Digital Disease Detection:

Application of Machine Learning in Community Health Informatics

Ekkarat Boonchieng, Ph.D. *, and Khanita Duangchaemkarn, Pharm.D., *IEEE Member* †*
*Center of Excellence in Community Health Informatics, Faculty of Science, Chiang Mai University, Thailand
†Biomedical Engineering Program, Faculty of Engineering, Chiang Mai University, Thailand,
E-mail: *ekkarat@boonchieng.net,†d.khanita.TH@ieee.org

***Abstract* —Health informatics is a new research area which is interdisciplinary amongst information science, computer science and healthcare. The concept of health informatics is to develop a new way to manipulate healthcare data from various resources and devices by optimizing the method of data acquisition, data storage, data processing, and data visualization. Community health informatics can be described as the systematic application of information and computer science to obtain valuable data for solving health problems and providing it to health policy makers. The challenge of community health informatics is to maximize the efficiency and efficacy of big data analysis. This discussion paper aims to present the various applications of machine learning and software engineering approaches that implemented in digital disease detection.**

Keywords— machine learning; health informatics; digital epidemiology; digital disease detection

I. INTRODUCTION

A. Digital disease detection

The core of digital disease detection is to detect any potential signal of the outbreak earlier. In 2002, Eysenbach et al. defined a new term, "infodemiology" (the merging of two separate term "information" and "epidemiology"), which is described as a new research discipline and methodology that studies the determinants and distribution of health information [1]. Furthermore, "Digital disease detection" is a new term for the recently emerging communicable disease surveillance problem, as defined by Brownstein et al. in 2009 [2]. It describes the use of informal data from online resources, other than governmental, to provide a trend of global health that yields the near real-time results for communicable diseases detection.

Communicable diseases surveillance is a critical process in disease control. The procedure of communicable disease surveillance includes: the systematic collection, analysis, interpretation, and dissemination of information on a health-related event. One of the important public health functions of surveillance is outbreak detection. This function determines an abnormal increment of the disease incidences that are above the normal baseline. The surveillance system is very important for detecting the signal of emerging diseases, especially diseases with high rates of contagiousness [3]. Nowadays, there are many emerging communicable disease threats. For example: influenza, SARS, MERS Co-V, Ebola, or potential bioterrorism agents such as anthrax and small pox. These communicable diseases are rapidly spreading and sharing the same 'nonspecific' symptoms ("influenza-like symptoms"). Most of the time, we were not being able to detect the outbreak until it was announced by the governmental organization and sometimes it was too late to control.

The challenge of digital disease detection is to detect the signs of outbreak early on from available online big data which has dramatically increased in quantity and complexity. The old-fashioned data analysis seems to be less useful when the size of the data is increasing. Meanwhile, advancing techniques for big data analysis in public health have also been increasingly proposed and published.

B. The component of digital disease detection

There are three major components of the early warning system for epidemic detection. These are i) routine syndromic surveillance of the targeted disease; ii) disease model development based on historical epidemic data that gathered from official reports; and iii) epidemic forecasting or prediction of a near future event using predictive models or machine learning aided-models [5].

1) Component 1: Disease surveillance

In public health, there are many types of surveillance that are currently being used in every part of the world. Public health surveillance is a routine process to collect, analyze, and disseminate health determinant data for the purpose of public health [6]. Syndromic surveillance is one type of public health surveillance that is a novel approach using "pre-diagnostic data" together with statistical analysis to detect the pandemic earlier than traditional surveillance. Especially the unusual diseases with nonspecific symptom presentations, such as flu, SARS, or MERS Co-V which share the same nonspecific symptoms during the incubation period. The syndromic surveillance has also shown the importance of situation awareness, which means monitoring the effectiveness of epidemic responses and characterizing the affected populations.

The components of a good syndromic surveillance system are (i) an indicator or pre-diagnostic data; e.g. syndromes, medication sales, absenteeism, patient chief complaints; (ii) automated or partially automated and real-time acquisition of the entries; (iii) near-real-time and effective statistical algorithm to detect the unexpected elevations of signal in the indicator data entries, (iv) powerful tools for visualizing data and analyzing the results [7], [8].

2) Component 2: Predictive model development

Once the process of data acquisition is completed, disease models and parameters will be considered. In 2013, Nsoesie et al [9] published a systematic review on influenza detection. There are five approaches of model development for epidemic forecasting. These are (i) a time series model; (ii) approaches in meteorology; (iii) compartmental models; (iv) agent-based

978-1-5090-2033-1/16/$31.00 ©2016 IEEE

978-1-5090-2033-1/16/$31.00 ©2016 IEEE

models; and (v) metapopulational models. These different types of models are used for identifying the dynamics of the outbreak and describing the behavior of disease transmission.

3) Component 3: Epidemic forecasting

The epidemic forecasting can be either a statistical-based process or a machine learning-based process. There are 2 widely used statistical-based processes among digital disease detection studies which are (1) temporal detection algorithms, and (2) space-time detection algorithms. However, these processes are static forecasting approaches. The major limitation of the statistical-based process is that it cannot capture the dynamic change of the epidemic trend. Despite the fact that some parameters in the model can be changed in between single epidemic period, when predicting next consecutive epidemic periods, the result often presents the over or underestimation of the peak time or peak height behavior. Thus, the application of statistical-based analysis for real-time epidemic forecasting is limited.

C. Machine learning in health informatics

Machine learning is an interdisciplinary technique that combines informatics, statistics, data science and computer science together which is a useful technique in health informatics. Machine learning helps us to extract useful features from a lot of data to solve or predict health-related events, including medical decision support, forecasting, ranking, classification, clustering, detecting anomalies, or sentimental analysis. In the biomedical domain, almost all of the data sets are unstructured data and are usually multidimensional. Manual data manipulation is often impossible for processing a lot of data in health informatics. The ultimate challenge of machine learning is to discover structural patterns or temporal patterns in the data lake which cannot normally be seen or differentiated by human experts [4].

Machine learning is an essential tool for the outbreak forecasting process. This approach has proved to be the most effective analytical method to predict the peak time, peak height, daily/weekly activities, magnitude, and duration of the outbreak. The machine learning-based process also shows outstanding outcome when combining it together with the parameterization and optimization processes. This hybrid approach has proved to be effective and efficient in solving the epidemic forecasting problem.

The objective for this paper is to systematically review the use of machine learning techniques in community health informatics to solve the digital disease detection problem.

II. Article Selection and Evaluation

The scope of this systematic review included studies that predicted the communicable disease dynamic at the community or local level using the machine learning approach at any component of the early warning system. First, we searched PubMed and ScienceDirect database for articles on communicable disease forecasting. A search for ("communicable diseases"[MeSH Terms] OR ("communicable" [All Fields] AND "diseases"[All Fields]) OR "communicable diseases"[All Fields] OR ("communicable"[All Fields] AND "disease"[All Fields]) OR "communicable disease"[All Fields]) AND ("forecasting"[All Fields] OR "forecasting" [MeSH Terms]) on PubMed retrieved 608 articles. A ScienceDirect search for "communicable disease forecasting" retrieved 359 records. A Google Scholar search for "communicable disease forecasting" AND "machine learning" retrieved 4,870 results. Next, we screened for "communicable disease", "infectious disease", "outbreak", "epidemic" and "forecasting" or "prediction" in all the titles and abstracts. After we excluded all irrelevant articles, 56 full papers were eligible for analysis. Lastly, we excluded the articles that had not used machine learning techniques in their methodology and the outcome of the study is not intent to solve the problem on epidemic forecasting or digital disease detection. The study is therefore based on the remaining 21 articles. We grouped and presented studies based on how machine learning technique was used and in which digital disease detection processes. The article searching diagram is shown in Fig. 1.

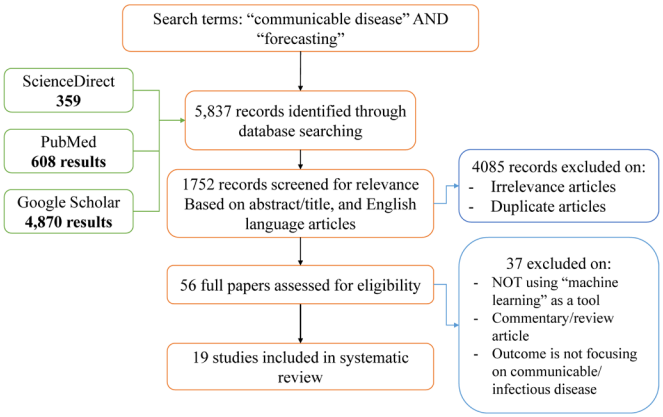

Fig. 1 Article searching diagram

III. Results and Discussions

We would like to make the statement that the machine learning technique can be described in various ways and has multiple terminologies. By using the term "machine learning", we might not have discovered all of the published articles in these fields.

The description for each ML technique used in eligible studies is shown in Table I.

A. Infectious diseases variation

Among 13 studies, the infectious diseases of interest consisted of influenza-like diseases (flu, H1N1, swine flu); dengue fever, hepatitis B, and some studies reported more than one specific disease. These diseases are of global health concern by many international health organizations.

Four studies forecast the event in the United States, 3 in China, 1 in India and Thailand, and 2 of them predicted global outbreak using online data sources.

B. Data aquisition

Data types used in each study are varied from the unstructured online sources to the official infectious disease incidence reports from government organizations. The information accumulated from Twitter and Wikipedia have also shown promising results in term of model accuracy [10], [11]. There are some attempts to include climate factors to develop a predictive model for some tropical infectious diseases such as the Dengue fever epidemic model by Kesorn et al. [12] and Sang

TABLE I. SUMMARY OF STUDIES CHARACTERISTICS

AUTHOR	YEAR	LOCATION	DISEASE	DESCRIPTION	DATA TYPE	ML ALGORITHM	OUTCOME
Freifeld et al.	2008	Global	Multiple	The system classifies alerts by location and disease and then overlays them on an interactive geographic map using text processing algorithms.	Unstructured electronic information sources	word-level N-gram approach	Location and Disease Classifier Performance over the One Month Period
Torii et al.	2011	Global	Multiple	Automated article detection using machine learning text classifier.	Unstructured electronic information sources	Support Vector Machine (SVM)	Classifier performance over 15 days
Nsoesie et al.	2011	USA	Influenza	Comparing partial epidemic curve.	Daily infected or influenza-like illness (ILI) cases	six supervised classification methods	Accuracy and consistency of the Classification Methods
Signorini et al	2011	USA	H1N1, Swine flu	Tracking public interest with data from twitter compare with CDC Weekly Reported and Estimated ILI%. (Nationwide)	ILI-related tweet	SVM	disease activity
Spratt et al	2013	USA	Dengue	Applying supervised classification to multidimensional data sets to improve modelling performance.	Clinical laboratory, cytokine and proteomic analyses were	SVM, CART, MARS, RF	Model performance
Ch et al.	2013	India	Malaria	Proposing novel method based on coupling the Firefly Algorithm (FFA) and Support Vector Machines (SVM) for malaria incidences forecasting.	Monthly averages of rainfall, temperature, relative humidity and malarial incidences	Coupling FFA with SVM	Model performance and Prediction errors
Chen et al.	2014	USA	Flu	Development of temporal topic model to capture hidden states of a user from twitters and aggregate states in a geographical region for better estimation of trends.	Twitter	Hidden Flu-State from Tweet Model (HFSTM) with EM-based clustering	Model accuracy
Generous et al.	2014	Global	Multiple	Training a statistical estimation model against ground truth data and then apply the model to generate estimates when the true data are not available.	Wikipedia	ailment topic aspect model (ATAM)	Model accuracy
Nsoesie et al.	2014	USA	Influenza	Developing a new approach to classify and forecasting epidemic curve.	Outbreak curves from seasonal influenza epidemics and 2009 H1N1 outbreak	Dirichlet process model	Model accuracy
Sang et al.	2014	China	Dengue	Predicting Local Dengue Transmission in Guangzhou, China, through the Influence of Imported Cases, Mosquito Density and Climate Variability	Weather variables, Breteau Index, imported DF cases and the local dengue transmission	Principal component analysis (PCA)	Model performance and Prediction errors

TABLE I (CONT) SUMMARY OF STUDIES CHARACTERISTICS

AUTHOR	YEAR	LOCATION	DISEASE	DESCRIPTION	DATA TYPE	ML ALGORITHM	OUTCOME
Zhang et al.	2014	China	Multiple	Evaluating and compare the performances of four different time series methods.	Infectious disease data collected through a national public health surveillance system in mainland China	SVM	Model performance and Prediction errors
Gan et al.	2015	China	Hepatitis B	Investigates the use of a hybrid algorithm combining grey model (GM) and back propagation artificial neural networks (BP-ANN) to forecast hepatitis B in China.	The incidence data of hepatitis B are collected from the Ministry of Health of the People's Republic of China from the years 2002 to 2012	BP-ANN	Forecasting accuracy
Kesorn et al.	2015	Thailand	Dengue	Improving a dengue surveillance system in areas with similar climate.	Climate factors, mosquitoes density, and Dengue morbidity incidence	SVM	Prediction performance

et al [13], and the Malaria epidemic model by Ch et al. [14]. Climate factors such as rainfall density, temperature relativity, and humidity are already known to be activators of mosquitoes breeding. These could result in Dengue fever and malaria epidemics.

C. Machine learning technique in model

There are several methods to develop a model for outbreak detection. The Machine learning technique has been selected to enhance the model's performance through every step of digital disease detection.

HealthMap is a good example system for digital disease detection. The Word-level N-gram approach was used to identify the diseases and the location displayed on the web browser. The system allows people to explore the occurred events around the world [15], [16]. Generous et al. are also trying to improve the epidemic forecasting model by using ailment topic aspect model (ATAM). The algorithm trains a statistical estimation model against ground truth data and then applies the model to estimate the data for some parameters when the true data is not available [11]. This method can be used in multiple languages, not only in the English language.

1) Data acquisition

Retrieving data from online resources and social media is most beneficial in real-time or near real-time analysis. However, if the data filtering process is not good enough, there is a higher chance of getting a rubbish output as well. Support vector machine (SVM) technique is widely used for text processing. The better the text classifier performs, the better the model can predict accurately.

2) Predictive model development

Generally, development of the predictive model is based on conventional mathematic modeling. Nsoesie et al. has reported that many predictive models have been used for epidemic forecasting, such as compartmental model, metapopulational model, agent-based model, autoregressive model, etc. [9]. However, the robustness of the model is varied because of the undetected uncertainty parameters and the dynamics of human behavior. It is still a challenge for researchers working in this field to improve performance and accuracy. Ch et al. implement coupling the firefly algorithm (FFA) to SVM for malaria incidence forecasting. In this study, FFA was used as an optimization algorithm and then selected the best parameters for SVM analysis [14].

In real-time epidemic forecasting, the faster the system is able to predict, the more effectively we can obtain information about the situation. The frequently used analysis is a regression model such as autoregressive moving average (MVA) [12], and linear or non-linear regression models. The Dirichlet process method has been proposed for epidemic curves predicted by Nsoesie et al [17]. This process is able to compare the patterns of the epidemic curves in order to recognize the severity and the duration of the outbreak. Moreover, other machine learning algorithms have been selected and proposed for epidemic forecasting supervised classification approach [18].

D. Machine learning technique and its epidemic forecasting performance

For the past several years, studies being conducted in this field are based on the data history. By splitting that data into two sets, a training data and a testing data set, the simulation obtained

from the training data set is used for a learning process to predict the testing data set. Since they are all retrospective predictions, the accuracy in real-time remains unassessed. Among machine learning techniques that are used in epidemic forecasting, the SVM technique happens to give a promising result in the model development process. Meanwhile, the AdaBoost technique, which doesn't seem to perform well in epidemic predictions, has the most precise prediction in the very early stages of digital disease detection, according to Santillana et al. [19].

IV. Conclusion

The Machine learning technique has played an important role in epidemic forecasting over the past several years. The selected studies presented in this review article have shown that by using the machine learning technique, the challenges of digital disease detection problems can be overcome. However, real-time monitoring and forecasting is the next big challenge for the researchers who are working in this field.

References

[1] G. Eysenbach, 'Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet.', *J. Med. Internet Res.*, vol. 11, no. 1, p. e11, Jan. 2009.

[2] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, 'Digital Disease Detection — Harnessing the Web for Public Health Surveillance', *N. Engl. J. Med.*, vol. 360, no. 21, pp. 2153–2157, 2009.

[3] J. Buehler, R. Hopkins, J. M. Overhage, D. Sosin, and V. Tong, 'Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks Recommendations from the CDC Working Group depar', *Morb. Mortal. Wkly. Rep.*, vol. 53, no. RR-5, 2004.

[4] A. Holzinger, 'Interactive machine learning for health informatics: when do we need the human-in-the-loop?', *Brain Informatics*, Mar. 2016.

[5] M. F. Myers, D. J. Rogers, J. Cox, A. Flahault, and S. I. Hay, 'Forecasting disease risk for increased epidemic preparedness in public health.', *Adv. Parasitol.*, vol. 47, pp. 309–30, 2000.

[6] S. S. Morse, 'Public health surveillance and infectious disease detection.', *Biosecur. Bioterror.*, vol. 10, no. 1, pp. 6–16, Mar. 2012.

[7] J.-P. Chretien, H. S. Burkom, E. R. Sedyaningsih, R. P. Larasati, A. G. Lescano, C. C. Mundaca, D. L. Blazes, C. V Munayco, J. S. Coberly, R. J. Ashar, and S. H. Lewis, 'Syndromic surveillance: adapting innovations to developing settings.', *PLoS Med.*, vol. 5, no. 3, p. e72, Mar. 2008.

[8] O. P. Wójcik, J. S. Brownstein, R. Chunara, and M. a Johansson, 'Public health for the people: participatory infectious disease surveillance in the digital age', *Emerg. Themes Epidemiol.*, vol. 11, no. 1, p. 7, 2014.

[9] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V Marathe, 'A systematic review of studies on forecasting the dynamics of influenza outbreaks.', *Influenza Other Respi. Viruses*, vol. 8, no. 3, pp. 309–16, May 2014.

[10] L. Chen, T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, 'IEEE Xplore Full-Text HTML : Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models', in *Data Mining (ICDM), 2014 IEEE International Conference on*, 2014, pp. 755–760.

[11] N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky, 'Global disease monitoring and forecasting with Wikipedia.', *PLoS Comput. Biol.*, vol. 10, no. 11, p. e1003892, Nov. 2014.

[12] K. Kesorn, P. Ongruk, J. Chompoosri, A. Phumee, U. Thavara, A. Tawatsin, and P. Siriyasatien, 'Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machine and the Aedes aegypti Infection Rate in Similar Climates and Geographical Areas.', *PLoS One*, vol. 10, no. 5, p. e0125049, Jan. 2015.

[13] S. Sang, W. Yin, P. Bi, H. Zhang, C. Wang, X. Liu, B. Chen, W. Yang, and Q. Liu, 'Predicting local dengue transmission in Guangzhou, China, through the influence of imported cases, mosquito density and climate variability.', *PLoS One*, vol. 9, no. 7, p. e102755, 2014.

[14] S. Ch, S. K. Sohani, D. Kumar, A. Malik, B. R. Chahar, A. K. Nema, B. K. Panigrahi, and R. C. Dhiman, 'A Support Vector Machine-Firefly Algorithm based forecasting model to determine malaria transmission', *Neurocomputing*, vol. 129, pp. 279–288, Apr. 2014.

[15] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, 'HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports.', *J. Am. Med. Inform. Assoc.*, vol. 15, no. 2, pp. 150–7, Jan. 2008.

[16] M. Keller, M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein, 'Use of unstructured event-based reports for global infectious disease surveillance.', *Emerg. Infect. Dis.*, vol. 15, no. 5, pp. 689–95, May 2009.

[17] E. O. Nsoesie, S. C. Leman, and M. V Marathe, 'A Dirichlet process model for classifying and forecasting epidemic curves.', *BMC Infect. Dis.*, vol. 14, no. 1, p. 12, Jan. 2014.

[18] E. O. Nsoesie, R. Beckman, M. Marathe, and B. Lewis, 'Prediction of an Epidemic Curve: A Supervised Classification Approach.', *Stat. Commun. Infect. Dis.*, vol. 3, no. 1, Jan. 2011.

[19] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, 'Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance.', *PLoS Comput. Biol.*, vol. 11, no. 10, p. e1004513, Oct. 2015.