The Analysis of Determining Cost of Products and Forecasting Dengue Fever Hemorrhagic Incidents: A Machine Learning Approach

Tien Rahayu Tulili[1,*] Yohanes K Windi[2] Bambang Cahyono[1] Damar Nurcahyono[1]
Karyo Budi Utomo[1] Ahmad Rofiq Hakim[1]

[1]*Program Studi Teknologi Rekayasa Komputer, Politeknik Negeri Samarinda, 75116, East Kalimantan-Indonesia*
[2]*Program Studi Kesehatan Masyarakat, Poltekkes Surabaya, 405016, East Java-Indonesia*
[*]*Corresponding author. Email: tien.tulili@polnes.ac.id*

ABSTRACT
Dengue is a viral infection transmitted by Aedes mosquitos. This disease mostly spread in the tropical and sub-tropical countries and according to WHO, the dengue outbreaks has increased 30-fold over the last five decades. The disease is still an ongoing burden of throughout the world. In Indonesia, for example, the incident of dengue hemorrhagic fever (DHF) has shown up 8,056 cases spread in the last five years. One of the ways to help the government to mitigate any possible of the spread is by utilizing a nearly accurate forecast system in predicting the cases. This study aims to employ machine learning methods in predicting the cases occurred in East Kalimantan. Various kinds of data (such as climate, demographical and epidemiological data) are used in developing some machine learning models. Furthermore, identifying variables prior the models' development is done to achieve the best model of prediction; furthermore, a comparative study of the models built is discussed. Monthly dengue cases, incidence rate (IR), climate factors (rainfall, atmospheric pressure, the duration of the sun) and socio-economic conditions (population density, the number of inhabitants) from three different cities/districts (Samarinda, Balikpapan, and Berau) in East Kalimantan from 2007-2019 are gathered. Prior machine learning's modeling, all data are analyzed with Pearson Correlation method to identify which variables has a positive correlation with DHF cases. Several machine learning algorithms, those are: Neural Network, Deep Learning, Generalized Linear Model, Generated Boast Tree and KNN, implemented in the modelling and forecasting. The results showed that most climatic factors are negatively correlated to DHF cases in East Kalimatan. Furthermore, the selection of variables leveraged the performance of the models.

Keywords: *Forecast, Dengue Hemmorhagic Fever, Machine Learning, Deep Learning, Neural Network, Generalized Linear Model, And KNN.*

1. INTRODUCTION

Dengue is a viral infection transmitted by Aedes mosquitos. This disease mostly spread in the tropical and sub-tropical countries and according to WHO, the dengue outbreaks has increased 30-fold over the last five decades. Moreover, it is estimated that the new cases reach between 50 to 100 infections every year in more than 100 endemic countries. Therefore, the disease is still an ongoing burden of throughout the world[1]. Indonesia, which is geographically located in tropical area, has shown significant number of the dengue incidents particularly dengue hemorrhagic fever (DHF). The incidence rate is the highest among the Southeast Asia countries[2]. In the last five years, the incident of the disease has shown up 8,056 cases spread in Indonesia[3].

According to General Directorate for Disease Prevention and Control of Health Ministry of Republik Indonesia, data on July 2020 showed that the number of incidents reached to 71,633 [4]. Actions have been taken by the governments for years to decrease the number of the cases and the death, such as burying, draining, closing, and JUMANTIK program - Juru Pemantau Jentik (a person pointed to monitor mosquito larvae in each house). These actions have affected to the number of the cases particularly from 2017-2018. However, in

Copyright © 2021 The Authors. Published by Atlantis Press International B.V.
This is an open access article distributed under the CC BY-NC 4.0 license -http://creativecommons.org/licenses/by-nc/4.0/.

January 2019, there was 13,683 cases and this number tended to increase in the following few months. In one study found that for over five decades the trend of DHF incident tends to increase[5] in which the incidence peaks occurred in 1973, 1988, 1998, 2009, and 2016.

East Kalimantan as one of provinces in Indonesia reported the highest DHF incidence rate in 2017[6]. The incidents of DHF in April 2020 reached 1183 cases [7]. It cannot be denied that DHF cases is still a main problem in Indonesia, particularly in East Kalimantan.

Studies about infectious diseases and models of predicting the spreading of the diseases have been done. Some various data used in the studies, for example, geographical data[8][9][10], search query[11] and social media data[9][12][13], socio-economic data[14][15][16], entomological data[17], and official surveillance data[15]. Several studies to predict several infectious diseases by optimizing parameters involving clinical data such as surveillance data and non-clinical data such as temperature and humidity [18][10][16]. Some studies involved climate variables in the predictions [18][11][10][19]. As far as our knowledge that optimization in the preprocessing process is still a lack of interest in most of the research. This research will optimize the preprocessing process by adding and modifying variables positively related to DHF cases.

Therefore, in this study the various kinds of data are used such as climate and demographic conditions. The weather conditions include temperature, humidity, wind speed, rainfall, the duration of the sun. These variables used because of its unpredicted values and variety. Furthermore, the demographic conditions involve population density (per km2), the number of the inhabitants, land area, the number of outbreaks of DHF as well as the incidence rate (IR) of DHF. The climate, socio-economic, and epidemiological data gathered from three major cities-those are Samarinda, Balikpapan, and Berau - in East Kalimantan for over ten years from 2007 to 2019.

The findings of this study will help stakeholders, such as health ministry, hospitals, public health center, health research center, and environmental center; to have better preparation and adequate intervention (prevention) before the DHF spread in wider areas; and to reduce cost and hospitalization.

2. METHOD

Our study methodology can be seen on Figure 1. Firstly, all data taken from three different sources are preprocessed to remove all invalid data. Next, we investigate the correlation between variables particularly correlation each variable to DHF cases. We utilized Pearson matrix correlation as Pearson correlation recommended to implemented to see the correlation between numerical data. In the preprocessing phase, we implement two scenarios. In the first scenario, we take all variables but ignoring the IR; meanwhile in the second scenario, we propose new variables, that is IR and modified IR. The modified IR is the monthly one-year lag of IR incident. In the prediction modelling, data from 2007-2017 utilized as training data (three cities combination) meanwhile data from 2018-2019 utilized as testing data. We implement Z-transformation to normalize data both in training and testing phase. We applied five unsupervised numerical prediction of machine learning, those are Neural Network, Generalized Linear Model. Deep Learning, Gradient Boast Tree, and KNN. For each combination constructed, tuned, and predicted against the testing dataset was validated by the root mean squared (RMSE). This calculated to check the validity of the models and generally, the smaller RMSE value the better model predicted against unseen samples. The samples are categorized by cities.

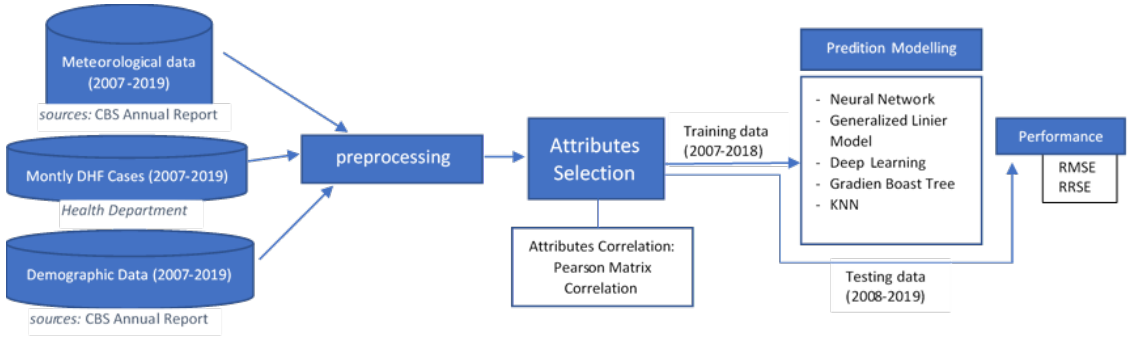

Figure 1 Research Methodology

The datasets used are taken from three different sources. All epidemiological data are taken from the local department of Ministry of Health of East Kalimantan. This includes the total number of dengue incidents per year per month per municipality from 2007 to 2019. The cities involved are three major cities in East Kalimantan– those are Samarinda, Balikpapan, and Berau in which each area has its own meteorology station. Samarinda and Balikpapan are categorized as a dense population city. The population density of Samarinda and Balikpapan is

1238.9 and 1297.74 inhabitants/km2 respectively, while Berau is 10.96 inhabitants/km2 only[20]. The incident rate (IR) of dengue epidemics calculated in this work is used in the model. The IR is calculated as the number of monthly DHF cases divided by the number of inhabitants multiply by 100,000. Two demographical data associated with dengue incidences were included in this study: population density (per km2) and the number of inhabitants. All data are from the annual report of Indonesia Central Bureau of Statistics (CBS). Furthermore, six climate variables included are minimum temperature and maximum temperature(celcius), humidity (%), atmospheric pressure (Mbs), wind velocity(knot), rainfall(mm), the duration of sunshine(mm). Most data are retrieved from the annual report of CBS and the remaining was taken from the database of BMKG.

In this research with employed two kind of scenarios. First, a scenario in which the IR is ignored; second, it is the vice versa. The scenario from either first or second group is divided into three different sub scenarios, those are utilizing eleven variables, the selected variables, and climatic variables only. The selected variables are resulted from the recommendation of Pearson matrix correlation.

3. RESULT AND DISCUSSION

The correlation between each variable to DHF cases can be seen on the figure 2. The Pearson matrix correlation shows eleven attributes to associated to itself shown in value as 1. As can be seen from the figure 2, DHF Incidents positively correlated to minimum temperature, wind pressure (atmospheric pressure), number of inhabitants (number Of Population), population density, and incidence rate. Conversely, the incidence variable negatively associated with maximum temperature, humidity, wind velocity, rainfall, and sun duration. These positive-correlated variables were utilized in our scenarios.

Attributes	MinTemp...	MaxTemp...	Humidity	windPressure	windVelocity	rainfall	sunDuration	numberOfPopulation	populationDensity	DHFIncidents	incidenceRate
MinTemperature	1	-0.726	0.199	0.029	-0.335	0.078	0.206	0.173	0.148	0.070	0.003
MaxTemperature	-0.726	1	-0.192	0.004	0.301	-0.195	0.004	-0.075	-0.147	-0.092	-0.069
Humidity	0.199	-0.192	1	-0.000	-0.085	-0.043	0.040	-0.074	-0.072	-0.039	-0.010
windPressure	0.029	0.004	-0.000	1	0.012	-0.006	0.145	0.123	0.106	0.038	0.032
windVelocity	-0.335	0.301	-0.085	0.012	1	-0.058	0.306	-0.098	-0.031	-0.037	0.033
rainfall	0.078	-0.195	-0.043	-0.006	-0.058	1	-0.176	0.049	0.091	-0.011	-0.050
sunDuration	0.206	0.004	0.040	0.145	0.306	-0.176	1	-0.186	-0.164	-0.105	0.002
numberOfPopulation	0.173	-0.075	-0.074	0.123	-0.098	0.049	-0.186	1	0.936	0.426	0.035
populationDensity	0.148	-0.147	-0.072	0.106	-0.031	0.091	-0.164	0.936	1	0.462	0.083
DHFIncidents	0.070	-0.092	-0.039	0.038	-0.037	-0.011	-0.105	0.426	0.462	1	0.807
incidenceRate	0.003	-0.069	-0.010	0.032	0.033	-0.050	0.002	0.035	0.083	0.807	1

Figure 1 Correlation between variables with Pearson Correlation

Table 1. RMSE of Scenario 1

		DEEP LEARNING	GENERALIZED LINEAR MODEL	GENERATED BOASTED TREE	KNN	NN
		RMSE				
SCENARIO 1	SUB SCENARIO 1: ALL VARIABLES WITHOUT IR					
	SAMARINDA	88.764	101.875	115.32	**0.959**	0.938
	BALIKPAPAN	53.745	59.183	58.812	**0.840**	1.433
	BERAU	77.919	34.224	39.810	0.866	**0.778**
	SUB SCENARIO 2: CLIMATIC VARIABLES ONLY					
	SAMARINDA	106.071	102.195	111.865	**1.145**	1.363
	BALIKPAPAN	63.655	54.893	51.313	**1.365**	2.939
	BERAU	114.245	59.628	51.766	**0.978**	1.367
	SUB SCENARIO 3: FILTERED VARIABLES					
	SAMARINDA	**0.965**	98.557	112.28	0.973	0.994
	BALIKPAPAN	**0.920**	52.295	56.832	0.957	0.978
	BERAU	**0.900**	34.203	35.783	1.053	1.076

Models built has successfully predicted the DHF cases in three cities: Samarinda, Balikpapan, and Berau although between the two cities and other small rural municipal have large differences in term of number of inhabitants and population density. The best model was Neural Network in which the DHF cases predicted in three cities by utilizing eleven variables. Meanwhile, in scenario 2 sub scenario 3, Deep Learning has given best results for two cities, Balikpapan and Berau. The best result of DHF prediction in Samarinda is when utilizing all variables into the modelling phase with the same algorithm, Deep Learning.

The result of the first scenario can be seen on Table 1. The performance of each model showed in the table were evaluated by calculating RMSE. As can be seen from table 1 that Neural Network, k-NN, and Deep Learning has showed small RMSE value, that is in the range 0-2; however, all models has showed better performances while the variables were filtered, particularly the Deep Learning that performed better significantly than those on the two other sub scenarios. The climatic variables which were popular implemented in some studies [18][10][16], has little leverage in developing a model for prediction. The combination of different variables such as demographic data and climatic data has improved the model performance as showed in the Table 1 as well as Table 2 in sub scenario 1 and 2.

It can be seen from Table 2, our proposed optimization has showed the better performance of the model compared to those on the table 1 (shown by asterisk on the table 2). The adding of IR and modified IR has improved most of the models. IR and modified IR could be considered as important variables in building a predictive model.

Despite the low performances of two models, those are GLM and GBT, the variable combination still leveraged the performance of the models as can be seen in Table 1 and Table 2 on sub scenario 1 and 3.

Table 2. RMSE of Scenario 2 with IR

		DEEP LEARNING	GENERALIZED LINEAR MODEL	GENERATED BOASTED TREE	KNN	NN
SCENARIO 2	RMSE					
	SUB SCENARIO 1: ALL VARIABLES WITH IR					
	SAMARINDA	0.990	105.848	128.878	0.957	***0.930**
	BALIKPAPAN	***0.184**	61.944	86.675	0.934	0.843
	BERAU	***0.193**	37.873	33.33	0.938	0.918
	SUB SCENARIO 2: CLIMATIC VARIABLES ONLY					
	SAMARINDA	***0.957**	118.130	133.315	0.976	1.181
	BALIKPAPAN	***0.401**	76.076	91.885	0.968	1.324
	BERAU	***0.426**	35.859	33.330	0.986	1.139
	SCENARIO 3: FILTERED VARIABLES					
	SAMARINDA	0.930	111.016	128.882	0.935	***0.924**
	BALIKPAPAN	***0.168**	70.215	86.685	0.892	0.837
	BERAU	***0.160**	37.385	33.348	0.930	0.914

4. CONCLUSION

In this research, we have implemented five unsupervised machine learning algorithms to predict the DHF cases in two dense population municipals and one rural population municipal. The climate variables have little impact on the performance models, while the combination with other kind of data such as demographic data and surveillance data has improved the models' performance. It is important to note that the correlation between variables and the DHF cases is also leverage the performance of the models particularly of those that positively correlated to the DHF cases. The IR and modified IR utilized in our research has given significant results particularly while implementing in Deep Learning model. These variables are also could be considered as the important variables which might contribute to the model performance.

However, the monthly data utilized in the modelling and testing phase can only be used for monthly prediction only. The number of training data is not big enough as well. The modified IR in the testing phase is equal to zero because the data is considered unseen. In the future, it is recommended to predict the value by using other approaches which can predict numerical prediction. Furthermore, the optimization of each models is recommended to improve the model performance and also to consider other kinds of parameter such as socio-economic data, the mobility of each citizen in the municipal, the water ponds or sanitation of houses[21], the data of JUMANTIK program, and big data[22].

ACKNOWLEDGMENTS

We thank to Politeknik Negeri Samarinda who supported us with the funding and as well as to Local

Health Department of Republik Indonesia of East Kalimantan who provided all data needed in the research.

REFERENCES

[1] "WHO | Dengue fever in Indonesia - update 4." [Online]. Available: https://www.who.int/csr/don/2004_05_11a/en/. [Accessed: 03-Oct-2020].

[2] WHO, "Global Strategy for Dengue Prevention and Control 2012–2020," *World Heal. Organiszation*, 2012, doi: /entity/denguecontrol/9789241504034/en/index.html.

[3] "Pusat Data dan Informasi - Kementerian Kesehatan Republik Indonesia." [Online]. Available: https://pusdatin.kemkes.go.id/article/view/19010400002/situasi-demam-berdarah-dengue-di-indonesia.html. [Accessed: 03-Oct-2020].

[4] Kementrian Kesehatan RI, "Hingga Juli, Kasus DBD di Indonesia Capai 71 Ribu," *Kementrian Kesehat. RI. (2020). Hingga Juli, Kasus DBD di Indones. Capai 71 Ribu. 2019–2020. https//www.kemkes.go.id/article/view/20070900004/hingga-juli-kasus-dbd-di-indonesia-capai-71-ribu.html*, 2020.

[5] A. Husnayain, A. Fuad, and L. Lazuardi, "Correlation between Google Trends on dengue fever and national surveillance report in Indonesia," *Glob. Health Action*, 2019, doi: 10.1080/16549716.2018.1552652.

[6] Kemenkes RI, "InfoDatin Situas Demam Berdarah Dengue," *Journal of Vector Ecology*. 2018.

[7] "Jumlah Penderita DBD di Kaltim Capai 1.183 Orang, 10 Meninggal Dunia." [Online]. Available: https://samarinda.kompas.com/read/2020/04/13/23224361/jumlah-penderita-dbd-di-kaltim-capai-1183-orang-10-meninggal-dunia. [Accessed: 03-Oct-2020].

[8] M. R. B. Pineda-Cortel, B. M. Clemente, and P. T. T. Nga, "Modeling and predicting dengue fever cases in key regions of the Philippines using remote sensing data," *Asian Pac. J. Trop. Med.*, 2019, doi: 10.4103/1995-7645.250838.

[9] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," *Int. J. Environ. Res. Public Health*, 2018, doi: 10.3390/ijerph15081596.

[10] T. W. Kesetyaningsih, S. Andarini, Sudarto, and H. Pramoedyo, "Determination of environmental factors affecting dengue incidence in Sleman District, Yogyakarta, Indonesia," *African J. Infect. Dis.*, 2018, doi: 10.2101/Ajid.12v1S.3.

[11] C. Alicino *et al.*, "Assessing Ebola-related web search behaviour: Insights and implications from an analytical study of Google Trends-based query volumes," *Infect. Dis. Poverty*, 2015, doi: 10.1186/s40249-015-0090-9.

[12] J. Gomide *et al.*, "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter," in *Proceedings of the 3rd International Web Science Conference, WebSci 2011*, 2011, doi: 10.1145/2527031.2527049.

[13] D. W. Seo and S. Y. Shin, "Methods using social media and search queries to predict infectious disease outbreaks," *Healthc. Inform. Res.*, 2017, doi: 10.4258/hir.2017.23.4.343.

[14] A. Aswi, S. M. Cramb, P. Moraga, and K. Mengersen, "Bayesian spatial and spatio-temporal approaches to modelling dengue fever: A systematic review," *Epidemiology and Infection*. 2019, doi: 10.1017/S0950268818002807.

[15] S. P. M. Wijayanti *et al.*, "The Importance of Socio-Economic Versus Environmental Risk Factors for Reported Dengue Cases in Java, Indonesia," *PLoS Negl. Trop. Dis.*, 2016, doi: 10.1371/journal.pntd.0004964.

[16] S. Mala and M. K. Jat, "Implications of meteorological and physiographical parameters on dengue fever occurrences in Delhi," *Sci. Total Environ.*, 2019, doi: 10.1016/j.scitotenv.2018.09.357.

[17] I. G. N. M. Jaya, A. S. Abdullah, E. Hermawan, and B. N. Ruchjana, "Bayesian Spatial Modeling and Mapping of Dengue Fever: A Case Study of Dengue Feverin the City of Bandung, Indonesia," *Int. J. Appl. Math. Stat.*, 2016.

[18] S. Sahay, "Climatic variability and dengue risk in urban environment of Delhi (India)," *Urban Clim.*, 2018, doi: 10.1016/j.uclim.2017.10.008.

[19] A. Appice, Y. R. Gel, I. Iliev, V. Lyubchich, and D. Malerba, "A Multi-Stage Machine Learning Approach to Predict Dengue Incidence: A Case Study in Mexico," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.2980634.

[20] BPS, "Provinsi Kalimantan Timur dalam Angka 2019," in *BPS Provinsi Kalimantan Timur*, 2019.

[21] L. S. Jayashree, R. Lakshmi Devi, N. Papandrianos, and E. I. Papageorgiou, "Application of Fuzzy Cognitive Map for geospatial dengue outbreak risk prediction of tropical regions of Southern India,"

Intell. Decis. Technol., 2018, doi: 10.3233/IDT-180330.

[22] S. I. Hay, D. B. George, C. L. Moyes, and J. S. Brownstein, "Big Data Opportunities for Global Infectious Disease Surveillance," *PLoS Med.*, 2013, doi: 10.1371/journal.pmed.1001413.