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

Intl. Journal on ICT
Vol. 8, No. 1, June 2022. pp. 10-19
doi: 10.21108/ijoict.v8i1.614

Performance Analysis of the Hybrid Voting Method on the Classification of the Number of Cases of Dengue Fever.

Muhammad Arief Rahman[1], Sri Suryani Prasetiyowati[2], Yuliant Sibaroni[3]

[123]*School of Computing, Telkom University*
Jl. Telekomunikasi No.1, Bandung, 40257, Indonesia

[*]marief@student.telkomuniversity.ac.id

Abstract

This study proposes the use of a hybrid classification method in machine learning algorithms. the machine learning algorithm is an algorithm used in the machine learning process based on data. for the hybrid method using the ensemble learning method, namely the voting method. the voting method is a combination algorithm for predicting a class. This study aims to improve the results of machine learning algorithm classification accuracy in the classification process for the distribution of the number of dengue cases in the city of Bandung. The machine learning algorithm used in this research is Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree (DT). The results obtained from the three algorithms are superior with KNN with an accuracy of 87%, SVM at 84%, and DT at 79%. by using the hard voting method approach, the accuracy increased to 91%. The proposed model can obtain better accuracy results from the three machine learning algorithms. The contribution of this research is to provide information that the hybrid classification of the number of dengue cases using a voting approach can increase the accuracy of the proposed model.

Keywords: Classification, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Hybrid Classifier

I. INTRODUCTION

Dengue hemorrhagic fever (DHF) is a health problem in Indonesia. With a total of 267 million people, Indonesia is at risk of contracting dengue fever. The main factor causing dengue fever is caused by the bite of the Aedes aegypti mosquito which carries the dengue virus. Other factors can be caused by high rainfall, humidity, and temperature. weather factors are very influential in the process of breeding these mosquitoes [1].

According to the Ministry of Health of the Republic of Indonesia, the number of DHF sufferers in Indonesia from the beginning of the year to July 2020 was 71,633 cases with 459 deaths, while in 2019 there were 919 deaths with 138,127 cases [2]. The region in Indonesia that has the highest number of cases is West Java with the highest ranking with 10,772 cases. The area of West Java that has the highest number of cases is the city of Bandung. The city of Bandung was recorded to have the highest number of cases, namely 4,424 cases.

To overcome the problem of the number of cases of dengue fever that occurred in the city of Bandung, a solution is given to classify the effect of weather on the number of cases that occur. Classification is useful for

Received on 24 Mar 2022. Revised on 2 June 2022. Accepted and Published on 21 June 2022.

grouping variables based on predefined classes [3]. one of them is the use of classification methods in machine learning algorithms. there has not been much research to improve the results of classification performance. usually only focuses on the accuracy of classification results obtained.

Previous studies used a hybrid stacked classifier with SVM, KNN, and C5.0. the results obtained on the individual classification of SVM obtained the best accuracy with 88.2%. while for hybrid SVM, KNN, and C5.0 using meta RF and GLM with 10 cross validations of 91.2% and 90.8%, respectively. in s study concluded that the variation of the classification combination model with the ensemble approach can increase the accuracy of the results [4].

In another similar study, DHF in Indonesia was predicted using the Decision Tree with an accuracy of 76% [5]. Another study used a comparison of the Support Vector Machine and C4.5 algorithms for the identification of pests and diseases in chili plants, with an accuracy of 82.33% and 89.25%, respectively [6]. Another study using k-Nearest Neighbor for heart disease obtained an accuracy of 81,31% [7].

Based on the research above, it can be concluded that there are still few studies that have developed the performance of classification algorithms using ensemble learning. In this study, the authors focus on the hybrid classifier using the hard voting ensemble method of three classification algorithms, namely Support Vector Machine, K-Nearest Neighbor, and Decision Tree. The ensemble voting approach can solve the problem of determining better predictive results. it can be seen from the research above that each classification algorithm has fairly good accuracy. by using the voting ensemble of the three classification algorithms, the prediction results can be more optimal and the accuracy obtained will increase. The contribution of this research is to provide information that the ensemble voting method can improve prediction results that are more optimal in the classification process.

II. LITERATURE REVIEW

Dengue fever research using the Support Vector Machine algorithm gets high accuracy results, the Support Vector Machine is a good classification and prediction algorithm but also requires good data preprocessing [8]. Another study on dengue fever compared the Backpropagation method, the Gaussian method, and the Support Vector Machine. The Support Vector Machine has the largest error value compared to the other 2 methods [9].

Dengue fever prediction research uses the Decision Tree algorithm and Support Vector Machine. Decision Tree is more efficient, the processing time is fast, etc. However, this method produces a lot of false positives. Support Vector Machine is better than Decision Tree based on accuracy, sensitivity, specificity, and area under the curve. Support Vector Machine can produce high accuracy when selecting features correctly [10].

Another study predicts heart disease using the K-Nearest Neighbor algorithm and the Support Vector Machine. comparing the use of data normalization method by not using data normalization in two different classification algorithms. the results obtained by normalizing the data can increase the accuracy of the classification process [7].

Another research on breast cancer classification uses ensemble learning from several classification methods. This study compares the values of accuracy, precision, recall, and f1-score of each classification algorithm used. choose the three best classifications and then optimize by comparing the probability of the highest, minimum, and average votes. the election of the most votes managed to reach a peak in the resulting accuracy compared to using advanced algorithms carried out by previous studies [11].

In thyroid prediction research, it is proposed to develop an ensemble method. This method combines Bagging, Boosting, Stacking, and Voting. each method will predict the results of a model that is built. Voting will do the voting of the three methods. The results obtained from the four methods are that Boosting has a weaker level of accuracy than Bagging and Stacking. Voting from Bagging, Boosting, and Stacking got good accuracy with a small number of wrong predictions. [12].

In infectious disease prediction research, three basic classification methods are used with a combination of ensemble voting methods. In this case, the classification of each classification algorithm method is carried out, and then the majority vote is carried out. The ensemble voting method has the highest efficiency, which can be seen from the confusion matrix in the classification of infectious diseases [13].

In this case study, the classification of dengue fever uses a hybrid of three basic classification methods. The voting method can improve accuracy results by overcoming the weaknesses of each classification algorithm used. Voting also minimizes errors from the classification algorithm [14].

The ensemble learning selection method is a classification process with combined classification using a selection system that will determine the selected model. each model will be free to choose the class to be used and free to perform the desired test. The model chosen is the one that has the maximum sound from the test and class [4].

III. RESEARCH METHOD

This research was carried out with several processes described in the research methodology in Fig. 1.

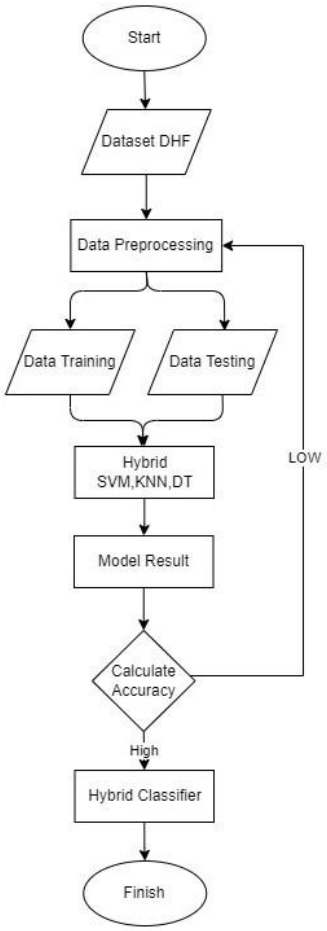

Fig. 1. System Design Hybrid Classifier.

A. Dataset

In this study, the Bandung city of dengue fever dataset was used from 2012 to 2018. This dataset has attributes including year, sub-district, number of cases, rainfall, humidity, temperature, and class. The dataset will be divided into 80% training data and 20% testing data. training data is used to train the model to be built while testing data is used to test the model that has been built. The variables used in this research are shown in the Table 1.

TABLE 1
Research Variable

Variable	Explanation
X_1	Years
X_2	Rainfall
X_3	Humidity
X_4	Temperature
X_5	Sub-district
Y	Class based on number of case

B. Data Preprocessing

Data preprocessing is a data cleaning phase that will eliminate error values in the data set. Various kinds of preprocessing techniques and methods have been developed. choosing the right technique is the essence of data preprocessing [15].

The techniques and methods used include making class labels based on the number of cases of dengue fever as shown in the Table 2.

TABLE 2
CLASS LABELING

Class	Label Class	Range
High	2	Cases > 55
Medium	1	Cases <= 55
Low	0	Cases < 20

Another technique uses the data normalization method to convert the input values to min and max to normalize the distribution and increase the success rate [9]. The formula is shown in (1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

C. Support Vector Machine

Support Vector Machine is an algorithm that can solve binary class problems. Support Vector Machine has many kernels. Each kernel has a different value [8].

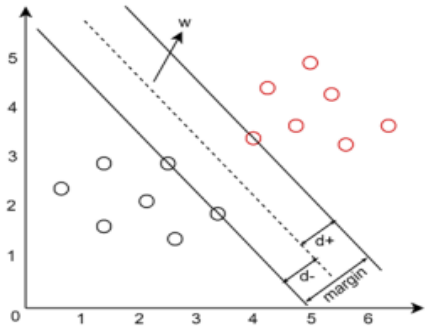

Fig. 2. Decision boundary in SVM with maximal margin

Figure 2 explains that there are two black and red circles. The red circle is the first class and the black circle is the second class. W is a hyperplane (decision boundary) that serves as an intermediary between the two classes. margin is the distance between the hyperplane and the closest data for each class [16].

In this study the author uses the RBF kernel with the formula below:

$$K(X, X^1) = \exp\left(-\frac{||x - x^1||^2}{2\sigma^2}\right) \quad (2)$$

D. K-Nearest Neighbor

K-Nearest Neighbor is an algorithm that classifies by calculating the distance of a neighbor numbered K. The distance can be calculated by various methods. K-Nearest neighbor does not undergo a training and understanding process.

The K value in K-Nearest Neighbor must be odd or more than one but must not exceed the training data. the greater the value of K, the smaller the value of the classification noise generated. the euclidean method is used in calculating the neighboring distance [7]. the euclidean formula below:

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (3)$$

X_{1i} = Data Sample
X_{2i} = Testing Data
i = Data Variable
d = Distance
p = Data Dimension

E. Decision Tree

Decision Tree is a picture of a tree that has roots, branches, and leaves. The top of the decision tree is called the node. Each node will show the attribute, the decision branch, and each leaf will show the category result. the decision Tree works by testing the attributes that are on the internal node, the test results will be on the branch and the results will be on the leaf node [17].

The conceptual rule of a decision tree is a model sequentially uniting a series of tests efficiently and cohesively where the feature value will be compared with the test threshold value. Entropy is used to measure the value of the random datasets. Entropy has values of 0 and 1. Entropy is said to be good if its value is close

to 0. Meanwhile Information gain is a metric of reciprocal information. The higher the information gain value the better [18]. Entropy formula and gain information below.

$$Entropy(S) = \sum_{i=1}^{c} p_i \, log_2 \, p_i \quad (4)$$

$$Gain(S, A) = \sum_{v \in V(A)} \frac{|S_i|}{|S|} \, Entropy(S_i) \quad (5)$$

S = set of cases
Si = number of cases on the partition
A = attribute
$|S_r|$ = number of cases in S

F. Hybrid Classification

Ensemble learning using the hard voting method can be used to improve the classification results [4]. After applying the three classification algorithm methods, the most votes will be selected for each algorithm to predict the final result of the test. In majority voting, class y labels are predicted by majority voting for each classifier C:

$$y = mode \{C_1(x), C_2(x), .., C_n(x)\} \quad (6)$$

y is the final class from the results of the classification prediction voting, C_i is the result of class prediction by classification.

G. Confusion Matrix

The confusion matrix is a process for assessing the work results and whether they have achieved the desired goals. confusion matrix can evaluate the performance of the model that has been built [16].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (7)$$

$$Precision = \frac{TP}{FP+TP} * 100\% \quad (8)$$

$$Recall = \frac{TP}{FN+TP} * 100\% \quad (9)$$

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

IV. RESULTS AND DISCUSSION

In this research analysis, we use several classification algorithms and hybrid classification methods to improve the accuracy of the results obtained. By using the Decision Tree algorithm, the authors get the problem of training data that has overfitting constraints. By adding some parameters the problem of overfitting on the Decision Tree can be overcome.

Another algorithm uses K-Nearest Neighbor with a value of K=3. K value is taken based on its accuracy performance. other algorithms Support Vector Machine uses the RBF kernel. After completing the classification stage, the majority vote will be selected using the hard voting method. Accuracy results are shown in the Table 3 and Figure 3.

TABLE 3
ACCURACY CLASSIFIERS

Classification	Accuracy
DT	79%
SVM	84%
KNN	87%
Hybrid	91%

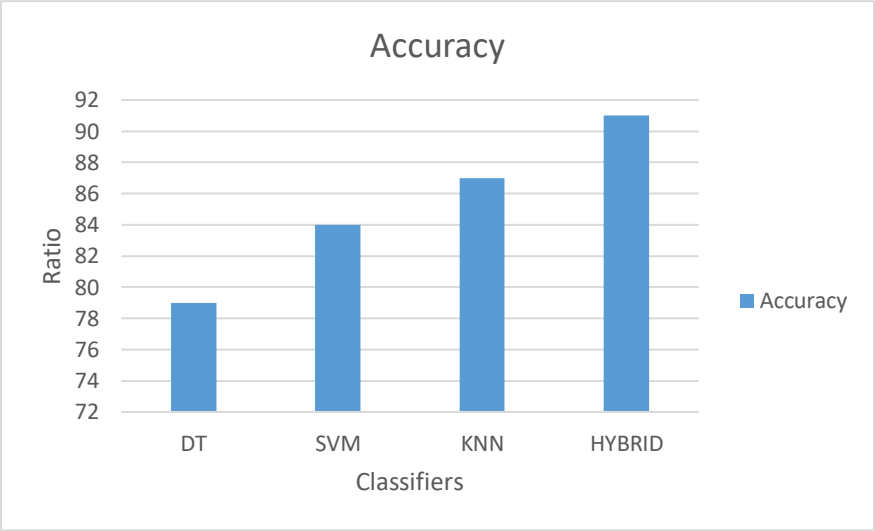

Fig. 3. Accuracy Classifiers

Based on the results obtained, Decision Tree has an accuracy of 79%, Support Vector Machine has an accuracy of 84%, K-Nearest Neighbor has an accuracy of 87% and Hybrid Classification has an accuracy of 91%. The high and low accuracy obtained by the classification algorithm affects the number of datasets and the number of variations in the dataset values. Class imbalance also affects machine learning algorithms in the classification learning process. the oversampling method is used in unbalanced class handling. This method affects increasing accuracy results compared to not oversampling.

In the Decision Tree algorithm, there is an overfitting constraint. the author adds parameters such as tree depth, number of leaf nodes, and splits. overfitting occurs due to a lack of variation in the dataset so the Decision Tree algorithm learns the data too well. this also affects the accuracy of results obtained. for the K-Nearest Neighbor algorithm, the selection of the K value greatly affects the accuracy of results. The author compares the K values from 3 to 10 to find the most optimal K value in getting the best accuracy results. for the Support Vector Machine algorithm, the selection of the right SVM kernel also affects the accuracy of results obtained.

The voting method based on the most votes can improve the results of classification accuracy in the distribution of the number of dengue cases. this method outperforms the 3 classification methods of machine learning DT, SVM, and KNN in the classification accuracy results. the voting method based on the most votes

combines several classification prediction results from several classification methods which cause better prediction results because it overcomes the problem of prediction errors from the classification algorithm.

In previous research in the use of hybrid classification on the dataset of DHF patients. the method used is voting from the KNN 91%, NB 93%, and DT 92% algorithms which get a hybrid accuracy of 95% [13]. In this study, hybrid classification using a dataset of DHF patients was able to improve the performance of the classification results.

Similar studies use hybrid classification in the dataset on the number of dengue cases. the method used is voting from the NB 74%, KNN 79%, and ANN 86% algorithms which get a hybrid accuracy of 90% [14]. In this study, the accuracy obtained from the classification algorithm is still not good, which causes the hybrid accuracy to be not optimal.

TABLE 4
Precision-Recall-F1-score

Metrics	DT	SVM	KNN	Hybrid
Precision	81%	85%	87%	91%
Recall	81%	83%	86%	90%
F1-Score	81%	83%	86%	90%

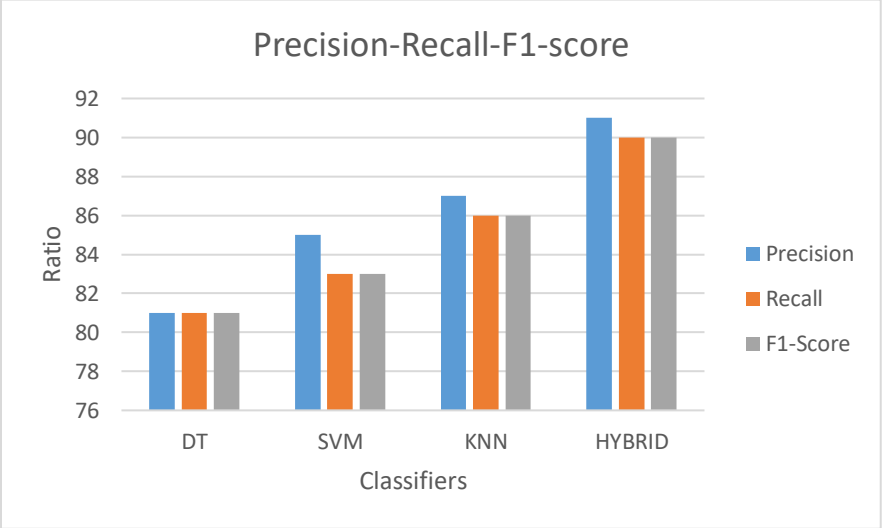

Fig. 4. Precision-Recall-F1-score

From the results of the evaluation of the Table 4 and Figure 4, the precision, recall, and F1-score values are taken from the average of each class used. hybrid classification got the best score with 91% precision, 90% recall, and 90% F1-score. Compared to other algorithms, hybrid classification can minimize the error rate that occurs.

V. CONCLUSION

In this study, the classification of the spread of dengue fever is based on the influence of weather on the number of dengue cases in the city of Bandung. it can be concluded that the use of the voting method based on the most votes can improve the classification accuracy results. The proposed model can achieve an accuracy of 91% in classifying the spread of dengue cases. the voting method will be more optimal if the comparison of the classification algorithm used has good accuracy. for further research on the spread of dengue, there is still the ability to improve accuracies such as the use of various classification methods and the incorporation of several ensemble combination methods using voting. voting is currently the best ensemble method for improving the performance of results.

VI. REFERENCES

[1] C. E. Kosasih, M. Lukman, T. Solehati, and H. S. Mediani, "Effect of dengue hemorrhagic fever health education on knowledge and attitudes, in elementary school children in West Java, Indonesia," *Linguist. Cult. Rev.*, vol. 5, no. S1, pp. 191–200, 2021, doi: 10.21744/lingcure.v5ns1.1349.

[2] KEMENKES RI, "Profil Kesehatan Indonesia Tahun 2020," *Kementrian Kesehatan Republik Indonesia*, 2021. .

[3] M. Nilashi *et al.*, "Journal of Soft Computing and Decision Support Systems Disease Diagnosis Using Machine Learning Techniques: A Review and Classification," *JSCDSS*, vol. 7, no. 1, pp. 19–30, 2020, [Online]. Available: http://www.jscdss.com.

[4] S. Rani and N. S. Gill, "Hybrid model for twitter data sentiment analysis based on ensemble of dictionary based classifier and stacked machine learning classifiers-SVM, KNN and C5.0," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 4, pp. 624–635, 2020.

[5] A. Z. Abdullah, B. Winarno, and D. R. S. Saputro, "The decision tree classification with C4.5 and C5.0 algorithm based on R to detect case fatality rate of dengue hemorrhagic fever in Indonesia," *J. Phys. Conf. Ser.*, vol. 1776, no. 1, pp. 0–10, 2021, doi: 10.1088/1742-6596/1776/1/012040.

[6] M. Irfan, N. Lukman, A. A. Alfauzi, and J. Jumadi, "Comparison of algorithm Support Vector Machine and C4.5 for identification of pests and diseases in chili plants," *J. Phys. Conf. Ser.*, vol. 1402, no. 6, 2019, doi: 10.1088/1742-6596/1402/6/066104.

[7] D. A. Anggoro and N. D. Kurnia, "Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 1689–1694, 2020, doi: 10.30534/ijeter/2020/32852020.

[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," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 434, no. 1, 2018, doi: 10.1088/1757-899X/434/1/012070.

[9] I. M. Y. A. D. Dala, I. K. G. Putra, and B. P. Wira, "Forecasting Cases of Dengue Hemorrhagic Fever Using the Backpropagation, Gaussians and Support-Vector Machine Methods," *RESTI*, vol. 5, no. 2, pp. 335–341, 2021, doi: https://doi.org/10.29207/resti.v5i2.2936.

[10] R. Sanjudevi and D. Savitha, "DENGUE FEVER PREDICTION USING CLASSIFICATION TECHNIQUES," *IRJET*, vol. 6, no. 2, pp. 558–563, 2019.

[11] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using an Ensemble Machine Learning Method," *J. Imaging*, vol. 6, no. 6, 2020, doi: 10.3390/JIMAGING6060039.

[12] D. C. Yadav and S. Pal, "Thyroid prediction using ensemble data mining techniques," *Int. J. Inf. Technol.*, vol. 14, no. 3, pp. 1273–1283, 2022, doi: 10.1007/s41870-019-00395-7.

[13] T. Sajana, M. Syamala, L. Phaneendra Maguluri, and C. Usha Kumari, "A hybrid approach for classification of infectious diseases," *Mater. Today Proc.*, 2021, doi: 10.1016/j.matpr.2020.11.727.

[14] F. N. Inayah, S. S. Prasetiyowati, and Y. Sibaroni, "Classification of Dengue Hemorrhagic Fever (DHF) Spread in Bandung using Hybrid Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Network Methods," *IJoiCT*, vol. 7, no. 1, pp. 10–20, 2021, doi: 10.34818/ijoict.v7i1.562.

[15] M. Soni, Y. Barot, and S. Gomathi, "A review on Privacy-Preserving Data Preprocessing," *J. Cybersecurity Inf. Manag.*, vol. 5, no. 2, pp. 16–30, 2020, doi: 10.54216/jcim.040202.

[16] M. M. Muzakki and F. Nhita, "The spreading prediction of Dengue Hemorrhagic Fever (DHF) in Bandung regency using K-means clustering and support vector machine algorithm," *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, vol. 0, no. c, pp. 453–458, 2018, doi: 10.1109/ICoICT.2018.8528782.

[17] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.

[18] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.