IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 13, 2018, accepted September 4, 2018, date of publication September 19, 2018, date of current version October 17, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2871241

Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes

P. SIRIYASATIEN[1], S. CHADSUTHI[2], K. JAMPACHAISRI[3], AND K. KESORN[4]

[1]Parasitology Department, Medicine Faculty, Vector Biology and Vector Borne Disease Research Unit, Chulalongkorn University, Bangkok 10330, Thailand
[2]Physics Department, Science Faculty, Naresuan University, Phitsanulok 65000, Thailand
[3]Mathematics Department, Science Faculty, Naresuan University, Phitsanulok 65000, Thailand
[4]Computer Science and Information Technology Department, Science Faculty, Naresuan University, Phitsanulok 65000, Thailand

Corresponding author: K. Kesorn (kraisakk@nu.ac.th)

The work of P. Siriyasatien was supported by the National Science and Technology Development Agency, Thailand, through the Research Chair Grant. The work of K. Kesorn was supported in part by the Thailand Research Fund and in part by the Office of Higher Education Commission Thailand under Grant MRG5980224.

ABSTRACT Dengue infection is a mosquitoborne disease caused by dengue viruses, which are carried by several species of mosquito of the genus *Aedes*, principally *Ae. aegypti.* Dengue outbreaks are endemic in tropical and sub-tropical regions of the world, mainly in urban and sub-urban areas. The outbreak is one of the top 10 diseases causing the most deaths worldwide. According to the World Health Organization, dengue infection has increased 30-fold globally over the past five decades. About 50–100 million new infections occur annually in more than 80 countries. Many researchers are working on measures to prevent and control the spread. One avenue of research is collaboration between computer science and the epidemiology researchers in developing methods of predicting potential outbreaks of dengue infection. An important research objective is to develop models that enable, or enhance, forecasting of outbreaks of dengue, giving medical professionals the opportunity to develop plans for handling the outbreak, well in advance. Researchers have been gathering and analyzing data to better identify the relational factors driving the spread of the disease, as well as the development of a variety of methods of predictive modeling using statistical and mathematical analysis and machine learning. In this substantial review of the literature on the state of the art of research over the past decades, we identified six main issues to be explored and analyzed: 1) the available data sources; 2) data preparation techniques; 3) data representations; 4) forecasting models and methods; 5) dengue forecasting models evaluation approaches; and 6) future challenges and possibilities in forecasting modeling of dengue outbreaks. Our comprehensive exploration of the issues provides a valuable information foundation for new researchers in this important area of public health research and epidemiology.

INDEX TERMS Dengue, endemics, forecasting, prediction, surveillance system, survey, review.

I. INTRODUCTION

Although dengue has been known for more than 200 years, it was only in 1950 that the first dengue viruses were isolated. The incidence of dengue infection has increased approximately 30 times since 1950. According to the World Health Organization (WHO), and mentioned by Stanaway *et al.* [1], about 20,000 people per year die from this disease. Since 1970, epidemics of dengue infection have occurred in more than 128 countries worldwide [2] with a significant expansion of the spread of the disease to new areas of the world. This has greatly increased the challenge for national and international public health authorities in order to take effective action to prevent the spread of the dengue infection.

Citing the WHO report, outbreaks of dengue are likely to increase every year (Figure 1). In the period 2009 – 2017, the growth rate in the number of patients was greater than in any other period since 1950, with more than 3 million people affected. The number of countries identified as having dengue outbreaks was the highest ever: more than 80 countries.

Dengue viruses often have a relationship with other mosquito–spread viruses, such as Yellow Fever, Zika, Japanese Encephalitis, and West Nile viruses. Dengue virus is classified into 4 serotypes: DEN-1, DEN-2, DEN-3, and DEN-4. Usually, anyone infected by one serotype will gain lifelong immunity to that virus serotype, but protection against infection from the other dengue virus is only partial or temporary. A subsequent infection of a patient from

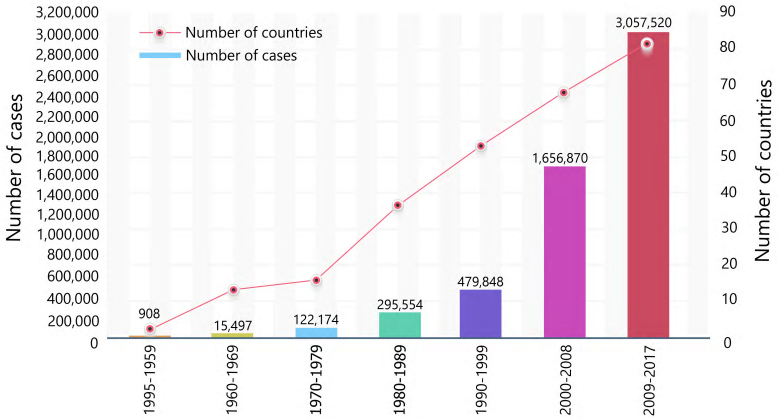


FIGURE 1. The number of reported dengue cases between 1950 – 2017.

another serotype of the virus is called a secondary infection. Within 7 to 10 days, the mosquito-borne virus can be passed on to other insects and humans, causing a fast spread of the virus in that vicinity. *Aedes* mosquito is usually active in the early morning and early evening before dark. It likes the temperature to be about 28°C to 35°C. The incubation period of the disease in humans is 3 to 15 days but most likely 5 to 6 days.

The patient will start feeling dengue infection symptoms such as fever (about 40°C or 104°F), with headache, muscle bone and joint pain, eye pain, loss of appetite, vomiting, nausea, stomach ache, rash, and thrombocytopenia. Clinical presentation of dengue infection can cause Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF), or Dengue Shock Syndrome (DSS) which is the most severe form (WHO, 2009). Currently, there are no anti-viral drugs available for the treatment of dengue infection [3]. If the symptoms of the dengue infection reach the radical level (leakage of plasma blood leading to organ failure), this becomes life-threatening for both children and adults. Therefore, definitive clinical diagnosis in the early stages of infection and clinical treatment by doctors and nurses with appropriate experience will contribute to the survival of the patient.

II. CURRENT DENGUE FORECASTING RESEARCH SITUATIONS

WHO has defined a strategy for the integrated monitoring and control of the spread of dengue infection in several regions of the world with the goal of reducing the disease occurrence and deaths from DSS over the period 2012 to 2020. WHO also has supported research related to monitoring of potential dengue outbreaks areas and countries that may otherwise be ignored by providing capital support to the researchers to develop innovative surveillance systems for the prevention, control and forecasting of dengue outbreaks more efficiently, and assist in the preparation of plans to manage the outbreaks [4].

This has enabled many researchers in the past decade to develop systems for forecasting dengue outbreaks by using a

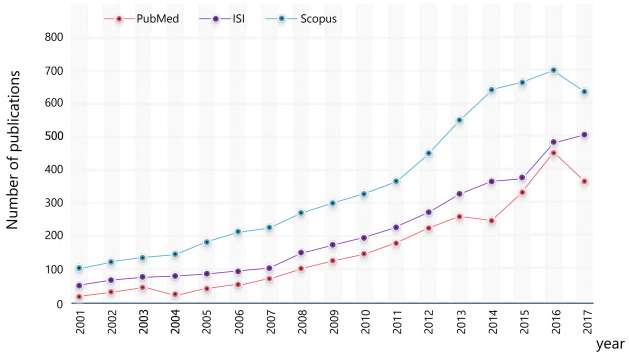


FIGURE 2. The number of research articles about dengue outbreak prediction, published in three reputable databases: PubMed, Web of Science, and Scopus.

variety of techniques, using data collected by querying published articles and research reports, and academic forecasts about dengue infection from reputable publication databases and journal forums such as PubMed, Web of Science, and Scopus. These are popular databases for researchers in which to publish their research on forecasting dengue epidemics. This information can be extracted from these databases by simple queries like ''Dengue and Forecasting,'' or ''Prediction'' or ''Surveillance.'' A more comprehensive query that we used in our research, being discussed in this paper, was ''dengue AND (forecasting OR prediction OR surveillance).'' Results from this query included publications over the period from 2001 to 2017. Figure 2 shows the volume of research published and the growth trend in the number of publications. From the survey, we found that the highest number of research articles, on forecasting of dengue outbreaks, was published in the database of PubMed in 2016, which was approximately twice that published in Scopus and Web of Science. Figure 2 reflects the increasing importance of research in the development of forecasting models of dengue epidemics, illustrating the severe nature of the problem of dengue infection and disease spread, as is also indicated in the report of WHO [2]. The growing threat of dengue outbreaks,

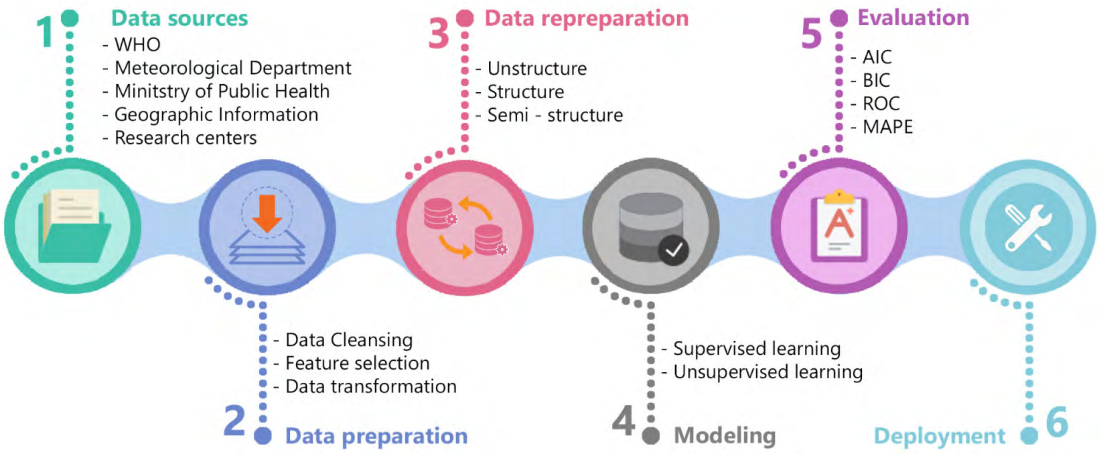


FIGURE 3. A data science process for dengue forecast model construction.

extending into more locations around the world, is the reason for researchers to develop more tools of higher efficiency for forecasting of the dengue infection.

To the best of our knowledge, only one previous publication exists that reviews dengue outbreaks forecasting [5], which, while being similar to our current article, was more limited in scope. Their work focused only on five issues: (1) the type of surveillance (active/passive), (2) study design and objectives, (3) development and delivery of tools and methods, (4) outbreak definition, and (5) dependent variables. In addition, they only focused on 24 existing works, whereas, in our study, we collected and analyzed research articles related to the forecasting of outbreaks of dengue infection in the past 20 years, during which period more than 100 systems for this purpose were developed, according to our findings. Our review encompassed the data science processes applied in those 100 systems (Figure 3) thus being a significantly more comprehensive and diverse study in which we identified in greater depth the main issues and components of such systems. In this current paper, we focus on six aspects which greatly affect the efficacy of forecasting models for dengue outbreaks:

1) *Data sources* refer to the available sources of data, from multiple databases. These data sources have, in the past, usually been provided by the Public Health Ministry of a country, or the Department of Meteorology, or specific hospitals, or WHO. However, new sources of data, in a new variety of forms, are now available, and which often come from field studies published by other researchers on social media and blogs. Of relevance here are search engines accessing the Internet, and searching social networks such as Facebook, Twitter and others. Section III will explain in more detail these various types of data sources.

2) *Data preparation* is the process of preparing the data prior to creating a forecasting model, and includes the selection of attributes that are key to cleansing and transforming the data into the appropriate format appropriate for the data to be applied to create a well-performing forecasting model. *Dengue factors* are the factors that have an impact on the spread of dengue infection both directly and indirectly, such as temperature, relative humidity, and population density. Researchers are, and have been, trying to find the factors that have a relationship or linkage to the severity of outbreaks and apply them to use in forecasting models, to increase performance of those models. Details about data preparation techniques are described in section IV.

3) *Data representation* refers to the data in a variety of formats available for use in forecasting models. New formats for representing data include, importantly, ontologies, represented by new languages such as OWL (Ontology Web Language) and RDF (Resource Description Framework) which are more recently available in standard form, enabling their use in different tasks and applications. Various data representation approaches are shown in section VI.

4) *Modeling evaluation* refers to standard measures of evaluating and comparing the effectiveness and efficiency of forecasting models, and their predictive accuracy. Different measures usually provide different information to a user and, thus, various measures should be used to evaluate the efficiency of any forecasting model. This includes statistical and mathematical methods using for data preparation processes e.g. scaling, normalization, and data cleansing. Details of these will be described in subsequent sections. Please read more details in section VII.

5) *Evaluation* refers to standard measures of evaluating and comparing the effectiveness and efficiency of forecasting models, and their predictive accuracy. Different measures usually provide different information to a user and, thus, various measures should be used to evaluate the efficiency of any forecasting model. However, the deployment process is out of scope of this paper so we do not describe this process in detail. Section VIII will describe more details of forecasting model evaluation schemes.

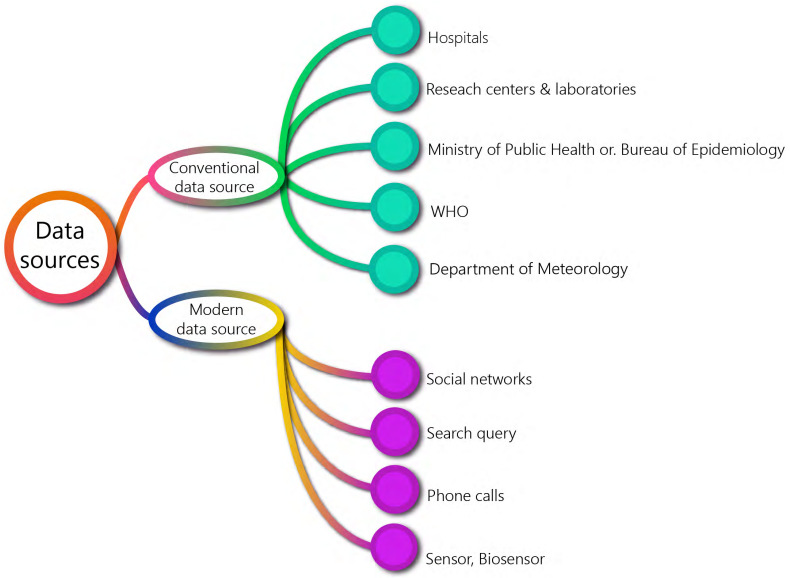


FIGURE 4. Traditional and modern data sources for dengue prediction system.

III. DATA SOURCES FOR DENGUE FORECASTING MODELS

In the creation of forecasting models, the dataset is an important element to ensure relevant and usable outcomes. The process of forecasting will necessarily use historical information, and requires a large amount of data, sufficient to support learning to inform the forecasting models. In addition, there must be sufficient data to allow valid testing of predictions against observed historical events. So, the volume and variety of data is important as is the correctness of the data from reliable sources [6].

Typically, and traditionally, data comes from government institutions, such as the Ministry of Public Health, Meteorological Department, Ministries of Agriculture or Lands Administration, and other formal institutions such as hospitals. However, there is no single source that can provide data covering all aspects of forecasting model construction. Now, in the information age, enormous volumes of data are available from new data sources on the Internet, and social networks, now being termed Big Data. Here, we refer to data sources as being conventional data sources, or modern data sources (Figure 4).

A. CONVENTIONAL DATA SOURCES

Conventional data and data sources typically include medical and epidemiological data from traditional health data centers (such as hospitals), environmental and weather data from meteorological departments, and demographic and geographical data from other appropriate government sources. Finding information from these primary data sources is not difficult as almost all countries have departments or ministries of ''Public Health'' or some such, that maintain centralized data centers for health data, which can provide large amounts and variety of data such as the diagnostics, laboratory tests, medication, and ancillary clinical data [7].

1) HOSPITALS

As data from hospitals are usually local to that country, and even to a specific area of the country, the main advantage accruing from these data sources is that the data is usually recognized as valid with high reliability and is useful in many ways for constructing a knowledge base [8], [9], or decision support systems and policy making [9]. The localized data is also more specific and therefore usable for forecasting models [10], [11]. Therefore, it is not surprising that several researchers have used data from hospitals as their main dataset for their forecasting model for dengue outbreaks. For example, Tien *et al.* [10] used the information from local hospitals in South Vietnam to find the significant factors relating to the number of dengue patients with dengue virus serotypes DEN-3 or DEN-4. Shaukat *et al.* [11] also used information from the district headquarters hospital in Jhelum in Pakistan that showed a similar relationship. Importantly, both of these researchers used data from similar, conventional sources, which allowed them to come to similar conclusions. However, the main disadvantage of using data from local hospitals is that the forecasting model is therefore not sufficiently general to apply to other areas where an epidemic could have a different spreading pattern, and using a forecasting model based on data from other areas could give poor accuracy.

2) RESEARCH CENTERS AND LABORATORIES

Some dengue factors such as the rate of infection of mosquitoes and mosquito larvae incidence, cannot be officially collected because of the complexity and high cost of doing so, and also must be done by experts [12]. As a result, there are no agencies that monitor and collect this information. As well, this data may result from a collaboration between institutions and research teams and, therefore, it may not be publicly available. Nonetheless, it could be

very useful for dengue outbreaks prediction because it might be highly correlated with the spreading of the outbreak if made available. For a recent study, Kesorn *et al.* [13] used mosquito infection rates information from the Vector Biology and Vector Borne Disease Research Unit, Parasitology Department, Medicine Faculty, Chulalongkorn University, Thailand, wh0 collected *Ae. aegypti* larvae and the adult mosquitoes, using human bait, twice per season between 2007 and 2012 and were able to visually identify their species. The dengue virus detection was conducted by accumulating all the *Ae. aegypti* larvae and mosquitoes in cryogenic vials (five larvae or mosquitoes/pool/vial) in liquid nitrogen. Dengue virus-infected mosquito rates were obtained from a previous report by Chompoosri *et al.* [12]. Kesorn *et al.* [13] found a significant correlation between the number of human dengue cases and the mosquito infection rate, and used the information on those factors to construct the forecasted dengue epidemic model, which greatly enhanced the prediction power of their prediction model.

A recent report by Lau *et al.* [14] demonstrated that the NS1 antigen kit was effective and had low cost for detecting dengue viruses in mosquitoes. They also found that dengue infection in female mosquitoes was correlated to dengue epidemics in the area where the mosquitoes were collected. An early study of Klovdahl *et al.* [25] conducted on human immunodeficiency virus (HIV) epidemics and the relationship between the hepatitis B virus (HBV) and HIV, and the researchers developed new perspectives on the link between communicable diseases and vulnerable populations.

3) MINISTRY OF PUBLIC HEALTH OR A BUREAU OF EPIDEMIOLOGY

As indicated above, most countries have a Ministry of Public Health or a Bureau of Epidemiology, however called, and most previous research has been based on data about dengue cases collected by these government institutions. Such government health organizations have the duty to monitor the disease and undertake health surveillance in the country and abroad according to international standards, and would usually have a strong network around the country to effectively monitor disease outbreaks that threaten people's health in that country in a thorough and timely manner. Those organizations are considered as important data sources that collect facts about, and report on, dengue incidents including the number of cases and mortality rates in each area to indicate the severity of the outbreak. In order to forecast outbreaks of the disease, researchers need to seek information from these government agencies for primary data. For example, Hii *et al.* [15] used weekly data from the Singapore Ministry of Public Health Singapore as a primary source of data on dengue prophylaxis. They showed that forecasts of outbreaks could be made 16 weeks in advance with high accuracy. Similarly, Shi *et al.* [16], who exploited data from the same source as [15], showed similar forecasting outcomes but used different techniques to create a forecasting model, called Least Absolute Shrinkage, and Selection Operator (LASSO). Kesorn *et al.* [13] and Siriyasatien *et al.* [17] made use of outbreaks information from the Ministry of Public Health in Thailand in order to forecast outbreaks in similar climate areas. Johansson *et al.* [18] also used data from the Mexican Health Secretariat for the period 1985 to 2012 to develop a model of dengue epidemics in Mexico. These studies used local data from their countries, making it easier to acquire the data. The main disadvantage of the studies, however, was that their models may not be able to work effectively for other countries, as the training data does not allow the model to learn the outbreaks pattern of different places. Therefore, some researchers preferred to use datasets from international organizations that overcame these restrictions, enabling the forecasting model to learn and estimate the epidemic severity in other venues.

4) WORLD HEALTH ORGANIZATION

Beside exploiting data from local hospitals or government organizations, some researchers used data from international organizations available publicly such as WHO which is a United Nations specialized agency responsible for coordinating public health programs internationally. WHO is an important organization able to collect health-related information from countries worldwide. Therefore, several researchers have employed data from WHO or other international organizations. For example, Brady *et al.* [19] collected data from public health organizations around the world to create a map of dengue outbreaks in 128 countries. Likewise, Bhatt *et al.* [3] used data from various international data sources, including from WHO, to study the distribution of dengue globally and in individual nations and realized that factors influencing dengue outbreaks included rainfall and temperature, data that was readily available from Department of Meteorology, and information on the degree of urbanization in those locations.

5) DEPARTMENT OF METEOROLOGY

Several researchers extensively used weather data when it was realized that it is important in forecasting dengue infections. Climatic factors, including rainfall, temperature and relative humidity, affects the presence of the pathogen and increases the dengue virus density and are therefore important factors in dengue epidemics. Ramadona *et al.* [20] examined the effects of tropical weather conditions on dengue infection in Singapore by taking into account the epidemiology of the disease and the details of the dengue virus strain, and found that relative humidity and average temperature are the two primary factors that affect dengue infection, being more significant than just rainfall.

Information on climatic conditions has also been demonstrated as a significant factor in dengue forecasting models in, for example, in Mexico, in both Mexico City and Puebla City which are located 2,000 meters above sea level, in a subtropical highland climate, now understood to be suitable for dengue outbreaks distribution. The annual rainfall is about 820 millimeters, which is highest between June and October.

The average annual temperature varies from 12°C to 16°C, and is rarely above 30°C. Notwithstanding these obvious differences with truly tropical countries, these cities have as high a risk of dengue virus as areas in a warm climate [21]. Chikungunya fever, for example, which is a viral illness caused by the Chikungunya virus (CHIKV) occurring in Africa and Southeast Asia, and which has now spread to Western countries, and to the regions between the Tropic of Cancer and the Arctic Circle. Another important example is Nepal, a country also above an elevation of 2,000 meters above sea level, therefore having a cool or temperate climate, but which has recently been affected by global warming which has caused substantial population movements to and between cities. This has led to greatly increased numbers of infected people in the mountain areas, which had never previously experienced dengue cases. It is evident that the changes in the environment and in the climate have impacted Nepal, and the significant migration of rural and remote populations to and from urban areas, has contributed to the rapid growth of dengue infection in mountainous regions which had never previously had dengue outbreaks, yet CHIKV is now endemic in these areas [22]. In addition, the *Ae. albopictus* are more likely to be carriers of the dengue virus than the *Ae. aegypti*, as the eggs of *Ae. albopictus* tolerate lower temperatures and can survive in the dry season [23]. These climate conditions being suitable for *Ae. albopictus*, outbreaks of dengue have occurred in countries with similar climates, such as the United States and Europe. The first cases of CHIKV transmission were reported for the first time in Europe in 2007, in localized outbreaks in north-eastern Italy [24] where 197 cases were recorded, confirming that mosquito-borne outbreaks caused by *Ae. albopictus* are plausible in Europe.

B. MODERN DATA SOURCES

New sources of data able to be used for forecasting dengue epidemics have become available due to advances in information technology, with vast amounts of information becoming available on the Internet. Epidemiological researchers are now paying close attention to this relatively new, and powerful, source of data.

1) SOCIAL NETWORKS

Today, Internet technology enables remote communication between patients and doctors, not just face-to-face in a clinic or hospital, indicating that social networking can also be used to connect doctors with patients, patients with patients, or doctors with doctors, enabling easy coordination of or sharing of one another's knowledge. Researchers are trying to leverage on those social networks to forecast dengue outbreaks. Interested people, using a search engine, can search these new sources of information. Studies have found that most dengue forecasting has been based on environmental data, climate and weather data, and geographical data, as has been described earlier. Now there are studies on the use of the information resources available on social networks such as Facebook, Twitter, Flickr, Myspace, Friendster, Digg, or Meetup. Matthews *et al.* [26] studied the relationships of social networks, mobile networks and the outbreaks of dengue for analysis of spatial information of outbreaks. This research used mobile phones and image processing techniques to diagnose outbreaks. Currently, information from Twitter users, a.k.a tweets, is being used to forecast Brazil's urban and national epidemics [27]. These researchers used tweets about dengue infection from local people to create a forecasting model and compared that model to other models generated by Google Trends (www.google.com/trends) and Wikipedia access logs (http://stats.grok.se). They found that data from Twitter sources highly correlated with the dengue outbreaks.

2) SEARCH QUERY FROM THE INTERNET

In addition to information from social networks, many researchers [19], [28]–[31] have used dengue infection search data to generate forecasting models using Time Series Analysis. Althouse *et al.* [28] used dengue infection search data from Google Insights to model dengue outbreaks forecast in Singapore and Thailand, searching with keywords in three categories: nomenclature, signs/symptoms, and treatment, and the efficiency of the model in retrieving data was evaluated using R^2 and the Pearson correlation. This work showed that searching the data in this way achieved higher forecasting efficiency, clearly indicating that data now available on the Internet is beneficial for monitoring areas with poor surveillance systems. Chan *et al.* [30] and Milinovich *et al.* [31] used Google Insights to search for information to use in forecasting other infectious diseases, and achieved greater prediction accuracy. Based on their experiment, they recommended the use of Google Insights for effective surveillance. This recommendation would also extend to other sophisticated search engines.

In epidemiological surveillance, information from a website or email can also be used to forecast a dengue epidemic. Hoen *et al.* [32] used electronic information such as HealthMap (http://www.healthmap.org), ProMED-mail, and other electronic resources to identify new dengue-endemic zones. The findings show that the proposed approach is more reliable and faster than the passive case reporting-based surveillance systems in which reports were often delayed for weeks or months.

The Online Toolkit for the Analysis of Point Patterns (OnTAPP) [33] is a tool for identifying and storing epidemiological data on dengue outbreaks. OnTAPP accesses the data via the Internet and by conducting spatial analysis using the epidemiological data at the individual level, the information is displayed in spatial data models.

3) PHONE CALLS

There are, however, restrictions that apply to some areas where dengue reports are available online. For areas with little available data, such methods are also less reliable. Some researchers have attempted to use other data to replace local outbreak reports which may often not be available in

a timely fashion. For example, Rehman *et al.* [34] used data recorded from telephone calls into the department responsible for monitoring dengue outbreaks and occurrences. In Pakistan, for instance, telephone calls to the local authorities regarding dengue outbreaks provided more than 300,000 items of data able to be used to forecast the number of dengue cases in particular locales, with predictions able to be made 2-3 weeks in advance of the outbreaks. This clearly showed that this ''informal'' data collected from established patterns of telephone calls in each city, enabled the efficient prediction of the number of patients in specific locations.

A study of population migration patterns using mobile data found a correlation between population migration and dengue outbreaks in the community [35]. This was a new and novel way to gain information, so mobile phone user location information can be seen as a viable way to track and anticipate or treat outbreaks.

4) SENSOR AND BIOSENSOR

Sensor technology is now a viable option for the collection of data over mobile networks and the Internet. Under the technology umbrella now known as the Internet of Things (IOT), individual sensors are now addressable over the Internet. This now allows sophisticated use of sensors in wearable computing, remote sensing, and geographic information systems (GIS), which allows the gathering of vast quantities of information to be obtained pervasively but unobtrusively [36]. Biosensing technology to prevent mosquito-borne outbreaks of Zika, Dengue outbreaks, and West Nile virus is also the subject of on-going research. Stanciu *et al.* [37] have developed an amperometric biosensor that utilizes functionalized nanoparticles that specifically bind to the target virus' DNA or RNA. By relying on an agent that responds only to the specific virus of interest, the device registers a change in resistance when the binding occurs and this change is employed by the sensor to detect the presence of the specific virus. The sensor can then determine the presence of the virus in the blood or mosquito sample and the concentration of the virus.

The ultimate goal of this research into bio-sensing is to develop a personal device that is simple to use, to sense the presence of the virus, which thereby allows prompt detection of the virus, and removes the need for hospitalization of the patient, as well as sending an alert to public health officials immediately. This will have a significant impact in developing countries. The technology of the device is that it will operate through a low-power wireless network, using thin-film rechargeable batteries and thin-film photovoltaics to power and harvest energy from the environment, without requiring human intervention to maintain functionality and performance.

Scientists from the Amity University in Uttar Pradesh, and Maharshi Dayanand University in Haryana, synthesized the biosensor by depositing nanoparticles of zinc oxide, palladium, and platinum on a fluorine-doped tin oxide electrode and then coated the electrode with a DNA probe and tested its efficiency in detecting dengue viruses [38]. A significant reduction in current was observed when the biosensor's DNA probe bound to a complementary target DNA, made using RNA from dengue viruses. The biosensor can be regenerated by dipping it in a sodium hydroxide solution for five minutes, after which time its ability to show similar current response when exposed to a target DNA is restored. This device can be used as a lab-on-chip disease diagnosis platform at a point of care for detecting all individual serotypes of dengue virus on a paper-based substrate.

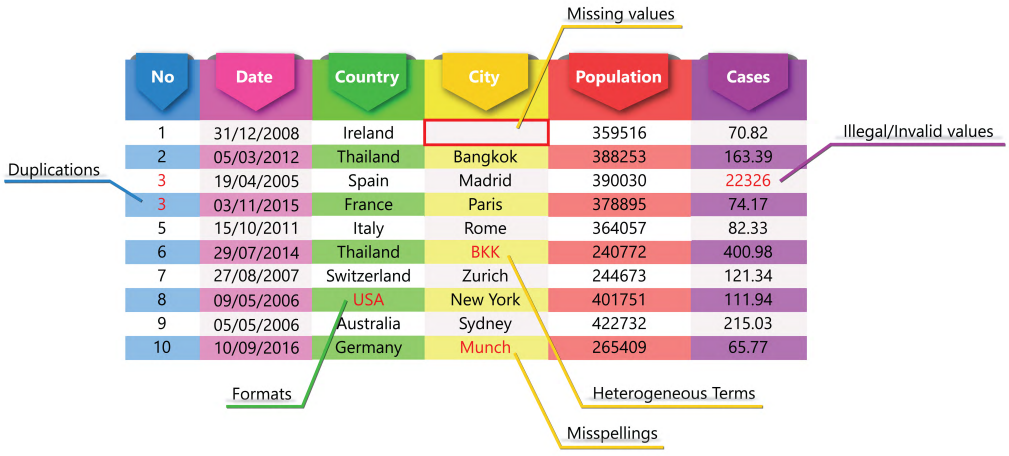
IV. DATA PREPARATION

Data preparation or data pre-processing is the process of preparing data for introduction into the forecasting process. The purpose is to reduce noise and increase the accuracy and consistency of data [39], using complex and diverse methods. This is because a forecasting model learns data patterns by using all the values of the dataset. If a dataset is ''dirty'' with several incorrect and incomplete data and other issues, the forecasting model will not learn as effectively. Therefore, it is necessary to first fix the issues and then apply the statistics and computational methods or Machine Learning algorithms. Typically, the pre-processing consists of four main steps (Figure 5): 1) Normalization or Standardization 2) Data cleansing 3) Feature selection and 4) Data transformation.

FIGURE 5. Major tasks of data preprocessing to enhance data quality before using in a forecasting model.

A. NORMALIZATION AND STANDARDIZATION

To ensure that the analysis of the data to find the relationship between variables is performed on reliable data, the data must be standardized before use. However, many researchers have shown little interest in this aspect of data preparation [40]. Data normalization or standardization is a process of scaling data, which is the adjustment of data on the various factors being included in the model that have a lower bound and upper bound of data that are on greatly different



No	Date	Country	City	Population	Cases
1	31/12/2008	Ireland		359516	70.82
2	05/03/2012	Thailand	Bangkok	388253	163.39
3	19/04/2005	Spain	Madrid	390030	22326
3	03/11/2015	France	Paris	378895	74.17
5	15/10/2011	Italy	Rome	364057	82.33
6	29/07/2014	Thailand	BKK	240772	400.98
7	27/08/2007	Switzerland	Zurich	244673	121.34
8	09/05/2006	USA	New York	401751	111.94
9	05/05/2006	Australia	Sydney	422732	215.03
10	10/09/2016	Germany	Munch	265409	65.77

FIGURE 6. Examples of errors found in collected data which directly affected the predictive performance of the models.

scales or ranges of measurement. For example, rainfall is measured in millimeters, infected mosquitoes are measured as a number and a percentage of a population of mosquitoes, and patient cases are measured as a number (which may be much lower than the number of mosquitos, for example).

Scaling is essential, to allow comparisons and associations, by adjusting the values of those data into the same range: $[-1, 1]$ or $[0.0, 1.0]$ (Figure 6). Unless this is done, one particular data value may dominate other data with smaller numerical values and that will introduce correlation errors between features or factors [41]. The most popular scaling methods are Z-score [42], Decimal Scaling [43], and Min-Max Normalization [44]. Kesorn *et al.* [13] conducted normalization of a set of data to compare dengue factors with the number of cases. After normalization and plotting of a line chart using the normalized data, some factors showed a high correlation with the number of cases. Principal Component Analysis (PCA) [45] is one of the more popular methods of normalization, and is used to reduce the number of correlated variables [46]. In that study, the main components of the variables were used and the PCA score was used to represent the correlated variables. Ahmed and Siddiqui [47] applied PCA to spatial data to identify which of the environmental factors most affect the occurrence of dengue infections in Pakistan.

B. DATA CLEANSING

The data cleansing process involves finding and eliminating incomplete, incorrect, and inconsistent data. Inconsistent data from various sources may contain misspellings, and erroneous data that comes from a single source may encounter misspelled data. In the case of data integration from multiple sources, data is often redundant with different formats, making it necessary to do data consolidation or integration of the data from the various sources by eliminating the redundant and unused data [48].

Data Cleaning/Cleansing or Data Scrubbing is a very important process in predictive modeling. If the analysis of the relationship of variables is done on poor quality data, models derived from the analysis are not reliable. Van den Broeck *et al.* [40] discusses the importance of data cleansing because researchers must work on eliminating data errors. A researcher should identify the data cleansing process used in medical information, including the type and error rate of the data, so that they are fully aware of the implications of the data. The majority of data errors are caused by aggregation of data from multiple sources and integration of this information has several problems, which directly affect the predictive power of the model. The study of problems that occur in input data was undertaken in [48] and their results are shown in Figure 7 and explained as follows.

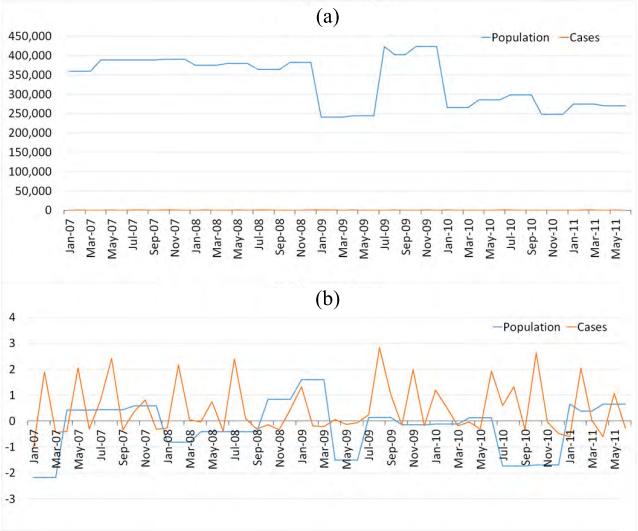


FIGURE 7. Comparison of data (A) before and (B) after data scaling to adjust data into the same range and avoid data domination problem.

First, missing data refers to the situation where some of the study variables in the available dataset are null, meaning the value is missing. This affects the accuracy of the data analysis [49]. As reported by Woo *et al.* [50], it was found that

89% of the datasets he identified had missing data, with 21% of the respondents having problems dealing with incomplete data. There are several reasons for data loss problems, such as the respondents in data collection surveys and activities not wanting to disclose information, or not understanding what was required. Data may also be missing due to data input or import errors.

A further problem is that of synonyms; different words with the same meaning, or, similarly, heterogeneous terms: where the same disease or condition can be caused, or contributed to, by several factors, which are named differently. Most of these kind of errors are caused by being loaded from many sources, which use different terminology. We could also consider the variety of codes used to be synonyms. For example, ''Bangkok'' may be coded as ''BKK'' whereas other sources use the full name ''Bangkok,'' in upper case, or ''Bangkok'' in mixed case.

Incorrect values are also very possible, or the value is invalid for the particular data, such as the age of the patient where the value is ''150,'' which is not possible. These are generally termed ''illegal values.'' As well, with text, there is always the potential for spelling problems, misspellings and localized spelling. For example, ''Bangkok'' entered as ''Bankkok,'' or ''Los Angeles'' entered as ''LosAngeles'' (no space). Data duplication is another problem, where the same set of data appears in more than one database. Of importance, also, is the use of different formats for the same data e.g. dates (''09/11/2001'' would be read as November 9th, 2001 by a British system), could cause data inconsistency resulting in poor learning of the data patterns by the prediction model (Figure 7). These latter problems can be easily solved by using today's computer technology with data analysis applications such as spreadsheets having removal functions for redundant data, or using SQL queries with the UNIQUE function in Relational Databases. In addition, data visualization techniques can be used to test the correctness of the data, as well as statistical techniques such as Outlier Analysis to see the data points that fall outside cluster sets, or using computer code to validate the data. Also, in this regard, ontologies have become a popular technology in many research fields. Ontologies can effectively solve many of these problems. For example, WordNet [51] can be used to effectively solve the ambiguity of synonyms, polysemy, and misspellings [52].

Of those problems, missing data is the most difficult problem to resolve. Missing data values results in less reliable forecasts [53]. Therefore, several researchers have studied how to deal with data loss. This common problem in data collection can be classified in three ways. From [54], the first category is termed data Missing Completely At Random (MCAR) which means that a missing data point is a completely random occurrence. The second type is termed data Missing at Random (MAR) which means (rather obliquely) that for a data point to be missing, that is not related to the missing data, but it is related to some of the observed data. The third type is data Not Missing at Random (NMAR), which is a characteristic of lost data, not based on other information.

In the dengue outbreaks domain, the lost data is often MCAR. The missing data randomly occurs in the observations where the variables such as geographic information (region, province, and district), climate data (rainfall, wind speed, relative humidity, and temperature) are independent and unassociated. If a high amount of missing data occurs, the prediction accuracy and reliability of the model could be low. Therefore, researchers have studied how to effectively handle this problem and have identified solutions.

Three approaches have been devised to handle the problem of missing data [54], [55]. The first approach is to manage the lost data by deleting the lost data records. The method, termed Listwise Deletion [56] and Pairwise Deletion [57], makes no attempt to generate values for the missing data, but deletes the entire record of lost data. This method is appropriate when the size of the dataset is large and the remaining data is adequate for a forecasting model to learn.

The second approach is statistical analysis [58] where values for the missing data are inferred by calculating the mean of the available numerical data, or replaces lost data with the mode of the categorical data. These methods are considered as the easiest way to substitute the lost data. However, the disadvantage of this method is that the values created to replace the missing data are not necessarily close to what the missing data may have been if present, thus making the forecasting model less efficient. Donner [59] used the regression technique to estimate missing values, and Andridge and Little [60] used the Hot Deck Imputation method, which imputes the missing data from an observed response from a ''similar'' unit.

The third approach to missing data imputation or inference is by applying Machine Learning, such as the use of Neural Networks [61], [62], Genetic Algorithm (GA) [63]–[65], K-Nearest Neighbor (KNN) [66] and Particle Swarm Optimization (PSO) [67], to impute lost data.

Another method of data substitution called Multiple Imputation [68] which is a statistically based technique comprised of three steps: imputation, analysis and pooling: (1) Create N datasets, each with imputed values for the original missing data. Imputed values can be different in each of the N completed datasets. (2) Analyze each of the N completed datasets. This step results in N analyzes, and (3) Integrate (pool) the results of the analysis of the N datasets into a single set (final set). This technique is an effective method but the computational cost is high because of the need to construct the N datasets and analyze all of them [39].

Most researchers perform data cleansing by eliminating outliers. Outliers are values that fall outside the range of the standard deviation (SD) (Figure 8). Calculating the mean and standard deviation of the data is the most popular method for detecting outliers [69]. In addition, some researchers incorporate Chebyshev's theorem [70] in their statistical detection of outlier data. Sarfraz *et al.* [71] applied statistical

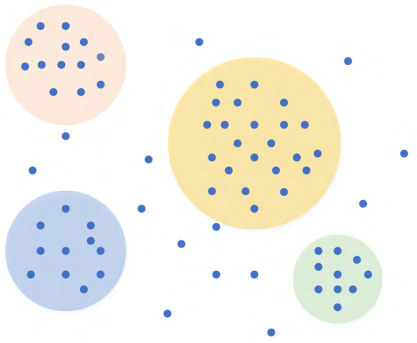


FIGURE 8. Clustering technique to detect outlier data that fall outside of the cluster sets.

methods for outlier detection and elimination as a data cleansing process in several studies dengue infection, ensuring thereby that their data was complete and reliable before starting further analysis. Clustering algorithms have also been used for eliminating the outliers [72]. Clustering algorithms combine similar data into a group, ensuring, thereby, that the information in any group is free of outliers. Clustering is an iterative process, which may therefore not be a suitable method for large datasets due to its high computational cost. The association rule technique [73] identifies erroneous data. Maletic and Marcus [74] and Marcus *et al.* [75] used the ordinal association rule technique that has 2 processing steps: (1) Identifies data relationships using the Apriori algorithm [73], and (2) Prune out the outliers identified as not conforming to the rules of association.

C. FEATURE SELECTION

Typically, the forecasting techniques will focus on as many factors as possible, adding a greater variety of dimensions that potentially make the forecasts more effective. The factors used to model dengue infection prediction include climatic conditions (weather, rainfall, temperature, humidity, and seasonality), demographics, population movements and migration patterns, geographical features, and rates of dengue infection in mosquitoes in particular locations. The inclusion of this variety of factors, however, does not ensure better forecasting performance. Factors that are not correlated with the dependent variables will lower predictive accuracy as they could merely be noise and irrelevant for the prediction model.

Normally, forecasting models based on Machine Learning algorithms can process many variables effectively, but including only those factors that are highly correlated will make the model more accurate and require less computational processing. Both manual selection and automatic selection of factors could be applied [39]. Manual selection can be performed with the assistance of medical experts while the automatic approach usually uses feature selection algorithms such as statistical methods or machine learning methods. Some data mining techniques can also be used for feature selection. For example, Buczak *et al.* [76] used fuzzy association rule mining to find relationships in epidemiological data, environmental data, and economic and social data and then eliminated some factors that proved less important to the model, which increased the efficiency of the forecasting model. Most of the literature concluded that climatic factors such as rainfall and season play a greater role in the disease transmission during outbreaks than other factors. However, Siriyasatien *et al.* [17] used the Pearson Correlation Coefficient to select factors that highly correlate to dengue cases. They found that season and the dengue virus infection rate of female mosquitoes were significantly correlated with the number of cases. In addition, they also discovered the model that consisted of those two factors achieved higher prediction accuracy over forecasting models that included additional factors of rainfall, temperature and relative humidity.

Selecting useful and relevant variables for a particular problem of interest is called *variable selection* or *attribute selection*. It is different from the dimensionality reduction approaches, such as PCA (Principle Component Analysis) or LDA (Linear Discriminant Analysis). Although both approaches decrease the number of features, dimensionality reduction seeks to construct new combinations of features from the original features, resulting in fewer features remaining that are then easier to interpret and understand in a simpler, faster and more cost-effective process. Feature selection is useful by itself and can be used as a *filter method*, a *wrapper method*, or an *embedded method* [77]–[79]. Statistical measures are applied to each feature in the filter method and used to identify either useful features to be kept or, alternatively, finding irrelevant and redundant features that can be removed. This results in improved accuracy of the predictive model. Filter methods are often implemented in the univariate case by considering each feature independently or with respect to the dependent variable. For instance, Information Gain, Euclidean Distance, the Pearson Correlation Coefficient, and Chi-squared test are measures of association.

Wrapper methods create different combinations of features using searching processes and these new combinations are evaluated and compared based on model accuracy. Some examples of wrapper methods include sequential forward selection, sequential backward elimination and recursive feature elimination. The embedded method incorporates feature selections as part of the training process. Features that best contribute to the model accuracy are simultaneously identified as the model is being formed. The common example of embedded methods are regularization methods, also called penalization methods, such as LASSO and ridge regression, which allow bias into the model by adding some constraints in the optimization process to mitigate the over-fitting problem, but with fewer features, creating a less complex model.

D. DATA TRANSFORMATION

When a data cleansing process has completed its task, the next step is transforming the data to make it suitable for processing by various techniques because the original value may not be appropriate in some cases. For example, dates of birth could not be used directly and need to be transformed to age

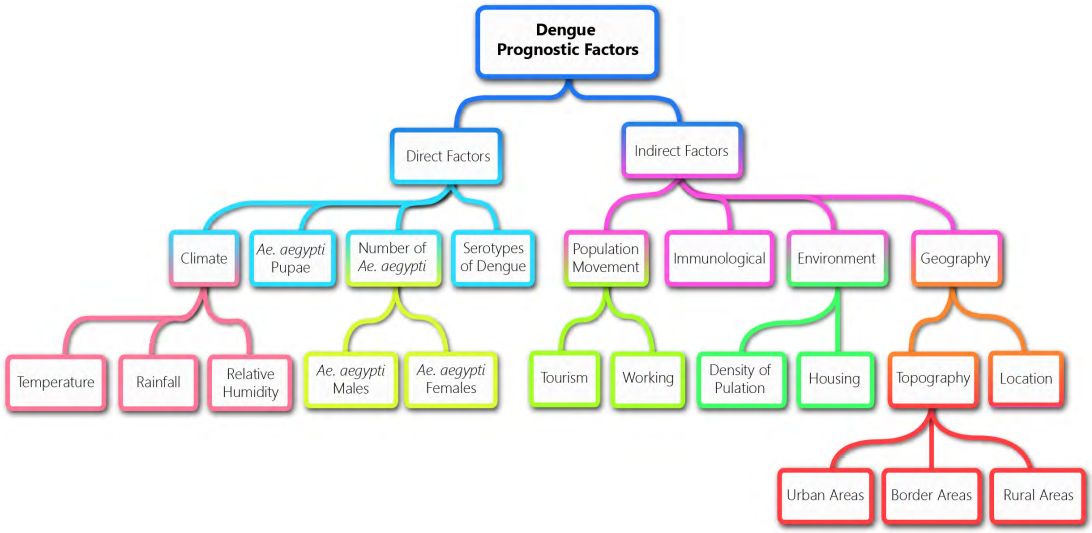


FIGURE 9. Dengue prognostic factors which are typically used in forecasting models, usually proven by several researchers as having a positive correlation with dengue cases occurrence.

which is more informative. Therefore, data transformation is a process to map data to new values or types for further use in other steps.

Data conversion can be done in several ways. For example, 1) Discretization and 2) Concept Hierarchy Generation. Discretization is the method suitable for numerical data such as patient age. This data will be represented as interval labels e.g. 1-7 years, 8-14 years, or 15-20 years or as a conceptual label e.g. youth, adult, or senior. Conceptual hierarchy generation is also a method for nominal data used to transform text, such as location information, into higher-level concepts including perhaps city, province, or country, so simplifying this data to make the forecasting model more efficient [43]. Data transformation may use clustering techniques that are very popular discretization methods for numerical data, in which cluster data is based on the distance between data points. Decision trees are another popular scheme to discretize numerical data using the top-down splitting approach. In addition, Correlation analysis can also be deployed for the discretization task. The ChiMerge technique [80] works based on the chi-square statistic for discretization numerical data. The main difference between the ChiMerge approach and the Decision Tree is that the ChiMerge approach works as a bottom-up approach to form a larger interval. It considers that members of two datasets that have distribution similarity will be merged together until the conditions for predefined stopping criterion is met.

V. DENGUE FACTORS

Researchers around the world have conducted research on dengue prediction in many countries and various factors have been used to study the correlation between those factors and the number of cases. The objective of their studies is to find the factors that are significantly correlated to the number of cases and deploy those factors in the forecasting model. As a result, the forecasting model can effectively work on predicting the severity of outbreaks in the future. Therefore, policy makers or responsible staff can be warned in advance and can prepare protocols and all relevant resources for the coming epidemic. In this report, we categorize dengue factors into two main groups as shown in Figure 9: 1) Direct Factors and 2) Indirect Factors.

A. DIRECT FACTORS

Direct factors refer to the factors that directly affect the mosquitoes' live cycle, especially that of *Ae. aegypti*. This includes the insects themselves as well as the spawn and larvae which will increase the number of carrier mosquitoes and, consequently, increase the opportunity of disease outbreaks occurring, and diffusing. In addition, the transmission of dengue virus serotypes from each mosquito parent, which may be different in each, to the larvae, can mean there are two serotypes in individual larvae that can more easily spread the dengue virus and the chance of an outbreak is higher than where the mosquito is infected with only one serotype. Therefore, eradicating the mosquito larvae at the early stage before they become adult mosquitoes will reduce the risk of a dengue outbreak in the area.

1) CLIMATE

Rainfall is a factor that affects the incubation period of mosquitoes, which can only complete their lifecycle in still or standing water [81]. Conversely, unusual weather conditions of drought and higher than the usual temperature, such as occurs in the El Nino phenomenon, adversely affects the number of mosquito breeding habitats, and therefore populations of mosquitoes. However, this may be significantly offset by the situation in drought and high temperatures with the human

practice of storing or hoarding water for consumption, making available many water containers full of still water, which provides a good venue for mosquito living and breeding. In addition, global warming is also affecting mosquito breeding cycles, with mosquitoes that usually hibernate during cold weather being more active in the warmer temperatures now being experienced in those locations. Global warming also enables the spread of insects such as mosquitoes by extending their geographical range.

These direct factors contribute to dengue outbreaks, and public health authorities should pay more attention to the presence of mosquito larvae and extend the monitoring of mosquito larvae for virus infection. Focks and Chadee [82] and Seng *et al.* [83] conducted research into mosquito larvae population densities which strongly suggested that the number of mosquito larvae in a vicinity is an effective predictor of dengue epidemics. There is little doubt that both the mosquito population size and climate factors are significant factors in the occurrence of dengue epidemics. Given these findings, Pham *et al.* [84] suggested that monitoring and controlling the number of adult mosquitoes, and mosquito larvae at the specific times of high rainfall and high temperature would successfully reduce the opportunity for the disease to spread.

The increase in global temperature may lead to the spread of epidemic areas and to a greater incidence of the disease in insect vectors. An increasing proportion of mosquitoes infected with dengue directly affects the number of dengue patients. Naish *et al.* [85] discovered that the ecosystem conditions and climate factors directly affect the incidence of mosquito vectors and of the dengue virus, both of which are sensitive to changes in temperature, rainfall and humidity. Gharbi *et al.* [86] found that as the temperature rises so does the extent of the spread of dengue infection. McLennan-Smith and Mercer [87] showed that increasing the temperature from the usual range of 26°C-28°C up to 30°C reduces the incubation period of DEN-2 and DEN-4, allowing the faster spread of the dengue virus. The monsoon season, accompanied as it is by higher temperatures and significantly higher rainfall, is always accompanied by an increase in the incidence of dengue outbreaks. Karim *et al.* [88] found that the high relative humidity experienced during the monsoon season accelerates the growth rate of the larvae and results in a higher survival rate and longer life of the mosquitoes, leading to the greater growth and spread of the carrier mosquito population and, consequently, the spread of the dengue virus. This poses a very serious and higher risk of dengue becoming endemic in the area.

Research from Taiwan [89] also found that a month with an average temperature above 18°C created a high-risk of dengue infection, and the risk was higher in urban areas than in rural areas. They also found that an increase in average temperature of 1°C results in a rise in the mosquito population, with an associated rise in the risk of dengue infection by 195%. By this measure, therefore, if Taiwan experiences a temperature > 18°C, the population in the risk area will have double the risk of a dengue outbreak. Likewise, Hii *et al.* [15] identified the interesting situation that an increase in the average weekly temperature and in the amount of rainfall could be used as an early warning prediction three months prior to the outbreak, giving authorities the opportunity to intervene to reduce the spread of the disease. Elsewhere, in Texas in the USA, Reiter *et al.* [90] found that, given the high temperatures experienced in that state, more people used air-conditioners in their homes, which would be closed to outside conditions. This lowered the risk of dengue virus infection of the population as it greatly reduced the contact between humans and mosquitoes. Brady *et al.* [19] also found that temperature was one of the most prominent factors contributing to the prediction power of outbreaks. As a follow-on to Brady research, Tazkia *et al.* [91] developed the Dengue Early Warning System (DEWS) to calculate the probability of outbreaks of dengue, based on climate and environmental factors, which demonstrated a prediction accuracy up to 97.05%.

However, research by Karim *et al.* [88] found that various dengue strains, the body's immune system, and environmental variables could be other factors associated with dengue outbreak, which clearly implies the need for multi-dimensional analysis to support predictions.

Relative humidity is one of the factors that correlate with dengue outbreaks [92]. Xu *et al.* [93] used absolute humidity to predict dengue outbreaks and confirmed that absolute humidity is a factor that should be used as one of the factors in a prediction model of dengue outbreaks. Humidity levels influence mosquito hybridization, dispersal, feeding habits, egg production, larvae survival rates, mosquito breeding success and hatching rates, resulting in a greater and more widespread dengue virus transmission. These lifecycle factors are affected by increases, and therefore also decreases, in temperature and similar changes in relative humidity. For example, if the relative humidity decreases, it may cause an increase in dengue outbreaks because mosquitoes are more active in seeking their hosts in a low relative humidity environment. This information is consistent with the research by Xu *et al.* [93], who examined the impact of weather factors on dengue infection in Singapore and found that relative humidity and temperature are important determinants in dengue epidemics.

Rainfall is a factor which determines the survival of mosquito larvae and eggs of mosquitoes [94]. Both the laying of eggs, and the survival of larvae, depend on, and need, standing or still water, which is obviously usually available where there is an abundance of rainfall. According to Runge-Ranzinger *et al.* [5], dengue outbreaks are most severe in the rainy season, and an increase in average daily rainfall provides increased the number and availability of habitats conducive to larvae survival and growth with the consequence of an increase in the mosquito population. Wongkoon *et al.* [92] concluded that the severity of dengue epidemics is highly associated with the number of rainy days and consequently the number of larvae. However, Hsu *et al.* [95] found that too much rainfall results in

the decline of dengue epidemics, which may be due to the volume of water creating flows of fast-moving water which is disruptive to larvae survival and growth and the thinning of the larval population. This assumption is consistent with several studies [96]–[98] where it was found that, with high rainfall, the mosquito eggs will be washed away with the resulting lowering of the number of mosquitoes and, thus, a decreased likelihood of, or severity of, a dengue outbreak. Sumanasinghe *et al.* [99] deployed rainfall, temperature, and population data together with the Support Vector Regression (SVR) technique to model dengue prediction in Thailand. The results showed that those factors are highly correlated with the number of cases, which can effectively enhance the power of the forecasting model. Obviously seasonal conditions are influential factors in the severity of dengue epidemics.

2) MOSQUITO DENSITY

In addition to environmental and climate factors, several researchers have examined mosquito carrier populations using a mosquito larvae density index. This consists of the percentage of houses in an area infested with larvae (House Index: HI), the percentage of water-holding containers infested with larvae (Container Index: CI), the number of positive containers per 100 houses inspected (Breteau Index: BI), and the number of pupae per 100 houses inspected (PI). These indices can be used to estimate the predicted dengue-transmitting vector population, referred to as the Aburas Index [100]. In that study, the number of cases was found to be significantly correlated with the mosquito density, and this was used as an important factor to effectively control an epidemic in Jeddah, Saudi Arabia. If the CI increases, the severity of dengue outbreaks will also increase. In the study by Strickman and Kittayapong [101], the risk of dengue outbreaks was predicted by counting mosquito larvae and pupae in the water-holding containers from 10 households in Chachoengsao Province, Thailand. The monthly count of pupae per household during the period May and June effectively indicated the risk of dengue infection in those two months. Clearly, eradication of mosquito larvae or pupae by inspecting and clearing domestic water containers of larvae and pupae every 7 days, or by emptying stagnant water from any containers, both in close proximity to houses, and within a 50-meter radius, which is the mosquito's flying radius, is a simple and effective means of reducing the spread of the disease.

However, the number and variety of factors that contribute to the occurrence of dengue infection are not universally agreed. The study by Ma *et al.* [102], for example, using data for the period 1998 to 2002 in Singapore, found that the number of mosquito larvae/pupae was not correlated to the incidence of disease which was more due to ecological factors, social and economic conditions or the habitat of the low-income population. We feel that this does not contradict other research, such as reported above but must be seen as indicating the complexities of disease prediction.

Female *Ae. aegypti* are important carriers of dengue and yellow fever virus by their ingestion of human blood [103]. Thongrungkiat *et al.* [104] studied the mechanism of dengue virus transmission from both male and female mosquito larvae to eggs in breeding sites in Bangkok. They found that, while male mosquitoes had a higher rate of dengue virus infection than in female mosquitoes, with DEN-4 being the most found, both male and female larvae are important factors that significantly affect the spread of the outbreaks. Ponlawat and Harrington [105] stated that understanding the biology of mosquito breeding is essential for studying the behavior of mosquitoes, gene flow, and the structure of mosquito populations, and that genetic control can be used to study the effects of mosquito age, size, and sperm production levels of male *Ae. aegypti* and *Ae. albopictus*. That study illustrated that larger male mosquitoes, older than 25 days, can produce large amounts of sperm. If these male mosquitoes breed with large-sized female mosquitoes, dengue infection and yellow fever epidemics are more likely.

3) DENGUE VIRUS SEROTYPES

There are various serotypes of the dengue virus, each of which can cause dengue epidemics. Studies over a period of 10 years showed that dengue epidemics that occurred in the northern region in Thailand were caused by two strains of dengue. During the period 2003-2005 the serotype DEN-2 was the main cause of epidemics in that region, while between 2006 and 2009 the serotype DEN-1 was prominent. The serotype DEN-2 was again active in 2010-2011, and the serotype DEN-1 in 2012. This demonstrated that these were the two major strains found in the region and that the incidence of each fluctuated year to year. As previously discussed, anyone infected with one serotype will gain lifelong immunity to that virus serotype, but protection against infection from the other dengue virus serotype is only partial or temporary. In the situation of the fluctuating incidence of DEN-1 and DEN-2 as just stated, the populations of northern Thailand had no immunity from period to period, as the other serotype became predominant, resulting in continual infections being experienced [106] and this correlated with the density of the dengue virus in the region.

The epidemiological data for Thailand over the past 30 years showed that both the DEN-1 and DEN-4 serotypes have co-existed, but the DEN-1 strain was usually predominant, and most infections were from the DEN-1 virus strain. However, this was not a constant situation, with both DEN-1 and DEN-4 incidence fluctuating. Limkittikul *et al.* [107] found that the distribution of dengue serotypes varies from season to season, at different times, and in different places. The incidence of the different strains of dengue in each region of Thailand and other countries in Southeast Asia is clearly affected by the various factors of population density, population movements, the number of infected mosquitoes, environmental changes; natural or man-made, social infrastructure growth and change, as well as the general state of health of the people in each area. These data can be used

for surveillance and control of dengue infection or for future vaccination.

Apart from DEN-1, DEN-2, DEN-3, DEN-4, Mustafa *et al.* [108] recently reported a fifth serotype of dengue virus (DEN-5) in the state of Sarawak in East Malaysia. It is possible that the new virus could derive from the DEN-4. Additionally, only one case has been admitted and, thus, the disease caused by DEN-5 has never been used in any prediction models at the present.

4) BITE RATE

According to WHO [109], *Ae. aegypti* and *Ae. albopictus* usually have high biting activity from early morning to twilight. Chompoosri *et al.* [12] found that the biting behavior of *Ae. aegypti* varies between seasons in different provinces in Thailand. The highest bite rates occurred in summer, with 30 individual mosquitoes/person/hour while *Ae. albopictus* had the highest biting rate in winter at 18 individual mosquitoes/person/hour [110]. Other research, from India, showed that *Ae. aegypti* and *Ae. albopictus* in rural and urban areas have the highest bite rates from 3 a.m. to 6 a.m. and the lowest bite rate is during 12 a.m. to 3 a.m.. Different severity levels of the epidemic depend significantly on the biting rate of the mosquitoes, which is one of the factors significantly associated with disease outbreaks and is a factor that could be used in the prediction of outbreaks. The bite rate of mosquitoes will vary from season to season and depends on weather and other climatic conditions, and mosquito density.

B. INDIRECT FACTORS

Indirect factors refer to factors that do not directly affect the number of mosquito larvae, but which may be related to the dengue epidemic. Such factors include topography, location, population movement, environment, and immunity.

1) GEOGRAPHY

Topography is an important factor for dengue vector urbanization because it could provide an appropriate environment for mosquito growth, reproduction, and virus transmission. A study by the Center for Disease Control and Prevention [111] indicated that border areas have a higher risk for dengue outbreaks than other areas. For example, dengue outbreaks occurred in July 2005 in the Brownsville area of Texas, which is near the Mexican border, and the dengue outbreaks continued in Tamaulipas, Mexico into August 2005, where 223 cases were reported. The outbreaks have also spread into new areas along the Mexican border because people crossed the Mexican border to visit their families, to buy consumer goods, to work, and to seek health care services. Furthermore, the lack of good infrastructure development or planning for accommodation to support the dramatic increase of population over the past decade has resulted in many people in Colonias in Tamaulipas not having good sanitation services. In addition, the city has been affected by global warming that causes higher temperatures and storms, resulting in widespread flooding in the area, which has no proper method to disperse or drain or treat the water. These factors have resulted in Mexico being a high-risk area for dengue outbreaks [112].

Rural areas in developing countries have also reported a rapid spreading of dengue infection due, it would appear, to the lack of good public health resources. Severe dengue outbreaks occurred simultaneously in rural areas of many countries in 2010-1011. Apart from those areas, the new locations where dengue appeared for the firstl time were Croatia and France in 2010, the Madeira Islands and Portugal in 2012, Florida, the United States and Yunnan, China in 2013. Consequently, the mortality rate in those areas is increased [113]. Dilip *et al.* [114] reported in September 2012 that dengue epidemic in the rural areas of West Bengal of India that Gopalpur village had the highest number of cases at 37% of the population, followed by Badalpur, 26%, Tajpur, 23%, and Ramchandrapur, 14%. Of these cases, 19% had migrated from areas where the disease is endemic and there were no plans for the prevention of breeding of mosquitoes. A study of dengue surveillance from 2006 to 2008 in Cambodia [115] found that the incidence of dengue infection in pre-school children has also increased in both urban and rural areas rates from 1.5 to 211.5 people per 10,000 people. In 2007, an outbreak of DEN-3 occurred in rural areas of Cambodia where communities do not have proper water resources management and the disease is spread by people in rural areas travelling between their villages and Phnom Penh.

Those examples indicate that location is one of the factors influencing the epidemic of dengue; urban and rural areas. In urban areas, which are developed areas with good infrastructure, where the living environment is hygienic, and most people have good economic and social status, and importantly where water management is superior to that in rural areas and therefore should be free of dengue disease, dengue disease is still endemic in cities in Mexico, Colombia, Ecuador and Brazil. One significant source of mosquito breeding and propagation is the practice of collecting and storing water for consumption in large containers, which are difficult to clean regularly [116] and, being standing water, is ideal for mosquito breeding and propagation. Travel from rural areas to urban areas is now a common cause of dengue epidemic in urban areas [117] together with greater population density. In Cambodia, water containers which are usually uncapped or uncovered have been reported to be the most common cause of dengue among children over 7 years old [115].

Given this reported relationship between rural and urban areas, if urban epidemics can be forecast in a timely fashion, this can be extrapolated into warnings that are more effective, and will allow daily surveillance of the disease. In 457 villages within Kaohsiung, Taiwan [118], during 2009-2012, researchers used information on the occurrence of dengue outbreaks for the past 30 days and were able to effectively predict the outbreaks of dengue in the villages on the next day.

2) SPATIAL AND SPATIAL-TEMPORAL INFORMATION

Some researchers studied the influence of spatial analysis on dengue prediction and showed that this information can improve the performance of the prediction model. For example, Yu *et al.* [119] proposed a spatial-temporal dengue prediction model characterized by population density, environmental conditions, and infrastructure factors in Kaohsiung City (Taiwan). The models effectively predicted the spatial diffusion of the dengue fever epidemic. However, the model can be influenced by various external factors, such as virus serotypes and human intervention. Some researchers have suggested that spatial information was not adequate to enhance the prediction model. For example, Costa *et al.* [120] introduced a spatial-temporal model to predict dengue fever in Rio de Janeiro. They found that it is difficult to predict the disease diffusion using only temporal or spatial models as *Ae. aegypti* is strongly sensitive to local climate triggers. This view is consistent with the results of Thiruchelvam *et al.* [121] who discovered that dengue incidences were localized, and therefore results may not be able to be applied to other areas or timeframes/seasons. The best feedback models for a locale were based on feedback from that localized region only and did not benefit from data from neighboring regions.

3) POPULATION MOVEMENT

Population movements and migration between locations are two of the factors that affect the spread of epidemics of dengue, whether it is traveling for tourism or work, or for other reasons. Population movement also caused the spread of the CHIKV to Europe and the United States, which has drawn global awareness. International travel is now considered as one of the significant factors for the quick spread of the outbreak, thereby closely associated with an increase in population, and increased tourism, transportation, population movement, and the importation of merchandise from overseas.

Ten cases of dengue were identified as having occurred during trips to Puerto Rico, the Virgin Islands and the Hawaiian Islands between May 27, 2001, and January 30, 2002, all areas where the disease is endemic [122].

A report by Baaten *et al.* [123] found a total of 1,207 tourists from the Netherlands were infected with the dengue virus during a short trip to a dengue-endemic area. Ratnam *et al.* [124] reported that tourists are at risk of dengue infection in the proportion of 14.6/1,000 people/month, with an increase in risk if international travel was undertaken during the rainy season. Following a tour by travelers from Papua New Guinea, the number and severity of outbreaks of DEN-2 increased in Cairns in northern Queensland, and the Torres Strait, between 2003 and 2004. Approximately 900 cases including 3 severe cases and 1 death are confirmed. It can, therefore, be concluded that travel is one of the factors leading to the spread of dengue infection in new areas by the introduction of new dengue pathogens from one area to another [125]. As international travel increases, the situation will obviously deteriorate with the possibility of infections from newly introduced serotypes of dengue virus increasing [126], [127]. Therefore, information on travel patterns can be used to assess the probability of travelers being infected with dengue virus after returning from their tour as well. In addition, Reiner *et al.* [128] used population movement data, together with data on population densities, particularly on domestic and family situations, to predict the probability of dengue epidemics in the northeast of Peru. In this situation, family reunions at particular times, such as religious festivals and social occasions when family members returned from other cities, tended to be accompanied by dengue outbreaks.

Population movements, as a cause of the spread of dengue outbreaks, also have a significant impact economically, particularly at the local level. A study by WHO [113] found that in eight countries, the loss of working days of dengue patients was on average nearly 15 days and the average cost of treatment was $514, while hospitalized in-patients lost nearly 19 days of work at an average cost of $1,491. This will clearly influence the cost of providing health services. As well, companies and enterprises can be badly affected due to working days lost, with workers succumbing to the disease.

In addition to population movement, trading is another factor indirectly related to the spreading of dengue outbreaks. Hawley *et al.* [129] reported that dengue spread from Asia to North America and Europe due to international trade in items such as used tires and bamboo home decorative items, which are sold all over the world. These items can be the habitat of *Ae. albopictus* [113].

4) ENVIRONMENT

Urban growth provides many breeding areas for dengue vectors. Mosquitoes tend to prefer to feed on humans rather than animals, and therefore can adapt well to the urban environment. Mosquitoes can fly over a range of 100-500 meters to find food and to spawn, increasing the risk of mosquito-borne infection because houses are within the mosquitoes' flight range. One factor in determining the dengue epidemic is the relationship between the biology of mosquitoes and the behavior of people living in urban areas. The density of residences is associated with dengue in some parts of Southeast Asia. Machado-Machado [94] found that in Bangkok and Manila, the *Ae. aegypti* has replaced *Ae. albopictus* as rapid growth was occurring in populated urban areas, together with increased population migration enabled by modern transportation. These are risk factors leading to the geographic expansion and the increasing density of mosquito populations [92].

It has also been found that the majority of outbreaks occur in communities where housing conditions are unhygienic and, or the population is malnourished. This latter situation implies lower immunity to and a greater risk of disease infection [113]. Toan *et al.* [130] found 73 cases of dengue in Hanoi, Vietnam, in an area where people were living in

rented houses located near a drain with stagnant water. The risk to these patients of dengue was higher than elsewhere and people living in unhygienic conditions or where wastewater is discharged directly into the ponds, are more likely to suffer dengue outbreaks as well. In Kaohsiung and Fongshan, Taiwan [131] ditches flow through the city creating a major risk factor for dengue outbreaks every year. At the mouth of the Ai River, both sides of the river are densely populated business areas, with dengue outbreaks occurring every three years resulting in an infection rate of 1/10,000 people, except in the Cianjhen district where there was a rate of dengue infection double that of other districts. In one particular area, the Sinsing district, the occurrence of dengue infection is more frequent, occurring every 1-2 years. Therefore, this study can be used as a guideline for government agencies to assess the frequency of the disease and to allocate budgets for disease prevention and control in particular areas.

Cheong *et al.* [132], in a study in Selangor, Malaysia of the period 2008-2010, found that agricultural land use methods, and water and forests, are closely related to the incidence of dengue because the right environment to live and reproduce is a causal factor of the expansion of the disease in the area. The most violent outbreaks of dengue in cities, such as occurred in Kaohsiung, Taiwan between January 1998 and December 2011, were attributed to the nature of the residential, agricultural, and forestry environments [95].

5) IMMUNOLOGY

In Thailand, there are four varieties of dengue serotypes [133], all of which share a certain antigen, thus providing a cross-reaction and cross-protection between the dengue serotypes, for a short period. When a person is infected with a dengue virus, they will be immune to that particular dengue virus type forever. This is termed long-lasting homotypic immunity [134] and also provides cross-protection against other serotypes (heterotypic immunity) for a short period of between 2 and 9 months, after which the infected person can be newly infected with another serotype of dengue infection different from the first. This is termed a secondary dengue infection [135]. This situation means that people who live in areas where the dengue virus is endemic may be infected 3 or 4 times, once for each of the four dengue serotypes.

In Bangkok, the proportion of the DEN-1, DEN-2, DEN-3, or DEN-4 will vary year to year. Viral and epidemiological studies have concluded that this variability is the major cause of dengue infection in areas where the dengue virus is both endemic and multiple serotypes exist, or where sequential outbreaks of different types of dengue occur: infection with one virus serotype does not give any, or lasting immunity against other serotypes. Especially in densely populated areas, recurrent infections and repeated infections are therefore common. DEN-2 is a high-risk for dengue infection, especially if the secondary dengue infection follows the first infection caused by DEN-1. Patients with low levels of antibodies against dengue virus, such as children first exposed to the dengue virus, are more likely to experience recurrent dengue virus infections. Typically, children have a passive dengue antibody from their mother, but this is insufficient to prevent dengue virus infection (non-neutralizing antibody). They usually only have enhanced antibodies. This is similar to the phenomenon that occurs in repeated infections in older children, called ''Antibody-dependent enhancement.'' Repeated infections of dengue caused by previous infections of various dengue serotypes lead to the occurrence of enhanced antibodies. This may be a factor encouraging the increase in the number of new serotypes of dengue virus in the monocytes of white blood cells.

Some researchers [136], [137] found that patients with dengue virus infection often do not have good nutrition [138]. Nutritional deficiency affects the immune response of the body when it is infected with dengue virus. It has been found that a change of T and B cells secrete chemokines and causes T-cell degeneration [139], [140] which in turn causes autoantibodies against platelets and mucosal cells, resulting in thrombocytopenia [141], [142]. Oki and Yamamoto [143] study of epidemics of dengue using population immunity levels in a mathematical model, found that the number of epidemics occurring is highly associated with the number of female mosquitoes per person in the local population (MPP). Therefore, population immunity can be used to predict MPP more effectively than ever before. The study also discovered that when the population immunity declined, the severity of the dengue outbreak would increase.

The development of a dengue vaccine has been in progress since 1929 and continued to evolve over the next 25 years [144], [145]. Prior to 1952, a vaccine had been developed from DEN-2, cultured in rat brains, but it was not successful because it was not safe for humans [146]. Vaccines have also been cultured in other parts of animals, for example, the primary dog kidney cell (PDK) and the development of recombinant subunit proteins and DNA vaccines. The ideal dengue vaccination vaccine must provide safe, preventive and immunogenic treatments for four serotypes. In late 2015 and early 2016, a vaccine for dengue, first named ''Dengvaxia'' (Chimeric Yellow Fever-Dengue Tetravalent Dengue Vaccine: CYD-TDV) was developed by Sanofi Pasteur [147], and has already been used in many countries for patients aged 9 to 45 years who live in areas at risk for outbreaks. However, WHO advises that the vaccine should only be used in areas with a high prevalence of dengue infection and the study by Aguiar *et al.* [148] suggested that to achieve the most effective results Dengvaxia should be used to vaccinate people who have already been infected by at least one dengue virus. The vaccine has been licensed and used in Brazil, the Philippines, Singapore, Mexico, Indonesia, Thailand, Paraguay, Peru, Costa Rica, El Salvador and Guatemala. However, in the face of recent reports on severe adverse effects experienced in some cases, Dengvaxia has been suspended [149]. Other vaccines are still in the third phase of development and are still only in the initial trial stage in the laboratory. At this time, the efficacy of dengue vaccines is still under review.

VI. DATA REPRESENTATION

The data representation of health information is constantly evolving, electronic data collection is increasing, and diverse forms of data collection are available. According to various studies, three data models are extant: 1) unstructured data, 2) semi-structured data and 3) structured data. Exploiting the information structured in these various ways is a huge challenge [150]. Appropriate dengue information collection will enable applications that access and analyze data more accurately, quickly and more effectively in supporting decision-making in policy setting, control and disease prevention. Dengue prediction has been reported in three data representational models, as shown in Figure 10.

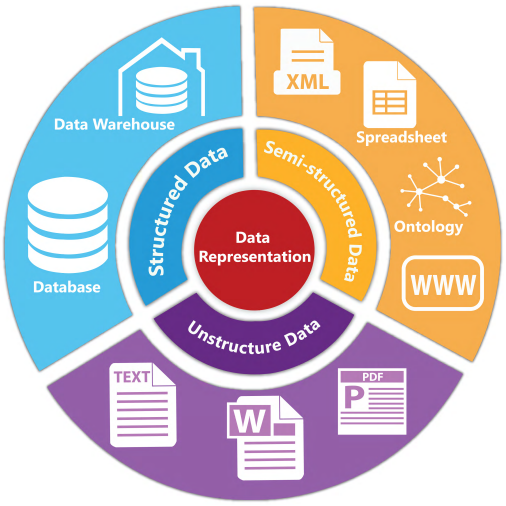


FIGURE 10. Popular data representation techniques for dengue information: Unstructured, Semi-structured, and Structured data.

A. UNSTRUCTURED DATA

Unstructured data is stored as ''freehand'' text or, in modern times, photographs, graphical images, videos and natural language, voice, recordings. Low-cost, large-volume disk storage and the availability of Cloud-based data storage (section IX-D) has overcome the problem of large volumes of data storage needed for particular types of unstructured data. Many applications are now available, some as freeware, many at a low cost, to access unstructured data, albeit often in relatively simple and unsophisticated ways. This means that some types of unstructured data require complex processing systems, such as for Natural Language Processing and Text Mining. The emergence of Big Data technology (section IX-C), however, has led researchers to turn their attention to unstructured data processing methods, as this is the data storage mode of ''big data.'' Jovanović and Bagheri [151] attempted to annotate unstructured data to enhance the efficiency of accessing and processing. Their approach included special annotation schemes for semantic-based text management, curation, indexing, and searching. This method is called *biomedical semantic annotation*, which can be adapted to different areas of biomedicine.

B. SEMI-STRUCTURED DATA

Semi-structured data is data that is not stored in a well-structured form, as in an RDBMS, for example. A relatively new form of data storage is where the data content is stored together with tags or markers indicating data type and other data characteristics. These tags and markers are referred to as the markup language, and allow a variety of data types, data structures, and data types (say text and images) to be stored together. XML and HTML are examples of markup languages that are used in this way. The main advantage of this form of data is that it is universally accessible by any software that comprehends the markup language. For example, XML datasets can be loaded into spreadsheets or report generators and presented in a structured way.

1) SPREADSHEETS

One powerful example of this is where geographic information systems can use the ArcMap v9.3.1® to generate population, demographic and dengue incidence maps based on data in a spreadsheet. This was done in a study of the population of Saudi Arabia [152], with maps showing the number of dengue infections in that population. Clinical and laboratory data was collected in spreadsheet format and analyzed by the SPSS program or NORM program to model the shock predictions of dengue infection [153].

2) HIERARCHICAL DATA FORMAT

Hierarchical Data Format (HDF) developed at the National Center for Supercomputing Applications and supported by The HDF Group, is designed to store and organize large amounts of data. The structure of HDF5 (Figure 11), the most recent iteration of HDF, comprises (1) Groups: folder-like elements within an HDF5 file that might contain other groups or datasets, and (2) Datasets: the actual data contained within the HDF file. Datasets are often stored within groups in the file. HDF is used to represent information contained in websites on the Internet, and to render that data visually on computer screens. The information of interest here includes epidemiological, economic and climatic, and other, data that appears on the websites of various agencies. These data are analyzed in prediction models for predicting the likelihood, severity, and timing of dengue outbreaks in specific locations.

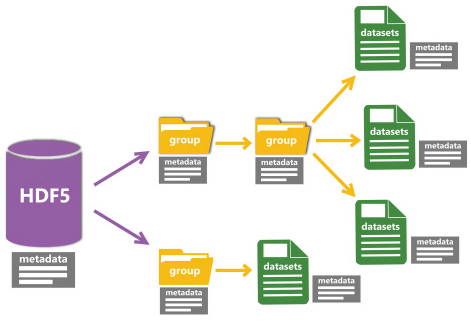


FIGURE 11. An example HDF file structure containing groups, datasets and associated metadata.

Buczak *et al.* [81] developed an epidemic prediction system that improved the efficacy of dengue prevention measures, using mean temperature and rainfall data recorded at Singapore Changi Airport, that was available in HDF format and originally obtained from the NASA Tropical Rainfall Measuring Mission (TRMM). The HDF format enables the system to manage extremely large, complex and heterogeneous data collections.

3) ONTOLOGY

One of the most popular forms of data representation today is the Ontology, which is a powerful tool now becoming more widely used to replace simple data storage systems. Ontologies can be used in many different areas of medical science, powerfully allowing the sharing of information and facilitate efficient semantic retrieval. Mitraka *et al.* [154] created the IDODEN ontology for dengue information, which includes the cognitive features of dengue disease covering all aspects of the disease, such as biology, epidemiology, clinical characteristics and entomology of dengue infection. Herdiani *et al.* [155] developed the Dengue Hemorrhagic Fever Ontology (DHFO) to collect data on dengue infection in Indonesia. They used an ontology to collect data from many sources, including other dengue ontologies, that are publicly available. This dengue ontology can be used for decision making in dengue control activities and for formulating preventative policies. This ontology is shareable with other researchers.

Lozano-Fuentes *et al.* [156] created a Vector Surveillance Model Ontology (VSMO) incorporating a database management system for more efficient data management together with a decision support system for dengue outbreaks, giving greater ease of data retrieval and data extraction, and exporting of data into various formats, such as .txt or .csv files, for further use. Typically, ontology storage is based on RDF (Resource Description Framework), RDFS (RDF Schema), and OWL (Ontology Web Language). Dias *et al.* [157] developed a decision support system for the control of dengue epidemics, based on data collected in the form of ontologies developed elsewhere, which featured the introduction of an ontology together with the application of a geographic information system. Effective management of epidemics in affected areas is a feature of that system.

C. STRUCTURED DATA

Structured data is stored in a structured format, such as in a Relational Database, in which the data is organized in a way that allows ease of access data. Structured data can be easily and efficiently manipulated by an appropriate Database Management System (DBMS) and accessed using the SQL language.

1) RELATIONAL DATABASE

In Relational databases, the data is normalized into rows (record) and columns (attribute) with limited redundancy. Such database systems are used to store disease information relevant to disease prediction and are the most widely used forms of data storage. The advantages of having data stored in this manner are:

- *Consistency*: available data in single or multiple files are consistent and data conflicts are avoided. Consistency implies same names, same data types, same field lengths with data content drawn from the same data domain.
- *Shareability*: the ability to share information, so that where data is either stored in a single database or where it is stored in different databases, is still shareable, primarily due to (1) above.
- *Nonredundant data*: the storage of any data ''fact'' only once in a single place which enhances shareability, and there is no duplication of data across accessible and shareable databases, especially if attention has been paid to concepts of Normalization of the data.
- *Integrity*: the absence of errors in the data giving users confidence that the data is good.
- *Standardization*: the ability to define data according to the same standard across diverse databases. Standardization includes well defined and known standards for data names, data types, field lengths and with like data content drawn from the same data domain.
- *Security*: the ability to protect the data from unauthorized access and possibly malicious attack.

However, databases are often subject to certain limitations:

- *Expense*: cost of acquisition, upgrading and maintenance database systems, which may be substantial. Total Cost of Ownership (TOC) includes the direct and indirect costs of ''owning'' a database, calculated over the life of the database.
- *Complexity*: although in recent times the complexity of owning, administering and accessing a substantial database system has decreased, it still presents certain complexities regarding data storage, database design, querying, etc.
- *Risk of system failure*: the risk of system disruption is greater where data is stored in a centralized database system, and especially if data back-ups are not available. Therefore, system failures either in the database itself or in the communications with the database will result in disruption to the entire system.

The Dengue Database (DENVDB) [158] is a free online database developed by the National University of Singapore, consisting of protein sequences, and data on virus strains, the year and the country of dengue outbreaks. The system facilitates data searches based on these data. Similarly, WHO has created a database called DengueNet [159] that can be shared and monitored for dengue epidemics worldwide. The data available is used in collaborations between WHO Regional and Country Offices, Ministries of Health, WHO Collaborating Centers and Laboratories. DengueNet contains a wide range of dengue data, including incidence, case fatality rates (CRF), frequency and distribution of dengue cases, the number of fatalities, and circulating virus serotypes.

TABLE 1. Examples of dengue databases.

Name	Collecting Period	Attributes	Sample size (records)	Support functions
Dengue virus database (DENVDB) [158]	1945-2009	protein sequences ,strain, year, and country	7,427 (26,247 sequences)	Search, Variability Analysis, Blast, Download facility
DengueNet [242]	1955-2011	incidence, frequency and distribution of dengue fever and dengue hemorrhagic fever cases, case fatality rates (CRF), number of fatalities and circulating virus serotypes.	n/a	Data query, Data entry, Interactive maps, Maps and resources, Reports
Dengue virus occurrence database [151]	1960-2012	OCCURRENCE_ID, SOURCE_TYPE, LOCATION_TYPE, ADMIN_LEVEL, GAUL_AD0, GAUL_AD1, GAUL_AD2, POINT_ID, UNIQUE_LOCATION, X:, Y:, YEAR, COUNTRY, REGION	8,309	Download
Dengue Drug Target Database (DengueDT-DB) [197]	n/a	Drugs Targets (Target Name, GenBank ID Protein, SwissProt ID, Gene Sequence, GenBank ID Gene, Synonyms, Protein Sequence, Number of Residues, Molecular Weight, Theoretical pI, Function, Process, Component, Specific Function, Pathway map, Reaction, Pfam Domain Function, Transmembrane Regions, PDB file Link, 3D Structure Image, Download PDB File, References) Drugs (Drug Bank Accession Number, Drug Type, Generic Name, Molecular Weight, KEGG ID, PDB Link) Genes (Dengue Virus Serotype, Database ID, Name of Gene, Strain, Accession No, Protein Id, Property) Attenuated Vaccine (Dengue Serotype, S/No., Accession No., Patent, Reference)	13 dengue drugs Targets, 80 dengue virus genes, 2 drugs information and genome sequences, 4 Attenuated Vaccine	Search, Download
OnTAPP [50]	n/a	n/a	659	Search, Upload, Download, Visualization (Mapping)
Dengue virus antibody database [44]	n/a	mAbs (Name, Host, Isotype, Immunogen, Event, History, PubMedID) Activity (mAb, Assay Type, Neutralizing, Units, Values, PubMedID) Epitope (mAb, Epitope Type, Method, Serotype, Protein, Domain, Sequence, PubMedID)	410 unique mAbs, 595 activity, and 545 epitope mapping	Search, Download
Virus Variation Resource [84]	n/a	Sequence type, Serotype, Disease Type, Host, Region/Country, Genome region,	Total 550,000 nucleotide seq.	Search, Download

However, Roberto *et al.* [160] found that dengue data collected by WHO was only 84% consistent with observed data from specific countries, indicating inaccurate data on more than half a million cases. Therefore, the WHO data on dengue epidemics was not useful. Messina *et al.* [161] developed a database for the 8,309 outbreak areas from 1960 to 2012 in four continents, including Africa, the Americas, Asia, and Oceania. This database includes 14 attributes, constituting a complete database of the current situation regarding dengue outbreaks in those areas. The Dengue Drug Target Database [162] is a database that collects information about dengue viruses, dengue virus genes, dengue virus targets, and drugs for drug development, epidemiology and comparative genomics and provides data on 13 new or existing drug targets, 80 genes of dengue virus and information on 4 drugs and 6 attenuated vaccines.

Chaudhury *et al.* [163] developed the DENV-Ab DB database containing 410 unique mAbs, 595 activity records, and 545 epitope mapping records. This database contains information on the origin of the data, immunogens, host immune history, and selection methods, as well as binding/neutralization data against all four dengue virus serotypes, and epitope mapping at the domain or residue level to the dengue virus E protein. Hatcher *et al.* [164] developed a system for virus validation in which viral sequence data is stored. The system includes modules on the influenza virus, Dengue virus, West Nine virus, Ebola virus, MERS coronavirus, Rotavirus, and Zika virus. Each module is based on GenBank and each can produce additional information, such as annotated genes and proteins, parse sample descriptor, and can map data into a controlled vocabulary. This system makes it easy to access viral sequence data and efficiently analyze genomic relationships. We summarize and compare the existing dengue databases in various aspects (Table 1).

TABLE 2. Comparison of data representation matrix.

Aspects	Unstructured data	Semi-structured data	Structured data
Technology	- Character and binary data	- XML/RDF/OWL	- Tables
Flexibility	- Very flexible, absence of schema	- Flexible, tolerant schema	- Schema-dependent, rigorous schema
Scalability	- Very scalable	- Schema scaling is simple	- Scaling DB schema is difficult
Robustness	- Low	- New technology, not widely spread	- Very robust
Cost	- Low	- Low	- High
Query performance	- Only textual supported	- Query over anonymous node is possible	- Structured Query supported
Transaction management	- No transaction management	- Adapted from RDB, not matured	- Matured transaction management
Version management	- Versioned as a whole	- Versioned over triples or graphs is possible	- Versioned over tuples, rows, or tables

2) DATA WAREHOUSE

A Data Warehouse (DW) is a large data storage system containing data from diverse sources in sufficient quantity and quality to allow sophisticated analysis e.g. Online Analytical Processing (OLAP). There are significant differences between a data warehouse and a database. According to the definition by Bill Inmon [165], the primary characteristics of a data warehouse is the combination of data from various sources. For example, (in our situation) source A and source B may have different ways of identifying factors affecting dengue outbreaks, but in a data warehouse, there will be only a single way of identifying these factors, having synthesized or standardized the various data from multiple sources. The second important characteristic is that it is Time-Variant, meaning that historical data is kept in the data warehouse, either as a transformed form of data from other databases, or as archival data from other databases, and may be versions of the same data from different timeframes. Data warehouses are non-volatile, inasmuch as once data is in the data warehouse, it should never be altered. In addition, the relationship between tables in a data warehouse is typically in the form of a ''star'' or ''snowflake'' schema.

A database, on the other hand, is data storage at the current operational level, and the data is usually and systematically removed from the database in regular archiving processes. In other words, the data in the database is raw data limited by time and locale, but a data warehouse is the implementation of collected, historical data from many operational databases, and provides a more comprehensive dataset, broader in scope and time, for the purpose of analyzing historical trends and predicting future possibilities. Data warehousing capabilities, however, have not yet been widely used in dengue infection research. Wisniewski *et al.* [166] usage of a computer to monitor blood infections in patients is one example of the creation of a database of clinical laboratory data that may be termed a data warehouse. Trick *et al.* [167] also used a data warehouse to link patient data from various agencies to help monitor hospital bloodstream infections. Kim *et al.* [168] developed a surveillance network for Emerging Infectious Diseases in the Caribbean (ARICABA) in the form of a data warehouse consisting of system monitoring and surveillance, the main benefit of this system is that it updates data through automatic data collection, processing and transmission from multiple sources. However, no such data warehouse-based dengue surveillance system currently exists. Table 2 compares data representation in various aspects [169].

VII. MACHINE LEARNING TECHNIQUES FOR FORECASTING MODEL CONSTRUCTION

Forecasting and prediction in any data processing situation is based on historical data, with data mining techniques used to synthesize the relationships between variables. Data mining and Machine Learning techniques can be categorized into two groups: Supervised Learning and Unsupervised Learning. Supervised Learning, also called Classification, is a technique that requires reliable sources in sufficient quantities to enable the Machine Learning or data mining algorithms to be trained to create the predictive models. Answers provided by the learning algorithms must fall into one of a set of prestated classifications or classes; thus ''supervised'' learning or Classification. The ability to pre-determine the number of answer classes is an advantage of Supervised Learning, providing structure and constraints.

Unsupervised learning, also called Clustering, is a technique that does not need to know how many clustering results are possible or probable; the algorithm applies *ad hoc* learning, or heuristics, to categorize data, by analogy. The advantage of unsupervised learning is that it does not require subject matter specialists, but clusters similar data automatically. The disadvantage is that the results may not be accurate with apparently analogous data grouped incorrectly. A further problem of significance is that there are many Machine Learning and data mining algorithms available, and the researchers may be unable, for whatever reason, to choose an appropriate algorithm for their purpose.

A. SUPERVISED LEARNING

Supervised learning methods consist of several sub-methods that are suitable for different applications and encompass Machine Learning and data mining algorithms.

One forecasting method is the Decision Tree which is a simple and popular method being applied to a wide variety of applications.

1) DECISION TREE

A Decision Tree consists of a Root node, and Intermediate and Leaf nodes, which are uni-directionally linked, with the direction indicated in the node. The decision tree is the process of analyzing associations and dependencies between variables and structuring them in a tree model, which effectively classifies the data into 'is-a' and ''has-a'' type relationships, amongst others. The nodes in the tree represent the item sets that travel from the root node to the intermediate node in the tree branch. The final result is found in the leaf nodes. For example, Weka provides a decision tree algorithm, called C4.5, for performing classification tasks. Tanner *et al.* [170] deployed the C4.5 algorithms to create a dengue classifier model and used K-fold cross-validation for model validation to solve the model over-fitting problem, and a further measurement of the sensitivity and specificity of the model was applied. Lee *et al.* [171] developed a simple decision tree to decide between inpatient and outpatient treatment regimens for dengue infection. However, a limitation in using decision trees, is that they are only suitable for nominal or categorical data prediction, and cannot be used for numerical data forecasting. Recently, however, a modification of the decision tree concept to support numerical data prediction has been developed called *decision tree regression* model, which does extend the usefulness of this approach.

2) REGRESSION ANALYSIS

Regression analysis is a statistical method that analyzes relationships between two or more variables to find the correlation coefficient between the variables. The correlation coefficient is used to predict the value of the variables in the desired regression equation. Researchers often use Regression Analysis to create numerical values, and also use it to select variables for predictive modeling (Feature Selection). Interesting examples are [15] and [93] where a multivariate Poisson regression model is used to predict how long the government must control dengue epidemics to prevent the spread of the disease in order to prevent further damage to public health. In that research, regression analysis is used, together with geographic data, for model construction to estimate the risk level of dengue and represent the relationship between the number of cases, economic factors, and social variables involved. Chan *et al.* [118] deployed logistic regression to predict the probability of a dengue outbreak within 24 hours in a small area in Kaohsiung City, Taiwan. Linear regression was used to predict the age of dengue virus-infected mosquitoes [172], who developed a method to shorten the mosquito life cycle and thereby prevent the spread of the disease. Siriyasatien *et al.* [17] also used multivariate Poisson regression to find the relationship between the factors that cause dengue disease and to form a correlation of those factors. The key factor in that research was the rate of dengue infection in mosquitoes, which makes their predictive models different from other work. Shaowei Sang *et al.* [46] used Poisson regression analysis to determine the relationship between dengue outbreaks and other variables such as weather, the Breteau Index, and dengue cases, in the period March 2006 to December 2012. This allowed the generation of predictive models using Time Series and Poisson Analysis and to evaluate the model using residual testing, pseudo-R^2 and Akaike Information Criterion (AIC).

3) ARTIFICIAL NEURAL NETWORK

The Artificial Neural Network (*ANN*) is a technique that imitates the behavior of the human brain, which contains several nodes connected together. The output of one node will be the input of another node and connected together as a network which is similar to neurons. The purpose of *ANN* techniques is to forecast or estimate the result of interest. *ANN* needs large test datasets to learn and adjust the sigmoid values of the nodes to obtain results that are close to the test data. In the study of [173], the researcher presented a method to filter people who are at risk of dengue outbreaks using Bioelectrical Impedance Analysis (BIA) with Multilayer Feedforward Neural Network (*MFNN*) using back-propagation, and evaluated the performance of the presented model using a sum-square error. The study found that the accuracy of the model was about 96.86%. Ibrahim *et al.* [174] developed a system for predicting the day that the dengue-induced fever in a patient would reach its zenith. This is a high-risk day in which the patient may go into shock. The system also used patient information for the prediction using *MFNN*, and achieved 90% accuracy.

Deep Learning is another popular technique that is based on the *ANN* method. Deep Learning is an algorithm that attempts to create a model to represent the semantics of the data at a high level. The difference between Deep Learning and *ANN* is that Deep Learning has more hidden layers than there are in *ANN*. For example, Google has partnered with a researcher, Andrew Ng, to establish the ''Google Brain'' project, which concentrated on developing models that are able to learn their own features, and computational efficiency [175]. Deep learning has also been applied in many other tasks such as wind speed estimating [176], weather forecasting [177], [178], Diabetes prediction [179], and plant diseases [180]. To the best of our knowledge, there is no published research using Google Brain for dengue forecasting task at the time of writing this article.

4) SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is one of the techniques for Supervised Learning. The principle that the data is represented by a vector in $N \times N$ dimensions and the learning algorithm is used to create a line or plane that can divide the vector data into as many segments as possible. The SVM has been tested and found to be a classification algorithm that yields the best result for feature selection [181]. SVM was used in conjunction with the Radial Basis Function (RBF) for

predicting the human mortality rate of dengue infection [13]. The experimental evidence showed an improvement in prediction up to 88.37%.

5) k-NEAREST NEIGHBOR

$k - NN$ is another method used to create dengue prediction models. The output from k-NN can be both categorical and numerical data. k-NN classifies data based on the distance among data in the vector space. By using the Nearest Neighbor Index [182], spatial information of the risk areas of dengue infection have been analyzed using dengue hemorrhagic data from 1998 to 2004, indicating dengue movement patterns from rural communities to urban communities in Trinidad.

B. UNSUPERVISED LEARNING

Unsupervised learning is used to derive inferences from datasets which have no specified-group responses. Clustering is one of the most common techniques in the category. Neighborhoods are a way to cluster groups by finding the distance between one data point and another. Examples of this algorithm include K-means, Hierarchical clustering, Self-organizing maps, Gaussian mixture models, and Hidden Markov models. Based on our survey, unsupervised learning techniques have not been used extensively for dengue outbreaks forecasting purposes. Most of the state-of-the-art methods have used K-means clustering for such a task.

K-means clustering is used to classify distinct groups of genes [183] to analyze the relationship between the genetics of the virus found in *Ae. aegypti* genes and the genetics of the virus found in patients. The study found that responsive genes and the post-infection period in *Ae. aegypti* midgut tissues are highly correlated. Research by Nguyen *et al.* [184], in Tien Giang, Vietnam, has improved the effectiveness of dengue analysis and prediction by using the K-means algorithm in conjunction with *ANN*. This algorithm is called the K-mean clustering-based perceptron, and the method was found to improve the prediction accuracy when compared with other methods. The K-means pre-identifies the number of nodes in the adaptive network and passes this value as a parameter to the process. Specifically, the results showed that this method has greater predictive efficiency than the conventional *ANN* method alone. K-means was also used to identify the hotspots and localized regions of high incidences of dengue in Malaysia [185]. The K-medoids algorithm [186], which is related to the K-means algorithm, used dengue mosquito species data to diagnose dengue outbreaks in India in order to predict the number and age groups of potential dengue patients. In contrast to the K-means algorithm, K-medoids selects data points as centroids (medoids or exemplars) and works with a generalization of the Manhattan Norm to define the distance between data points. K-means endeavors to minimize the sum of squared error, while K-medoids reduces the sum of dissimilarities between points in a cluster, and a point indicated as the centroid of that cluster.

C. TIME SERIES ANALYSIS

Time Series Analysis is the analysis of the movement of the chosen data points to find patterns or features of interest. Time series methods are the most popular method used extensively for modeling dengue prediction [15], [18], [86], [187]–[191]. The data used for this method is collected by period over time, such as the number of patients in each month over the past several years. The time series components consist of four parts: (1) Long-Term Trend (T): represents the movement or change of data over the long-term. (2) Seasonal variation (S): refers to seasonal variations identified in the same season or seasons each year over a number of years. The analysis of seasonal variations is measured in the form of seasonal indexes. (3) Cyclical Variation (C): refers to a cyclical movement in which cyclical movements are similar to seasonal fluctuations but have a longer duration. (4) Irregular Variation (I): Uncertainty variability, therefore being unpredictable, such as natural disasters; earthquakes, volcanic eruptions and so forth. One well-known algorithm for the Time Series Analysis is Auto-Regressive, Integrated, Moving Average (ARIMA) technique [192].

D. ASSOCIATION RULES

Association Rules Analysis is the analysis of the relationships between item sets in a database. Once a relationship is established, the rule is then used to predict the number of cases in the future. The best-known algorithm used in the Association Rules Analysis is the Apriori algorithm [193]. Another useful algorithm is the FP-growth algorithm [194] which is a technique for finding rules without creating candidate item sets, using FP-trees or Frequency Pattern Trees. This technique requires less memory than the Apriori algorithm. More sophisticated Association rules have been used for predicting epidemics. For example, Buczak *et al.* [76] deployed Fuzzy Association Rule Mining, which is a process used to extract rules that show the relationships between variables, such as economic conditions, social conditions, and weather conditions, to create predictive models for dengue epidemics. This work can predict the situation of dengue outbreak about four weeks in advance with high accuracy.

In addition to the methods mentioned above, there are other approaches that can be applied to the prediction of dengue outbreaks, including mathematical modeling and computational modeling.

E. COMPUTATIONAL APPROACH

The vector-borne disease model uses a mathematical computational approach, which is an efficient method, to study disease-spreading dynamics and to control or to prevent disease outbreaks. In order to build and test the model, the computational results show some information of the experiment. A lattice model, and a scale-free network on a regular lattice approach, allow researchers to investigate the transmission dynamics of vector-borne diseases in space and time including human mobility and vector movement. The role of human

mobility and vector movement has also been considered as a driving force of disease spread. The understanding of disease transmission can lead to better approaches to decrease disease transmission. The computational results of the model can be used to inform effective policies.

The lattice model can be used to determine how the population is divided into subgroups and how the position of individuals (actual on the earth's surface) is translated into the x and y coordinates (grid site). The models also provide the coupling between adjacent grid sites or interactions between individuals, which could determine the effect of spatial separation. The individuals are placed onto the square lattice topology with periodic boundary conditions. Each lattice site can represent a host or a vector. The number of hosts or vectors may also vary in response to the size of cities and habitats.

In the simple transition rule, every individual can interact with its four nearest neighbors to transmit the disease. In the lattice model, the status of the host population can be one of Susceptible, Exposed, Infectious, or Recovered (SEIR) [195]. The adult female vector population can also be divided into three compartments: Susceptible, Exposed, or Infectious (an SEI compartmental model) [196]. The probability of changing the status of the host and the vector may be a constant value, depending on the climate data [197]. However, in the epidemiological chain for vector-borne diseases, the transmission rule in lattice model must be infected mosquito → susceptible human → infected human → infected mosquito. Thus, the evolution rule may contain the human or vector movement for disease transmission.

In general, mosquitoes may fly within an area of about a 50m radius from where they were born. However, wind conditions or other external factors may extend this radius to 200m for *Ae. aegypti*. The lattice model may include vector movement as a variable factor. The probability of a particular vector biting a human decreases as this radius distance increases. As well, mosquitoes can move to neighboring sites and secondary neighboring sites [198] while the central site refers to the site where a mosquito constantly stays and the set of neighboring sites immediately adjacent to the central site is called the first neighborhood ring. The next set of adjacent sites is called the second neighborhood ring, and so on. Every mosquito will fly within the n^{th} neighborhood ring. Each mosquito will select a human-occupied site and the human occupants inside the sites in the neighborhood ring are chosen by uniformly random selection. This model can describe the mosquitoes' flight distance indicating a number of neighborhood rings to seek a human target.

In another approach, using Lévy flight distribution, Botari and co-workers studied flying mosquitoes on a large scale with probability $P(r)$ to find the target site within the range r, which are characterized by an exponent, σ (1). A time-dependent probabilistic rule, $\lambda(t)$ (2) is used to indicate the limits of the disease spread (because of the seasonality of the dengue epidemic) [199].

$$P(r) \sim r^{-(1+\sigma)} \tag{1}$$

where the exponent σ is a free parameter that controls the shape of the distribution, which also used to investigate long-range interactions.

This model fits the registered cases of dengue in Rio de Janeiro, between 2006 to 2008, and was able to explain the abnormally high number of cases in 2008 very well. The mosquitoes, in principle, fly randomly to another site located within the range r according to a Lévy flight distribution with probability λ(t) (2). The probability of spreading the disease λ(t) for each of n mosquitoes depends on the temperature (and rainfall index).

$$\lambda(t) = n_1 + n_2 \sin(2\pi t/T + \varphi) \tag{2}$$

where n_1 indicates the probability related to the average temperature, and n_2 refers the increase or decrease of the probability related to variations in temperature. T denotes the period of the sine function, set as $T = 365$ (one year), and φ is related to an arbitrary phase.

The mosquitoes may not only move from a site to a neighboring site by Lévy flight distribution, but they also move to a neighboring site at a diffusion rate D [200]. The rate for a vector to move from site j to a neighbor site k is D/z_j, where z_j is the number of sites that are neighbors of site j. Vector diffusion will be the mechanism for spreading infection in the host population.

As we discussed previously, human mobility intuitively plays a key role in the transmission dynamics of infectious diseases [201]. The importance of human mobility in pandemics of influenza through a meta-population model of influenza has been studied by [202]. Human movement contributes to the probability of pathogens being introduced to new geographic locations. The effect of human mobility has been applied in the lattice model on the dispersion of dengue epidemics Barmak and co-workers used a truncated Lévy distribution (proved by [203]), to study the mobility patterns of human populations for dengue epidemic prediction [195], [196]. The distribution of the length of human displacements (the probability of a human travelling a distance, $P(r)$ (3) is:

$$P(r) \propto (r + r_0)^{-\sigma} \exp(-r/\kappa) \tag{3}$$

where r_0, σ, and κ are parameters that characterize the distribution. This article summarizes the comparison of various algorithms as shown in Appendix.

We also investigated the popular approaches for dengue forecasting models, using in the literature. The term ''dengue fever,'' together with an algorithm name, was used for the publications in the SCOPUS database. Figure 12 shows that the most popular technique used for dengue prediction over the past ten years (2008-2017) is the regression technique, followed by the Time series analysis approach. This is because these two approaches are not too complicated to understand and are suitable for predicting the output variable e.g. a number of cases, while the Neural Network (including the Deep learning approach) is in the top three, as it is well

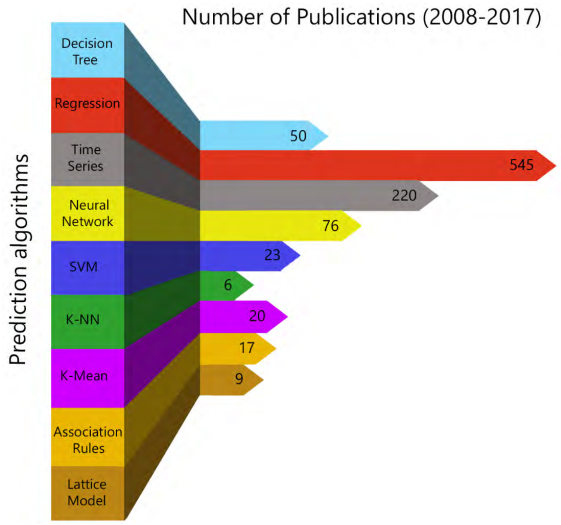


FIGURE 12. Numbers of publications in the SCOPUS database on dengue prediction algorithms in the past ten years (2008-2017).

known for prediction performance but it might be difficult to interpret the prediction results and parameter tuning.

VIII. FORECASTING MODEL EVALUATION METHODS

It is imperative that the forecasting process employed by a model be carefully evaluated before the actual implementation. Inadequate model testing may result in inefficient models, resulting in erroneous forecasts. Several research studies have shown significant predictive efficiency, but a problem that may occur is that of over-fitting. For example, model over-fitting refers to a model that learns and remembers only the training data and usually obtains low accuracy when it is applied to other datasets. This problem arises when the researcher tries to adjust (tuning) the model to predict the best training data using parameters that are suitable for the training data only. This is a common problem in developing Machine Learning algorithms which leads to the situation that the forecasting model works well only with the training data, but when the model is used on test data, the model is less efficient. Also known as a high variance model, over-fitting can be avoided by minor adjustments to the parameters, without causing the contrary situation of under-fitting. Alternatively, a mathematical method, called Regularization [204], can be used to assist in adjusting the parameters while ensuring that appropriate parameters are used and that noise is filtered from data. If these outcomes are achieved the result is the construction of an appropriate forecasting model. Figure 13 illustrates various approaches of evaluation methods for the dengue forecasting models.

A. AKAIKE INFORMATION CRITERION

This process consists of a technical review to ensure over-fitted events are avoided, and the model's predictions are accurate. There are several ways to measure the performance of a model, with the appropriate measurement process depending on the particular method being evaluated. One approach to predictive model performance measurement is the Akaike Information Criterion (AIC), which compares the performance of different models using the same set of data for each model. AIC is considered to be the best evaluation process for this purpose. AIC does not, however, provide information about the quality of a model, only the quality relative to other models. AIC is more suitable for measuring a model with a large dataset, and it is often selected as the best model for large datasets [205]. Gharbi *et al.* [86] and Wu *et al.* [206] used AIC to evaluate the most appropriate forecasting model for dengue epidemics. AIC has been incorporated with pseudo-R^2 to evaluate predictive models and to select the best model for forecasting dengue outbreaks [46]. The AIC can be computed as (4):

$$AIC = 2k - 2\ln(\hat{L}) \tag{4}$$

where k is the number of parameters to be estimated (degrees of freedom) and $\hat{L}$ is the maximum value of the likelihood function for the model [207].

B. BAYESIAN INFORMATION CRITERION

Bayesian Information Criterion (BIC) applies Bayesian statistics as the method for evaluating the efficiency of a forecasting model. The analytical approach is similar to AIC, but BIC penalizes large models, which have a higher number of attributes. As a result, BIC is less likely to select a large model as the best model [205]. Bhatnagar *et al.* [187] and Brasier *et al.* [208] adopted BIC in combination with MAPE (section E) to select the best dengue forecasting models which were constructed using Time Series technique. BIC formula [209] is shown as (5):

$$BIC = \ln(n)k - 2\ln(\hat{L}) \tag{5}$$

where n is the number of data points in the observed data; k and $\hat{L}$ are similar to (4).

Typically, both AIC and BIC should be used in the model selection process to avoid over-fitting or under-fitting problems. The model with the lowest AIC and BIC will be chosen as the best fitting model for dengue incidence as illustrated in Siriyasatien *et al.* [17] and Sitepu *et al.* [210].

C. RECEIVER OPERATING CHARACTERISTICS

Receiver Operating Characteristics (ROC) is also used to analyze the performance of predictive models by determining the relative positive rate of occurrence (or sensitivity) against the false positive rate (or specificity) when different threshold values are set. Hii *et al.* [15] and Soundravally *et al.* [211] used ROC to analyze the predictive value of the model. An ideal model will have 100% sensitivity and 100% specificity. Therefore, the higher the overall accuracy of the prediction model, the ROC curve will be closer to the upper left corner, [212]. Figure 14 demonstrates an example of ROC curve [213].

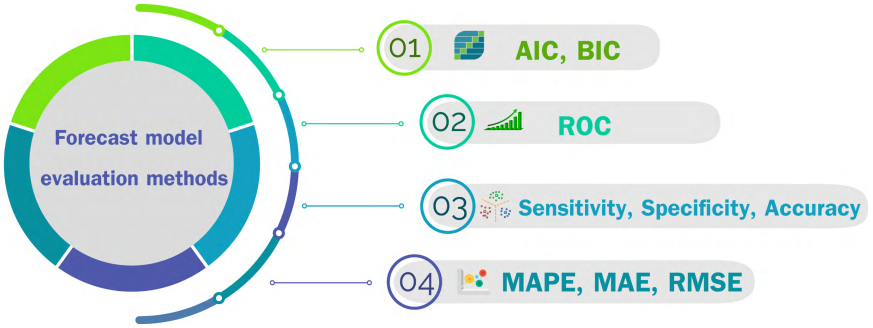


FIGURE 13. Dengue forecast model evaluation schemes using by researchers in the past decades.

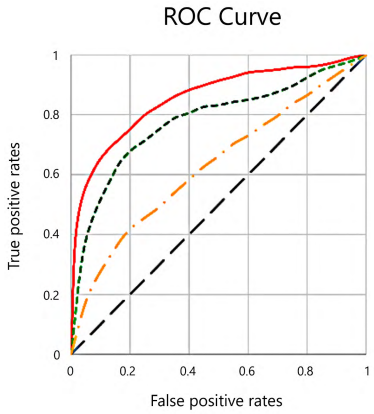


FIGURE 14. Example of ROC curve to investigate the prediction performance by determining the relative positive rate of occurrence (or sensitivity) against the false positive rate (or specificity).

D. SENSITIVITY, SPECIFICITY, AND ACCURACY

Sensitivity is the process of analyzing the performance of a binary classification by finding the ratio of the predicted results to the amount of information in that group. For example, the ratio of the number of cases being diagnosed correctly, and the total number of observed dengue cases. Sensitivity is also known as True Positive, Hit rate, or Recall. *Specificity*, also called True Negative or Precision, is the ratio of the number of cases that are correctly identified and the total number of cases that are identified as dengue cases. Most dengue forecasting studies focused on enhancing models to increase sensitivity and specificity, resulting in more efficient and reliable models.

Accuracy considers both true positive and false negative and is simply a fraction of correctly predicted observations to the total observations. However, high accuracy does not indicate that the forecasting model is best. Accuracy is a good measure only when you have symmetric datasets where values of false positive and false negatives are almost the same. Therefore, a researcher has to consider other parameters to evaluate the performance of a model. The effective forecasting model should have high *Sensitivity* and *Specificity*. da Costa [214] studied the key variables in the diagnosis of dengue by detecting the NS1 Antigen of the dengue virus and evaluated the performance of his works using both *Sensitivity* and *Specificity*. It was found that these variables have a *Sensitivity* in the range of 66-74, while the *Specificity* is almost 100%. The formula for calculating the *Sensitivity*, *Specificity*, and *Accuracy* (6), (7), and (8), as shown at the bottom of this page, respectively:

E. MEAN ABSOLUTE PERCENTAGE ERROR

Another model of performance forecasting is based on the approximation of errors between the observed values and the forecast values. Approximation errors fall into three types: Relative error, Absolute error, and Percent error, which is the relative error expressed as a percentage. Absolute error is the difference between the observed value and the estimated value, whereas the relative error is the absolute error divided by the observed value.

Mean Absolute Percentage Error (MAPE) or Mean Absolute Percentage Deviation (MADE) are used to analyze the efficiency of the forecasting model, taking into account the balance between the accuracy of the forecasting model and the number of variables. Wongkoon *et al.* [215] used MAPE to analyze the performance of the prediction model by incorporating AIC to select the best model that was constructed

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{6}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{7}$$

$$Auccuracy = \frac{True\ Positive + True\ Negative}{Ture\ Positive + False\ Positive + False\ Negative + True\ Negative} \tag{8}$$

using the ARIMA technique. MAPE is computed by (9):

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|}{n} \times 100 \quad (9)$$

where A_t is the observed value; F_t is the forecast value; and n is the number of data points. While MAPE is one of the most popular measures for forecasting errors, many studies have discussed its limitations and potential for misleading results. First, the measure is not defined when the observed value is zero, $A_t = 0$ [216]. In addition, MAPE puts a heavier penalty on negative errors, $A_t < F_t$, than on positive errors [217].

F. MEAN ABSOLUTE ERROR

Another measure of forecasting error is the Mean Absolute Error (MAE), which calculates the mean of the forecast error calculated from the comparison between the predicted values from the predictive model and the observed value. Sitepu *et al.* [210] used MAE together with MAPE to analyze the efficiency of the forecasting model, based on the actual results compared to the predicted results. MAE is shown as Equation (10):

$$MAE = \frac{\sum_{i=1}^{n} |A_i - F_i|}{n} \quad (10)$$

G. ROOT-MEAN-SQUARE-ERROR

Root-Mean-Square-Error (RMSE) or Root-Mean-Square Deviation (RMSD) is a measure of the differences between the observed data and those predicted values by a model, so-called *residuals* when sample data is used in the computation, or *prediction errors* when out-of-sample data is used. The RMSE is utilized to compare the prediction errors of various models for the same dataset. However, it cannot compare errors between datasets using the same model, as it is scale-dependent [216]. The main drawback of RMSE is it is sensitive to outliers [218]. RMSE formula is shown in Equation (11):

$$RMSE = \left(\frac{\sum_{i=1}^{n} (A_i - F_i)^2}{n} \right)^{1/2} \quad (11)$$

Development of a new forecasting system for dengue occurrence and outbreaks should be evaluated for the effectiveness of the system by comparing it with existing systems. However, almost all dengue forecasting systems typically use local data, which use different datasets and factors. Therefore, comparing the predictive efficiency of new systems with existing systems is difficult (validation against established systems). In addition, there is no central standard for assessing dengue prognoses. Researchers should evaluate the effectiveness of the prediction system carefully, trying to compare the new system with that of the most fully developed system appropriate to their country and compare their systems to other local systems developed in their country, based on reliable theories.

MAE and RMSE both measure average prediction error in the same unit of variables (scale-dependent measure), regardless of direction. Their values can range from 0 to ∞. The main difference between MAE and RMSE is that MAE is less sensitive to outliers. It assigns a proportional weight to each error, a natural measure of average error [219], and is usually used to compare different forecasting methods on a single dataset. On the other hand, RMSE is highly sensitive to large errors since it gives a high weight to large differences. Due to the square in RMSE, it is good for expressing larger deviations of errors. Similar to MAE, RMSE is used as a relative measure to compare different forecasting models for the same data but tends to be larger than MAE. RMSE is more appropriate than MAE for the comparison of model performance when the error is normally distributed [220]. Alternatively, MAPE is scale-independent and is often used to compare the performance of prediction among different datasets. However, MAPE can be calculated only for strictly positive and quantitative data. In addition, it is reasonable only for data where the change in percentage is invariant with respect to the change in scale. For example, changing the observed value from kilograms to pounds will give the same percentage, whereas changing the temperature from Fahrenheit to Celsius will yield different percentages [221].

IX. CHALLENGES AND FUTURE TRENDS OF DENGUE INFORMATION PROCESSING AND PREDICTION

The seriousness of the prevalence and proliferation of dengue infection around the world makes surveillance of the progress and occurrence of the disease an imperative. Predicting the outbreak of dengue in advance allows authorities to prepare well and in good time, thereby being better able to cope with the severity of outbreaks that may occur. Over the past several decades, researchers have been working to develop and prove methods for predicting outbreaks using various state-of-the-art approaches that have been discussed in this article. Particular points of interest identified and discussed by the various research publications include five main challenges (Figure 15), which will be described in the following subsections.

A. CHALLENGES FOR DYNAMIC FORECASTING MODELS

1) DYNAMIC FORECASTING MODELS AND DATA GATHERING TECHNOLOGY

New dengue and other pandemic data are constantly being generated and should be incorporated into the existing data to ensure that the predictive model has a full set of contemporary data from which to learn, making the predictive model contemporary and relevant. However, ensuring that new observations are incorporated into the existing body of data is essentially a manual task, which is time-consuming and inconvenient, and may not be comprehensive, resulting in ineffective forecasting model predictions. This is an

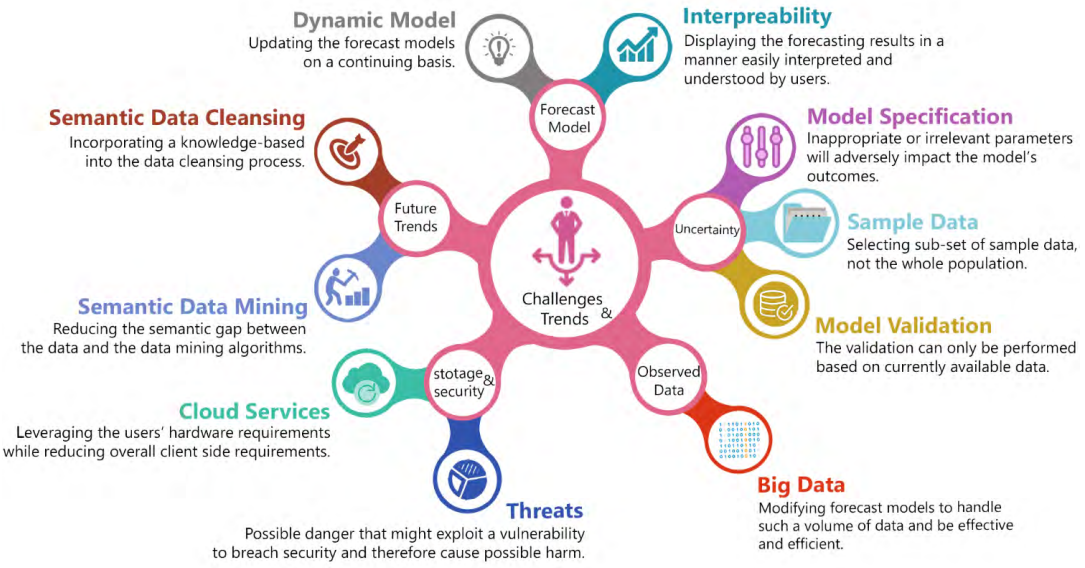


FIGURE 15. Challenges and future trends for dengue forecast model development in five different aspects.

important matter for researchers to address, to develop mechanisms for automatic updating of the data on a continual basis. This would optimize the effectiveness and efficiency of the forecasting model. This problem is exacerbated by the need to update the models on a continuing basis which would impose overheads that are too high. Infrequent or spasmodic updating can decrease the effectiveness and efficiency of the forecasting model and results in the policymaker's planning and management of the dengue epidemic being ineffective. Given that manual collecting of data is a laborious task, automatically collecting information using a new technology called the "Internet of Things (IoT)" is now possible. The IoT is the networking of mobile devices, handheld devices, and sensors of all types. The fact is that there are now many millions of devices linked by the IoT which gather and transmit data, making the IoT an important new technology that should be of great interest to epidemiological researchers.

One significant application of this technology is in the area of medicine, using what is now termed the Internet of Medical Things (IoMT). IoMT is an amalgamation of medical devices and applications that can connect to healthcare information technology systems using networking technologies. For example, Lakshmi and Karthik [222] developed a device for dengue identification and patient monitoring using sensor technology to measure various patient parameters to identify and diagnose dengue infection, patient immunity levels, blood pressure, heart rate and breathing rates. All the data coming from this personal sensing device are logged into a cloud service and is available to physicians for remote diagnosis and treatment by accessing the cloud service.

IoT technology has also been deployed by MediaTek in Taiwan to monitor dengue outbreaks and send alerts to possibly affected populations [223]. More than 200 smart bug zapper devices were installed in Tainan City, during a severe outbreak, to track the vectors, and send information on the number of mosquitoes killed. This data was uploaded to a cloud computing system and then analyzed using Big Data technology to find the distribution of the disease vectors, thereby identifying risk areas and allowing resources to be focused in those areas. Future developments of this particular device will enable the identification of different types of vectors, and image processing techniques may be deployed to enhance the ability of the smart bug zapper to gather more and more various data.

2) INTERPRETABILITY

One limitation of epidemic prediction models is that the interpretation of the forecast results may not be easy to understand by users who do not have the professional background, knowledge or expertise. Medical information is often difficult to understand, even by medical professionals, and Machine Learning techniques are difficult to understand by most people without education or training in what is, to many people, an arcane and complex topic. Predictive power is often inversely proportional to the complexity of the predictive techniques [39] and, therefore, the more complex the forecasting model, the more essential it is to have forecasting results displayed in a manner easily interpreted and understood by users, including presenting that information in a manner appropriate to the particular user's interests (Personalization). In any case, data presentation should be easily understood to be of maximum usefulness for decision making.

B. CHALLENGES TO UNCERTAINTY HANDLING

Validation of the predictive model is essential to ensure that the uncertainty inherent in the data is properly identified and understood. We have identified two main components

of uncertainty in a forecasting model, from [224]. The first aspect subject to uncertainty is related to the model specification itself, including the model's parameters. In this case, inappropriate or irrelevant parameters will adversely impact the model's outcomes. This can be measured using sensitivity analysis such as observing elasticities obtained in various specifications. Second, use of a selected subset of sample data, not the whole population, in the prediction model will also be problematic, therefore data selection criteria and size of the subset must be analyzed for correctness. These uncertainties are of particular interest and importance and require further investigation.

The impact of the forecasting uncertainties arising from the selection of sample data has not yet been made clear. With computer intensive numerical methods, however, it is now possible to compute some statistical measures without being bound by any particular analytical approach. These computer-based methods allow extensive simulations, which can be used for analysis. Different methods of calculating statistical errors are offered, such as Bootstrap [225] which has been used to validate some sources of inaccuracy in prediction models.

Another issue related to prediction uncertainty, and the validation of a predictive model, is that the validation can only be performed based on currently available data. This means that the model may not necessarily be valid for new observations and new data becoming available in the future. This problem has not been well addressed, and the problem of validly incorporating new observations is still unresolved. Therefore, a method for analyzing the chosen model and examining its predictive ability is still required.

One of the more well-known tools for model adequateness under changing conditions of the reality (that is, in circumstances different from those for which the model was developed) is the Chow test [226]. There are, however, limitations inherent in the Chow test, such as poor performance with small samples and poor support for non-linear models and non-continuous data [227]. Another tool to handle this uncertainty is the confidence interval which provides a good visual illustration of the uncertainty level of the proposed statistic. In this case, large intervals are usually associated with a high level of uncertainty.

Those uncertainties are challenges to accurate model construction, sample data selection, and when new observation becoming available. They should be handled properly using existing tools as they can lower the prediction performance of the model.

C. CHALLENGES RELATING TO BIG DATA

Current dengue prediction models have conventionally been tested using only local data. If local data from many sources around the world were to be combined, it is difficult to see how these current models can possibly handle such a volume of data and be effective, efficient, or even useful in their purpose. For example, over the last five years, the US healthcare system has generated some 150 exabytes of data (10^{18} bytes) [7]. With the substantial growth and widespread occurrence of dengue outbreaks, now seemingly occurring on a daily basis, future data volumes may reach zettabytes (10^{21} bytes) or yottabytes (10^{24} bytes) in the next few years. Since the amount of data about dengue epidemics is growing rapidly, Big Data technologies will play an important role in data analysis. The manner in which data can be, and is, processed will fundamentally change with new machine learning and data mining techniques that will surely be developed to cope with the large scale of data.

Today's forecasting models will not be able to handle such large and complex datasets. As well, most Machine Learning techniques currently used also cannot handle such volumes of data, meaning that the forecasting model may be less effective. Therefore, researchers must develop new technologies that support large datasets. Such Big Data technology must provide infrastructure, tools, and techniques to leverage big data effectively [150]. If this is achieved, it will bring potentially huge benefit to many people, from single-physician offices to multi-providers and hospital networks, better enabling them to predict likely outbreaks and detect outbreaks quickly, almost in real-time, thereby managing responses efficiently.

Big data is a new term that is being applied to datasets that are of large size (volume), can be used to mine large amounts of data within a predefined period of time (velocity), and come from a variety of sources and includes both unstructured and structured data (variety). However, the question is ''How big is Big Data?,'' given the often substantial difference in the amount of the data available from different scientific fields [228].

Some researchers have defined Big Data as being data that is more complex than traditional datasets, that are very large and are growing ever larger and is more widely applicable than traditional data which supported forecasting but is country-specific or localized. Big Data is of considerable, and growing, interest for predicting situations such as dengue outbreaks and other epidemics, which rely for their accuracy on reported data from many countries around the world, collected from many sources and identify more important factors involved in the outbreak of dengue. Global-level real-time forecasting is key to epidemiological forecasting, and for alerting appropriate agencies and supporting closer cooperation between agencies.

Effective forecasting requires data mining techniques that support Big Data processing. Current algorithms may not support highly multi-dimensional data, a shortcoming that is leading to the development of new techniques that support and allow high-volume data processing and multi-dimensional analysis, now generally termed Big Data Mining.

Big Data Mining refers to the process of discovering associations within and between datasets, patterns, anomalies, and significant data structures from large amounts of often multi-dimensional data. Therefore, one challenge of Big Data mining is to develop a mining algorithm that effectively

handles highly multi-dimensional data and can accurately predict future events.

Dengue Infection Predisposition in the past several decades has presented a number of factors that were associated both directly and indirectly with many epidemics. If the appropriate Health Department, or authority responsible for public health, uses all these factors in their forecasting models, it will make the models more effective in predicting dengue outbreaks. However, this comes with a high computational cost, with long calculation times, even with the statistical or mathematical methods for removing some of the uncorrelated attributes or features, which have been described earlier for feature selection,. Dimension reduction is another approach to resolving the curse of the dimensionality problem by using some mathematics approaches e.g. the Singular Value Decomposition (SVD) [229]. However, these can result in less efficient forecasting. It is recommended that computer scientists and public health researchers should pay greater attention to improving Machine Learning algorithms for high-dimensional modeling of communicable disease occurrences, and provide stronger support for Big Data mining technologies.

D. CHALLENGES RELATING TO DATA STORAGE, CLOUD SERVICE, SECURITY AND PRIVACY

Due to the exponential growth of available data for dengue forecasting systems, and the fact that these forecasting systems exploit information from many heterogeneous sources, machine learning techniques and data storage become important issues as they require potentially hugely greater computational power to handle the greater volume of data, and more complex models. Consequently, storing data in a local hard drive may not be practical. Researchers in this area recognize that the provision of adequate computing infrastructure that can effectively process the enormous and rapidly growing data is essential [230]. Cloud platforms have the potential to support a new paradigm of data-intensive processing. Cloud computing leverages the users' hardware requirements while reducing overall client-side requirements and complexity, and cost. It is capable of scaling-up greatly to store and process enormous volumes of data, on an as-required basis, making it a versatile and economical solution to the users' processing requirements. There are typically four services provided in the cloud-based model for the storage and analysis of dengue and biomedical data [232].

1) DATA AS A SERVICE (DaaS)

This service enables on-demand data access to up-to-date data that are accessible by a wide range of devices, from mobile devices to substantial computers, which are connected through the Internet.

2) SOFTWARE AS A SERVICE (SaaS)

This service provides a large diversity of software tools for data analysis e.g. data mining and machine learning tools. Therefore, the installation of several software packages on a local PC is no longer required, and the 'cost-of-ownership' is low.

3) PLATFORM AS A SERVICE (PaaS)

It offers a versatile software implementation environment including powerful database processing facilities and web servers to users for developing, testing, and deploying cloud applications.

4) INFRASTRUCTURE AS A SERVICE (IaaS)

This service facilitates a full computer infrastructure e.g. operating system, RAM, CPU, powerful database management and storage facilities, and other computing resources.

Recently, a fifth type of cloud service has been defined and introduced by Lai *et al.* [233], called the *Knowledge as a Service (KaaS)*, which provides the interoperations among members in a knowledge network. This service relies on data generated collaboratively by domain experts.

As cloud services achieve and gain popularity, the security problem becomes of concern for new model deployment. Cloud service architecture differs from traditional architectures and requires different and more sophisticated protection mechanisms. Table 3 illustrates the security challenges of cloud services derived, in part, from [231]. The level of service required in such an important and sensitive area as disease prediction introduces unique security challenges. Zissis and Lekkas [231] suggested that a combination of Public Key Infrastructure (PKI), Lightweight Directory Access Protocol (LDAP), and Single-Sign-On (SSO) can address most of the identified threats in cloud computing shown in Table 3.

Inevitably, cloud computing will be able to efficiently and securely support dengue forecasting systems. In cost-benefit terms, cloud computing is now the preferred option, being able to overcome the limitations of traditional forecasting systems; data volumes and processing requirements, and its dynamic characteristics are a significant improvement over traditional countermeasures.

E. FUTURE DIRECTION OF DENGUE DATA PROCESSING

1) SEMANTIC DATA CLEANSING

Data cleansing is a method for detecting data errors before they are used in predictive models, as has been described in section IV (B). However, those methods only achieve a certain degree of accuracy. Researchers have continually sought to improve data quality and develop interesting approaches to this problem. One such approach termed Semantic Data Cleansing is an ontology-based technique in which the semantics inherent in the data are defined. As stated in [234] "To support the sharing and reuse of formally represented knowledge among AI systems, it is useful to define the common vocabulary in which shared knowledge is represented. A specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects—is called an ontology."

TABLE 3. Security challenges of a cloud services for a cloud based dengue forecasting.

Service levels	Data as a Service (DaaS)	Software as a Service (SaaS)	Platform as a Service (PaaS)	Infrastructure as a Service (IaaS)	Knowledge as a Service (KaaS)
Users	- End users, developers, and organizations who want to access data provided by a cloud provider.	- End users, organizations who subscribe to services for data analyses offered by a cloud provider.	- Developers and organizations who want to utilize a programming environment, tools that are supported by a cloud provider.	- End users, developers, organizations that to use computing fundamental resources on cloud infrastructure e.g. storage, network, or deploying applications.	- Members in the same domain who are experts who want to collaborate and share knowledge through a cloud service.
Security requirements	- User authentication - Data security - Access control - Service availability - etc.	- Access control - Software security - Service availability - Communication security - etc.	- Access control - Application security - Data security - etc.	- Cloud management control security - Communication security - Secure images - Service availability - etc.	- Data security - Communication security - Software security - etc.
Threats	- Interception - Data modification (change, delete) - Impersonation - etc.	- Privacy breach - Session hijacking - Impersonation - etc.	- Programming flaws - Source code modification - Software interruption (delete) - Impersonation - etc.	- Session hijacking - Exposure in network - Connection flooding - Network disruption - Impersonation - etc.	- Data modification (change, delete) - Exposure in network - Connection flooding - Network disruption - Impersonation - etc.

The standard languages used in ontology development (Ontology Language) in order to share and exchange information on the Internet include OWL [235] and RDF [236]. OWL has become the standard language for the development of ontology technologies for using in accordance with the guidelines of the semantic web. OWL has been developed based on the RDF using the format of XML (Extensible Markup Language) and URI (Uniform Resource Identifier). The primary benefit of developing ontologies is to store knowledge in such a way that it can be reused and shared with others in a manner that allows semantic searching, where the ontology structure can be used to analyze and discover relationships inherent in the information content, which is not supported by standard database management systems (RDBMS). This semantic searching ability in ontologies allows meaningful information in both images and text to be identified [237]. Therefore, we can integrate ontologies with the data cleansing processes to enhance the efficiency of data imputation and transformation processes.

In addition to data cleansing, ontologies are also used in other ETL (Extract, Transform, and Load) processes, by bringing knowledge into the process which is consistent with the direction of the semantic web that is required for the computer to be able to understand, learn, and intelligently handle the *unknown* data. To the best of our knowledge, there has been little research activity in which ontologies have been applied in the ETL process over the past decade. Gonçalves *et al.* [238] proposed the use of ontologies for automatic data acquisition from websites. Mate *et al.* [239] proposed the Semantic ETL technique to incorporate ontology into ETL to extract and manipulate data from a medical database. Ontologies have also been used in several processes of data cleansing. For example, several researchers [240]–[242] have used ontologies to detect errors of data such as consistency checking and duplicate detection, thereby improving the quality of the data prior to importing into a database. In addition, researchers have used an ontology to help predict lost data. For example, microarray data [243] are the ontological genes developed by chemistry experts to validate and substitute lost data from the chemical structure.

Apart from quality data enhancement, ontologies are also deployed in data mining to improve the prediction power of disease outbreaks. Ontologies and associated technologies can more effectively assist technical staff in discovering outbreak information than by searching a relational database using SQL. The search for information through an ontology is called a semantic search. There are currently various biomedical ontologies being developed and being shared with other researchers, such as the Unified Medical Language System [244], Gene Ontology [245], and National Center for Biomedical Ontology [246].

2) SEMANTIC DATA MINING

Data mining is a process of data analysis to identify patterns of data and allow forecasting of events that are likely to occur in the future. The most commonly analyzed data structure is the single table format usually used by supervised and unsupervised learning techniques. However, the major challenge of traditional methods, and these statistical and Machine Learning approaches, is the loss of meaning inherent or implied in the interrelationships with the other data. This renders the analysis of data less effective and less relevant in different contexts. Data mining or more particularly ''semantic data mining'' is one direction that many researchers should pursue in order to allow the computer to learn and analyze data more effectively. Semantic data mining

provides several advantages for searching data and analyzing the meanings potentially contained in the data.

Ontologies represent data more effectively than conventional techniques that are not able to reflect the domain knowledge of their distinctions [247] and it is a popular approach among researchers to be used for encoding domain knowledge. Ontologies can also reduce the semantic gap between the data and the data mining algorithms and help to guide the mining processes or reduce the search space, and, finally, ontologies provide a formal way for representing the data mining flow. Kriegel *et al.* [248] and Dietterich *et al.* [249] suggested that the introduction of domain knowledge into the data mining process is the topmost open issue in future data mining research and practice, and ontologies support this concept of semantic data mining by understanding the relationships between those concepts and enabling them to be effectively used by data mining algorithms [53].

Initial systems that include ontologies to systematically describe the process of Machine Learning and data mining include CAMLET [250], IDA system [251], Exposé Ontology [252], OntoDM [253], EXPO [254], DAMON [255], KDDOTO [256], DMWF [257], DMOP [258], and KD Ontology [259]. Data mining algorithms have been used in conjunction with algorithms for Machine Learning, such as association rules [260], [261], classification [262], [263], and clustering [264], [265].

However, the limitations of using ontologies to support data mining are that such systems rely heavily on domain knowledge to construct the ontology. This makes such systems unable to be used effectively in other domains and the ontology lacks the flexibility to incorporate the data mining algorithms effectively.

In the field of medical and epidemiological research, there is currently little research information available on the use of ontologies for semantic data mining. What research is available includes work on gene clustering [266] using Gene Ontology and medical document clustering [267]. Other work has resulted in the discovery of semantic associations between biological data with the help of formal ontologies and has identified potential errors in the ontologies [268]. Ontologies can be used in conjunction with the Deep Learning algorithm to predict future behavior [247] based on self-motivation, social influence, and environmental events, and the predictive models have been shown to be very reliable. Such research into integrating knowledge with Machine Learning is of great interest to many epidemiological and computer scientists, and future work should develop models for more accurate predictions.

X. CONCLUDING REMARKS

This article analyzes the major components that can be used in a dengue prediction model. We have attempted to identify the factors directly related to the probability of a dengue epidemic occurring, particularly climate factors, the rate of mosquito bites, rainfall, and the rate of dengue infection in mosquitoes as the important factors contributing to severe outbreaks. As noted and discussed, the severity of outbreaks is likely to be reduced when rainfall is too high, causing the destruction of mosquito eggs in fast-moving streams and rivers, resulting in reduced mosquito populations and consequential diminution of epidemic risk. Excessive temperature may also affect the survival rate of mosquitoes and reduce the spread of the disease. If additional information is available about *Ae. aegypti* and *Ae. albopictus* bites and rates of mosquito infections in different areas, seasons or periods, this is useful information for dengue surveillance programs, as has been reported in [269]. Such information should be used in conjunction with other epidemiological surveillance factors such as mosquito population size, mosquito age, and weather conditions, as well as background information in the past to be used to assess the risk of allergy epidemic emotions [270].

Another surveillance method for dengue outbreaks surveillance is to ascertain information posted by tourists on social networks. The immediacy of this information quickly alerts the international community about the prevalence of the disease in an area [271]. Dengue outbreaks are a risk for all travelers who should, therefore, study the seasonal risk factors at their intended destinations, and seek and receive accurate advice about preventive measures during their travel.

International travel plays a significant role in the transmission of antibiotic-resistant pathogens. Each country should incorporate demographic information and population movement data into an integrated plan to reduce the threat and risk of communicable disease spreading in the globalized world [272]. Tourists are travelling in greater numbers than ever before, and migrant workers, migrants and refugees, together with an overall expansion of the world population, influence this spread of disease in a significantly greater fashion that in the past. Social growth and political unrest, and disparity of social and economic conditions between places of origin and destination, all influence health conditions, and influence the epidemiology of major communicable diseases around the world. Epidemiological agencies can use demographic information about the differences in the prevalence of the disease to integrate these population movements into predictive models, enabling the health challenges related to current and future migration of the population to be resolved efficiently. Work being undertaken around the world shows that epidemiological relationships resulting from health differences due to increased migration and leading to the outbreak of global epidemics, have affected national and international health planning policies [273].

We also addressed some challenges for dengue information processing and prediction including 1) Dynamic forecasting models, 2) Uncertainty handling, 3) Big data, 4) Storage and security, and 5) Semantic data processing. We do believe that these challenges will make a significant contribution to this research area in the future and researchers should pay attention to them.

TABLE 4. Summary and comparison of forecast algorithms deploying in dengue surveillance systems.

Types	Algorithm	Input	Output	Pros	Cons
Supervised Learning	- C4.5 [217] - [125] - ID3	- nominal	- Trees - Rules - Nominal	- Can select the most discriminatory features. - Can be used in Rule Generation problem - Classify data without much calculations - Dealing with noisy or incomplete data	- High error rate while ratio of training set by the number of classes is small - Exponential calculation growth while problem is getting bigger. - Need to discrete data for some particular construction algorithm.
	Regression [43] [94] [188]	- numeric	- numeric - nominal	- Simple model equation	- Difference data distribution pattern may get similar equation
	Artificial Neural Network[99] [100]	- numeric	- Numeric - Nominal	- Easy to do incremental learning - Can learn to ignore irrelevant attributes	- Difficult to reason the output - Parameters tuning
	SVM [111]	- Numeric	- Numeric - Nominal - Ordinal	- Learning result is more robust - Over-fitting is not common - Lower computational cost - Works well with fewer training samples	- Choice of the kernel - Long training time - Not easy to incorporate domain knowledge
	k-NN [198]	- Numeric - Nominal	- Numeric - Nominal	- Very easy to understand and implement - Flexible to feature - Effectively handles multi-class cases	- High computational cost - Lack of formal method to select the best *k* value
	Deep learning [15]	- Numeric	- Numeric - Nominal	- High performance compared to others - Reduces the need for feature engineering - An architecture that can be adapted to new problems relatively easily	- Requires a large amount of data - Computationally expensive to train - What is learned is not easy to comprehend (difficult to understand)
Unsupervised Learning	*K*-means [45]	- Numeric - Nominal	- Cluster	- Low computing time during training phase.	- Slow for large training dataset - Poor performance on noisy data - No explicit generalization - Lack of formal method to select the best *k* value - High computation in classification phase.
Rule Discovery	Apriori [5]	- Numeric	- Rules	- Easy to implement and parallelize - Uses large itemset property	- Require many database scan
	FP-Growth [81]	- Nominal - FP-Tree	- Rules	- Much faster than Apriori	- Repeatedly scan entire dataset
	Fuzzy Association Rule Mining [33]	- Numeric	- Rules	- Discover both positive and negative associations	- High computational costs - Membership function must to be given - Cause iterative database scans
Time series Analysis [188] [122] [129]	ARIMA	- Numeric	-Predict trend across time	- Predict one event from data in the past - Might not accurate if new factor is presented	Time series Analysis [188] [122] [129]
Computational approach	Lattice model [21]	- Numeric	- Numeric	- Flexible to data choices, e.g. land use, demography, climate - Correspond to epidemic situation - High performance	- High computational cost - Sensitive to input data

We acknowledge a significant limitation in the information presented in this article. The information is not universally comprehensive, as our survey of relevant research articles included only English language publications, whereas it is certainly true that a significant body of research has been published in other languages. Nonetheless, all articles cited come from reputable and searchable databases available on the Internet. As well, our coverage of the literature did not

include articles published in conference proceedings where the publishers did not allow the proceedings to be downloaded, or were unavailable to the public, for study purposes.

APPENDIX

See Table 4.

ACKNOWLEDGMENTS

Authors would like to thank all PhD students in the class of 2013 for their efforts in collecting relevant articles cited in this article. We also acknowledge the contribution of Ms Duangporn Limthamrong of the Naresuan University Division of Continuing Education and especially Mr Roy I. Morien of the Naresuan University Graduate School for their checking and editing of English grammar and expression in this article.

AUTHORS' CONTRIBUTION

K. Kesorn designed the article structure, gathered data, and wrote all sections of the manuscript. K. Jampachaisri partially wrote section XIII (C, G) and S. Chadsuthi partially contributed to section VII (E). P. Sriryasatien contributed to the discussion in section V (Dengue vaccine). All authors have read and approved the final manuscript and declare that no competing interests exist. The funder had no role in the study design, data collection and analysis, decision to publish, or the preparation of the manuscript.

REFERENCES

[1] J. D. Stanaway *et al.*, "The global burden of dengue: An analysis from the Global Burden of Disease Study 2013," *Lancet Infectious Diseases*, vol. 16, no. 6, pp. 712–723, 2016.

[2] WHO. (2017). *Dengue and Severe Dengue*. Accessed: Oct. 13, 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs117/en/

[3] S. Bhatt *et al.*, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, 2013.

[4] WHO. (2012). *Global Strategy for Dengue Prevention and Control 2012–2020*. Accessed: Nov. 16, 2017. [Online]. Available: http://www.who.int/denguecontrol/9789241504034/en/

[5] S. Runge-Ranzinger, O. Horstick, M. Marx, and A. Kroeger, "What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends?" *Tropical Med. Int. Health*, vol. 13, no. 8, pp. 1022–1041, 2008.

[6] A. A. Bharambe and D. R. Kalbande, "Techniques and approaches for disease outbreak prediction: A survey," in *Proc. ACM Symp. Women Res.*, New York, NY, USA, 2016, pp. 100–102.

[7] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big data for health," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[8] Y. Hagar, D. Albers, R. Pivovarov, H. Chase, V. Dukic, and N. Elhadad, "Survival analysis with electronic health record data: Experiments with chronic kidney disease," *Stat. Anal. Data Mining*, vol. 7, no. 5, pp. 385–403, 2014.

[9] G. N. Forrest, T. C. Van Schooneveld, R. Kullar, L. T. Schulz, P. Duong, and M. Postelnick, "Use of electronic health records and clinical decision support systems for antimicrobial stewardship," *Clin. Infectious Diseases*, vol. 59, pp. S122–S133, Oct. 2014.

[10] N. T. K. Tien, D. Q. Ha, T. K. Tien, and L. C. Quang, "Predictive indicators for forecasting epidemic of dengue/dengue haemorrhagic fever through epidemiological, virological and entomological surveillance," *Dengue Bull.*, vol. 23, pp. 44–50, Dec. 1999.

[11] K. Shaukat, N. Masood, S. Mehreen, and U. Azmeen, "Dengue fever prediction: A data mining problem," *J. Data Mining Genomics Proteomics*, vol. 6, no. 3, pp. 1–5, 2015.

[12] J. Chompoosri, U. Thavara, A. Tawatsin, S. Anantapreecha, and P. Siriyasatien, "Seasonal monitoring of dengue infection in *Aedes aegypti* and serological feature of patients with suspected dengue in 4 central provinces of Thailand," *Thai J. Vet. Med.*, vol. 42, no. 2, pp. 185–193, 2012.

[13] K. Kesorn *et al.*, "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, 2015.

[14] S. M. Lau *et al.*, "A new paradigm for *Aedes* spp. surveillance using gravid ovipositing sticky trap and NS1 antigen test kit," *Parasite Vectors*, vol. 10, p. 151, Mar. 2017.

[15] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, "Forecast of dengue incidence using temperature and rainfall," *PLoS Neglected Tropical Dis.*, vol. 6, no. 11, p. e1908, 2012.

[16] Y. Shi *et al.*, "Three-month real-time dengue forecast models: An early warning system for outbreak alerts and policy decision support in Singapore," *Environ. Health Perspect.*, vol. 124, no. 9, pp. 1369–1375, 2016.

[17] P. Siriyasatien, A. Phumee, P. Ongruk, K. Jampachaisri, and K. Kesorn, "Analysis of significant factors for dengue fever incidence prediction," *BMC Bioinform.*, vol. 17, no. 166, pp. 1–9, 2016.

[18] M. A. Johansson, N. G. Reich, A. Hota, J. S. Brownstein, and M. Santillana, "Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico," *Sci. Rep.*, vol. 6, Sep. 2016, Art. no. 33707.

[19] O. J. Brady *et al.*, "Refining the global spatial limits of dengue virus transmission by evidence-based consensus," *PLoS Neglected Tropical Diseases*, vol. 6, no. 8, p. e1760, 2012.

[20] A. L. Ramadona, L. Lazuardi, Y. L. Hii, Å. Holmner, H. Kusnanto, and J. Rocklöv, "Prediction of dengue outbreaks based on disease surveillance and meteorological data," *PLoS ONE*, vol. 11, no. 3, p. e0152688, 2016.

[21] S. Lozano-Fuentes *et al.*, "The dengue virus mosquito vector *Aedes aegypti* at high elevation in México," *Amer. J. Tropical Med. Hygiene*, vol. 87, no. 5, pp. 902–909, 2012.

[22] M. Dhimal, I. Gautam, H. D. Joshi, R. B. O'Hara, B. Ahrens, and U. Kuch, "Risk factors for the presence of chikungunya and dengue vectors (*Aedes aegypti* and *Aedes albopictus*), their altitudinal distribution and climatic determinants of their abundance in central nepal," *PLoS Neglected Tropical Diseases*, vol. 9, no. 3, p. e0003545, 2015.

[23] WHO Regional Office for South-East Asia. (2011). *Comprehensive Guidelines for Prevention and Control of Dengue and Dengue Haemorrhagic Fever: Revised and Expanded Edition*. Accessed: Nov. 16, 2017. [Online]. Available: http://www.searo.who.int/entity/vector_borne_tropical_diseases/documents/SEAROTPS60/en/

[24] R. Romi, "History and updating on the spread of *Aedes albopictus* in Italy," *Parassitologia*, vol. 37, nos. 2–3, pp. 99–103, 1995.

[25] A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow, "Social networks and infectious disease: The Colorado Springs study," *Soc. Sci. Med.*, vol. 38, no. 1, pp. 79–88, 1994.

[26] J. Matthews, R. Kulkarni, M. Gerla, and T. Massey, "Rapid dengue and outbreak detection with mobile systems and social networks," *Mobile Netw. Appl.*, vol. 17, no. 2, pp. 178–191, 2011.

[27] C. de Almeida Marques-Toledo *et al.*, "Dengue prediction by the Web: Tweets are a useful tool for estimating and forecasting dengue at country and city level," *PLoS Neglected Tropical Diseases*, vol. 11, no. 7, p. e0005729, 2017.

[28] B. M. Althouse, Y. Y. Ng, and D. A. T. Cummings, "Prediction of dengue incidence using search query surveillance," *PLoS Neglected Tropical Diseases*, vol. 5, no. 8, p. e1258, 2011.

[29] W. Anggraeni and L. Aristiani, "Using Google trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia," in *Proc. Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, Surabaya, Indonesia, 2016, pp. 114–118.

[30] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein, "Using Web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance," *PLoS Neglected Tropical Diseases*, vol. 5, no. 5, p. e1206, 2011.

[31] G. J. Milinovich, S. M. R. Avril, A. C. A. Clements, J. S. Brownstein, S. Tong, and W. Hu, "Using Internet search queries for infectious disease surveillance: Screening diseases for suitability," *BMC Infectious Diseases*, vol. 14, no. 1, p. 690, 2014.

[32] A. G. Hoen, M. Keller, A. D. Verma, D. L. Buckeridge, and J. S. Brownstein, "Electronic event–based surveillance for monitoring dengue, Latin America," *Emerg. Neglected Diseases*, vol. 18, no. 7, pp. 1147–1150, 2012.
[33] E. M. Delmelle, H. Zhu, W. Tang, and I. Casas, "A Web-based geospatial toolkit for the monitoring of dengue fever," *Appl. Geogr.*, vol. 52, pp. 144–152, Aug. 2014.
[34] N. A. Rehman, S. Kalyanaraman, T. Ahmad, F. Pervaiz, U. Saif, and L. Subramanian, "Fine-grained dengue forecasting using telephone triage services," *Sci. Adv.*, vol. 2, no. 7, p. e1501215, 2016.
[35] V. R. Louis *et al.*, "Modeling tools for dengue risk mapping—A systematic review," *Int. J. Health Geograph.*, vol. 13, no. 1, p. 50, 2014.
[36] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 3, pp. 579–590, May 2013.
[37] L. Stanciu. (2017). Novel technology could provide a faster, inexpensive way to detect, monitor dengue fever, Zika virus. Purdue University. Accessed: Nov. 16, 2017. [Online]. Available: https://www.purdue.edu/newsroom/releases/2017/Q1/novel-technology-could-provide-a-faster,-inexpensive-way-to-detect,-monitor-dengue-fever,-zika-virus.html
[38] C. Singhal, C. S. Pundir, and J. Narang, "A genosensor for detection of consensus DNA sequence of dengue virus using ZnO/Pt-Pd nanocomposites," *Biosensors Bioelectron.*, vol. 97, pp. 75–82, Nov. 2017.
[39] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting breast cancer recurrence using machine learning techniques: A systematic review," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 52:1–52:40, 2016.
[40] J. Van den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst, "Data cleaning: Detecting, diagnosing, and editing data abnormalities," *PLoS Med.*, vol. 2, no. 10, p. e267, 2005.
[41] L. Al Shalabi and Z. Shaaban, "Normalization as a preprocessing engine for data mining and the approach of preference matrix," in *Proc. Int. Conf. Dependability Comput. Syst.*, Szklarska Poreba, Poland, May 2006, pp. 207–214.
[42] Y. Wang and H.-J. Chen, "Use of percentiles and Z-scores in anthropometry," in *Handbook Anthropometry*, New York, NY, USA: Springer, 2012, pp. 29–48.
[43] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann, 2000.
[44] Y. Kumar Jain and S. Kumar Bhandare, "Min Max normalization based data perturbation method for privacy protection," *Int. J. Comput. Commun. Technol.*, vol. 2, no. 8, pp. 45–50, 2011.
[45] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.
[46] S. Sang *et al.*, "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.
[47] S. A. Ahmed and J. S. Siddiqui, "Principal component analysis to explore climatic variability and dengue outbreak in lahore," *Pakistan J. Statist. Oper. Res.*, vol. 10, no. 2, pp. 247–256, 2014.
[48] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, Dec. 2000.
[49] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002.
[50] A. M. Wood, I. R. White, and S. G. Thompson, "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals," *Clin. Trials*, vol. 1, no. 4, pp. 368–376, 2004.
[51] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[52] C. Leacock and M. Chodorow, "Combining local context with wordnet similarity for word sense identification," in *Proc. WordNet, A Lexical Reference Syst. Appl.*, 1998, pp. 265–283.
[53] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," in *Proc. 9th Int. Conf. Semantic Comput.*, Anaheim, CA, USA, Feb. 2015, pp. 244–251.
[54] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, 2006.
[55] A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 37, no. 5, pp. 692–709, Sep. 2007.
[56] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," in *Proc. Int. Conf. Data Sci. Eng.*, Cochin, India, Aug. 2016, pp. 1–5.
[57] H. W. Marsh, "Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes," *Struct. Equation Model., Multidisciplinary J.*, vol. 5, no. 1, pp. 22–36, 1998.
[58] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, pp. 1–8, 2016.
[59] A. Donner, "The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values," *Amer. Statistician*, vol. 36, no. 4, pp. 378–381, 1982.
[60] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," *Int. Stat. Rev.*, vol. 78, no. 1, pp. 40–64, 2010.
[61] N. A. Setiawan, P. A. Venkatachalam, and A. F. M. Hani, "Missing data estimation on heart disease using artificial neural network and rough set theory," in *Proc. Int. Conf. Intell. Adv. Syst.*, Kuala Lumpur, Malaysia, Nov. 2007, pp. 129–133.
[62] P. Rey-del-Castillo and J. Cardeñosa, "Fuzzy min–max neural networks for categorical data: Application to missing data imputation," *Neural Comput. Appl.*, vol. 21, no. 6, pp. 1349–1362, 2012.
[63] D. V. Patil and R. S. Bichkar, "Multiple imputation of missing data with genetic algorithm based techniques," *Int. J. Comp. Appl.*, vol. 6, no. 2, pp. 74–78, 2010.
[64] C. Leke, B. Twala, and T. Marwala, "Modeling of missing data prediction: Computational intelligence and optimization algorithms," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, San Diego, CA, USA, Oct. 2014, pp. 1400–1404.
[65] C. T. Tran, M. Zhang, and P. Andreae, "Multiple imputation for missing data using genetic programming," in *Proc. Annu. Conf. Genetic Evol. Comput.*, New York, NY, USA, 2015, pp. 583–590.
[66] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," *Neurocomput.*, vol. 72, nos. 7–9, pp. 1483–1493, 2009.
[67] M. H. N. Beirami, M. H. N. Ghavifekr, and R. P. Khajei, "Predicting missing attribute values using cooperative particle swarm optimization," *J. Basic Appl. Sci. Res.*, vol. 3, no. 1, pp. 885–890, 2013.
[68] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ, USA: Wiley, 1987.
[69] J. I. Maletic and A. Marcus, "Data cleansing: A prelude to knowledge discovery," in *Data Mining and Knowledge Discovery Handbook*. Secaucus, NJ, USA: Springer-Verlag, 2005.
[70] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. New York, NY, USA: Wiley, 1994.
[71] M. S. Sarfraz, N. K. Tripathi, T. Tipdecho, T. Thongbu, P. Kerdthong, and M. Souris, "Analyzing the spatio-temporal relationship between dengue vector larval density and land-use using factor analysis and spatial ring mapping," *BMC Public Health*, vol. 12, p. 853, Oct. 2012.
[72] D. Yu, G. Sheikholeslami, and A. Zhang, "*FindOut*: Finding outliers in very large datasets," *Knowl. Inf. Syst.*, vol. 4, no. 4, pp. 387–412, 2002.
[73] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 1993, pp. 207–216.
[74] J. I. Maletic and A. Marcus, "Data cleansing: Beyond integrity analysis," in *Proc. 5th Conf. Inf. Qual.*, Boston, MA, USA, 2000, pp. 200–209.
[75] A. Marcus, J. I. Maletic, and K.-I. Lin, "Ordinal association rules for error identification in data sets," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA, 2001, pp. 589–591.
[76] A. L. Buczak *et al.*, "Prediction of high incidence of dengue in the philippines," *PLOS Neglected Tropical Diseases*, vol. 8, no. 4, p. e2771, 2014.
[77] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[78] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[79] K. K. Sheena and G. Kumar, "Analysis of feature selection techniques: A data mining approach," in *Proc. Int. Conf. Adv. Emerg. Technol. (ICAET)*, 2016, pp. 17–21.
[80] R. Kerber, "ChiMerge: Discretization of numeric attributes," in *Proc. 10th Nat. Conf. Artif. Intell.*, Palo Alto, CA, USA, 1992, pp. 123–128.

[81] A. L. Buczak, P. T. Koshute, S. M. Babin, B. H. Feighner, and S. H. Lewis, "A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data," *BMC Med. Inform. Decis.*, vol. 12, no. 1, p. 124, 2012.

[82] D. A. Focks and D. D. Chadee, "Pupal survey: An epidemiologically significant surveillance method for *Aedes aegypti*: An example using data from Trinidad," *Amer. J. Tropical Med. Hygiene*, vol. 56, no. 2, pp. 159–167, 1997.

[83] C. M. Seng, T. Setha, J. Nealon, and D. Socheat, "Pupal sampling for *Aedes aegypti* (L.) surveillance and potential stratification of dengue high-risk areas in Cambodia," *Tropical Med. Int. Health*, vol. 14, no. 10, pp. 1233–1240, 2009.

[84] H. V. Pham, H. T. Doan, T. T. Phan, and N. N. T. Minh, "Ecological factors associated with dengue fever in a central highlands Province, Vietnam," *BMC Infectious Diseases*, vol. 11, no. 1, p. 172, 2011.

[85] S. Naish, P. Dale, J. S. Mackenzie, J. McBride, K. Mengersen, and S. Tong, "Climate change and dengue: A critical and systematic review of quantitative modelling approaches," *BMC Infectious Diseases*, vol. 14, no. 1, p. 167, 2014.

[86] M. Gharbi *et al.*, "Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors," *BMC Infectious Diseases*, vol. 11, no. 1, p. 166, 2011.

[87] T. A. McLennan-Smith and G. N. Mercer, "Complex behaviour in a dengue model with a seasonally varying vector population," *Math. Biosci.*, vol. 248, pp. 22–30, Feb. 2014.

[88] M. N. Karim, S. U. Munsh, N. Anwar, and S. Alam, "Climatic factors influencing dengue cases in Dhaka city: A model for dengue prediction," *Indian J. Med. Res.*, vol. 136, no. 1, pp. 32–39, 2012.

[89] P.-C. Wu, J.-G. Lay, H.-R. Guo, C.-Y. Lin, S.-C. Lung, and H.-J. Su, "Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan," *Sci. Total Environ.*, vol. 407, no. 7, pp. 2224–2233, 2009.

[90] P. Reiter *et al.*, "Texas lifestyle limits transmission of dengue virus," *Emerg. Infectious Diseases*, vol. 9, no. 1, pp. 86–89, 2003.

[91] R. A. K. Tazkia, V. Narita, and A. S. Nugroho, "Dengue outbreak prediction for GIS based early warning system," in *Proc. Int. Conf. Sci. Inf. Technol.*, Yogyakarta, Indonesia, Oct. 2015, pp. 121–125.

[92] S. Wongkoon, M. Jaroensutasinee, and K. Jaroensutasinee, "Weather factors influencing the occurrence of dengue fever in Nakhon Si Thammarat, Thailand," *Tropical Biomed.*, vol. 30, no. 4, pp. 631–641, 2013.

[93] H.-Y. Xu *et al.*, "Statistical modeling reveals the effect of absolute humidity on dengue in Singapore," *PLoS Neglected Tropical Diseases*, vol. 8, no. 5, p. e2805, 2014.

[94] E. A. Machado-Machado, "Empirical mapping of suitability to dengue fever in Mexico using species distribution modeling," *Appl. Geogr.*, vol. 33, pp. 82–93, Apr. 2012.

[95] W.-Y. Hsu, T.-H. Wen, and H.-L. Yu, "Analysis of impact of geographical environment and socio-economic factors on the spatial distribution of Kaohsiung dengue fever epidemic," in *Proc. EGU Gen. Assem. Conf. Abstr.*, vol. 15, 2013, p. 9056.

[96] S. Thammapalo, V. Chongsuwiwatwong, D. McNeil, and A. Geater, "The climatic factors influencing the occurrence of dengue hemorrhagic fever in Thailand," *Southeast Asian J. Tropical Med. Public Health*, vol. 36, no. 1, pp. 191–196, 2005.

[97] P. Arcari, N. Tapper, and S. Pfueller, "Regional variability in relationships between climate and dengue/DHF in Indonesia," *Singapore J. Tropical Geogr.*, vol. 28, no. 3, pp. 251–272, 2007.

[98] H. Halide and P. Ridd, "A predictive model for dengue hemorrhagic fever epidemics," *Int. J. Environ. Health Res.*, vol. 18, no. 4, pp. 253–265, 2008.

[99] N. Sumanasinghe, A. R. Mikler, J. Muthukudage, C. Tiwari, and R. Quiroz, "Data driven prediction of dengue incidence in Thailand," in *Proc. Recent Adv. Inf. Commun. Technol.*, 2017, pp. 95–107.

[100] H. M. Aburas, "ABURAS index: A statistically developed index for dengue-transmitting vector population prediction," *Proc. World Acad. Sci., Eng. Technol.*, vol. 23, pp. 151–154, Aug. 2007.

[101] D. Strickman and P. Kittayapong, "Dengue and its vectors in Thailand: Calculated transmission risk from total pupal counts of *Aedes aegypti* and association of wing-length measurements with aspects of the larval habitat," *Amer. J. Tropical Med. Hygiene*, vol. 68, no. 2, pp. 209–217, 2003.

[102] S. Ma, E. E. Ooi, and K. T. Goh, "Socioeconomic determinants of dengue incidence in Singapore," *WHO Regional Office South-East Asia*, vol. 32, pp. 17–28, Dec. 2008.

[103] L. C. Harrington, J. D. Edman, and T. W. Scott, "Why do female *Aedes aegypti* (Diptera: Culicidae) feed preferentially and frequently on human blood?" *J. Med. Entomol.*, vol. 38, no. 3, pp. 411–422, 2001.

[104] S. Thongrungkiat, L. Wasinpiyamongkol, P. Maneekan, S. Prummongkol, and Y. Samung, "Natural transovarial dengue virus infection rate in both sexes of dark and pale forms of *Aedes aegypti* from an urban area of Bangkok, Thailand," *Southeast Asian J. Tropical Med. Public Health*, vol. 43, no. 5, pp. 1146–1152, 2012.

[105] A. Ponlawat and L. C. Harrington, "Age and body size influence male sperm capacity of the dengue vector *Aedes aegypti* (Diptera: Culicidae)," *J. Med. Entomol.*, vol. 44, no. 3, pp. 422–426, 2007.

[106] P. Veeraseatakul, S. Saosathan, and S. Chutipongvivate, "Pattern of dengue serotypes in four provinces of northern Thailand from 2003–2012," *Dengue Bull.*, vol. 38, pp. 11–19, Dec. 2014.

[107] K. Limkittikul, J. Brett, and M. L'Azou, "Epidemiological trends of dengue disease in Thailand (2000–2011): A systematic literature review," *PLoS Neglected Tropical Diseases*, vol. 8, no. 11, p. e3241, 2014.

[108] M. S. Mustafa, V. Rasotgi, S. Jain, and V. Gupta, "Discovery of fifth serotype of dengue virus (DENV-5): A new public health dilemma in dengue control," *Med. J. Armed Forces India*, vol. 71, no. 1, pp. 67–70, 2014.

[109] WHO. (2017). *The Mosquito*. Accessed: Nov. 27, 2017. [Online]. Available: http://www.who.int/denguecontrol/mosquito/en/

[110] U. Thavara *et al.*, "Biology of dengue vectors and serotypes of dengue virus in infectious cycle in Thailand," *Bull. Dept. Med. Sci.*, vol. 57, no. 2, pp. 186–196, 2015.

[111] Centers for Disease Control and Prevention, "Dengue hemorrhagic fever—US-Mexico border, 2005," *Morbidity Mortality Weekly Rep.*, vol. 56, no. 1, p. 785, 2007.

[112] K. Knowlton, G. Solomon, and M. Rotkin-Ellman, "Mosquito-Borne dengue fever threat spreading in the Americas," *Natural Resour. Defense Council*, 2009. Accessed: Sep. 22, 2018. [Online]. Available: https://www.nrdc.org/sites/default/files/dengue.pdf

[113] WHO. (2014). *A Global Brief on Vector-Borne Diseases*. Accessed: May 23, 2014. [Online]. Available:http://www.who.int/campaigns/world-health-day/2014/global-brief/en/

[114] D. K. Biswas, R. Bhunia, and M. Basu, "Dengue fever in a rural area of West Bengal, India, 2012: An outbreak investigation," *WHO South-East Asia J. Public Health*, vol. 3, no. 1, pp. 46–50, 2014.

[115] S. Vong *et al.*, "Dengue incidence in urban and rural Cambodia: Results from population-based active fever surveillance, 2006–2008," *PLoS Neglected Tropical Diseases*, vol. 4, no. 11, p. e903, 2010.

[116] J. Quintero *et al.*, "Ecological, biological and social dimensions of dengue vector breeding in five urban settings of Latin America: A multi-country study," *BMC Infectious Diseases*, vol. 14, no. 1, p. 38, 2014.

[117] S. Karl, N. Halder, J. K. Kelso, S. A. Ritchie, and G. J. Milne, "A spatial simulation model for dengue virus infection in urban areas," *BMC Infectious Diseases*, vol. 14, no. 1, p. 447, 2014.

[118] T.-C. Chan, T.-H. Hu, and J.-S. Hwang, "Daily forecast of dengue fever incidents for urban villages in a city," *Int. J. Health Geograph.*, vol. 14, no. 1, p. 9, 2015.

[119] H.-L. Yu, J. M. Angulo, M.-H. Cheng, J. Wu, and G. Christakos, "An online spatiotemporal prediction model for dengue fever epidemic in Kaohsiung (Taiwan)," *Biometrical J.*, vol. 56, no. 3, pp. 428–440, 2014.

[120] A. C. C. Costa, C. T. Codeço, N. A. Honório, G. R. Pereira, C. F. N. Pinheiro and A. A. Nobre, "Surveillance of dengue vectors using spatio-temporal Bayesian modeling," *BMC Med. Inform. Decis. Making*, vol. 15, p. 93, Nov. 2015.

[121] L. Thiruchelvam, S. C. Dass, R. Zaki, A. Yahya, and V. S. Asirvadam, "Correlation analysis of air pollutant index levels and dengue cases across five different zones in Selangor, Malaysia," *Geospatial Health*, vol. 13, no. 1, p. 613, 2018.

[122] "Travel-associated dengue infections—United States, 2001–2004," Centers Diseases Control Prevention, Atlanta, GA, USA, Morbidity Mortality Weekly Rep. 54(22), 2005.

[123] G. G. G. Baaten, G. J. B. Sonder, H. L. Zaaijer, T. van Gool, J. A. P. C. M. Kint, and A. van den Hoek, "Travel-related dengue virus infection, The Netherlands, 2006–2007," *Emerg. Infectious Diseases*, vol. 17, no. 5, pp. 821–828, 2011.

[124] I. Ratnam, K. Leder, J. Black, and J. Torresi, "Dengue fever and international travel," *J. Travel Med.*, vol. 20, no. 6, pp. 384–393, 2013.

[125] J. N. Hanna *et al.*, "Multiple outbreaks of dengue serotype 2 in north Queensland, 2003/04," *Austral. New Zealand J. Public Health*, vol. 30, no. 3, pp. 220–225, 2006.

[126] D. Teichmann, K. Göbels, M. Niedrig, and M. P. Grobusch, "Dengue virus infection in travellers returning to Berlin, Germany: Clinical, laboratory, and diagnostic aspects," *Acta Tropica*, vol. 90, no. 1, pp. 87–95, 2004.
[127] WHO. (2009). *Dengue Guidelines for Diagnosis, Treatment, Prevention and Control: New Edition*. Accessed: Nov. 16, 2017. [Online]. Available: http://www.who.int/rpc/guidelines/9789241547871/en/
[128] R. C. Reiner, Jr., S. T. Stoddard, and T. W. Scott, "Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal," *Epidemics*, vol. 6, pp. 30–36, Mar. 2014.
[129] W. A. Hawley, P. Reiter, R. S. Copeland, C. B. Pumpuni, and G. B. Craig, Jr., "Aedes albopictus in North America: Probable introduction in used tires from northern Asia," *Science*, vol. 236, no. 4805, pp. 1114–1116, 1987.
[130] D. T. T. Toan, L. N. Hoat, W. Hu, P. Wright, and P. Martens, "Risk factors associated with an outbreak of dengue fever/dengue haemorrhagic fever in Hanoi, Vietnam," *Epidemiol. Infection*, vol. 143, no. 8, pp. 1594–1598, 2015.
[131] C.-H. Chiu, T.-H. Wen, L.-C. Chien, and H.-L. Yu, "A probabilistic spatial dengue fever risk assessment by a threshold-based-quantile regression method," *PLoS ONE*, vol. 9, no. 10, p. e106334, 2014.
[132] Y. L. Cheong, P. J. Leitão, and T. Lakes, "Assessment of land use factors associated with dengue cases in Malaysia using boosted regression trees," *Spatial Spatio-Temporal Epidemiol.*, vol. 10, pp. 75–84, Jul. 2014.
[133] H. J. Wearing and P. Rohani, "Ecological and immunological determinants of dengue epidemics," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 31, pp. 11802–11807, 2006.
[134] B. L. Innis, "Dengue and dengue hemorrhagic fever," in *Kass Handbook of Infectious Diseases: Exotic Virus Infectious*, J. S. Porterfield, Ed. London, U.K.: Chapman & Hall, 1995, pp. 103–146.
[135] A. Fox *et al.*, "Immunological and viral determinants of dengue severity in hospitalized adults in Ha Noi, Viet Nam," *PLoS Neglected Tropical Diseases*, vol. 5, no. 3, p. e967, 2011.
[136] S. B. Halstead, S. Nimmannitya, and S. N. Cohen, "Observations related to pathogenesis of dengue hemorrhagic fever. IV. Relation of disease severity to antibody response and virus recovered," *Yale J. Biol. Med.*, vol. 42, no. 5, pp. 311–328, 1970.
[137] W. Waidab, K. Suphapeetiporn, and U. Thisyakorn, "Pathogenesis of dengue hemorrhagic fever: From immune to genetics," *J. Pediatric Infectious Diseases*, vol. 3, no. 4, pp. 221–227, 2008.
[138] U. Thisyakorn and S. Nimmannitya, "Nutritional status of children with dengue hemorrhagic fever," *Clin. Infectious Diseases*, vol. 16, no. 2, pp. 295–297, 1993.
[139] L. Guglani and S. K. Kabra, "T cell immunopathogenesis of dengue virus infection," *Dengue Bull.*, vol. 29, pp. 58–69, Dec. 2005.
[140] R. Guabiraba and B. Ryffel, "Dengue virus infection: Current concepts in immune mechanisms and lessons from murine models," *Immunology*, vol. 141, no. 2, pp. 143–156, 2014.
[141] H.-Y. Lei, K.-J. Huang, Y.-S. Lin, T.-M. Yeh, H.-S. Liu, and C.-C. Liu, "Immunopathogenesis of dengue hemorrhagic fever," *Amer. J. Infectious Diseases*, vol. 4, no. 1, pp. 1–9, 2008.
[142] C. Pagliari *et al.*, "Immunopathogenesis of dengue hemorrhagic fever: Contribution to the study of human liver lesions," *J. Med. Virol.*, vol. 86, no. 7, pp. 1193–1197, 2014.
[143] M. Oki and T. Yamamoto, "Climate change, population immunity, and hyperendemicity in the transmission threshold of dengue," *PLoS ONE*, vol. 7, no. 10, p. e48258, 2012.
[144] A. B. Sabin and R. W. Schlesinger, "Production of immunity to dengue with virus modified by propagation in mice," *Science*, vol. 101, no. 2634, pp. 640–642, 1945.
[145] S. Hotta, "Experimental studies on dengue: I. Isolation, identification and modification of the virus," *J. Infectious Diseases*, vol. 90, no. 1, pp. 1–9, 1952.
[146] R. W. Schlesinger and J. W. Frankel, "Adaptation of the 'new guinea B' strain of dengue virus to suckling and to adult Swiss mice," *Amer. Soc. Tropical Med. Hygiene*, vol. 1, no. 1, pp. 66–77, 1952.
[147] World Health Organization, "Dengue vaccine: WHO position paper, July 2016—Recommendations," *Vaccine*, vol. 35, no. 9, pp. 1200–1201, 2017.
[148] M. Aguiar, N. Stollenwerk, and S. B. Halstead, "The impact of the newly licensed dengue vaccine in endemic countries," *PLoS Neglected Tropical Diseases*, vol. 10, no. 12, p. e0005179, 2016.
[149] B. Berkrot and M. Serapio. (Dec. 4, 2017). *Trouble Mounts for Sanofi Dengue Vaccine Over Safety Concerns*. Accessed: Aug. 1, 2018. [Online]. Available: https://www.reuters.com/article/us-sanofi-dengue/trouble-mounts-for-sanofi-dengue-vaccine-over-safety-concerns-idUSKBN1DY26Y
[150] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 3, pp. 1–10, 2014.
[151] J. Jovanović and E. Bagheri, "Semantic annotation in biomedicine: the current landscape," *J. Biomed. Semantics*, vol. 8, no. 1, p. 44, 2017.
[152] H. M. Khormi and L. Kumar, "Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study," *Sci. Total Environ.*, vol. 409, no. 22, pp. 4713–4719, 2011.
[153] N. T. Huy *et al.*, "Development of clinical decision rules to predict recurrent shock in dengue," *Crit. Care*, vol. 17, no. 6, p. R280, 2013.
[154] E. Mitraka, P. Topalis, V. Dritsou, E. Dialynas, and C. Louis, "Describing the breakbone fever: IDODEN, an ontology for dengue fever," *PLoS Neglected Tropical Diseases*, vol. 9, no. 2, p. e0003479, 2015.
[155] A. Herdiani, L. Fitria, H. Hayurani, W. C. Wibowo, and S. Sungkar, "Hierarchical conceptual schema for dengue hemorrhagic fever ontology," *Int. J. Comput. Sci.*, vol. 9, no. 4, pp. 53–58, 2012.
[156] S. Lozano-Fuentes, A. Bandyopadhyay, L. G. Cowell, A. Goldfain, and L. Eisen, "Ontology for vector surveillance and management," *J. Med. Entomol.*, vol. 50, no. 1, pp. 1–14, 2013.
[157] T. Dias, W. Pinheiro, and R. Salles, "A semantic platform of support decision to manage dengue epidemics," in *Proc. Congr. Comput. Intell., 11th Brazilian Congr. Comput. Intell. (BRICS)*, Recife, Brazil, 2013, pp. 687–692.
[158] National University of Singapore. (2017). *Dengue Virus Protein Sequence Database (DENVDB)*. Accessed: Mar. 31, 2018. [Online]. Available: http://proline.bic.nus.edu.sg/denvdb/
[159] WHO. (2017). *DengueNet*. Accessed: Mar. 31, 2018. [Online]. Available: http://www.who.int/denguenet
[160] I. Ruberto, E. Marques, D. S. Burke, and W. G. Van Panhuis, "The availability and consistency of dengue surveillance data provided online by the World Health Organization," *PLoS Neglected Tropical Diseases*, vol. 9, no. 4, p. e0003511, 2015.
[161] J. P. Messina, O. J. Brady, D. M. Pigott, J. S. Brownstein, A. G. Hoen, and S. I. Hay, "A global compendium of human dengue virus occurrence," *Sci. Data*, vol. 1, May 2014, Art. no. 140004.
[162] R. K. Shardiwal and S. Magar. (2017). *Dengue Drug Target Database*. Accessed: Mar. 31, 2018. [Online]. Available: http://www.bioinformatics.org/dengueDTDB/Pages/main.htm
[163] S. Chaudhury, G. D. Gromowski, D. R. Ripoll, I. V. Khavrutskii, V. Desai, and A. Wallqvist, "Dengue virus antibody database: Systematically linking serotype-specificity with epitope mapping in dengue virus," *PLoS Neglected Tropical Diseases*, vol. 11, no. 2, p. e0005395, 2017.
[164] E. L. Hatcher *et al.*, "Virus variation resource—Improved response to emergent viral outbreaks," *Nucleic Acids Res.*, vol. 45, pp. D482–D490, Jan. 2017.
[165] W. H. Inmon, *Building the Data Warehouse*, 2nd ed. New York, NY, USA: Wiley, 1996.
[166] M. F. Wisniewski, P. Kieszkowski, B. M. Zagorski, W. E. Trick, M. Sommers, and R. A. Weinstein, "Development of a clinical data warehouse for hospital infection control," *J. Amer. Med. Inform. Assoc.*, vol. 10, no. 5, pp. 454–462, 2003.
[167] W. E. Trick *et al.*, "Computer algorithms to detect bloodstream infections," *Emerg. Infectious Diseases*, vol. 10, no. 9, pp. 1612–1620, 2004.
[168] W. L. Kim, C. AnneDucharme, and B. J.-M. P. Bucher, "Development and implementation of a surveillance network system for emerging infectious diseases in the caribbean (ARICABA)," *Online J. Public Health Inform.*, vol. 3, no. 2, pp. 1–17, 2011.
[169] R. Sint, S. Schaffert, R. Ferstl, and S. Stroka, "Combining unstructured, fully structured and semi-structured information in semantic Wikis," in *Proc. 6th Eur. Semantic Web Conf.*, Heraklion, Greece, 2009, pp. 73–87.
[170] L. Tanner *et al.*, "Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness," *PLoS Neglected Tropical Diseases*, vol. 2, no. 3, p. e196, 2008.
[171] V. J. Lee, D. C. Lye, Y. Sun, and Y. S. Leo, "Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore," *Tropical Med. Int. Health*, vol. 14, no. 9, pp. 1154–1159, 2009.

[172] L. E. Hugo *et al.*, "Adult survivorship of the dengue mosquito *Aedes aegypti* varies seasonally in central Vietnam," *PLoS Neglected Tropical Diseases*, vol. 8, no. 2, p. e2669, 2014.

[173] F. Ibrahim, T. Faisal, M. I. M. Salim, and M. N. Taib, "Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network," *Med. Biol. Eng. Comput.*, vol. 48, no. 11, pp. 1141–1148, 2010.

[174] F. Ibrahim, M. N. Taib, W. A. B. W. Abas, C. C. Guan, and S. Sulaiman, "A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN)," *Comput. Meth. Programs Biomed.*, vol. 79, no. 3, pp. 273–281, 2005.

[175] Google. (2017). *Google Brain Team's Mission, Research at Google*. Accessed: Mar. 31, 2018. [Online]. Available: https://research.google.com/teams/brain/about.html

[176] A. Ghaderi, B. M. Sanandaji, and F. Ghaderi, "Deep forecast: Deep learning-based spatio-temporal forecasting," in *Proc. Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 1–6.

[177] M. Hossain, B. Rekabdar, S. J. Louis, and S. Dascalu, "Forecasting the weather of Nevada: A deep learning approach," in *Proc. Int. Joint Conf. Neural Netw.*, Killarney, Ireland, 2015, pp. 1–6.

[178] A. G. Salman, B. Kanigoro, and Y. Heryadi, "Weather forecasting using deep learning techniques," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst.*, Depok, Indonesia, 2015, pp. 281–285.

[179] A. Ashiquzzaman *et al.*, "Reduction of overfitting in diabetes prediction using deep learning neural network," in *Proc. 7th iCatse Int. Conf. IT Converg. Secur.*, Seoul, South Korea, 2017, pp. 35–43.

[180] A. Fuentes, S. Yoon, S. C. Kim, and D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," *Sensors*, vol. 17, no. 9, p. 2022, 2017.

[181] A. R. Kavsaoğlu, K. Polat, and M. Hariharan, "Non-invasive prediction of hemoglobin level using machine learning techniques with the PPG signal's characteristics features," *Appl. Soft Comput.*, vol. 37, pp. 983–991, Dec. 2015.

[182] K. D. Sharma, R. S. Mahabir, K. M. Curtin, J. M. Sutherland, J. B. Agard, and D. D. Chadee, "Exploratory space-time analysis of dengue incidence in Trinidad: A retrospective study using travel hubs as dispersal points, 1998–2004," *Parasite Vectors*, vol. 7, no. 1, p. 341, 2014.

[183] C. Chauhan *et al.*, "Comparative expression profiles of midgut genes in dengue virus refractory and susceptible *Aedes aegypti* across critical period for virus infection," *PLoS ONE*, vol. 7, no. 10, p. e47350, 2012.

[184] H. L. Nguyen *et al.*, "Specific K-mean clustering-based perceptron for dengue prediction," *Int. J. Int. Inf. Database Syst.*, vol. 10, nos. 3–4, pp. 269–288, 2017.

[185] N. Mathur, V. S. Asirvadam, S. C. Dass, and B. S. Gill, "Visualization of dengue incidences for vulnerability using K-means," in *Proc. IEEE Int. Conf. Signal Image Process. Appl.*, Kuala Lumpur, Malaysia, 2015, pp. 569–573.

[186] P. Manivannan and D. P. Isakki, "Dengue fever prediction using K-medoid clustering algorithm," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 1, pp. 77–84, 2017.

[187] S. Bhatnagar, V. Lal, S. D. Gupta, and O. P. Gupta, "Forecasting incidence of dengue in Rajasthan, using time series analyses," *Indian J. Public Health*, vol. 56, no. 4, pp. 281–285, 2012.

[188] C. C. Ho and C.-Y. Ting, "Time series analysis and forecasting of dengue using open data," in *Advances in Visual Informatics*. Cham, Switzerland: Springer, 2015, pp. 51–63.

[189] A. Lal, T. Ikeda, N. French, M. G. Baker, and S. Hales, "Climate variability, weather and enteric disease incidence in New Zealand: Time series analysis," *PLoS ONE*, vol. 8, no. 12, p. e83484, 2013.

[190] H. Lin *et al.*, "Time series analysis of Japanese encephalitis and weather in Linyi City, China," *Int. J. Public Health*, vol. 57, no. 2, pp. 289–296, 2011.

[191] F. A. Siregar, T. Makmur, and S. Saprin, "Forecasting dengue hemorrhagic fever cases using ARIMA model: A case study in Asahan district," in *Proc. IOP Conf. Ser., Mater. Sci. Eng.*, vol. 300, no. 1, 2018, p. 012032.

[192] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[193] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Databases*, San Francisco, CA, USA, 1994, pp. 487–499.

[194] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2000, pp. 1–12.

[195] D. H. Barmak, C. O. Dorso, M. Otero, and H. G. Solari, "Dengue epidemics and human mobility," *Phys. Rev. A, Gen. Phys.*, vol. 84, no. 1, p. 011901, 2011.

[196] D. H. Barmak, C. O. Dorso, and M. Otero, "Modelling dengue epidemic spreading with human mobility," *Phys. A, Stat. Mech. Appl.*, vol. 447, pp. 129–140, Apr. 2016.

[197] S. Chadsuthi, S. Iamsirithaworn, W. Triampo, and D. A. T. Cummings, "The impact of rainfall and temperature on the spatial progression of cases during the chikungunya re-emergence in Thailand in 2008–2009," *Trans. Roy. Soc. Tropical Med. Hygiene*, vol. 110, no. 2, pp. 125–133, Jan. 2016.

[198] L. C. de Castro Medeiros, C. A. R. Castilho, C. Braga, W. V. de Souza, L. Regis, and A. M. V. Monteiro, "Modeling the dynamic transmission of dengue fever: Investigating disease persistence," *PLoS Neglected Tropical Diseases*, vol. 5, no. 1, p. e942, 2011.

[199] T. Botari, S. G. Alves, and E. D. Leonel, "Explaining the high number of infected people by dengue in Rio de Janeiro in 2008 using a susceptible-infective-recovered model," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 83, no. 3, pp. 1–4, 2011.

[200] A. G. Dickman and R. Dickman, "Computational model of a vector-mediated epidemic," *Amer. J. Phys.*, vol. 83, no. 5, pp. 468–474, 2015.

[201] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza, and A. Vespignani, "Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic," *PLoS ONE*, vol. 6, no. 1, p. e16591, 2011.

[202] L. A. Rvachev and I. M. Longini, Jr., "A mathematical model for the global spread of influenza," *Math. Biosci.*, vol. 75, no. 1, pp. 3–22, 1985.

[203] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[204] A. Neumaier, "Solving Ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Rev.*, vol. 40, no. 3, pp. 636–666, 1998.

[205] J. J. Dziak. *Sensitivity and Specificity of Information Criteria*. Accessed: Jun. 27, 2012. [Online]. Available: http://methodology.psu.edu/media/techreports/12-119.pdf

[206] P.-C. Wu, H.-R. Guo, S.-C. Lung, C.-Y. Lin, and H.-J. Su, "Weather as an effective predictor for occurrence of dengue fever in Taiwan," *Acta Tropica*, vol. 103, no. 1, pp. 50–57, 2007.

[207] M. Snipes and D. C. Taylor, "Model selection and Akaike information criteria: An example from wine ratings and prices," *Wine Econ. Policy*, vol. 3, no. 1, pp. 3–9, 2014.

[208] A. R. Brasier *et al.*, "A three-component biomarker panel for prediction of dengue hemorrhagic fever," *Amer. J. Tropical Med. Hygiene*, vol. 86, no. 2, pp. 341–348, 2012.

[209] E. Wit, E. van den Heuvel, and J.-W. Romeijn, "'All models are wrong...': An introduction to model uncertainty," *Statistica Neerlandica*, vol. 66, no. 3, pp. 217–236, 2012.

[210] M. S. Sitepu *et al.*, "Temporal patterns and a disease forecasting model of dengue hemorrhagic fever in Jakarta based on 10 years of surveillance data," *Southeast Asian J. Tropical Med. Public Health*, vol. 44, no. 2, pp. 206–217, 2013.

[211] R. Soundravally *et al.*, "Association between proinflammatory cytokines and lipid peroxidation in patients with severe dengue disease around defervescence," *Int. J. Infectious Diseases*, vol. 18, pp. 68–72, Jan. 2014.

[212] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clin. Chem.*, vol. 39, no. 4, pp. 561–577, 1993.

[213] T. Netsuwan and K. Kesorn, "Unify framework for crime data summarization using RSS feed service," *Walailak J. Sci. Technol.*, vol. 14, no. 10, pp. 769–781, 2017.

[214] V. G. da Costa, A. C. Marques-Silva, and M. L. Moreli, "A meta-analysis of the diagnostic accuracy of two commercial NS1 antigen ELISA tests for early dengue virus detection," *PLoS ONE*, vol. 9, no. 4, p. e94655, 2014.

[215] S. Wongkoon, M. Jaroensutasinee, and K. Jaroensutasinee, "Development of temporal modeling for prediction of dengue infection in Northeastern Thailand," *Asian. Pacific J. Tropical Med.*, vol. 5, no. 3, pp. 249–252, 2012.

[216] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[217] S. Makridakis, "Accuracy measures: Theoretical and practical concerns," *Int. J. Forecasting*, vol. 9, no. 4, pp. 527–529, 1993.

[218] C. J. Willmott and K. Matsuura, "On the use of dimensioned measures of error to evaluate the performance of spatial interpolators," *Int. J. Geograph. Inf. Sci.*, vol. 20, no. 1, pp. 89–102, 2006.
[219] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.
[220] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.
[221] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Melbourne, VIC, Australia: OTexts, 2013.
[222] K. Lakshmi and S. Karthik, "Dengue identification and patient care monitoring using Internet of medical things," *J. Health Inf. Manage.*, vol. 1, no. 2, pp. 1–3, 2017.
[223] C. Whitlow. (2016). *Mediatek Dedicates IoT Platform to Dengue Fever Prevention in Southern Taiwan*. Accessed: Nov. 16, 2017. Tech News Hunter. [Online]. Available: http://technewshunter.com/apple/mediatek-dedicates-iot-platform-to-dengue-fever-prevention-in-southern-taiwan-31188
[224] M. B. Hugosson, "Quantifying uncertainties in a national forecasting model," *Transp. Res. A, Policy Pract.*, vol. 39, no. 6, pp. 531–547, 2005.
[225] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, 1st ed. New York, NY, USA: Chapman & Hall, 1993.
[226] G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, vol. 28, no. 3, pp. 591–605, 1960.
[227] G. Ghilagaber, "Another look at chow's test for the equality of two heteroscedastic regression models," *Qual. Quantity*, vol. 38, no. 1, pp. 81–93, 2004.
[228] S. Bansal, G. Chowell, L. Simonsen, A. Vespignani, and C. Viboud, "Big data for infectious disease surveillance and modeling," *J. Infectious Diseases*, vol. 214, pp. S375–S379, Dec. 2016.
[229] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky, M. Granzow, Eds. Norwell, MA, USA: Kluwer, 2003, pp. 91–109.
[230] M. Vukićević, S. Radovanović, M. Milovanović, and M. Minović, "Cloud based metalearning system for predictive modeling of biomedical data," *Sci. World J.*, vol. 2014, Apr. 2014, Art. no. 859279.
[231] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Gener. Comput. Syst.*, vol. 28, no. 3, pp. 583–592, 2012.
[232] L. Dai, X. Gao, Y. Guo, J. Xiao, and Z. Zhang, "Bioinformatics clouds for big data manipulation," *Biol. Direct*, vol. 7, p. 43, Nov. 2012.
[233] I. K. W. Lai, S. K. T. Tam, and M. F. S. Chan, "Knowledge cloud system for network collaboration: A case study in medical service industry in China," *Expert Syst. Appl.*, vol. 39, no. 15, pp. 12205–12212, 2012.
[234] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.
[235] N. Shadbolt, T. Berners-Lee, and W. Hall, "The semantic Web revisited," *IEEE Intell. Syst.*, vol. 21, no. 3, pp. 96–101, Jan. 2006.
[236] K. S. Candan, H. Liu, and R. Suvarna, "Resource description framework: Metadata and its applications," *SIGKDD Explor. Newsl.*, vol. 3, no. 1, pp. 6–19, 2001.
[237] S. Poslad and K. Kesorn, "A multi-modal incompleteness ontology model (MMIO) to enhance information fusion for image retrieval," *Inf. Fusion*, vol. 20, pp. 225–241, Nov. 2014.
[238] R. S. Gonçalves, S. W. Tu, C. I. Nyulas, M. J. Tierney, and M. A. Musen, "An ontology-driven tool for structured data acquisition using Web forms," *J. Biomed. Semantics*, vol. 8, no. 1, p. 26, 2017.
[239] S. Mate *et al.*, "Ontology-based data integration between clinical and research systems," *PLoS ONE*, vol. 10, no. 1, pp. 1–20, 2015.
[240] S. Brüggemann and F. Grüning, "Using ontologies providing domain knowledge for data quality management," in *Networked Knowledge—Networked Media: Integrating Knowledge Management, New Media Technologies and Semantic Systems*, T. Pellegrini, S. Auer, K. Tochtermann, and S. Schaffert, Eds. Berlin, Heidelberg: Springer, 2009, pp. 187–203.
[241] D. Cherix, R. Usbeck, A. Both, and J. Lehmann, "Lessons learned—The case of CROCUS: Cluster-based ontology data cleansing," in *Proc. Semantic Web, ESWC Satell. Events*, Crete, Greece, 2014, pp. 14–24.
[242] J. T. Wong and J. L. Hong, "Data cleaning utilizing ontology tool," *Int. J. Grid Distrib. Comput.*, vol. 9, no. 7, pp. 43–52, 2016.
[243] A. Mohammadi and M. H. Saraee, "Estimating missing value in microarray data using fuzzy clustering and gene ontology," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Philadelphia, PA, USA, Nov. 2008, pp. 382–385.
[244] B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett, "The unified medical language system: An informatics research collaboration," *J. Amer. Med. Inform. Assoc.*, vol. 5, no. 1, pp. 1–11, 1998.
[245] M. Ashburner *et al.*, "Gene ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, no. 1, pp. 25–29, 2000.
[246] Stanford University. (2017). *The National Center for Biomedical Ontology*. Accessed: Nov. 12, 2017. [Online]. Available: https://www.bioontology.org/
[247] N. Phan, D. Dou, H. Wang, D. Kil, and B. Piniewski, "Ontology-based deep learning for human behavior prediction with explanations in health social networks," *Inf. Sci.*, vol. 384, pp. 298–313, Apr. 2017.
[248] H.-P. Kriegel, K. M. Borgwardt, P. Kröger, A. Pryakhin, M. Schubert, and A. Zimek, "Future trends in data mining," *Data Min. Knowl. Discovery*, vol. 15, no. 1, pp. 87–97, 2007.
[249] T. G. Dietterich, P. Domingos, L. Getoor, S. Muggleton, and P. Tadepalli, "Structured machine learning: the next ten years," *Mach. Learn.*, vol. 73, no. 1, p. 3, 2008.
[250] A. Suyama, N. Negishi, and T. Yamaguchi, "CAMLET: A platform for automatic composition of inductive learning systems using ontologies," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, Singapore, 1998, pp. 205–215.
[251] A. Bernstein, F. Provost, and S. Hill, "Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 503–518, Apr. 2005.
[252] J. Vanschoren and L. Soldatova, "Exposé: An ontology for data mining experiments," in *Proc. 3rd Gener. Data Mining, Towards Service-Oriented Knowl. Discovery*, Barcelona, Spain, 2010, pp. 31–44.
[253] P. Panov, S. Džeroski, and L. N. Soldatova, "Representing entities in the OntoDM data mining ontology," in *Inductive Databases and Constraint-Based Data Mining*. New York, NY, USA: Springer, 2010, pp. 27–58.
[254] L. N. Soldatova and R. D. King, "An ontology of scientific experiments," *J. Roy. Soc. Interface*, vol. 3, no. 11, pp. 795–803, 2006.
[255] M. Cannataro and C. Comito, "A data mining ontology for grid programming," in *Proc. 1st Int. Workshop Semantics Peer-Peer Grid Comput.*, Budapest, Hungary, 2003, pp. 113–134.
[256] C. Diamantini, D. Potena, and E. Storti, "Ontology-driven KDD process composition," in *Proc. Int. Symp. Adv. Intell. Data Anal. VIII*, Lyon, France, 2009, pp. 285–296.
[257] J.-U. Kietz, F. Serban, A. Bernstein, and S. Fischer, "Towards cooperative planning of data mining workflows," in *Proc. Workshop 3rd Gener. Data Mining, Towards Service-oriented Knowl. Discovery*, Bled, Slovenia, 2009, pp. 1–12.
[258] M. Hilario, A. Kalousis, P. Nguyen, and A. Woznica, "A data mining ontology for algorithm selection and meta-mining," in *Proc. Workshop 3rd Gener. Data Mining*, Bled, Slovenia, 2009, pp. 76–87.
[259] M. Zakova, P. Kremen, F. Zelezny, and N. Lavrac, "Automating knowledge discovery workflow composition through ontology-based planning," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 2, pp. 253–264, Apr. 2011.
[260] V. Svátek, J. Rauch, and M. Ralbovský, "Ontology-enhanced association mining," in *European Web Mining Forum*. Berlin, Germany: Springer, 2006, pp. 163–179.
[261] C. Marinica and F. Guillet, "Knowledge-based interactive postmining of association rules using ontologies," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 784–797, Jun. 2010.
[262] N. Balcan, A. Blum, and Y. Mansour, "Exploiting ontology structures and unlabeled data for learning," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1112–1120.
[263] M. Allahyari, K. J. Kochut, and M. Janik, "Ontology-based text classification into dynamically defined topics," in *Proc. IEEE Int. Conf. Semantic Comput.*, Newport Beach, CA, USA, Jun. 2014, pp. 273–278.
[264] W. Song, C. H. Li, and S. C. Park, "Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9095–9104, 2009.
[265] L. Jing, M. K. Ng, and J. Z. Huang, "Knowledge-based vector space model for text clustering," *Knowl. Inf. Syst.*, vol. 25, no. 1, pp. 35–55, 2010.

[266] K. Ovaska, M. Laakso, and S. Hautaniemi, "Fast gene ontology based clustering for microarray experiments," *BioData Min.*, vol. 1, no. 1, p. 11, 2008.
[267] X. Zhang, L. Jing, X. Hu, M. Ng, J. X. Jiangxi, and X. Zhou, "Medical document clustering using ontology-based term similarity measures," *Int. J. Data Warehouse*, vol. 4, no. 1, pp. 62–73, 2008.
[268] H. Liu, D. Dou, R. Jin, P. Lependu, and N. Shah, "Mining biomedical ontologies and data using RDF hypergraphs," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, Dec. 2013, pp. 141–146.
[269] N. S. Korgaonkar, A. Kumar, R. S. Yadav, D. Kabadi, and A. P. Dash, "Mosquito biting activity on humans & detection of Plasmodium falciparum infection in *Anopheles stephensi* in Goa, India," *Indian J. Med. Res.*, vol. 135, no. 1, pp. 120–126, 2012.
[270] D. M. Bustamante and C. C. Lord, "Sources of error in the estimation of mosquito infection rates used to assess risk of arbovirus transmission," *Amer. J. Tropical Med. Hygiene*, vol. 82, no. 6, pp. 1172–1184, 2010.
[271] E. Schwartz *et al.*, "Seasonality, annual trends, and characteristics of dengue among ill returned travelers, 1997–2006," *Emerg. Infectious Diseases*, vol. 14, no. 7, pp. 1081–1088, 2008.
[272] D. W. MacPherson *et al.*, "Population mobility, globalization, and antimicrobial drug resistance," *Emerg. Infectious Dis.*, vol. 15, no. 11, pp. 1727–1731, 2009.
[273] B. D. Gushulak and D. W. MacPherson, "The basic principles of migration health: Population mobility and gaps in disease prevalence," *Emerg. Themes Epidemiol.*, vol. 3, no. 3, pp. 1–11, 2006.

P. SIRIYASATIEN received the B.Sc. degree in entomology from Kasetsart University, the Medical degree from Chulalongkorn University, the DTM&H degree from Mahidol University, Bangkok, and the PhD. degree from the University of Liverpool, U.K. He is currently the Head of the Parasitology Department, Medicine Faculty, Vector Biology and Vector Borne Disease Research Unit, Chulalongkorn University, where he is also an Associate Professor. His work focuses on the biology and control of medically important arthropods including mosquito vectors of dengue, Chickungunya and Zika, bed bugs, head louse, and flies. He also works on leishmaniasis, an emerging parasitic disease in Thailand.

S. CHADSUTHI received the Ph.D. degree in physics from Mahidol University, Thailand. She is currently an Assistant Professor with the Department of Physics, Faculty of Science, Naresuan University, Thailand. Her current research interests include computational epidemiology, stochastic modeling, and Monte Carlo simulations. She has been receiving research grants as a new scholar from the Thailand Research Fund (TRF) and Office of Higher Education Commission (2018–2019).

K. JAMPACHAISRI received the M.Sc. degree in statistics from the University of Wisconsin–Madison, USA, and the Ph.D. degree from the University of California at Riverside, USA. She is currently an Assistant Professor of statistics with the Mathematics Department, Faculty of Science, Naresuan University, Thailand. Her main areas of research interest are Bayesian statistics, linear model, multivariate analysis, and biostatistics.

K. KESORN received the Ph.D. degree in electronic engineering from the Queen Mary College, University of London, U.K. He is currently an Associate Professor with the Department of Computer Science and Information Technology, Science Faculty, Naresuan University, Thailand. His current research interests include semantic multimedia retrieval, knowledge-based modeling for multimodal information retrieval, semantic data processing, and data mining in biomedical information. He has participated in several national research projects and has been a reviewer for many world-class journals from various publisher, such as the IEEE, ACM, Springer, and Elsevier. He has been receiving research grants from the Thailand Research Fund, the Office of Higher Education Commission, and the Science Faculty, Naresuan University.