OPEN ACCESS
ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 10, No. 2, Dec 2024. pp. 225-241
doi: 10.21108/ijoict.v10i2.1026

Class Prediction of Dengue Fever Spread in Bandung Using Decision Tree with Time-Based Feature Expansion

Iqlima Putri Hawa [1,*], Sri Suryani Prasetiyowati [2], Yuliant Sibaroni [3]

[1,2,3]*School of Computing, Telkom University*
Bandung, Indonesia
[*] iqlima@student.telkomuniversity.ac.id

Abstract

In Indonesia, dengue hemorrhagic fever (DHF) has become a serious community health concern due to fluctuating incidence rates influenced by several factors. It requires comprehensive control strategies to prevent the rise of the incidence. This study seeks to classify the future spread of DHF in Bandung City, accompanied by optimal factors that influence the increase in its spread. This study proposes using Decision Tree to predict a classification of DHF spread with implementation of spatial time-based feature expansion. The developed scenario is to build a target class with class prediction model based on the previous time period. From the developed scenario, the selected model has optimal performance to form a class prediction model in the future. The used classes its selves are formed by ranging the incidence rate (IR) into low, medium and high class. The data used includes spatial-temporal information such as population, education level, rainfall, temperature, and blood type from 2017 to 2021. The results obtained show that the performance of Decision Tree using time-based feature expansion is more than 90%, with visual predictions that help identify high risk areas. The contribution of this study is to inform the public and health institution regarding DHF spread for the future and influential factor so that the government can provide policies as early as possible to prevent DHF spread.

Keywords: Incidence Rate, DHF, Prediction, Decision Tree, Feature Expansion, Classification

I. INTRODUCTION

DENGUE Hemorrhagic Fever is one of the diseases concerned due to its constant appearance and fluctuating cases. Based on Ministry of Health's annual data, in 2022 West Java became the province with the highest DHF cases. The number of cases reached 36.608 from a total of 143.266 DHF cases in Indonesia [1]. In 2023 West Java also had a major number of DHF cases, reaching 19.328 from a total of 114.720 DHF cases in Indonesia [2]. As for Bandung, the capital city of West Java, the DHF cases fluctuate throughout each year. Based on several health center data from Bandung between 2017 and 2022, the lowest cases are in 2017 with 1,786 cases and highest in 2022 with 5,205 cases [3]. This data shows how fluctuating and concerning the DHF cases in West Java is, as well as Bandung as its capital city.

Therefore, this research intends to predict and classify DHF incidence in sub districts of Bandung city for 2022, 2023, 2024 and 2025 in a form to contribute in preventing the spread of DHF. To achieve the best result of prediction using a Decision Tree, this research examines the best feature to use based on their accuracy with the technique of feature selection and feature expansion where this dataset are the features from 2017 until 2021, that are separated into several models. The prediction of DHF incidence rate classification for 2022, 2023,

Received on 10 Dec 2024. Revised on 20 Dec 2024. Accepted and Published on 23 Jan 2025

2024 and 2025 are being conducted using the finest model with an outstanding accuracy in order to give an accurate prediction. This research will benefit health sectors and other stakeholders by giving the preemptive solution to minimize DHF incidence by informing features that have major influence on DHF rate and the class prediction that shows the severity of DHF cases in these sub districts.

Although there are a lot of studies regarding DHF incidence prediction, there are rarely studies where the researcher delves on to predict the DHF incidence rate for years ahead using a Decision Tree based on spatial temporal of a certain location. Regarding health issue prediction, these studies [4], [5], and [6] have attempted to make predictions using several methods including Decision Tree. Study [4] examines DHF case prediction using Support Vector Machines (SVM) with accuracy of 84%, K-Nearest Neighbor (KNN) obtaining accuracy of 87%, and Decision Tree (DT) obtaining accuracy of 79%. Implementing hybrid classification approach using the hard voting technique, the three methods of classification accuracy could be improved reaching accuracy of 91%. Study [5] examines prediction of breast cancer prediction using Decision Tree (DT) and Adaptive Boosting (Ad boost), where the dataset is highly imbalanced. To enhance the Decision Tree's performance in identifying malignant observations, Adaptive Boosting is employed. Obtaining accuracy of 92.53% for Adaptive Boosting and 88.80% for Decision Tree. Study [6] examines stroke probability prediction with Decision Tree and Naive Bayes. The study obtained the decision tree by calculating Gini coefficients of each feature to select the division. Each Decision Tree and Naive Bayes model prediction gives accuracy of 88% and 79%. These studies show how Decision Trees could give great accuracy with the right improvement. It also shows that selecting the best method to build the Decision Tree and handle imbalance data are also crucial. In this research, feature selection and feature expansion are being implemented in an attempt for a Decision Tree to produce an accurate prediction.

Regarding spatial-temporal analysis with time-based feature expansion, these studies [7], [8], [9] predict climate and disease spread based on several features that influence the cause. Study [7] uses Naive Bayes to predict classification regarding Bandung City’s dengue fever cases and Java Island’s monthly rainfall distribution by using feature expansion obtaining accuracy more than 97% for both predictions. Study [8] conducts research to predict classification of COVID-19 transmission and dengue fever. Using SVM time-based feature expansion, each class prediction of DHF and COVID-19 transmission obtain accuracy of 90% and 93%. Study [9] conduct a class prediction of DHF spread by implementing Random Forest while also adapting feature expansion, resulting three optimum model three models with accuracy of 97%, 93%, and 93%. These studies show how feature expansion succeeds to improve the accuracy of a model with overall accuracy more than 90%, even better than these studies [4], [5], [6] where the model didn't use time-based feature expansion.

Despite the growing DHF case overtime in West Java especially in Bandung, there is a chance to prevent DHF spread by implementing class prediction. As it is previously shown, studies [4], [5], [6], [7], [8] and [9] succeed in undertaking classification regarding health issues with a satisfying result. The studies also showed that Decision Tree is one of the methods that is oftentimes being used to handle classification with a good result, while the usage of time-based is capable of enhancing the accuracy of the model. In this research, Decision Tree are being chosen for how often this method used for classification task in health-related predictions with a reliable result. Spatial-temporal analysis with time-based feature expansion is being chosen because it captures the interaction between space and time, allowing for more comprehensive understanding of DHF spread patterns in specific location and time period while also improving the accuracy of the Decision Tree model. Although the usage of Decision Tree and time-based feature expansion for class prediction already conducted in several studies, not many studies combining these two methods to give an elaborate analysis regarding class prediction provided with visualization of the certain location and time. This method combination also emerging possibilities of creating more accurate and precise results compared to the other studies. Hence, this research carries out a class prediction of DHF spread using Decision Tree in 30 sub districts of Bandung city for 2022, 2023, 2024 and 2025. Using a dataset from the previous 4 years, time-based feature expansion is being implemented to enhance accuracy. The takeout of this research are the features that influence the DHF spread the most, the accuracy of each model as a comparison to choose the best model, the result of each sub district class prediction and the visualization based on the class prediction to give a better understanding. By this take

out, there is a hope that this research will give a direct solution to inform the public and health institution regarding DHF spread for the future and influential factor so that the government can provide policies as early as possible to prevent DHF spread.

II. LITERATURE REVIEW

Dengue Hemorrhagic Fever is a critical state of dengue fever with symptoms of critical high fever, muscle aches, rash, hemorrhagic episodes, and circulatory shock [10]. The dengue virus was mainly spread by the bites of infected Aedes aegypti mosquitoes, causing the fever [11]. A lot of factors could influence the rise of DHF cases. To start with the environment, DHF cases could spread in tropical or subtropical areas. Since climate change indicators such as rainfall, humidity and temperature can have an impact on Aedes aegypti mosquito breeding, the spread of the virus fluctuates based on the climate change. By 2020, DHF rate increased up to 953,476, which mainly take place in tropical countries [12]. Diversities in a population such as blood type, gender, age, and education create a certain demographic that explains its influence regarding DHF cases in some areas. There are studies that delve into the relation between blood type and dengue fever. Blood type B patients were more likely to contract dengue virus infections than blood type AB patients [13]. Whereas people with blood type O have the worst outcomes in dengue hemorrhagic fever [14]. Several areas in Indonesia such as Kediri [15], Blitar [16], and Ternate [17] show that DHF incidence mostly occurs to men and people with the age range of 5-14 years old.

One of the most commonly used techniques for classifier representation in supervised learning is the Decision Tree. Decision Trees facilitate decision-making by the verification of particular qualities by each node in the decision tree [18]. The test features of each node are separated based on decision tree functions like the Gini index and entropy. An indicator of a criterion to lessen the possibility of misclassification is the Gini Index, also referred to as impurity. Whereas entropy, also referred to as information gain, reveals the degree of disorder in a set. Zero entropy means the points of each target class are equal [19]. There is a study that examines splitting choice of Decision Tree, where it concludes that since the Gini index has less bias concerning influences, it is better than entropy information. [20]. This study [21] uses both Gini index and entropy for their Decision Tree to detect breast cancer. It shows Decision Tree with implementation of feature selection using Gini index obtain accuracy of 87.83%, whereas accuracy using entropy is 86.77%.

Time-based feature expansion, including model selection and combination, is an effective approach for forecasting large sets of time series, with performance varying based on the nature of the time series [22]. Feature expansion is capable of improving the accuracy of classification as opposed to traditional machine learning algorithms for data classification. The cause of this is that classifiers can now take into account multiple dimensions due to the expanded features, which is not possible with low-dimensional data [23]. This study [24] compares two methods to make a classification system to predict the number of Social Welfare Service Recipients (SWSR). In this study, SVM uses time-based feature expansion, resulting accuracy value of 70% and 80%. Outperforming LSTM, which had accuracy values of 34.28% and 48.57%.

III. RESEARCH METHOD

DHF incidence rates in this study are being predicted using Decision Tree. The Decision Tree method predicts the incidence rate of 30 sub districts in Bandung for 4 years ahead. The results of the prediction are being visualized through the map of Bandung's sub districts based on the class. Fig.1 shows the steps that occur in the research to finally obtain the prediction and the spatial temporal in map form according to the predicted results.

A. Dataset

The datasets that are put into practice for this research contain DHF cases, climate, population, educational history, and blood type are gained from several sources such as Public Health Office, Meteorological, Climatological, and Geophysical Agency (BMKG) and Central Bureau of Statistic (BPS) of Bandung. Table I illustrates the description regarding the dataset. The dataset has 13 features that influence DHF cases and incidence rate of sub districts in Bandung between 2017 and 2021.

B. Preprocessing Data

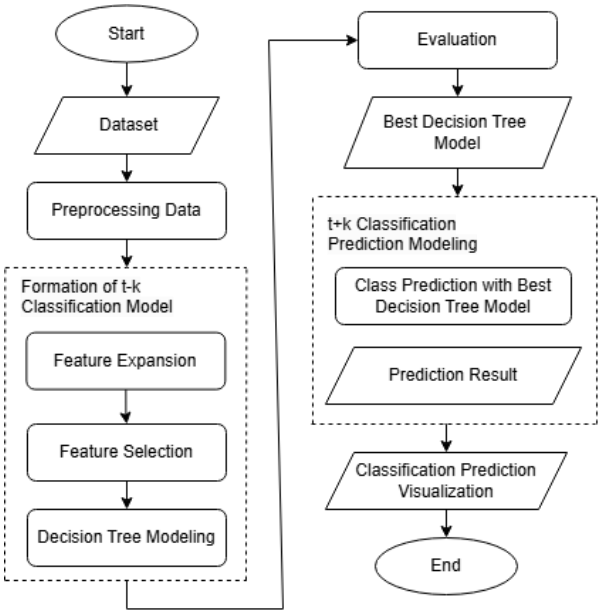


Fig. 1. Research Method

TABLE I
DATASET

Notation	Description
x1	Population
x2	Male Population Proportion
x3	Rainfall
x4	Temperature
x5	Humidity
x6	Blood Type A
x7	Blood Type B
x8	Blood Type AB
x9	Blood Type O
x10	Elementary School Graduates
x11	Junior High School Graduates
x12	High School Graduates
x13	College Graduates
y	Target/Incidence Rate (/100.000 population)

In order to proceed the research, the data that are being used need to be processed, from class labelling, scaling, and solving imbalance data problems. The datasets that are being used have incidence rate as the target class. Incidence rate (IR) shows the frequency of new cases of a health condition, in this case DHF, that specified population during a specific time period to provide insight of how fast the disease spread to a certain population. From Equation 1, it could be seen how the incidence rate equation measures DHF cases per 100,000 population, showing how the incidence rate from the data is being obtained [9]. Since this research intends to classify DHF incidence rate, class labeling is important to group these incidence rates by ranging it and label it into classes as it can be seen from Table II.

$$IR = \frac{Case}{Population} \times 100{,}000 \quad (1)$$

TABLE II
CLASS LABELING

Class	Label Class	Range
Low	0	IR<55
Medium	1	55 ≥ IR ≤ 100
High	2	IR > 100

In this research, normalization is implemented to ensure all the feature values in the dataset are normalized within the range of 0 to 1 using min max scaler. As can be viewed on Equation 2, the process of Min-Max scaling is influenced by the feature's minimum and maximum data values. x and x' indicates the original value and the scaled value, while i and n indicates the index of the dataset and the index of the feature [25] [26].

$$x'_{i,n} = \frac{x_{i,n} - min(x_n)}{max(x_n) - min(x_n)} \quad (2)$$

After examining the data, it turns out the number of certain classes is higher compared to the other class. Which means the dataset is imbalanced and could have an unfavorable impact on classification model performance by causing bias towards the majority class, reducing overall accuracy. This research used random oversampling (ROS) to handle imbalance data by raising the sample size of minority classes [27]. Fig. 2 illustrates how oversampling works.

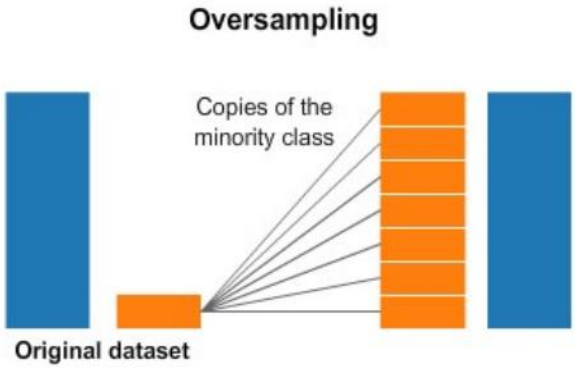


Fig. 2. Oversampling [28]

C. Feature Expansion

The foundation of feature expansion is the modeling of time-based class predictions. Feature expansion involves creating lagged features to capture temporal patterns, enhancing predictive accuracy and providing context for the case. In this research, the lagged features are based on the data frequency or the time interval in the dataset, using yearly time steps. Since this research attempt predict for 2022 $(t+1)$, 2023 $(t+2)$, 2024 $(t+3)$, and 2025 $(t+4)$, this research learns from the previous years. Hence, the time step selection being 1 year before $(t-1)$, 2 years before $(t-2)$, 3 years before $(t-3)$ and 4 years before $(t-4)$. The goal of this study is to use the best outcome of a combination of earlier $t-k$ models to build a predictive model of categorization for future $t+k$. This concept allows for the establishment of a classification model derived from the prior $t-k$ features, with a yt class target [29]. Table III explains how feature expansion in this research works, by using previous features in order to predict the class, using the target class to be studied by the model.

TABLE III
IMPLEMENTATION OF FEATURE EXPANSION

Time Step	Model	Data Feature	Target Class
1 year before	1A	2017	2018
	1B	2018	2019
	1C	2019	2020
	1D	2020	2021
2 years before	2A	2017, 2018	2019
	2B	2018, 2019	2020
	2C	2019, 2020	2021
3 years before	3A	2017, 2018, 2019	2020
	3B	2018, 2019, 2020	2021
4 years before	4A	2017, 2018, 2019, 2020	2021

D. Feature Selection

In order to choose the best model, it is necessary to select the best combination feature using SelectKbest with 'f_classif', since there are a lot of possible feature combinations as can be seen in Table IV. SelectKBest is a feature selection algorithm that identifies the most relevant features from a given dataset [30]. These selected combinations are tested by being implemented in the decision tree so the accuracy of each feature combination could be examined. Models with feature combinations that obtain the highest accuracy, which indicates optimal performance, are being chosen to predict the future classification.

TABLE IV
EXAMPLE OF MODEL 3B POSSIBLE FEATURE COMBINATIONS

Year	Data Feature	Feature
3	2018, 2019, 2020	xa2, xa3, xb10
3	2018, 2019, 2020	xa4, xb8, xc13
3	2018, 2019, 2020	xa1, xc6, xc9
4	2018, 2019, 2020	xa1, xa4, xb3, xc10
4	2018, 2019, 2020	xa1, xb9, xb13, xc2
…	…	…
39	2018, 2019, 2020	xa1, xa2, xa3, ..., xc13

E. Decision Tree Modeling

Decision Tree in classification is a method that applies distinct decision variables to represent path that every observation will follow in the tree, which aims to improve classification accuracy [31]. Nodes and branches are

the structure of each tree. Every node signifies a feature of the category being classified, while each subset specifies the possible values the node can assume [32]. This research uses Gini index as a crucial indicator of a data node's "impurity," enabling the algorithm to select the most suitable feature to split into at every node with one minimum samples leaf and two minimum samples split. The illustration of a decision tree could be seen on Fig. 3.

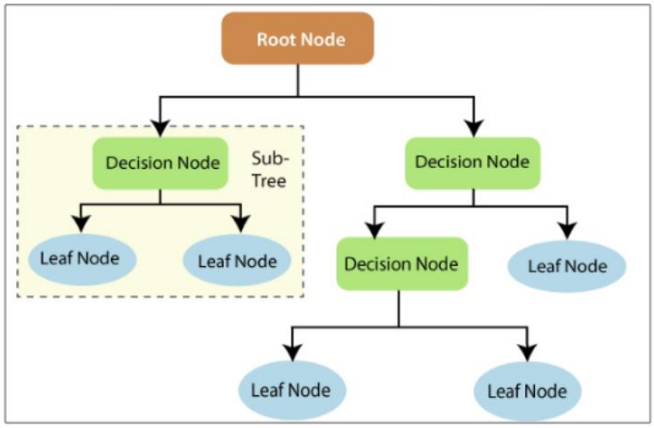


Fig. 3. Illustration of Decision Tree [31]

Gini index is a splitting criterion used in decision trees to measure dataset or node impurity and improve classification accuracy by determining the best feature for splitting the data at each node. The Gini index ranges from zero to one, zero signifies the data is pure (all instances in the dataset belong to a single class). One signifies the data is completely impure, with classes being evenly distributed across the dataset. Gini index in decision trees works by measuring data impurity to create splits that result in more homogeneous subsets, thereby improving classification accuracy. Gini chooses the minimum value for choosing the root node and for every decision we do it again on all features [33]. Equation 3 and 4 show how the Gini index works in a decision tree, with L signifying a dataset containing j distinct class labels. Where p_i represents the relative frequency of class i in L. When the dataset is divided based on attribute A into two subsets L_1 and L_2, with sizes N_1 and N_2 , GINI is computed using Equation 3 and the impurity reduction is determined using Equation 4.

$$GINI(L) = 1 - \sum_{i=1}^{j} {p_i}^2 \tag{3}$$

$$GINI_A(L) = \frac{N_1}{N} GINI(L_1) + \frac{N_2}{N} GINI(L_2) \tag{4}$$

The implementation of Decision Tree on the dataset begins at the root node, where the model finds the best feature and threshold to split the data into two subsets that minimize the Gini index by evaluating all possible splits (thresholds) and calculates the Gini index for each resulting subset using Equation 3. After the splits, Equation 4 is used to calculate the weighted Gini index to measure the quality of the splits. The feature and threshold used for the split are those that produce the lowest Gini index. The decision tree continues these splitting procedures for each node as recursive partitioning until the nodes are pure or certain criteria are satisfied, such as maximum tree depth or minimum sample size in leaf nodes. The prediction is performed then by traversing from the root to leaves. The classification for each node is then being determined by the majority class of the samples in that leaf node.

F. Evaluation

Confusion Matrix evaluates the accuracy of an ML model using a set of previously known target data. In addition, several other metrics, including sensitivity, specificity, precision, and f1 score, are generated related to this matrix [34]. The classification model with 3 classes in this research is evaluated based upon accuracy, f1 score, recall, and precision calculated from the values in the Confusion Matrix that can be seen on Fig. 4 using Equations 5, 6, 7, 8. These equations show the precision, recall, and f1 score for class C with total class q [29]. Each TP_c, TP_c, and FN_c are the number of TP, FP, and FN classifier's predictions for class C [35]. This research used standard k-fold cross validation with cross validation of 5-fold to evaluate the performance of a model.

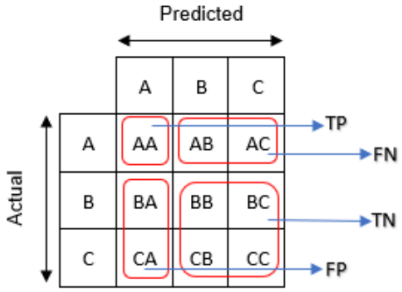


Fig. 4. Confusion Matrix for 3 Class [36]

$$Accuracy = \frac{\sum_{c=1}^{q} TP_c}{(TP_c + FP_c + FN_c + TN_c)} \quad (5)$$

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (6)$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (7)$$

$$F1\ score_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FN_c + FP_c} \quad (8)$$

IV. RESULTS AND DISCUSSION

A. Result

After selecting the feature and checking the accuracy, this research obtained the average values of confusion matrix of each model starting from Model 1A with time step of one year until Model 4A with time step of four years. These average values of confusion matric illustrate the performance of each model with different target class and data features the model studied when operating with the method of decision tree. The visualized result of each model confusion matrix can be viewed on Fig. 5, Fig. 6, Fig. 7, and Fig. 8.

Fig.5 shows the confusion matrix average values for Model 1 that visualized the average values of Model 1 accuracy, f1 score, precision and recall. Fig.6 shows the confusion matrix average values for Model 2 that visualized the average values of Model 2 accuracy, f1 score, precision and recall. Fig.7 shows the confusion matrix average values for Model 3 that visualized the average values of Model 3 accuracy, f1 score, precision and recall.

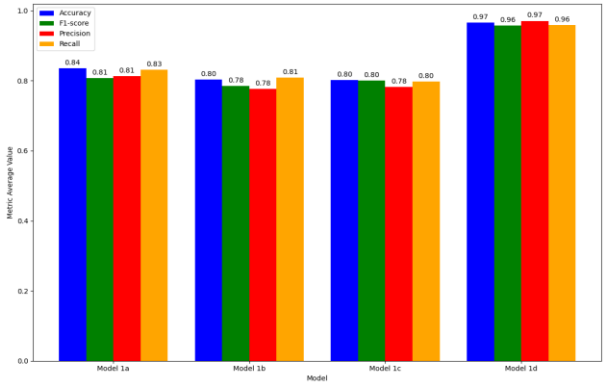


Fig. 5. Confusion matrix values in average of Model 1.

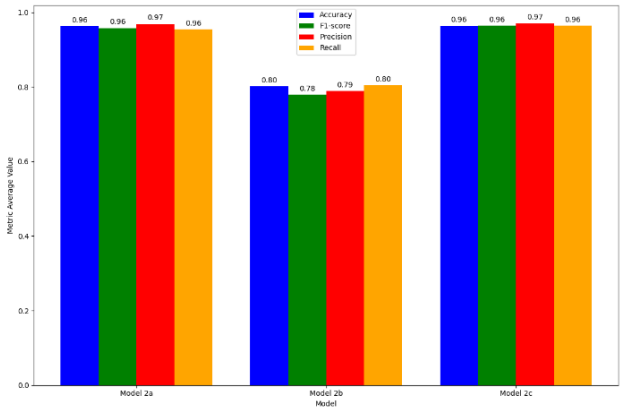


Fig. 6. Confusion matrix values in average of Model 2.

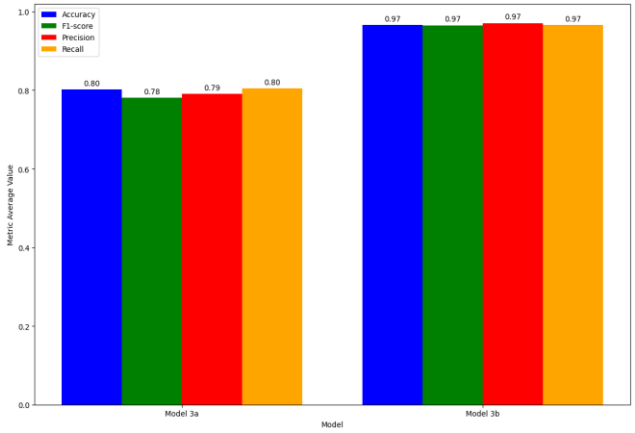


Fig. 7. Confusion matrix values in average of Model 3.

Fig.8 shows the confusion matrix average values for Model 4 that visualized the average values of Model 4 accuracy, f1 score, precision and recall. The selected model, as can be seen in Table V, has been concluded after the accuracy of each model with selected features are being examined. Table VI shows $t + k$ class prediction of DHF spread in Bandung sub district.

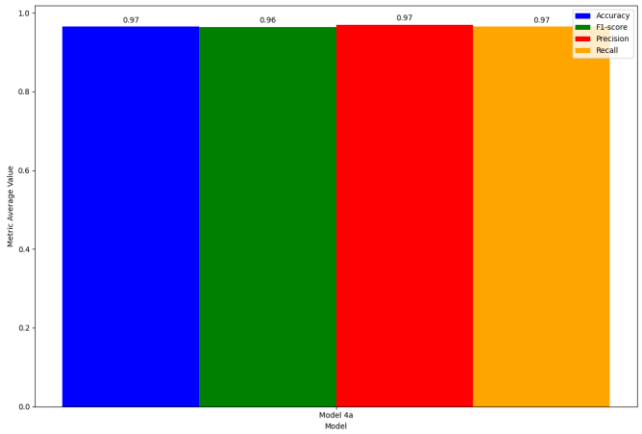


Fig. 8. Confusion values in average Values of Model 4.

TABLE V
SELECTED T-K MODEL AND FEATURE COMBINATION.

Selected Model	Selected Feature	Selected Target Class	Accuracy	F1 Score	Precision	Recall
1B	'xa1', 'xa2', 'xa3', 'xa4', 'xa6', 'xa8', 'xa9', 'xa10', 'xa11', 'xa12', 'xa13'	2019	0.90	0.80	0.87	0.89
2A	xa1, xa10, xa11, xb1, xb3, xb4, xb10, xb11	2019	0.98	0.96	0.92	0.98
3B	xa1, xa10, xa11, xb1, xb10, xb11, xc1, xc10, xc11, xc13	2021	0.96	0.96	0.96	0.96
4A	xa1, xa10, xa11, xa13, xb1, xb10, xb11, xb13, xc1, xc10, xc11, xc13, xd1, xd6, xd10, xd11, xd13	2021	0.96	0.96	0.96	0.96

TABLE VI
T+K CLASS PREDICTION OF BANDUNG SUB DISTRICT.

Sub District	t+k Class Prediction			
	2022	2023	2024	2025
Andir	1	1	2	2
Antapani	2	2	2	2
Arcamanik	2	2	2	2
Astana Anyar	2	2	2	2
Babakan Ciparay	1	1	1	1
Bandung Kidul	2	2	2	2
Bandung Kulon	1	1	1	1
Bandung Wetan	2	2	2	2
Batununggal	2	2	2	2
Bojongloa Kaler	2	2	2	2
Bojongloa Kidul	2	2	2	2
Buahbatu	2	2	2	2
Cibeunying Kaler	2	2	2	2

Cibeunying Kidul	2	2	2	2
Cibiru	2	2	2	2
Cicendo	2	2	2	2
Cidadap	2	2	2	2
Cinambo	2	2	2	2
Coblong	2	2	2	2
Gedebage	2	2	2	2
Kiaracondong	2	2	2	2
Lengkong	2	2	2	2
Mandalajati	2	2	2	2
Panyileukan	2	2	2	2
Rancasari	2	2	2	2
Regol	2	2	2	2
Sukajadi	2	2	2	2
Sukasari	2	2	2	2
Sumur Bandung	2	2	2	2
Ujung Berung	2	2	2	2

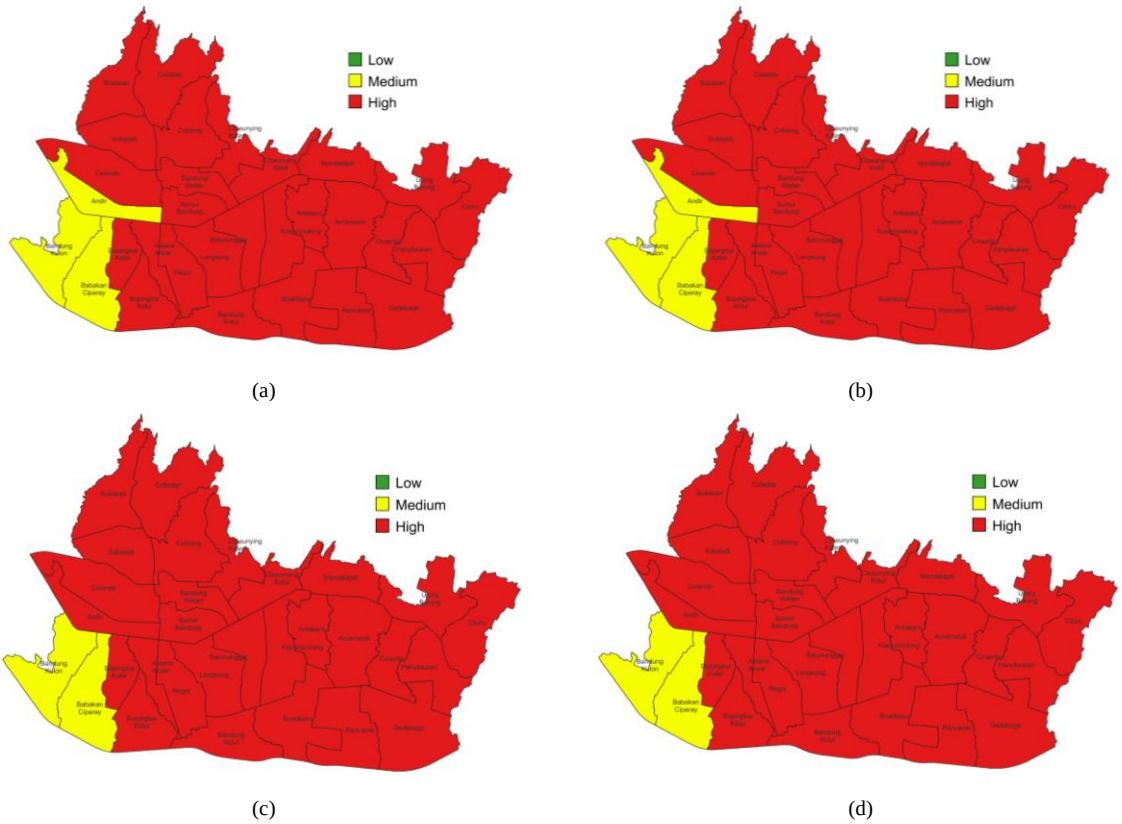
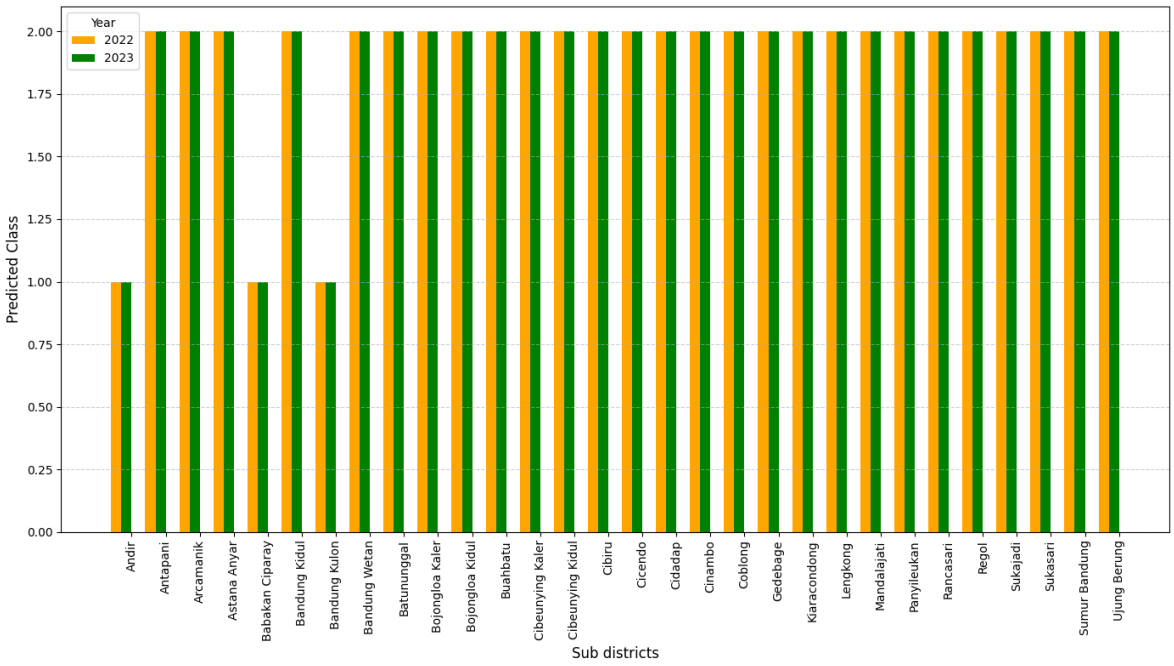
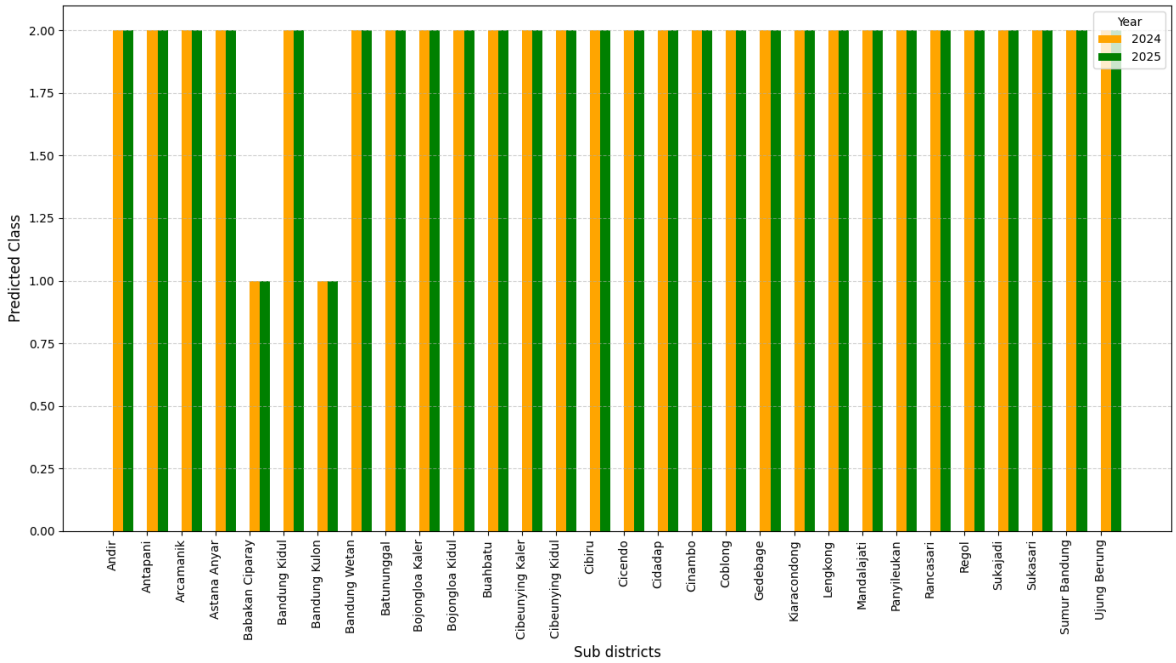


Fig.9 Visualization of DHF Incidence Rate Class Prediction in 4 years (a) 2022 (b) 2023 (c) 2024 and (d) 2025

In order to give more comprehend understanding regarding the class prediction of DHF spread in Bandung, we visualized the result of $t + k$ class prediction. Fig. 9 and Fig. 10 shows class prediction of DHF spread for each year of 2022 until 2025 in the form of maps and subplots.



(a)



(b)

Fig. 10. Subplots of DHF Incidence Rate Class Prediction in four years, (a) 2022 and 2023, (b) 2024 and 2025.

B. Discussion

Decision Tree analyzing the performance of each model is necessary to acquire an outstanding model that could produce an accurate prediction. Average confusion matrix values in this research are meant to explain the performance of each model that differentiates by the features and the target class that they are studied. Accuracy in a confusion matrix reflects the model's overall performance across all classes, represented as the percentage of correct predictions out of the total predictions made. F1 score in confusion matrix provides a fair evaluation of recall and precision, providing information of how well the model balances recall and precision. Precision in confusion matrix measures the proportion of correctly predicted positive cases (true positives) out of all cases that the model predicted as positive (sum of true positives and false positives), it focuses on how accurate the model's positive predictions are. Recall in confusion matrix measures the proportion of actual positive cases (true positives) that were correctly predicted as positive by the model (sum of true positives and false negatives), it focuses on how well the model captures all actual positive cases. Model 1 average values can be seen through Fig.5, where Model 1D has the best average values. Obtaining each of the average values of 0.97 accuracy, 0.96 f1 score, 0.97 precision and 0.96 recall. Model 2 average values can be seen through Fig.6, where both Model 2A and 2C have equally satisfactory average values. Obtaining each of the average values of 0.96 accuracy, 0.96 f1 score, and 0.97 precision and 0.96 recall. Model 3 average values can be seen through Fig.7, where Model 3B has the best average values. Obtaining each of the average values of 0.97 accuracy, 0.97 f1 score, and 0.97 precision and 0.97 recall. Model 4 average values can be seen through Fig.8, where it only has 1 model namely Model 4A. Obtaining each of the average values of 0.97 accuracy, 0.96 f1 score, and 0.97 precision and 0.97 recall. It can be said that these models perform exceptionally well because all of the average values are higher than 90. When it comes to accuracy, the models predict most cases correctly for every class. In terms of precision, they minimize false positives while successfully identifying the positive class. The models reduce false negatives by capturing the majority of true positive cases in terms of recall. Lastly, the model's ability to reduce false positives and false negatives is demonstrated by the f1 score, which shows a successful balance between precision and recall.

The selected model as shown in Table V, has been concluded after the accuracy of each model with selected features are being examined. The result shows that Model 1B is suitable to predict classification for 2022 with target class of 2019 and 11 features in total, resulting accuracy of 0.90. The model used selected feature of population, male population proportion, rainfall, temperature, blood type A, blood type AB, blood type O, elementary school graduates, junior high school graduates, high school graduates, and college graduates from 2018. Model 2A is suitable to predict classification for 2023 with target class of 2019 and 8 features in total, resulting accuracy of 0.98. The model used selected features of population, elementary school graduates, and junior high school graduates from 2017. While features from 2018 consist of population, rainfall, temperature, elementary school graduates, and junior high school graduates. Model 3B is suitable to predict classification for 2024 with target class of 2021 and 10 features in total, resulting accuracy of 0.96. The model used selected features of population, elementary school graduates and junior high school graduates from 2018, 2019,2020 and college graduates feature from 2020. Model 4A is suitable to predict classification for 2024 with target class of 2021 and 17 features in total, resulting accuracy of 0.96. The model used selected features of population, elementary school graduates, junior high school graduates, and college graduates from 2017, 2018, 2019, 2020 and blood type A from 2020. From these selected features and models, Model 2A with the finest accuracy of 0.98, compared to Model 3B and 4A with difference of 2% and Model 1B with difference of 8%. Looking from the same pattern of features in Model 3B and Model 4A that mainly the features are population, elementary school graduates, junior high school graduates and college graduates and both target class are 2021 seems to influence both models having the same accuracy of 0.96.

Finally, the class prediction of the spread of DHF using the Decision Tree as the method has been obtained as in Table VI, after selecting the feature and model to produce the accurate prediction. Later than, class prediction of DHF spread based on the sub district in Bandung are being visualized. It could be seen from Fig. 9 (a), and Fig. 9 (b), that prediction for 2022 and 2023 has the same visualization due to the same class target of 2019 that are being studied when selecting the best feature and model. This is also applied to prediction 2024

and 2025 from Fig. 9 (c) and Fig. 9 (d) for having the same visualization due to having the same class target of 2021 that are being studied when selecting the best feature and model. For 2022 and 2023, it seems that the DHF spread with medium intensity consist of Andir, Babakan Ciparay and Bandung Kulon. As for 2024 and 2025, the DHF spread with medium intensity consist of Babakan Ciparay and Bandung Kulon.

In order to comprehend and evaluate the effect of using different methods and preprocessing, this research examine the comparison of this research with Decision Tree without time-based feature expansion [37], also comparing research of time-based feature expansion with different methods such as Random Forest [9]. Study [37] uses Decision Tree and entropy as its measurement for potential split of each node with cross validation of 10-fold and obtaining the optimum accuracy of 87.72%. With time-based feature expansion, this research model can reach an accuracy value of 0.98 with Model 2A. Study [9] uses Random Forest as the method to predict DHF incidence rate. There are technical differences between this research and the one using Random Forest [9]. This research uses Random Oversampling (ROS) to handle imbalance data and standard k-fold cross validation of 5-fold, while study [9] using stratified k-fold cross validation with cross validation of 10-fold. This research resulted Model 2A as the best model with accuracy of 0.98, using features from 2017 and 2018. While study [9] resulted Model 2C with the highest accuracy of 96,67%, using features from 2017 and 2018. From this comparison it could be concluded that although using different methods and results, data from 2017 and 2018 seems more significant based on accuracy. Another pattern that could be seen from this research and study [9] is when visualizing the prediction for the future onto the map. Which is how the target classes that are being studied by the model could affect the prediction. In study [9], the visualization of predictions for 2023 and 2024 are the same. While the visualization for 2022 is different from the rest. This is caused by the model learning the same target class, which is target class of 2021. For this research, the visualization of prediction for 2022 and 2023 are the same due to target class studied by both models 2019. While the visualization for 2024 and 2025 are the same due to target class studied by both models are 2021.

V. CONCLUSION

Class prediction using time-based feature expansion succeed to perform class prediction of DHF spread in Bandung sub district for the future. The models are being developed by expanding feature from previous time of $t-k$ and target class of $t+k$. Using dataset of DHF spread in Bandung sub district from 2017 until 2021, the implementation of time-based feature expansion using decision tree shows a satisfying result of accuracy value up to more than 90%. Models with the finest accuracy are being achieved with feature expansion from previous time period.

ACKNOWLEDGMENT

The author wishes to thank Telkom University and both lecturers for the chance, support and funds that are being given. As a result, this author could finish this research.

REFERENCES

[1] "Ministry of Health of the Republic of Indonesia. 2023. Profil Kesehatan Indonesia 2022. Accessed from https://kemkes.go.id/id/profil-kesehatan-indonesia-2022.."

[2] "Ministry of Health of the Republic of Indonesia. 2024. Profil Kesehatan Indonesia 2023. Accessed from https://kemkes.go.id/id/profil-kesehatan-indonesia-2023.".

[3] "Jumlah Kasus Demam Berdarah Dengue (DBD) Menurut Puskesmas di Kota Bandung (2016-2022)".

[4] A. Rahman and S. S. Prasetiyowati, "Performance Analysis of the Hybrid Voting Method on the Classification of the Number of Cases of Dengue Fever," *International Journal on Information and Communication Technology (IJoICT)*, vol. 8, no. 1, pp. 10–19, Jul. 2022, doi: 10.21108/ijoict.v8i1.614.

[5] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, p. 184, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp184-190.

[6] Y. Liu, C. Zhang, X. Zheng, Y. Liu, and J. He, "Stroke Prediction Model Based on Decision Tree," *WSEAS TRANSACTIONS ON BIOLOGY AND BIOMEDICINE*, vol. 20, pp. 24–27, Mar. 2023, doi: 10.37394/23208.2023.20.3.

[7] S. S. Prasetiyowati and Y. Sibaroni, "Unlocking the potential of Naive Bayes for spatio temporal classification: a novel approach to feature expansion," *J Big Data*, vol. 11, no. 1, p. 106, Aug. 2024, doi: 10.1186/s40537-024-00958-x.

[8] R. N. Nurriadi, S. S. Prasetiyowati, and Y. Sibaroni, "Performance of Time-Based Feature Expansion Classification Method for Predicting Disease Spread," in *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, IEEE, Feb. 2024, pp. 120–125. doi: 10.1109/ICoSEIT60086.2024.10497460.

[9] Elqi Ashok, Sri Suryani Prasetiyowati, and Yuliant Sibaroni, "DHF Incidence Rate Prediction Based on Spatial-Time with Random Forest Extended Features," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 612–623, Aug. 2022, doi: 10.29207/resti.v6i4.4268.

[10] S. Parveen *et al.*, "Dengue hemorrhagic fever: a growing global menace," *J Water Health*, vol. 21, no. 11, pp. 1632–1650, Nov. 2023, doi: 10.2166/wh.2023.114.

[11] L. R. Alma *et al.*, "Rapid Survey of Mosquito Larvae Density with a History of Dengue Hemorrhagic Fever," *E3S Web of Conferences*, vol. 448, p. 05003, Nov. 2023, doi: 10.1051/e3sconf/202344805003.

[12] S. N. Ramadhani and Mohd. T. Latif, "Impact of Climate Change on Dengue Hemorrhagic Fever (DHF) in Tropical Countries: A Literature Review," *JURNAL KESEHATAN LINGKUNGAN*, vol. 13, no. 4, p. 219, Oct. 2021, doi: 10.20473/jkl.v13i4.2021.219-226.

[13] N. Iqbal, M. A. R. Afridi, Z. Ali, and A. Rafiq, "Association of different blood groups in patients with Dengue fever and their relationship with the severity of the illness," *Pak J Med Sci*, vol. 39, no. 5, Jul. 2023, doi: 10.12669/pjms.39.5.7275.

[14] M. R. Hashan, S. Ghozy, A. E. El-Qushayri, R. H. Pial, M. A. Hossain, and G. M. Al Kibria, "Association of dengue disease severity and blood group: A systematic review and meta-analysis," *Rev Med Virol*, vol. 31, no. 1, pp. 1–9, Jan. 2021, doi: 10.1002/rmv.2147.

[15] F. Amalia Febrianti, E. Qurniyawati, M. Atoillah Isfandiari, and N. Mohamed Gomaa Nasr, "AN EPIDEMIOLOGICAL OVERVIEW OF DENGUE HEMORRHAGIC FEVER (DHF) CASES IN KEDIRI REGENCY DURING 2017-202," *Jurnal Berkala Epidemiologi*, vol. 11, no. 3, pp. 215–223, Sep. 2023, doi: 10.20473/jbe.V11I32023.215-223.

[16] E. T. Suryani, "Profile of Dengue High Fever in Blitar City at 2015-2017," *Jurnal Berkala Epidemiologi*, vol. 6, no. 3, p. 260, Dec. 2018, doi: 10.20473/jbe.V6I32018.260-267.

[17] S. Tomia, U. K. Hadi, S. Soviana, and E. B. Retnani, "EPIDEMIOLOGY OF DENGUE HEMORRHAGIC FEVER CASES IN TERNATE CITY, NORTH MOLUCCAS," *Jurnal Veteriner*, vol. 21, no. 4, pp. 637–645, Dec. 2020, doi: 10.19087/jveteriner.2020.21.4.637.

[18] K. N. Babar, "Performance Evaluation of Decision Trees with Machine Learning Algorithm," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 05, pp. 1–5, May 2024, doi: 10.55041/IJSREM34179.

[19] D. Saha and A. Manickavasagan, "Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review," *Curr Res Food Sci*, vol. 4, pp. 28–44, 2021, doi: 10.1016/j.crfs.2021.01.002.

[20] X. Zhao and X. Nie, "Splitting Choice and Computational Complexity Analysis of Decision Trees," *Entropy*, vol. 23, no. 10, p. 1241, Sep. 2021, doi: 10.3390/e23101241.

[21] F. K. Nasser and S. F. Behadili, "Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers," *Iraqi Journal of Science*, pp. 4987–5003, Nov. 2022, doi: 10.24996/ijs.2022.63.11.34.

[22] L. Li, F. Li, and Y. Kang, "Forecasting large collections of time series: feature-based methods," Sep. 2023.

[23] D. Jung, J. Lee, and H. Park, "Feature expansion of single dimensional time series data for machine learning classification," in *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, IEEE, Aug. 2021, pp. 96–98. doi: 10.1109/ICUFN49451.2021.9528690.

[24] F. Y. Pritama, S. Suryani Prasetiyowati, and Y. Sibaroni, "Enhancing SVM Performance for Time-Based Classification Prediction Through Feature Expansion: A Comparative Analysis with LSTM," in *2024 12th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Aug. 2024, pp. 43–49. doi: 10.1109/ICoICT61617.2024.10698663.

[25] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl Soft Comput*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[26] B. G. Chepino, R. R. Yacoub, A. Aula, M. Saleh, and B. W. Sanjaya, "EFFECT OF MINMAX NORMALIZATION ON ORB DATA FOR IMPROVED ANN ACCURACY," *Journal of Electrical Engineering, Energy, and Information Technology (J3EIT)*, vol. 11, no. 2, p. 29, Aug. 2023, doi: 10.26418/j3eit.v11i2.68689.

[27] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," *International Journal of Artificial Intelligence Research*, vol. 4, no. 2, p. 86, Jan. 2021, doi: 10.29099/ijair.v4i2.152.

[28] E. Saputro and D. Rosiyadi, "Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes," *Bianglala Informatika*, vol. 10, no. 1, pp. 42–47, Mar. 2022, doi: 10.31294/bi.v10i1.11739.

[29] S. S. Prasetiyowati, A. Yahya, and A. A. Rochmawati, "Performance of Time-Based Feature Expansion in Developing ANN Classification Prediction Models on Time Series Data," *Intl. Journal on ICT*, vol. 9, no. 2, pp. 162–176, 2023, doi: 10.21108/ijoict.v9i2.868.

[30] D. P. -, S. N. -, A. A. C. -, and A. G. -, "Enhancing Heart Disease Prediction Through KBEST-PCA Fusion Feature Selection and Ensemble Modeling With Gaussian Naive Bayes Boosting," *International Journal For Multidisciplinary Research*, vol. 5, no. 4, Jul. 2023, doi: 10.36948/ijfmr.2023.v05i04.4378.

[31] R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales, "Optimal randomized classification trees," *Comput Oper Res*, vol. 132, p. 105281, Aug. 2021, doi: 10.1016/j.cor.2021.105281.

[32] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[33] S. Tangirala, "Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020, doi: 10.14569/IJACSA.2020.0110277.

[34] S. Mokhtari, K. K. Yen, and J. Liu, "Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning," *Int J Comput Appl*, vol. 183, no. 7, pp. 1–8, Jun. 2021, doi: 10.5120/ijca2021921347.

[35] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[36] K. Wabang, Oky Dwi Nurhayati, and Farikhin, "Application of The Naïve Bayes Classifier Algorithm to Classify Community Complaints," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 872–876, Nov. 2022, doi: 10.29207/resti.v6i5.4498.

[37] M. A. Rosid, A. S. Fitrani, Y. Findawati, S. Winata, and V. A. Firmansyah, "Classification Of Dengue Hemorrhagic Disease Using Decision Tree With Id3 Algorithm," *J Phys Conf Ser*, vol. 1381, no. 1, p. 012039, Nov. 2019, doi: 10.1088/1742-6596/1381/1/012039.