A Survey on Prediction and Detection of Epidemic Diseases Outbreaks

Ravi Kumar Suggala
Research Scholar,
Department of Computer Science and Engineering,
Centurion University of Technology and Management, Paralakhemundi, Odisha

Dr. Anurag
Pro Vice Chancellor,
Department of Computer Science and Engineering,
Centurion University of Technology and Management, Paralakhemundi, Odisha

Abstract- These days, the web can be utilized for diseases surveillance. In daily lives, early seasonal epidemics prediction like malaria, influenza may diminish their effect. In world the subtropical and tropical regions, the dengue outbreaks are endemic mainly in sub-urban and urban areas. One of the fundamental ten infections affecting the most deaths worldwide is the outbreak. The fundamental aim of this paper is to show a strategy to research and forecast the epidemic diseases spreading practices before they happen. In densely populated area, the various instances of epidemics widespread outbreaks are reported. These methodologies can limit outbreak to a little restricted region. Therefore, to decrease losses in human death form this would guarantee a superior way of dealing with stress is provided study the diseases spread and sufficient control mechanisms. Through this paper, we develop a prediction model that can expect the individual likeliness is influenced by a specific epidemic through surveying of early manifestations.

Keywords-Bio-surveillance, disease forecast, epidemic diseases, pathogen detection, tropical region, subtropical region.

I. INTRODUCTION

In world, the epidemics cause significant economic, health and the social impacts [1, 2]. The contagious diseases are the epidemic a disease [2, 3], when measurement of contagion entered to the level of outbreak, then to the whole nation it is possible to distribute [4]. To the entire population the epidemics diseases can manage to wipe out [5]. Some of the famous epidemic outbreaks are cholera, influenza, yellow fever, dengue, bird flu, and diphtheria, in the entire world these were manifested [6]. One of the aspects caused 43% in life lost globally is contagious disease and it cause critical world health issues. In current and past era India has noticed re-emerging disease outbreaks and emerging disease in diverse sectors of country. So, the data from various sectors needs to be analyzed to implement these measures. For instance, climatic, demographic, hospital records, queries of web search, social media content, etc, are used for getting the present medical scenario (disease information) along with past medical data for investigation. Figure 1 exhibits the analysis of traditional and modern data sources. Diphtheria may cause [8]. Usually, via contaminated air and direct contact the transmission were happened. However, by the irregular of vaccination, these diseases were caused. Towards prediction, detection, controlling, and identification via
classification of features, these factors may contribute. Also, between the features due to the integration involvement, the features’ clustering technique was presented and causes the adjustment of weight to the input. The result of features’ clustering method is approximate detection, prediction and less accurate on epidemic diseases.

To alleviate the negative impacts of pandemics, the infectious diseases granular and reliable forecasting facilitates targeted resource distribution and intervention planning. To forecast seasonal as well as emerging disease outbreaks there have been multiple efforts. The recent efforts relied on mathematical compartment models such as SIR [9] and SEIR. By considering the individual-level interactions across the social networks some studies improved model fidelity [10]. Rising the unknown parameters number, through data assimilation that should be adjusted however, on computational resources these individual-based models impose an additional burden. Furthermore, the accuracy performance of the epidemic model can drastically affected by the choice of calibration scheme. Thus in epidemic forecasting contexts a diversity of filtering approaches have been utilized [11, 12].

Most of the research works utilize the social communication sites like Twitter and Face Book. Because of few reasons it is a proficient asset to track patterns. High recurrence of presented messages at first, aides to perform investigation in minute-by-minute. For public the Twitter posts are accessible and more descriptive while contrasting with search engine logs. Additionally, more analysis is performed for client's profiles dissecting like demographic data [13]. The focus of this research is to survey the current procedures, methodologies, tools and frameworks of predicting the epidemic illnesses in machine learning strategies, data mining, time serious analysis, hybrid models and the social media data. Studied techniques assess various posts related to epidemic diseases for faster detection in an effort to maintain and accomplish healthier networks. Rest of the paper is sorted as follows: Section 1 includes the introduction part. Section 2 discusses the various predictions of epidemic outbreaks techniques. Finally, section 3 concludes the paper.
II. EPIDEMIC OUTBREAKS PREDICTION METHODS

There are numerous approaches to discover learning and predict epidemic disease. This section glances at different existing techniques. The surveys were selected to incorporate the current procedures and methods applied to different information for prior disease outbreak prediction. Examined techniques and methods are within the past recent years that fall in categories of data mining, time analysis, hybrid models, social media data and machine learning techniques. Figure 2 shows the various epidemic outbreaks prediction methods.

![Epidemic outbreaks prediction methods](image)

2.1. Prediction of Epidemic Outbreaks Using Machine Learning Techniques

In machine learning, several techniques like native Bayes, SVM, Neural network, random forest are used for prediction of epidemic diseases. Method for supervised learning is SVM (Support Vector Machine). The objective of flu related posts classifications the SVM is most commonly used [14-16].

Nadejda Lupolova et al. [14] have applied the SVM machine learning procedure to anticipate the bacterial (bovine E. coli O157) zoonotic potential. For human health more probable recognizing cow’s strains is danger. In addition, it was conceivable to produce a probability of host affiliation that show that only a minor (0.9) by utilizing an SVM classifier. At the point when tested across datasets the predictive limit was held. Public health management of illness the authors finding have profound implications because of dairy cattle mediations, for example, vaccination. Further, pathogen biology understanding the learning methods should be used broadly.

Using twitter posts, to distinguish influenza like disease in Portugal Santos et al. [15] applied SVM-based classifier. A dataset with 2704 posts was physically explained with 650 printed highlights to test and preparing. To prepare the classifier a set of clarified dataset was used. Utilizing feature choice process and Bag-of-Words feature...
representation, the classifier was actualized. In light of Mutual Information (MI) esteem it is utilized to elect the top highlights set.

In Beijing, from the networking sites (Chinese) to foresee influenza patterns Yang et al. [16] have applied a SVM-based strategy. From the sites of Chinese social networking the authors privilege that it is the main investigation to anticipate flu trend. For classification purposes the word weighting types such as Term Frequency (TF) weighting, TFIDF (Term Frequency-Inverted Document Frequency), Boolean and Inverted Document Frequency (IDF) weighting were considered by the TFIDF strategy.

Gregory F. Cooper et al. [17] have portrays a Bayesian framework, in a population that joins clinical diagnosis of individuals, in populace the epidemiological demonstrating of disease outbreaks. To build up a framework the authors applied this structure distinguishes the outbreaks of influenza detection from ED (Emergency Department) reports. Utilizing both the genuine information from outbreak and simulated data the authors assessed their strategy.

In view of artificial Beijing the EVD process of propagation and its intrusion were simulated by Peng Zhang et al. [18].

To demonstrate the interaction between the occurrence probability of 4 various infections and bio-security rehearses on Canadian Swine Farms Ruth Cox et al. [19] have made a apparatus based on BBN (Bayesian Belief Network). The authors utilized two datasets to construct the BBN. First datasets clarifies the bio security practices utilized in swine farms and the second datasets explains the disease occurrence and the animal health status. Also, to appraise the risks of disease outbreak at various spatial scales Yilan Liao et al. [20] have presented a BBN technique, in view of virus detection rates. Simulation result shows the presented technique is specific than other conventional techniques. If some data were inaccessible the displayed method makes vulnerability estimates.

2.2. Prediction of Epidemic Outbreaks Using Data Mining

For the dengue outbreak prediction Anna L Buczak et al. [21] have introduced a novel prediction strategy using FARM (Fuzzy Association Rule Mining). The acquainted prediction method was utilized to concentrate between climatic information, meteorological information, socio-political information, and clinical data from Peru and these connections were in the rules form. The authors automatically chose the best set of rules and then form the classifier.

Future the incidence of dengue was predicted as no outbreak (LOW) or outbreak (HIGH) the classifier was utilized.

In the Philippines, for predicting the incidence of dengue the author’s et al. [22] have presented FARM methods to excerpt affiliation rules from historical environmental, climate data and socio-economic data, and epidemiological indicating the patterns of future weather. The presented model predict low/high dengue incidence, proceed in four weeks in Philippines. To historical data incidence the threshold between low/high were resolved.

Linking the Bio-PEPA modeling with data mining techniques tested to mumps was exhibited by Dalila Hamami et al. [23], in both model parameter optimization and structure identification. In the information suggesting areas the authors recognize the tough features by preceding association rule learning with the clustering. The introduced model suggested the vaccination, seasonality and age with correlated features were included in this method.

For the HPAI (Highly Pathogenic Avian Influenza) analysis utilizing livestock disease information gathered over an extended period, Zhenshun Xu et al. [24] have introduced a data mining techniques with a model of data cube. Based on data cube model construction, to assess the spatial and temporal perspectives of the disease spread.
utilizing OLAP (Online Analytical Processing) actions with varying abstraction levels a multidimensional analysis of HPAI was performed. By applying sequential pattern mining and association rule mining the presented method provides useful data that achieves routes for potential sequential dissemination and site connectedness of HPAI outbreaks.

2.3. Prediction of Epidemic Outbreaks Using Time Series Analysis

The data gathered over particular time with internal structures like seasonal variety and auto correlation that is dissected to recognize the different patterns is called time series analysis. At different granularities levels the clinical information is regularly the time series gathered information. Further predictions are made based on these data parameters.

From C difficile infections to investigate ecological elements of clinical burdens Lawes et al. [25] presented a quasi-experimental study and non-linear time-series analysis. In all populaces a blended the intervention of persuasive-restrictive 4C antibiotic stewardship was initiated. A prediction model for DHF (Dengue Hemorrhage Fever ) and the dengue fever using Seasonal Auto Regression Integrated Moving Average (SARIMA) scheme was developed by Gupta et al. [26], for year 2011 to predict monthly DF/DHF occurrence. The developed model was an extension of ARIMA series in which over a time the pattern repeats seasonally. Their model was given as SARIM (P, D, and Q). With estimating criteria like MAPE (Mean Absolute Percentage Error), stationery R-squared and BIC (Bayesian Information Criteria) the authors used SARIMA(0,0,1)(0,1,1) technique in Rajasthan for predicting dengue fever.

By accumulating month occurrence data as revealed by nation -level surveillance system the authors [27] have presented the database for global first bacterial meningitis. Computing the gravity centre distribution of cases the authors supposed the activity of mean timing of disease. For bacterial meningitis most cases the authors introduced the synchrony exists in between pathogens. Their discoveries provide key insight to bacterial meningitis seasonal dynamics and include knowledge about host and meningitis global epidemiology and the earth.

In Nanjing, China to build up a climate -based model of forecasting for HFMD (Hand, Foot, and Mouth Disease) utilizing the HFMD surveillance information and climatic variables Sijun Liu et al. [28] have developed an Multivariate SARIMA (Seasonal Autoregressive Integrated Moving Average) technique. To approve the execution of prediction the authors used weekly data prediction model. Their results show 80% of sensitivity and 96.63% of specificity with high agreement rate between the observed values and the predicted values.

2.4. Prediction of Epidemic Outbreaks Using Hybrid Models

Sudheer et al. [29] have presented a novel technique dependent on coupling the SVM and Firefly algorithm (FFA) to estimate the malaria occurrences. SVM models execution relies upon the appropriate SVM parameters choice. In their work for decide the SVM parameters, the FFA have been utilized. The malarial incidences in Jodhpur and Bikaner area were anticipated by embracing the SVM-FFA model. The input variables are monthly averages of malarial incidences, temperature, rainfall and relative humidity. From primary health centers, monthly notifications of malaria cases time series have been obtained. Further, fro contrasted with other traditional techniques the outcome demonstrates that the SVM-FFA model provides more precise estimates.
Massimo Buscema et al. [30] have developed a TWC method (Topological Weighted Centroid) for distinguishing the sources of epidemic with German Escherichia coli outbreak application. To a real world example these novel mathematical tools was tested to point out the real source of outbreak. As indicated by four independent indexes such as sensitivity, searching area, specificity and peak distance the result of TWC algorithm were predominant in comparison with different techniques like LVM, NES, Mexican Prob, and Rossmo Algorithm.

In the regards of stochastic nonlinear filtering, Alex Skvortsov and Branko Ristic [31] have displayed a scheme for epidemic outbreak syndromic surveillance. With inhomogeneous mixing utilizing the model of epidemiological stochastic compartmental, the epidemic dynamics was modeled. In order to avert the shortages of asthma health care resource in progress Li Luo et al. [32] have introduced a search index, that was linked with weather, historical admissions and pollution data using the techniques of machine learning. Simulation result shows best area under curve in test set of 0.832 was accomplished by utilizing their technique.

2.5 Prediction of Epidemic Outbreaks Using Climate Factors

In French Guiana Antoine Adde et al. [33] have assessed the potential of utilizing atmospheric and oceanic conditions to predict the outbreaks of dengue fever. To recognize the climatic conditions the composite studies and diminished correlations were performed. At that point to build a forecast model a logistic regression was performed. Based on the temperatures of surface of Pacific Ocean Sea the author’s exhibit the model and it predict 80% of outbreaks in Azores High sea-level pressure. m meteorological records relative humidity, rainfall and temperature have been obtained. At that point when At regional scales and within the districts to investigate correlation maps, Andres Baeza et al. [34] have introduced a remote sensing data for the NDVI (Normalized Difference Vegetation Index). The introduced NDVI methods were used as a rainfall integrated measures. Analysis addresses the pairing between the climate variability and malaria incidence whether the irrigation has diminished. Also the introduced method reflects NDVI breakdown as useful risk indicator and increase in malaria transmission control.

2.6 Prediction of Epidemic Outbreaks Using Social Media

These days, from a huge range of web and social media sources the large amounts of health and emergency data are increasingly coming. Over the recent years, this data is useful for earlier detection of outbreak and disease surveillance, and few public web surveillance projects in this field have developed.

For monitoring and model the influenza activity in China, Q. Yuan et al. [35] have presented a technique in Baidu that utilized an data from internet search query. Main objective of their research was to present the comprehensive technique for keyword filtering and selection, influenza activity modeling and detection in China and composition of index. With Chinese influenza data case the sequential time series for elected composite index was correlated significantly. To determine the twitter usage as actual time method, Michelle Odlum et al. [36] have used Twitter of Ebola outbreak detection. To detect the similarities content analysis was conducted and develops clusters. On the tweet corpus trend analysis was performed. To build the clusters the presented method utilized the k-means algorithm.

In Google search volume epidemic-associated to predict the disease trends, X. Zhou et al. [37] have introduced a syndromic surveillance framework. To lessen the noise level with multiple alert levels the authors initially indicated the epidemic trend. Thus to monitor both regional and national levels of epidemic alert (U.S) the respective models
were built. For crime prediction Matthew S. Gerber [38] have presented the research by the spatiotemporally tagged tweets usage. To naturally recognize analysis fields across a major city in US the authors use statistical topic modeling and Twitter-specific linguistic analysis. The summary of the reviewed methods and techniques is shown in table 1.

Table 1: Summary of the Reviewed Methods and Techniques

<table>
<thead>
<tr>
<th>Method category</th>
<th>Author</th>
<th>Technique name</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning</td>
<td>Nadejda Lupolova et al. [14]</td>
<td>SVM</td>
<td>Predict bacterial (bovine E. coli O157) zoonotic potential</td>
</tr>
<tr>
<td></td>
<td>Carlos Santos et al. [15]</td>
<td>SVM based classifier</td>
<td>Detect flu-like illness in Portugal using twitter posts</td>
</tr>
<tr>
<td></td>
<td>Yang et al. [16]</td>
<td>SVM</td>
<td>Flu trend were predicted from Chinese social networking sites</td>
</tr>
<tr>
<td></td>
<td>Gregory F. Cooper et al. [17]</td>
<td>Bayesian framework</td>
<td>Detects the outbreaks of influenza from Emergency Department reports</td>
</tr>
<tr>
<td></td>
<td>Peng Zhang et al. [18]</td>
<td>Geodemographics and large scale machine learning</td>
<td>Analyze the propagation mechanism of Ebola virus disease</td>
</tr>
<tr>
<td></td>
<td>Ruth Cox et al. [19]</td>
<td>Bayesian Belief Network</td>
<td>Model the occurrence of four different diseases and bio-security practices on Canadian swine farms</td>
</tr>
<tr>
<td></td>
<td>Yilan Liao et al. [20]</td>
<td>BBN with virus detection rates</td>
<td>Assess disease outbreak risks at different spatial scales</td>
</tr>
<tr>
<td>Data mining</td>
<td>Anna L. Buczak et al. [21]</td>
<td>Fuzzy Association Rule Mining</td>
<td>Predict future dengue incidence either as outbreak or no outbreak</td>
</tr>
<tr>
<td></td>
<td>Buczak et al. [22]</td>
<td>FARM methods with rules extraction</td>
<td>Predict low/high dengue incidence within 4- weeks in Philippines</td>
</tr>
<tr>
<td></td>
<td>Dalila Hamami et al. [23]</td>
<td>Bio-PEPA modeling with data mining</td>
<td>Applied in both model parameter optimization and structure identification</td>
</tr>
<tr>
<td></td>
<td>Zhenshun Xu et al. [24]</td>
<td>data mining techniques with data cube scheme</td>
<td>Provides useful information for HPAI outbreaks</td>
</tr>
<tr>
<td>Time serious analysis</td>
<td>Lawes et al. [25]</td>
<td>Quasi and non-linear time serious analysis</td>
<td>Initiated Mixed Persuasive-Restrictive 4C antibiotic</td>
</tr>
<tr>
<td>Hybrid models</td>
<td>Climates factors</td>
<td>Social media</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Gupta et al. [26]</strong></td>
<td><strong>Adde et al. [33]</strong></td>
<td><strong>Q. Yuan et al. [35]</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SARIMA</strong></td>
<td><strong>Atmospheric and oceanic conditions</strong></td>
<td><strong>Internet search query data</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SARIMA</strong></td>
<td><strong>remote sensing data for NDVI methods</strong></td>
<td><strong>Twitter posts</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SARIMA technique</strong></td>
<td><strong>Syndromic-like observations algorithm</strong></td>
<td><strong>Detect and model the influenza activity in China</strong></td>
<td></td>
</tr>
<tr>
<td><strong>For HFMD Develop a forecasting model based on weather with 80% of sensitivity and 96.63% of specificity</strong></td>
<td><strong>Identify epidemic sources with German Escherichia coli outbreak application and measured sensitivity, searching area, specificity and peak distance.</strong></td>
<td><strong>Build clusters with k-means algorithm for Ebola outbreak detection</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Sijun Liu et al. [28]</strong></td>
<td><strong>Search index combined with environmental data and machine learning</strong></td>
<td><strong>disease-related Google search volume</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SARIMA</strong></td>
<td><strong>For estimation and early outbreak prediction of epidemic</strong></td>
<td><strong>Design syndromic surveillance system to predict the epidemic</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Paireau [27]</strong></td>
<td><strong>Topological Weighted Centroid algorithm</strong></td>
<td><strong>For estimation and early outbreak prediction of epidemic</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SARIMA technique</strong></td>
<td><strong>For HFMD Develop a forecasting model based on weather with 80% of sensitivity and 96.63% of specificity</strong></td>
<td><strong>For estimation and early outbreak prediction of epidemic</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Skvortsov and Branko Ristic [31]</strong></td>
<td><strong>Search index combined with environmental data and machine learning</strong></td>
<td><strong>For estimation and early outbreak prediction of epidemic</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Li Luo et al. [32]</strong></td>
<td><strong>SVM-FFA</strong></td>
<td><strong>Hybrid models</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Malarial incidences in Jodhpur and Bikaner area were predicted.</strong></td>
<td><strong>Malaria incidences in Jodhpur and Bikaner area were predicted.</strong></td>
<td><strong>Stewardship intervention</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Massimo Buscema et al. [30]</strong></td>
<td><strong>Identify epidemic sources with German Escherichia coli outbreak application and measured sensitivity, searching area, specificity and peak distance.</strong></td>
<td><strong>Stewardship intervention</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Measure MAPE, BIC and stationery R-squared technique for dengue fever prediction in Rajasthan**

**Developed the first bacterial meningitis global database**

**In advance avoid asthma resource shortages and result show best area under curve in test set of 0.832 was achieved**

**Increase in malaria transmission control**

**Detect and model the influenza activity in China**

**Build clusters with k-means algorithm for Ebola outbreak detection**

**Design syndromic surveillance system to predict the epidemic**
3. CONCLUSION

This paper analyses the various approaches and schemes that is utilized in epidemic outbreak prediction model. It can be concluded that, these models are based only on the clinical disease data. By taking social, political, geological and climatic data the presented models can be improved and yield more accurate forecasts. The information exchange became much fast and efficient while communicating through social channel increases. Through various mining techniques these channels can be tapped and obtained the real time feedback along with more accurate outbreak detection. So, an enhancement to the prediction model along with machine learning technique can be proposed as a future work.

REFERENCES


