ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 7, No. 1, Jun 2021. pp. 10-20
Doi: 10.34818/ijoict.v7i1.562

Classification of Dengue Hemorrhagic Fever (DHF) Spread in Bandung using Hybrid Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Network Methods

Fatri Nurul Inayah[1], Sri Suryani Prasetiyowati[2], Yuliant Sibaroni [3]

[1,2,3] School of Computing, Telkom University
Bandung, West Java, Indonesia

*fatrinrlinayah@student.telkomuniversity.ac.id

Abstract

In some tropical countries, cases of dengue fever are still quite high, one of them is in Indonesia. Information about the status of the high and low number of cases of an area is important, so that it can be known by the public correctly and accurately. The information on the classification of the number of dengue cases also can be used to assist the government in preventing the spread of the number of cases from spreading, it is necessary to use an appropriate classification method, with fairly high accuracy. In this study, the classification method used is a combination of Naïve Bayes, K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN) methods, with the hope that a predictive model for the classification of dengue fever can be built. The data used in this study is a dataset of the number of cases of the spread of dengue hemorrhagic fever in the city of Bandung in the period 2012-2018. The performance results obtained using the Naïve Bayes, K-Nearest Neighbor, Artificial Neural Network methods are 74%, 78%, 86%, respectively. To increase the accuracy of the classification results of the three methods, hybridization of the three methods is carried out. The results of the hybrid classifier with the voting method turned out to be able to increase accuracy, to 90%.

Keywords: Artificial Neural Network, Classification, Dengue Fever, Hybrid Classifier, K-Nearest Neighbor, Naïve Bayes.

I. INTRODUCTION

DENGUE Hemorrhagic Fever (DHF) is a type of infectious disease caused by the dengue virus where the virus comes from the bite of Aedes Aegypti and Aedes Albopictus and will continue to plague an area every year. This disease is still a priority in Indonesia because besides being a tropical country it is also caused by the pattern of life of the people themselves by not maintaining the cleanliness of the surrounding environment. Dengue fever cases in Indonesia until July 2020 reached 71,633 cases and the highest number of cases was in West Java province with 10,772 cases [1].

West Java, especially Bandung Regency, experienced fluctuation from 2014 to 2018, in 2014 there were 995 cases with 4 deaths, in 2015 there were 1013 cases with 4 deaths, in 2016 there were 3,470 cases with 10 deaths, in 2017 there were 1,015 cases with 4 deaths and in 2018 there were 1,778 cases with 11 deaths [2].

Received on 22 May 2021, Revised at 31 May 2021, Accepted at 18 Jun 2021, Publish online on 25 Jun 2021

By looking at the increasing number of cases of spread and death in Bandung Regency, it is necessary to handle the spread of dengue fever, one of which is by predicting its spread as an early preventive measure, and as information on the classification of the number of dengue cases can be used to assist the government in preventing the spread of the number of cases from spreading, it is necessary to use an appropriate classification method, with fairly high accuracy. In previous studies, classification techniques have been carried out with various methods, by describing the application of the J48 Algorithm for dengue fever prediction using the J48 Decision Tree method which obtained an accuracy value of 75.833% [3]. Similar research on the combination of the K-Nearest Neighbor and Naive Bayes methods for data classification obtained an accuracy value of 76.92% [4], and in other studies regarding the prediction of potential blood donors, the accuracy value of the Neural Network method was 80%. higher than using the K-Nearest Neighbor and Naive Bayes method [5].

By looking at the results of the accuracy of several methods that have been used by previous researchers, the writer found that the K-Nearest Neighbor and Naive Bayes combination method has higher accuracy results than not doing the combination, whereas in the Neural Network algorithm comparison, K Nearest Neighbor, and Naive Bayes resulted in a higher level of accuracy on the Neural Network. Thus, the authors apply a combination of several methods, namely Naive Bayes, K-Nearest Neighbor, and Artificial Neural Network in the classification of dengue fever and compare them without using a hybrid to see the comparison of predictions of the classification of the spread of dengue fever in terms of getting the accuracy and speed of diagnosis. These three methods have their respective advantages and disadvantages so that with this hybrid method it is possible to overcome the weaknesses of the three methods to obtain high accuracy values.

II. Literature Review

Several studies have conducted dengue fever prediction using classification techniques. In the classification technique, there are several algorithms including Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Network. The use of the Naïve Bayes algorithm is relatively easy without repeated parameter estimation and has good performance [6]. The K-Nearest Neighbor algorithm is a simple method based on analogy learning [4]. The Artificial Neural Network algorithm is capable of performing difficult and complex computations similar to that of the human brain [7].

In the same case in this study, the prediction of dengue fever in the study [3] describes the application of the J48 Decision Tree method gets an accuracy value of 75.833% for cross-validation fold 5 and 80% accuracy results using fold 10. In the study [8] that implements a comparison of the SVM, Naïve Bayes, and Random Forest method respectively 77.5%, 56%, and 84.1%. From the comparison result of the three algorithms, the Random Forest has a good classification.

A study [5] presented the prediction of potential blood donors by performing comparisons of several algorithms. The study was conducted to compare the highest accuracy value using the Neural Network algorithm, K-Nearest Neighbor, and Naïve Bayes. From the test results by measuring the performance of the three algorithms, the neural network algorithm has fairly good accuracy with an accuracy value of 80%.

K-Nearest Neighbor and Naïve Bayes algorithms were also used in the study [4] on data classification. This study was conducted using a combination of K-Nearest Neighbor and Naive Bayes methods to overcome the weaknesses of each of these algorithms to obtain a higher percentage of accuracy. The combination results obtained 76.92% higher than the Naive Bayes and K-Nearest Neighbor methods.

A study [9] presented the classification of the diagnosis of dengue fever. This study was also carried out using a combination of Naïve Bayes and K-Nearest Neighbor to prove that the combination of the two methods can produce a good performance. The classification system created produces an accuracy value of 95.4%, sensitivity 96.2%, and specificity of 94.4%. So, this hybrid method produces a higher accuracy value than the basic method.

A study [10] also used a combination of algorithms in the classification of dengue fever. This study uses a hybrid classification model is the combination of Naïve Bayes and Decision Tree with the voting method to get the best accuracy results. The voting method is applied to select a classifier that performs well from several classifiers so that the research results in an accuracy value of 92%.

This study used a hybrid classification method of Naive Bayes K-Nearest Neighbor and Artificial Neural Network for the classification of dengue fever. The Artificial Neural Network algorithm in the prediction of dengue fever results in fairly good accuracy [5]. The combined study of K-Nearest Neighbor and Naive Bayes, it provides a fairly good accuracy compared to the K-Nearest Neighbor method and the Naïve Bayes method [4]. The use of a combination of Naive Bayes and a Decision Tree using the voting method gives an accuracy of 92% [10], as well as study [9] which gives an accuracy of 95.4% using the hybrid Naïve Bayes-KNN method. So, the author tries to combine the Naive Bayes, K-Nearest Neighbor, and Artificial Neural Network algorithms for the classification of dengue fever and the effect of the voting method on the results of accuracy.

III. RESEARCH METHOD

Fig. 1 describes the system design in the method applied in this study.

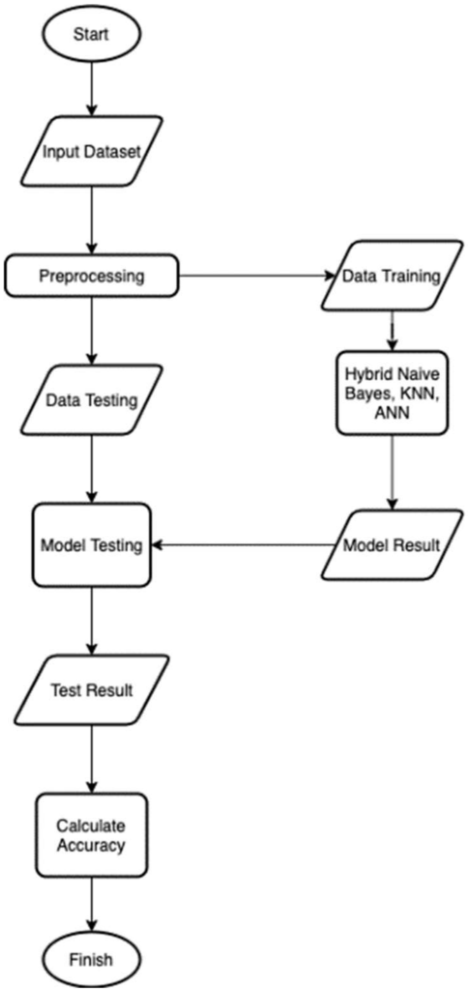

Fig. 1. System Design

A. Dataset

The data used in this study is a dataset of dengue fever in Bandung Regency in the period 2012 to 2018, with the number of cases in 2012 as many as 4.780 cases, in 2013 as many as 2.271 cases, in 2014 as many as 4.095 cases, in 2015 as many as 1.669 cases, in 2016 as many as 3.880 cases, in 2017 there were 1.787 cases, and in 2018 there were 2.827 cases. This data consists of several attributes including District, number of cases, rainfall, humidity, temperature, class (high, medium, and low). This dataset will be divided into 2 datasets, namely training data and testing data, this data distribution is used 80% for training data and 20% for testing data. Training data is used to determine the probability of classifying in decision making, while data testing is used for testing in this study in model building.

B. Preprocessing

1) Labeling Class:

TABLE I
CLASS LABELING

Class	Label Class	Range
High	2	Number of cases > 55
Medium	1	Number of cases <= 55
Low	0	Number of cases < 20

There are 3 categories in the class the number of cases with ranges in each class indicating categories is high (2), medium (1), and low (0).

2) Normalization: Normalization is the process of scaling attribute values in data, with the formula:

$$v' = \frac{v-min}{\max - min}\,(new\_\max - new\_\min\,) \; + \; new\_min \qquad (1)$$

C. Classification Process

Classification involves grouping objects based on existing groups [11]. In classification, there are target variable categories [12], for example, the classification of dengue fever spread can be separated into three categories, namely high-level spread, medium-level spread, and low-high spread.
Classification has two types of models, namely [13]:

1. Descriptive modeling, namely a classification model that functions as an explanatory tool to distinguish objects from different classes. Examples of descriptive modeling are; organizational structure, plant grouping table.
2. Predictive modeling, which is a classification model used to predict class labels for unknown records.

1) Naïve Bayes: Naive Bayes Classifier is a classification method based on Bayes theorem. British scientist Thomas Bayes proposed a classification method that uses probability and statistics. Thomas Bayes predicts the probability of the future based on previous experience, which is called Bayes' theorem [14].

The main feature of this Naïve Bayes classifier is a very strong (naïve) assumption that each condition or event is independent, where it is assumed that each attribute of the sample (sample data) is independent of one another based on class attributes. Naive Bayes Categorical is Naive Bayes with statistical data in the form of categories or definite data so that in the process have obtained definite results. Naive Bayes is a Bayesian method with the basic formula, In the Bayes theorem, if there are two separate events (e.g. A and B), then the Bayes theorem as follows [15]:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{2}$$

This equation is the Bayes theorem that will be used to perform classification calculations. In classification with continuous data or numerical data, the Gaussian distribution formula can be used [16].

$$P(X_i = x_i|C = c_i) = \frac{1}{\sqrt{2\pi\sigma ij}} exp \frac{(xi-\mu ij)2}{2\sigma 2ij} \tag{3}$$

In addition to the relatively easy Naive Bayes modeling without repeated and complicated parameter estimates, Naive classification Bayes is often very successful and widely used because they have better performance [6].

2) *K-Nearest Neighbor (KNN):* K-Nearest Neighbor (KNN) is a classification method for a set of data based on defending data networks. pre-defined (labeled). KNN is included in the supervised learning group, which classifies the results of the new query instance into the KNN according to the closest degree to the existing category. The new class of data will be selected based on the class group that is closest to the vector distance. To calculate the distance value between two points in the training data and the points in the testing data, the Euclidean Distance formula can be used as follows [13].

$$\mathrm{d(x,y)} = \sqrt{\textstyle\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4}$$

In the KNN algorithm, a test data is z = (x ', y'), where x 'is a vector/attribute test data, while y' is the test data class label unknown, then calculate the distance (or similarity) of the test data to each training data d (x', x), then take K the first nearest neighbor in Dz. After that, the number of data that participated in the existing class from the K neighbor is calculated. The class with the most data following it becomes the winning class which is given as the class label on the test data y' [17].

In the algorithm above, this KNN classifies the test data using a training set. In classifying the data, the first thing is to calculate the K value which indicates the number of K-Nearest Neighbor. By doing a data test, calculate the distance between all training data, and classify the distance. Then through coordination, the class label is assigned to most of the test data [18].

3) *Artificial Neural Network:* Artificial Neural Network (ANN) is a machine learning algorithm that can be used for estimation/regression and classification. ANN works to imitate the workings of the human brain in terms of (1) Knowledge obtained by the network from the environment, through a learning process; (2) The strength of the connection between units, called synaptic weights, serves to store the knowledge that has been acquired by the network [19].

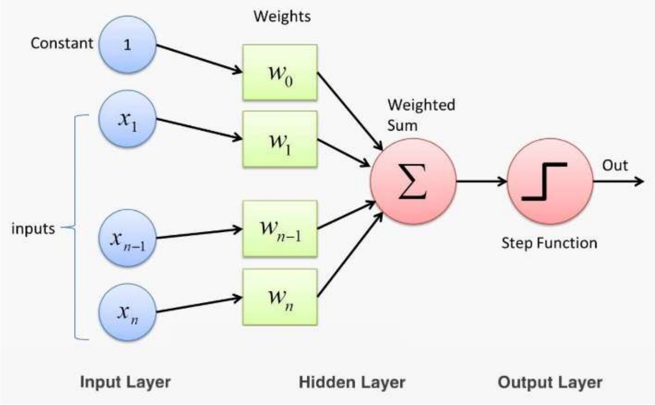

Fig. 2. ANN Architecture [19]

Figure 2 shows that an ANN network has three components, namely synapse (w_1, w_2, …, w_n), adder, and activation function.

The ANN algorithm aims to create an artificial system capable of complex and complex computations similar to that of the human brain, such as pattern recognition. By using a training set, the network connects the input to output via estimated parameters, making some generalizations outside of the training set. Networks are distinguished by their architecture, level of complexity, number of layers, loops, feedback loops, activation or transfer functions. So, it can be said that the Artificial Neural Network is a connection of simple elementary neuron objects consisting of input, weight, bias, activation, and output functions, where input is the variable used and the output is the result. Weights and bias are randomly assigned and then optimized to minimize errors. The activation function determines the neuron output [8].

4) *Hybrid Classification:* Hybrid is a combination of two or more systems in one function [20]. Hybrid model classification is a method that works by combining more than 1 model in this case using the voting method and combining Naive Bayes, K-Nearest Neighbor and Multi-Layer Perceptron Classifier to make predictions. The voting classifier is a meta-classifier that makes predictions by combing the predictions of several independent classifiers based on a predefined voting strategy [21]. The way this hybrid works is by ensuring that errors are made, by one classifier can be solved by another classifier. Hybrid model trains many models and predicts the outputs based on their highest probability of selecting a class as output [9].

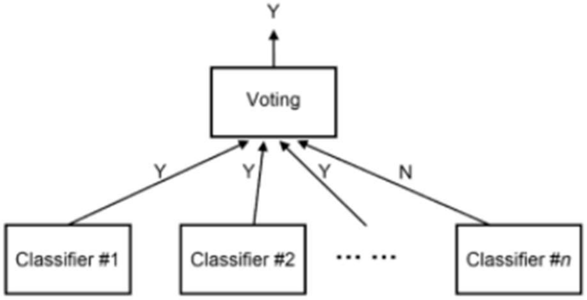

Fig. 3. Voting Classifier Architecture [21]

Based on figure 3 of the process architecture of the voting classifier, there are 2 classes (Y and N) in a dataset. In the process, there are several of the n classification to be combined, and from the result combination the final decision will be made through voting classifier. The final voting decision is the Y class because the number of classifications predicted for each Y class is greater than n class [21].

D. Accuracy

Calculating this accuracy value is the last stage of this process. In calculating the accuracy is done to describe the accuracy of the system in the classification of data correctly or in other words the comparison between the data classified correctly with the entire data. The accuracy value can be obtained by the formula:

$$\frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (5)$$

IV. RESULTS AND DISCUSSION

The research was done by 2 datasets, namely training data and testing data. Training data is used to determine the probability of classifying in decision making, while data testing is used for testing in this study in model building. For class data, there are 3 categories based on the range of the number of cases.

A. Dataset Scenario

This research has 3 categories of dataset. These categories aim to represent the class with range of the number of cases. Based on table 1, high class for the number of cases above 55 cases, medium class for the number of cases less than 55 cases, and low class for the number of cases less than 20 cases. The dataset categories are as follows.

TABLE II
DATASET CATEGORIES

Class	**Categories**
High	2
Medium	1
Low	0

B. Result of Performance Evaluation

Evaluating the results shows that the performance of the proposed model system has been verified in terms of certain parameters, such as accuracy, precision, recall, and f1-score.

TABLE III
ACCURACY ANALYSIS

Classification	**Accuracy**
Naïve Bayes	74%
KNN	79%
ANN	86%
Hybrid	90%

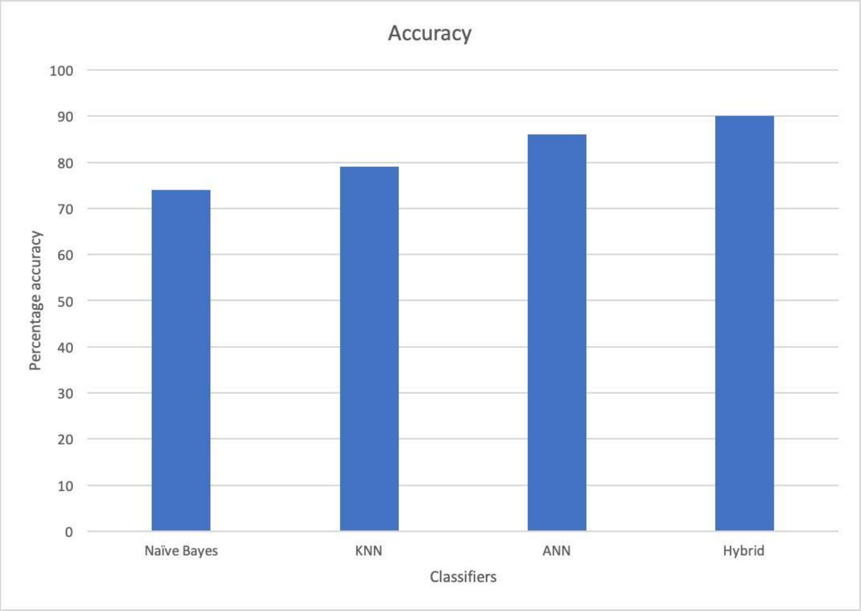

Fig. 4. Performance Evaluation for Hybrid Classifier with The Existing Methods

Based on Table 3 and Figure 4, the accuracy of Naïve Bayes, KNN, and ANN classifiers are compared with the hybrid classification. Naïve Bayes, KNN, and ANN classifiers have 74%, 79%, 86% accuracy which is lower performance and the hybrid classifier top- most performance. It is analyzed that when the hybrid classification is used accuracy is increased up to 90%. At this stage, voting method will ensure the error of each algorithm, therefore in this hybrid method, it will correct each other's error. Previous studies [10], presented classification techniques have been carried out with hybrid decision tree and Naïve Bayes classification for dengue fever using voting method but with a different classification method and only the Decision Tree and Naïve Bayes classifier were used in the hybrid classifier with an accuracy level of 92%.

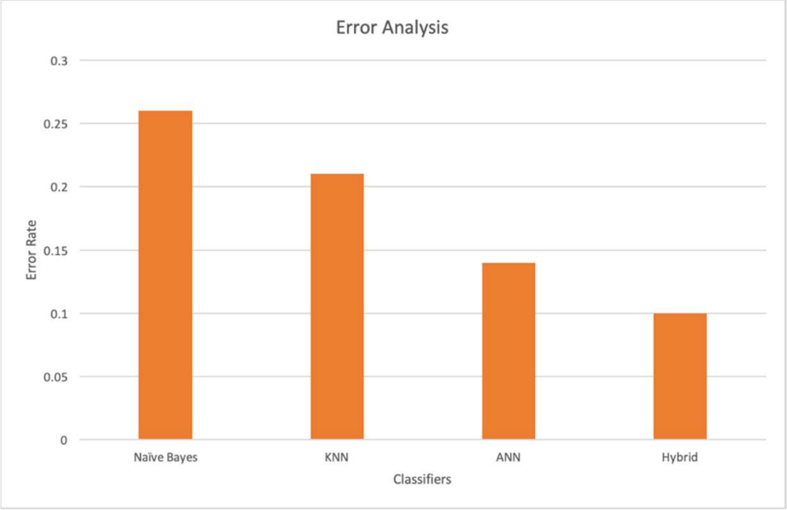

Fig. 5. Error Rate

Based on Figure 5, the error rate in classification results shows that the hybrid classifier is better with an error of 0.1 than with the existing methods. Error Analysis performance describes the error rate value of Naïve Bayes, KNN, and ANN classifier shows that the error rate is higher performance and the hybrid classifier is lower performance. Based on hybrid works, that errors made by one classifier can be solved and will ensure the error by another classifier, and train many models and predict output based on the highest probability of selecting a class as output so that it can produce the best accuracy.

TABLE IV
PRECISION-RECAL-F1-SCORE ANALYSIS

Metrics	Naïve Bayes	KNN	ANN	Hybrid
Precision	46%	77%	58%	94%
Recall	41%	62%	50%	82%
F1-score	43%	63%	53%	86%

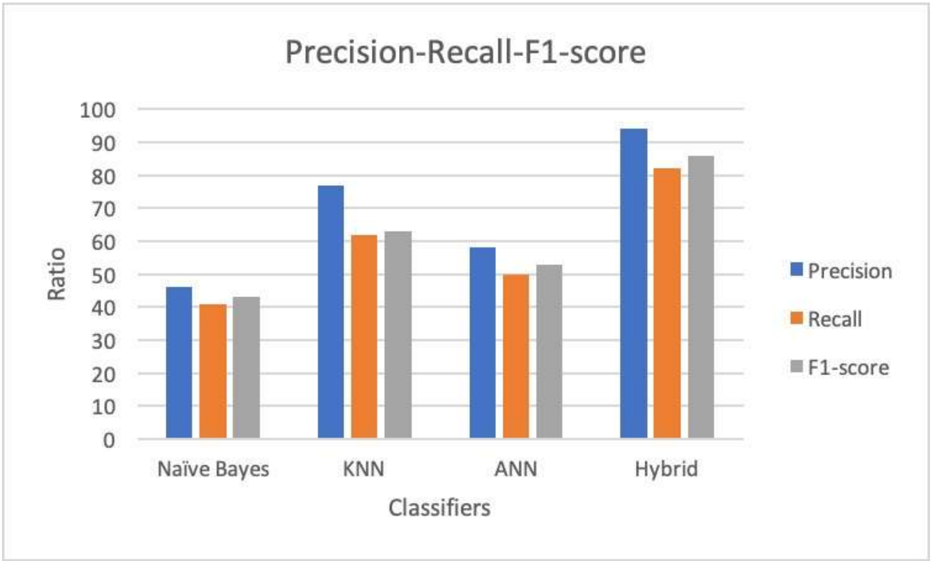

Fig. 6. Comparative Analysis of the Hybrid Classifier with The Existing Classifiers

The average of every metrics is taken as the overall system's performance. Analysis performance describes the precision, recall, and f1-score value of the Naïve Bayes, KNN, and ANN classifier are compared with the hybrid classification. The hybrid has 94% precision, that the value of hybrid classifier with voting method is high as compared to Naïve Bayes, KNN, and ANN classifiers. The precision, recall, and f1-score in the hybrid method has higher results because it combines the three algorithms and uses the voting method. The voting process is used to produce the final prediction result, and the voting process can select the most effective classifier from multiple classifiers and produce the prediction result.

V. CONCLUSION

In this study, can be concluded that the classification of dengue based on the range of cases uses the Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Network methods. For classification, the performance shown by the Naïve Bayes, K-Nearest Neighbor, Artificial Neural Network methods has less accuracy in the classification of dengue with an accuracy level of 74%, 78%, 86%, respectively. To increase the accuracy of the classification results of the three methods, hybridization of the three methods is carried out with voting method which renders an excellent performance accuracy of up to 90%. The outcomes confirmed that the proposed hybrid classifier accurately predictions the spread of dengue fever in Bandung regency.

REFERENCES

[1] Kementrian Kesehatan RI, Sekretariat Jendral, "Profil Kesehatan Indonesia Tahun 2020", Jakarta: Kementrian Kesehatan RI, 2020, Accessed: Jul 9, 2020. [Online]. Available: https://www.kemkes.go.id/article/view/20070900004/hingga-juli-kasus-dbd-di-indonesia-capai-71-ribu.html

[2] Dinas Kesehatan, "Profil Kesehatan Kabupaten Bandung Tahun 2018", Bandung : Dinas Kesehatan Kabupaten Bandung, 2018.

[3] S. Suciati, I. Mujiati, D. Ayu, V. Ristianah and W. A. Lestari, "Penerapan Algoritme J48 Untuk Prediksi Penyakit Demam Berdarah", *Telematika*, vol. 9, no. 2, pp. 1-10, 2016.

[4] P. P. M. K. Sari, E. Ernawati and P. Pranowo, "Kombinasi Metode K-Nearest Neighbor dan Naive Bayes Untuk Klasifikasi Data", *J. SEMNASTEKNIMEDIA*, vol. 3, pp. 37-41, 2015.

[5] D. R. S. N. Mandiri, "Komparasi Algoritma Neural Network, K-Nearest Neighbor Dan Naive Baiyes Untuk Memprediksi Pendonor Darah Potensial", *Speed-Sentra Penelitian Engineering dan Edukasi*, vol. 8, no. 3, 2017.

[6] K. Vembandasamy, R. Sasipriya and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 9, pp. 441-444, 2015.

[7] Z. Assaghir, A. Janbain, S. Makki, M. Kurdi and R. Karam, "Using Neural Network to Predict the Hypertension" *International Journal of Science & Engineering Development Research,* 2017.

[8] R. Arafiyah, F. Hermin, I. R. Kartika, A. Alimuddin and I. Saraswati, "Classification of Dengue Haemorrhagic Fever (DHF) using SVM, Naive Bayes and Random Forest", In *IOP Conference Series: Materials Science and Engineering,* vol. 434, no. 1, pp. 012070, Bristol, United Kingdom: IOP Publishing, 2018.

[9] E. A. Zurroh, "Klasifikasi diagnosa penyakit Demam Berdarah Dengue menggunakan metode Hybrid Naive Bayes-K Nearest Neighbor", *Doctoral dissertation, UIN Sunan Ampel Surabaya*, 2021.

[10] P. Taneja and N. Gautam,2019, "Hybrid Classification Method for Dengue Prediction", *International Journal of Engineering and Advanced Technology (IJEAT) ISSN*, 2249-8958, 2019.

[11] B. Santosa, and A. Umam, "Data Mining dan Big Data Analytics: Teori dan Implementasi Menggunakan Python & Apache Spark", Yogyakarta, Indonesia: Penebar Media Pustaka, 2018.

[12] E. T. L. Kusrini, "Algoritma data mining", Yogyakarta, Indonesia: Andi Offset*,* 2009.

[13] A. Wanto *et al*, "Data Mining: Algoritma dan Implementasi", Medan, Indonesia:Yayasan Kita Menulis, 2020.

[14] A. F. Watratan and D. Moeis, "Implementasi Algoritma Naïve Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 di Indonesia", *Journal of Applied Computer Science and Technology*, vol. 1, no. 1, pp. 7-14, 2020.

[15] D. C. Rini, Y. Farida and D. Puspitasari, "Klasifikasi Menggunakan Metode Hybrid Bayessian-Neural Network (Studi Kasus: Identifikasi Virus Komputer)", *Jurnal Matematika "MANTIK"*, vol. 1, no. 2, pp. 38-43, 2016.

[16] C. Fadlan, S. Ningsih and A.P. Windarto, "Penerapan Metode Naïve Bayes Dalam Klasifikasi Kelayakan

Keluarga Penerima Beras Rastra", *JUTIM (Jurnal Teknik Informatika Musirawas)*, vol. 3, no. 1, pp. 1-8, 2018.

[17] E. Prasetyo, "Fuzzy K-Nearest Neighbor in Every Class Untuk Klasifikasi Data", In *Seminar Nasional Teknik Informatika,* pp. 57-60, 2012.

[18] Asfaw, "T. Performance Comparison of K-Nearest Neighbors and Gaussian Naïve Bayes algorithms for Heart Disease Prediction", *International Journal of engineering Science Invention(IJESI),* vol. 8, no. 8, 2019.

[19] B. Santoso and A.I. Azis, "Machine Learning & Reasoning Fuzzy Logic Algoritma, Manual, Matlab, & Rapid Miner", Yogyakarta, Indonesia: Deepublish, 2020.

[20] F. Juliaristi, "Peramalan Banyak Kasus Demam Berdarah (DB) di Kota Surabaya Menggunakan Hybrid Integer-valued Autoregressive Integrated Moving Average (INARIMA) dan Radial Basis Function Neural Network (RBFNN)", *Doctoral dissertation*, *Institut Teknologi Sepuluh Nopember*, 2016.

[21] S. Misra, H. Li and J. He, "Machine learning for subsurface characterization", Houston, USA: Gulf Professional Publishing, 2019.