the selection steps is provided in **Figure S1**.

During the initial phase of screening, there was a **0.7% rate of disagreement** between reviewers, which was resolved through discussion.

All included studies are summarized in **Table 1**, organized alphabetically by the first author's last name. Articles from **Absar N [20] to Zhong R [73]** are peer-reviewed journal publications, whereas studies from **Ajith A [74] onward** represent conference proceedings and technical papers.

Discussion

Several infectious diseases are emerging and threatening the human health condition across the world. The burden of infectious diseases is certainly a planetary issue, annually causing millions of deaths [24]. Therefore, the study of infectious disease behavior has been a subject of scientific interest for many years; the early identification of emerging infectious disease outbreak patterns is critical and offers great advantages [36,43]. Indeed, as is evident from the studies under observation, accurate and reliable predictions of infectious diseases can be invaluable to public health organizations planning interventions to reduce or prevent disease transmission [38] and mitigate the negative impacts of diseases [35]. As seen by Ketu S. et al., the viral epidemic, in addition to exerting direct damage on people's lives, can affect a country's economy [42,43]. As reiterated by Roster K. et al., recent epidemic outbreaks, such as the COVID-19 pandemic and the zika epidemic in Brazil, have demonstrated the importance and difficulty of accurately predicting new infectious diseases [57]. A lack of knowledge about new infectious diseases and their consequences, along with complicated social and governmental factors, may influence the spread of every newly emerging disease [42]. It is, in fact, essential to try to estimate the future movement and pattern of a new disease [39], so that preventive measures such as closing schools, shopping centers, and theaters; closing borders; suspending public services; and stopping travel can be quickly implemented [42]. However, it is difficult to predict whether or how an infectious disease will emerge due to the episodic and rapid nature of its epidemic spread. In addition, collecting data on a specific infectious disease is not always easy. Knowledge about the transmission paths of

emerging diseases, the level and duration of immunity to reinfection, and other parameters needed to build realistic epidemiological models are often scarce. For these reasons, it is necessary to find appropriate and useful information sources and data and build up reliable prediction models with these [43,57]. Indeed, to develop increasingly effective control and prevention strategies, reliable computational tools that may help to understand disease dynamics and predict future cases are needed. Policymakers can use these computational tools to make decisions that are more informed [24]. Several approaches have been proposed in the literature to produce accurate and timely predictions and potentially improve public health response [32]. While time series forecasting and machine learning are less dependent on disease assumptions, they still require a lot of data, which might not be available at the beginning of an outbreak [57]. Complex dependencies and uncertainties must be taken into account when modeling the spread of infectious diseases across space and time. Machine-learning methods, especially neural networks, are useful for modeling these kinds of complex problems, even if they in some cases lack probabilistic interpretations [53]. Mechanical models provide only a partial solution to the unsolved problem of predicting the course of contagion dynamics. To remain mathematically or computationally tractable, these models must rely on simplifying assumptions, thus limiting the quantitative accuracy of their predictions and the complexity of the dynamics they can model [51]. Mathematical modeling is the most scientific technique for understanding the evolution of natural phenomena, including the spread of infectious diseases. Therefore, these modeling tools have been widely used in epidemiology to predict risks and inform decision-making [46]. These models provide decision-makers with additional information regarding infectious disease responses, despite their imperfections. These results could be useful insofar as informing decisions on planning, resource allocation, and social-distancing policies [37]. Deep learning offers a new and complementary perspective to build effective models of contagion dynamics on networks, as demonstrated by Murphy C. et al. [51]. The analysis found that models based on combining multiple machine-learning methods, incorporating information from different models that are based on multiple data sources, produced the most robust and most accurate results [47].