DISEASE SPREAD PREDICTION USING MACHINE LEARNING

HARSHA POOLLA[1], M.KAVYA[2], M.BHARGAVI[3]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2 & 3]

DEPARTMENT OF CSE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN,MAISAMMAGUDA, DHULAPALLY KOMPALLY, MEDCHAL RD, M, SECUNDERABAD, TELANGANA 500100

ABSTRACT. Since 1968, Dengue Harmonic Fever's incidence in Indonesia has continued to rise and has become a public health issue. Indonesia has the largest number of Dengue Harmonic Fever cases than 30 other epidemic countries worldwide. It is very important to carry out research related to dengue cases' prediction to prevent the spread of Dengue. This literature review is intended to determine the extent of the dengue prediction approach carried out by previous researchers, and a research gap will be obtained. The algorithm used to cluster articles is a modularity algorithm, using several open-source tools to process data. The online databases used are Google Scholar and Crossref by using keywords: journal, algorithm, prediction, and Dengue. The data are taken from the expansion of 1928-2020. This study's results are 200 articles that are suitable and divided into four clusters of important articles. Also, several important parameters were obtained in the prediction study of dengue fever, namely humidity, temperature, rainfall, and population density.

INTRODUCTION

The Covid-19 pandemic has become an intensive concern nationally and globally. However, Dengue Harmonic Fever (DHF) in Indonesia can't ignore as it becomes a public health problem and caused death every year. The DHF disease because of the Aedes aegypti mosquito and Aedes albopictus mosquito. Both of them are female and are capable of infected children and adults [1][2]. In Indonesia, DHF is a severe case since 1968, and every year it always increases. It makes Indonesia becomes the country with the highest DHF death cases in Southeast Asia [3]. In recent years, Dengue Harmonic Fever has infected all tropical and non-tropical countries worldwide and has become one of the leading public health problems in many countries [4]. In 2014, DHF broke out in Japan with 160 confirmed cases. In European countries such as Italy, France, Tanzania, and Medina in the same year, there was also an increase in dengue

cases [5]. In 2017, Sri Lanka had 186.000 DHF issues with 440 death of problems [6]. According to World Health Organization, Indonesia is the highest country for Dengue Harmonic Fever cases than 30 epidemic countries worldwide. In 2020, there were 71.633 cases reported in Indonesia. Thus, the DFH cases prediction model is vital to help the government take precautionary measures before real cases occur. The accuracy of the DHF cases prediction model will help (1) Carrying out early alertness for the explosion of DHF cases at a certain period, and they can make preparations for dealing with DHF cases (2) To control and prevent DHF cases during a Covid-19 pandemic, (3) Assisting medical officers and government or non-government agencies to take preventive measures before the occurs outbreak. Based on the description above, the research purpose in this article is to get the DHF prediction model citation review using a social network to map the DHF prediction model research area and to get the DHF research gap in the prediction model.

RELATED WORKS The systematic review is one way of mapping the research to be carried out. Many researchers conduct periodic reviews to mapping research gaps. Some studies that do systematic reviews are [7] use the Institute of Electrical and Electronics Engineers (IEEE) explore for mapping Dengue prediction literature review. The data article was taken from August 2014 - January 2015, and the result is 96 articles were Cluster into five areas. Dengue Fever (DF) epidemiology, DF Forecasting model, DF classification, and spatial model visualization. Paper [8] uses scientific publication data from Web of Science Core Collection - articles indexed in Science Citation Index Expanded (SCI-EXPANDED). The data are taken from 1945 – 2014. The study is searching by a combination of title, author, abstract, and author affiliation, including city and country of origin. The article results are normalized using fuzzy logic. Paper [9] uses Pubmed, Cochrane Library, ScienceDirect, and Herdin. The data was taken from 1 January 1958 – 31 December 2017. The articles checklist uses PRISMA and the result analysis based on objectives, methods, results, descriptive epidemiologic, and case reports. In this study, two article databases were compared, namely Crossref and google scholar. And the comparison results are clustered by a modular algorithm. The articles are taken from 1928 until 2020, the proposed workflow show in Figure 1

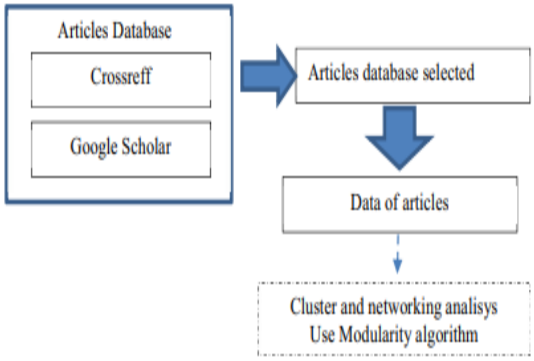

Figure 1. Proposed workflow

EXISTING SYSTEM

- For regression linear regression, support vector regressor were used.
- As the data was non-linear in nature linear regression didn't perform well by just scoring an accuracy of 48.49% with mean absolute error and mean percentage error being 190.04535 and -0.69568688 respectively.
- Support vector regressor performed well by fitting a polynomial curve and scoring an accuracy of 90.63% with mean absolute error and mean percentage error being 103.48891 and -0.38543662.

DISADVANTAGES

The existing system has more mean absolute error and mean percentage error.

PROPOSED SYSTEM

- The first step of this model describes data preprocessing which is an essential part and the performance of the algorithm heavily depends upon data preprocessing.
- In the second step, the prediction model proposing a novel twofold linear regression model to solve this problem which outperforms compared to all previous models.
- This model will achieve less mean absolute error which is minimum as compare to traditional machine learning techniques.

ADVANTAGE

- A novel twofold linear regression model is used to overcome the problem of mean absolute error.
- It has less mean absolute error compare to existing technique

MATERIAL AND METODOLOGY

Data Collection The data articles were taken from Crossref and google scholar. All items through several online databases, The Association for Computing Machinery (ACM) digital library, Elsevier, Institute of Electrical and Electronics Engineers (IEEE) explore, and Science Direct. **Methodology**

The study does in four methodologies. The first step is comparing all article data sources. The author compares four data sources that are SCOPUS, Google Scholar, Web of Science, and Crossref. Based on the comparison, Crossref was considered more effective because all articles on Crossref are included on several online databases. The papers were collected using the free open sources software. The second step, taking all articles relevant and related to DHF model predictions, uses some keywords: approach, parameters, and prediction model. The third step is database analysis using social network analysis with modularity algorithm proposed and using open-source network analysis software for taken the graph. The networking analysis graph is developed by connecting all keywords included in some articles. And the last step is clustering important articles into some clusters.

Main path article In this part, the principle of path analysis is identifying the process using a citation network model. The first step is to take the metadata from Crossref with open-source software. The Crossref is a choice because it is free to access, and some articles publish in some journals, including in it. Each item will extract into a network note where connect in the next step. The Network analysis develops with the following steps: (1) Article retrieval based on entering the paper title, keywords, publication journal name, and year of publication, (2) Storing article metadata in the form of a reference manager file and, (3) Calculating the amount of traversal for each quote on the network. The number of traversals is the frequency with which a particular quote appears on different paths in several articles, both the source article and the one cited by others. The traversal count value significantly indicates how many links with other papers [10].

Clustering of the article analysis The article clustering analysis is doing by identify groups of similar articles and specific research domains. Some of the items have the same research topic. The article clustering develops by the modularity algorithm proposed. This algorithm is carried out in two steps: iterating the notes on the graph and assigning each letter to it by increasing modularity to lead to its group. The second step is to create the supernotes from the Cluster in step one, and this process always repeats so that this algorithm relies on efficient and effective heuristics. VOSviewer uses a modularity algorithm to build article clustering, which shows on Equation 1 [10] [11] [12]:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

Where Q is the fraction of edges that fall within group 1 or 2, Ai j is the weight of the boundary between nodes i and j. ki and kj the sum of the consequences of the advantages attached to nodes i and j, respectively; 2m indicates the total sum of all the edge weights in the graph; ci and cj are the communities of the nodes i and j, and δ is a simple delta function.

CONCLUSION The results of the comparison and grouping of articles using this network analysis technique obtained several things, among others, (1) several parameters in the prediction study of DHF were temperature, rainfall, number of cases in recent years, population density, and land use, (2) There are four crucial clusters of articles in research and, (3) The research area that is still very interesting to discuss is prediction using a machine learning approach. The development of a dengue prediction model using machine learning in research is very recommend. Machine learning approach is considered simpler and more practical. So it is expected to be able to reduce the number of dengue fever spread in Indonesia. Future research will compare several prediction models using a machine learning approach to evaluate model predictions' accuracy.

REFERENCES

[1] G. J. Ebrahim, "Dengue and dengue hemorrhagic fever," J. Trop. Pediatr., vol. 39, no. 5, pp. 262– 263, 1993, doi: 10.1093/trope/39.5.262.

[2] T. Chakraborty, S. Chattopadhyay, and I. Ghosh, "Forecasting dengue epidemics using a hybrid methodology," Phys. A Stat. Mech. its Appl., vol. 527, p. 121266, 2019, doi: 10.1016/j.physa.2019.121266.

[3] W. Anggraeni et al., "Modified Regression Approach for Predicting Number of Dengue Fever Incidents in Malang Indonesia," Procedia Comput. Sci., vol. 124, pp. 142–150, 2017, doi: 10.1016/j.procs.2017.12.140.

[4] "Treatment, prevention, and global control strategy for dengue prevention and control 2."

[5] S. Polwiang, "Estimation of dengue infection for travelers in Thailand," Travel Med. Infect. Dis., vol. 14, no. 4, pp. 398–406, 2016, doi: 10.1016/j.tmaid.2016.06.002.

[6] V. S. Aryaprema and R. De Xue, "Breteau index as a promising early warning

signal for dengue fever outbreaks in the Colombo District, Sri Lanka," Acta Trop., vol. 199, no. August, p. 105155, 2019, doi: 10.1016/j.actatropica.2019.105155.

[7] A. Munir and A. Sari, “Sistematic Review: Model peramalan wabah penyakit demam berdarah,” Semin. Nas. Apl. Teknol. Inf. 2015, pp. 117–124, 2015.

[8] F. H. Nieto, "Ex post and ex-ante prediction of unobserved multivariate time series: a structuralmodel based approach," Journal of Forecasting, vol. 26, no. 1. Wiley, pp. 53–76, 2007, doi: 10.1002/for.1017.

[9] K. A. Agrupis, M. Slade, J. Aldaba, A. L. Lopez, and J. Deen, "Trends in dengue research in the Philippines: A systematic review," PLoS Negl. Trop. Dis., vol. 13, no. 4, pp. 1–18, 2019, doi: 10.1371/journal.and.0007280.

[10] A. Sivaprasad, N. S. Beevi, and T. K. Manojkumar, "Dengue and Early Warning Systems: A review based on Social Network Analysis," Procedia Comput. Sci., vol. 171, no. 2019, pp. 253– 262, 2020, doi: 10.1016/j.procs.2020.04.027.