Title:

Determining important features for dengue diagnosis using feature selection methods

Authors:

Yulianti Paula Bria[1], Paskalis Andrianus Nani[1], Yovinia Carmeneja Hoar Siki[1], Natalia Magdalena Rafu Mamulak[1], Emiliana Metan Meolbatak[1], Robertus Dole Guntur[2]

[1] Universitas Katolik Widya Mandira, Jl. San Juan No. 1 Penfui Timur, Kabupaten Kupang, Nusa Tenggara Timur 85361, Indonesia

[2] Universitas Nusa Cendana, Jl. Adisucipto Penfui, Kupang, Nusa Tenggara Timur 85001, Indonesia

Corresponding author:

Yulianti Paula Bria, Ph.D

Universitas Katolik Widya Mandira

Jl. San Juan No. 1 Penfui Timur, Kabupaten Kupang, Nusa Tenggara Timur 85361, Indonesia

Email: yulianti.bria@unwira.ac.id

Abstract

Objectives: This research aims to determine the important features including symptoms and risk factors for dengue diagnosis.

Methods: The dataset for this study is in the form of medical records collected from two hospitals in East Nusa Tenggara Province including Kewapante and Soe hospitals. Feature selection methods including feature importance, recursive feature elimination, correlation matrix from Pearson's correlation coefficient and KBest were leveraged to determine important features. Important features were also gathered from fifteen Indonesian medical doctors to confirm the results. To obtain the best significant features for dengue prediction, we used six machine learning techniques including logistic regression, k-nearest neighbors, eXtreme gradient boosting, random forests, Naïve Bayes and support vector machines.

Results. The random forest classifier yields the highest accuracy for the best combination of features with the accuracy of 0.93 (LR: 0.90 (0.04), KNN: 0.89 (0.04), XGBoost: 0.91 (0.03), RF: 0.93 (0.04), NB: 0.88 (0.09), SVM: 0.89 (0.04)) and precision of 0.90 (LR: 0.86 (0.22), KNN: 0.67 (0.14), XGBoost: 0.77 (0.13), RF: 0.90 (0.13), NB: 0.66 (0.20), SVM: 0.66 (0.18)). This study shows the significant features for dengue diagnosis including fever, fever duration, headache, muscle and joint pain, nausea, vomiting, abdominal pain, shivering, malaise, loss of appetite, shortness of breath, rash, bleeding nose, bitter mouth, temperature and age.

Conclusions. This beneficial information can help society in differentiating dengue from non-dengue diseases including malaria, typhoid fever, COVID-19 and other dengue-like symptoms diseases. This is pivotal to educate society to seek medical advice when dengue symptoms appear.

Keywords: Dengue fever, Feature selection, Significant dengue features, Dengue prediction, Dengue diagnosis

INTRODUCTION

Dengue infection is a life-threatening disease spread by female mosquitos, *Aedes aegypti*. This disease is one of the most prevalent diseases in many countries including Indonesia. In 2022, Indonesia contributed to 143,266 dengue cases with the mortality rate in the same year with 1,237 people [1]. Based on the report from Ministry of Health of Indonesia in the week-19 2023, Indonesia had 31,380 dengue cases which claimed 246 people [1]. This indicates that dengue eradication must be prioritized by the government and society without ignoring other priority health problems such as tuberculosis, malaria, stunting, etc.

Early-stage dengue diagnosis is challenging since dengue shares similar symptoms to other diseases including malaria, typhoid fever, and even COVID-19. Malaria, for example, shares the same symptoms with dengue fever such as fever, nausea, vomiting and headache [2].

Some countries have their own identified symptoms for dengue fever. Australia, for example, defines the combination of fever, headache, arthralgia, myalgia, rash, nausea and vomiting as dengue symptoms [3]. Whereas, Singapore uses the combination of fever, headache, backache, myalgia, rash, abdominal discomfort and thrombocytopenia for dengue symptoms [3]. In Indonesia, medical doctors refer to Dengue guideline for diagnosis, treatment, prevention and control [4] dan comprehensive guidelines for prevention and control of Dengue and Dengue Haemorrhagic Fever [5]. The guideline for dengue diagnosis and treatment is issued by Ministry of Health of Indonesia, which is used as a reference for medical personnel [6]. This is adopted from WHO dengue case classification [4].

The number of deaths in East Nusa Tenggara (NTT) province from dengue cases in 2022 was 29 out of 3,376 cases [7]. These cases spread all over NTT's districts. Most of the death cases were because of the severe conditions. People often visit the nearest medical centre when they identify rash or severe conditions because of the lack of knowledge of dengue symptoms and risk factors [8]. Understanding important features of dengue is beneficial to avoid the progression to severe condition, which can avoid death. This information is helpful to seek medical advice as soon as dengue symptoms appear. The

important features are pivotal to develop early-stage dengue detection tools to assist in dengue diagnosis from other dengue-like symptoms diseases such as malaria, typhoid fever and even COVID-19.

Even though there are some guidelines used to diagnose and treat dengue [4,5], different countries have different symptoms [3]. Therefore, it is essential to identify first significant symptoms that contribute most for dengue prediction in Indonesia, which will be done in this study. This study will also use the combination of symptoms and dengue risk factors that contribute most for dengue diagnosis.

To obtain significant features from datasets, we use feature selection methods. Feature selection methods are often used to minimize the number of input variables that are considered to be the most significant to a machine learning model to improve the model performance [9,10]. In recent years, numerous publications focus on the implementation of feature selection methods for disease prediction [9–13]. In the classification stage, most researchers use machine learning techniques such as BayesNet [9,10,13], support vector machine [9,11] and tree-based classifiers [9,10,13].

The use of feature selection for dengue fever has been implemented successfully by Ramasami et al. [14]. They focus on applying feature selection process and relative analysis to enhance the performance of dengue prediction models. In Indonesia, dengue prediction research has been focused on the use of machine learning techniques for predicting the dengue outbreak [15], predicting number of dengue incidents [16–18], forecasting model for dengue fever [19], and focusing spatial modelling for dengue fever [20–22]. To the best of our knowledge, this study is the first study to elaborate some feature selection methods to determine significant features for dengue diagnosis based on medical records collected. The results will be compared with the knowledge gathered from the fifteen Indonesian medical doctors' knowledge to confirm the results. This research also aims to provide important symptoms and factors for malaria diagnosis in Indonesia.

METHODS

Flow chart of Study

Figure 1 shows the approach to obtain significant features for dengue diagnosis.

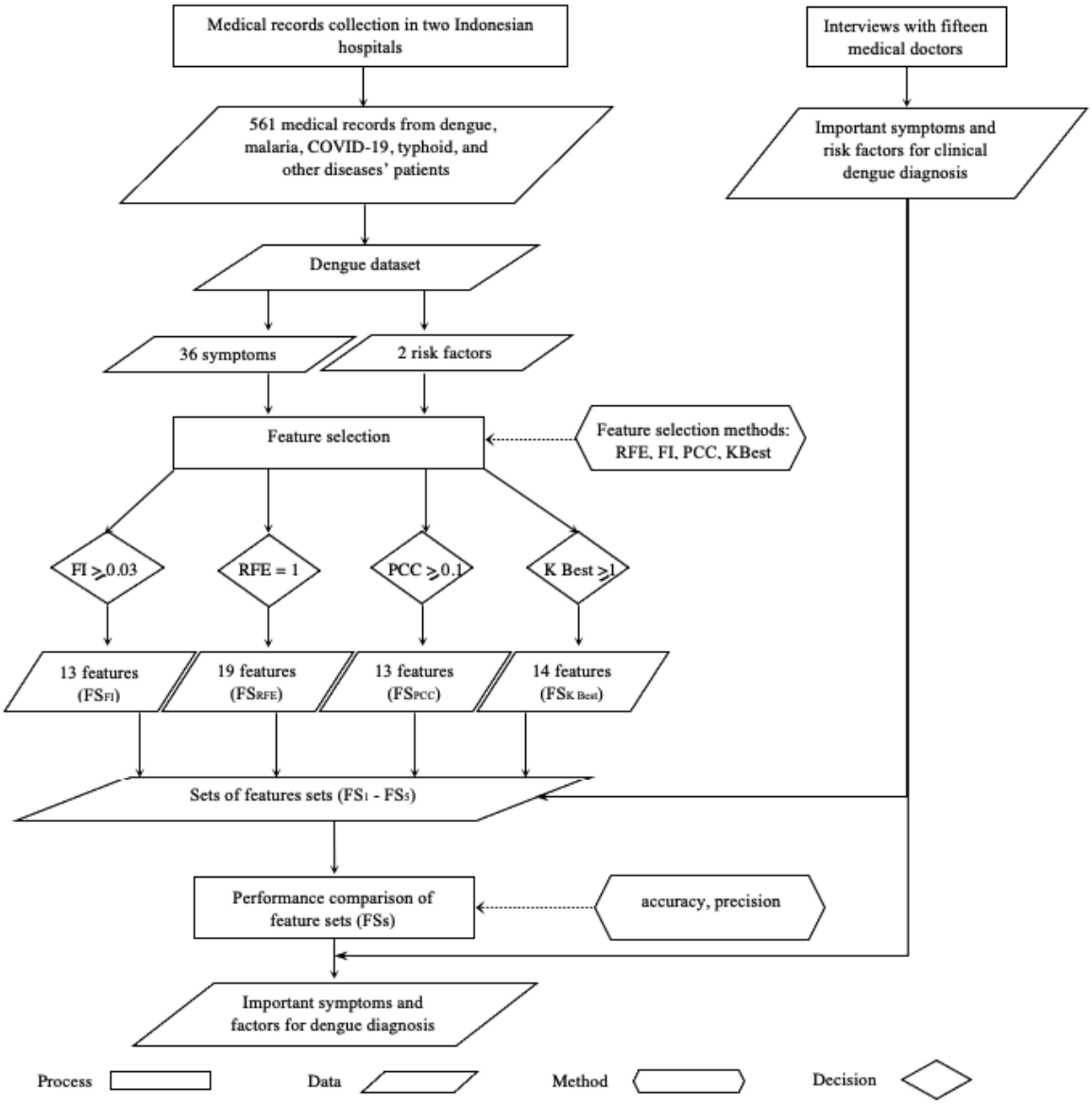

Figure 1. The approach for determining important features for dengue diagnosis

Data collection – medical records collection

To obtain the dengue dataset, we conducted the data collection in two hospitals in Kewapante Hospital, Maumere in Sikka District and Soe Hospital in South Central Timor District of NTT Province. Medical records were collected in the department of medical records of each hospital after obtaining the data

collection approvals from the hospital directors in each hospital. Medical records of patients diagnosed with dengue fever or other dengue-like symptoms diseases, such as malaria, typhoid fever, COVID-19, dyspepsia, pneumonia, and gastritis, were collected for the years 2017-2023. These two hospitals' medical records were paper-based, requiring manual recording using an Excel spreadsheet. The features recorded from the medical records collected are shown in the following list:

1) Age;

2) Gender;

3) Temperature;

4) All recorded symptoms;

5) Duration of fever;

6) Working diagnosis;

7) Laboratory test results;

8) Final diagnosis.

Table 1 shows the characteristics of collected medical records from the two Indonesian hospitals. The total medical records collected (*n*) is 561 records. The medical records consist of 473 non-dengue cases and 88 dengue cases. Features in the form of symptoms are indicated using S and features in the form of risk factors are indicated using F. The collected dataset will then be named as a dengue dataset, which has 36 symptoms and two risk factors. Most of the symptoms are binary in the form of 1 for Yes or Female and 0 for No or Male. The duration of fever (S_2), temperature (S_{25}) and age (F_1) are in the form of number. The target in the dataset is Diagnosis, which is the form of the binary value (1 for dengue and 0 for non-dengue diseases).

Table 1 Characteristics of collected medical records (*n*=561)

Notation	Feature	*n*	%
Symptoms			
S_1	Fever		
	Yes	318	56.68
	No	243	43.32
S_2	Duration of fever	Mean 2.38, SD 3.62	
S_3	Headache	158	28.16

	Yes	403	71.84
	No		
S_4	Arthralgia / Myalgia		
	Yes	24	4.28
	No	537	95.72
S_5	Nausea		
	Yes	290	51.69
	No	271	48.31
S_6	Vomiting		
	Yes	196	34.94
	No	365	65.06
S_7	Abdominal pain		
	Yes	129	22.99
	No	432	77.01
S_8	Shivering		
	Yes	33	5.88
	No	528	94.12
S_9	Body pain		
	Yes	23	4.10
	No	538	95.90
S_{10}	Heartburn		
	Yes	224	39.93
	No	337	60.07
S_{11}	Chest pain		
	Yes	44	7.84
	No	517	92.16
S_{12}	Dizziness		
	Yes	171	30.48
	No	390	69.52
S_{13}	Malaise		
	Yes	489	87.17
	No	72	12.83
S_{14}	Loss of appetite		
	Yes	289	51.52
	No	272	48.48
S_{15}	Sneezing		
	Yes	182	32.44
	No	379	67.56
S_{16}	Coughing		
	Yes	304	54.19
	No	257	45.81
S_{17}	Shortness of breath / fast breathing		
	Yes	140	24.96
	No	421	75.04
S_{18}	Rash		
	Yes	10	1.78
	No	551	98.22
S_{19}	Bleeding nose		
	Yes	3	0.53
	No	558	99.47
S_{20}	Bitter mouth		
	Yes	22	3.92
	No	539	96.08
S_{21}	Sore throat		
	Yes	11	1.96
	No	550	98.04
S_{22}	Blurry vision		
	Yes	2	0.36
	No	559	99.64
S_{23}	Seizure		
	Yes	7	1.25
	No	554	98.75
S_{24}	Diarrhea		
	Yes	74	13.19
	No	487	86.81
S_{25}	Temperature	Mean 37.06, SD 1.93	

S_{26}	Sweating		
	Yes	28	4.99
	No	533	95.01
S_{27}	Swallowing pain		
	Yes	27	4.81
	No	534	95.19
S_{28}	Pale		
	Yes	25	4.46
	No	536	95.54
S_{29}	Jaundice		
	Yes	1	0.18
	No	560	99.82
S_{30}	Anaemia		
	Yes	2	0.36
	No	559	99.64
S_{31}	Black water		
	Yes	2	0.36
	No	559	99.64
S_{32}	Constipation		
	Yes	37	6.60
	No	524	93.40
S_{33}	Flatulence		
	Yes	37	6.60
	No	524	93.40
S_{34}	Feeling anxious		
	Yes	3	0.53
	No	558	99.47
S_{35}	Bleeding coughing		
	Yes	25	4.46
	No	536	95.54
S_{36}	Loss of consciousness		
	Yes	5	0.89
	No	556	99.11
Non-symptom-related factors			
F_1	Age	Mean 31.66, SD 24.57	
F_2	Gender		
	Female	304	54.19
	Male	257	45.81

SD: standard deviation

Interview results with fifteen Indonesian medical doctors

To confirm the results from the significant features obtained from the feature selection process, we interviewed 15 Indonesian medical doctors about important symptoms and risk factors for dengue diagnosis. These 15 medical doctors were provided with the structured interview questions regarding symptom and risk factors for dengue diagnosis. The questions were in Bahasa Indonesia. Therefore, we translated it in English. Table 2 shows the summarized interview results from the 15 Indonesian medical doctors about symptoms and risk factors for clinical diagnosis of dengue fever.

Table 2 The summarized symptoms and risk factors from the fifteen Indonesian medical doctors

Symptom and risk factor	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	Total Y
Fever (S_1)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	15
Fever duration (S_2)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	-	Y	Y	13
Headache (S_3)	Y	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	-	-	Y	9
Arthralgia/joint pain and Myalgia/muscle pain (S_4)	Y	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	Y	-	-	9
Nausea (S_5)	Y	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	2
Vomiting (S_6)	Y	-	-	Y	-	-	-	-	-	-	Y	-	-	-	-	3
Abdominal pain (S_7)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	-	13
Shivering (S_8)	Y	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	-	-	-	8
Body pain (S_9)	Y	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	-	-	-	8
Heartburn (S_{10})	-	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	1
Chest pain (S_{11})	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Dizziness (S_{12})	Y	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	Y	-	-	8
Malaise (S_{13})	Y	Y	-	-	Y	Y	Y	Y	Y	Y		Y			Y	10
Loss of appetite (S_{14})	Y	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	-	-	-	8
Sneezing (S_{15})	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Coughing (S_{16})	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Shortness of breath (S_{17})	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	1
Rash (S_{18})	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y	Y	Y	14
Bleeding nose (S_{19})	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y	Y	Y	Y	14
Bitter mouth (S_{20})	Y	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	-	-	-	8
Sore throat (S_{21})	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Blurry vision (S_{22})	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Seizure (S_{23})	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
Diarrhea (S_{24})	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
High Temperature (S_{25})	Y	Y	Y	-	Y	Y	Y	Y	Y	Y	-	-	Y	-	Y	11
Orbital pain	-	Y	-	Y	Y	Y	Y	Y	Y	Y	-	-	-	-	-	8
Loss of consciousness (S_{36})	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	-	Y	Y	-	-	11
Age (F_1)	-	Y	-	-	Y	Y	Y	Y	Y	Y	Y	-	-	Y	-	9
Gender (F_2)	-	Y	-	-	Y	Y	Y	Y	Y	Y	-	-	-	Y	-	8
Endemic area (F_3)	-	Y	-	Y	Y	Y	Y	Y	Y	Y	Y	-	Y	-	Y	11

D: medical doctor; Y: Yes

Machine learning techniques used

In this study, we employ commonly used machine learning techniques in dengue and malaria prediction including, support vector machine (SVM) [23–25], random forest (RF) [25,24], eXtreme gradient boosting (XGBoost) [23], logistic regression (LR) [23,24], k-nearest neighbour (KNN) [26] to develop dengue classifiers that can accurately distinguishing dengue from non-dengue diseases.

Performance metrics used

To evaluate the performance of the classifiers, we use two performance metrics including accuracy and precision. The formula for the two performance metrics can be seen in Equations (1)–(2).

$$Accuracy = \frac{(TN+TP)}{(TN+TP+FN+FP)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Where,

TP is the number of dengue records that are correctly classified;

TN is the number of non-dengue records that are correctly classified;

FP is the number of non-dengue records classified as dengue;

FN is the number of dengue records classified as non-dengue.

Feature selection methods used

Feature selection filters redundant or irrelevant features [27]. By reducing the number of features, it will minimize the computational cost of the prediction and increase the performance of the machine learning classifier. The feature selection methods assess the relationship between each feature and the target feature and choose the input features that have the strongest correlation with the target feature [28]. The higher the score, the more the feature is related to the target feature.

In this study, feature selection methods used to determine important features are feature importance [29], recursive feature elimination (RFE) [30,31], correlation matrix from Pearson's correlation coefficient (PCC) [2,28] and KBest [27].

Ethical Statement

This research was approved by Human Ethics Committee of Widya Mandira Catholic University (reference number: 001/WM.H9/LPPM/SKKEP/X/2023).

RESULTS

Table 3 shows the four feature selection scores from FI, RFE, CM and KBest for features that meet the threshold. The threshold value for each feature selection method is used to obtain the most important

features from FI (>=0.030), RFE (1), PCC (>=0.100) and KBest (>1.000). From this first process of filtering, some features are eliminated.

Table 3. The number of occurrences of features in the four feature selection methods with their selection results

Feature (notation)	FI (>0.030): (FS_{FI})	RFE (1): (FS_{RFE})	PCC (>=0.100): (FS_{PCC})	KBest (>=1.000): (FS_{KBest})	Number of occurrences
age (F_1)	0.163 [a]	3	0.230 [c]	0.316	2
gender (F_2)	0.023	1 [b]	0.070	2.429 [d]	2
fever (S_1)	0.054 [a]	1 [b]	0.330 [c]	0.705	3
fever_duration (S_2)	0.141 [a]	3	0.160 [c]	0.153	2
headache (S_3)	0.037 [a]	1 [b]	0.080	3.478 [d]	3
muscle_joint_pain (S_4)	0.038 [a]	1 [b]	0.110 [c]	7.296 [d]	4
nausea (S_5)	0.032 [a]	1 [b]	0.140 [c]	0.116	3
vomiting (S_6)	0.023	3	0.080	4.056 [d]	1
abdominal_pain (S_7)	0.022	1 [b]	0.130 [c]	8.929 [d]	3
shivering (S_8)	0.013	2	0.120 [c]	8.349 [d]	2
body_pain (S_9)	0.009	1 [b]	0.010	0.053	1
heartburn (S_{10})	0.028	2	0.060	1.826 [d]	1
chest_pain (S_{11})	0.007	2	0.090	4.500 [d]	1
dizziness (S_{12})	0.016	1 [b]	0.080	3.906 [d]	2
malaise (S_{13})	0.069 [a]	1 [b]	0.020	0.350	2
loss_of_appetite (S_{14})	0.037 [a]	1 [b]	0.240 [c]	0.347	3
sneezing (S_{15})	0.029	2	0.150 [c]	0.133	1
coughing (S_{16})	0.062 [a]	1 [b]	0.200 [c]	0.242	3
shortness_of_breath (S_{17})	0.038 [a]	1 [b]	0.230 [c]	0.301	3
rash (S_{18})	0.013	1 [b]	0.200 [c]	0.236	2
bleeding_nose (S_{19})	0.078 [a]	1 [b]	0.380 [c]	0.955	3
bitter_mouth (S_{20})	0.070 [a]	1 [b]	0.060	2.149 [d]	3
sore_throat (S_{21})	0.000	1 [b]	0.060	2.088 [d]	2
blurry_vision (S_{22})	0.000	1 [b]	0.030	0.372	1
seizure (S_{23})	0.000	1 [b]	0.050	1.317 [d]	2
diarrhea (S_{24})	0.011	1 [b]	0.070	2.501 [d]	2
temperature (S_{25})	0.086 [a]	3	0.090	4.181 [d]	2
Total selected features	13	19	13	14	

a: selected feature for FI; b: selected feature for RFE; c: selected feature for PCC; and d: selected feature for KBest

Table 3 also shows the total number of significant features for each feature selection method. Feature importance from RF has 13 significant features. RFE selects 19 significant features. There are 13 significant features for PCC and 14 significant features for KBest respectively. These results show that each feature selection method has its own combination of significant features. In Colum 6 of Table 3, we total number of occurrences for each feature based on the given thresholds from the four feature selection methods. The higher the number of occurrences, the more significant the feature. There are

some features that are significant for three or four feature selection methods. Muscle_joint_pain (S_4), for example, is the only feature choosed by the four feature selection methods. This indicates that this feature is the most significant feature among other features. From Table 3, we generate FS_1 from selected features >=3, FS_2 from selected features >=2 and FS_3 from selected features >=1. FS_4 consists of FS_1 and selected features = 1. FS_5 consists of FS_{PCC} and the selected symptoms from 15 medical doctors.

In order to choose the significant features for dengue diagnosis based on various combination of features, we compare feature sets (FS_s) generated. Table 4 shows the performance comparison from various features sets generated. As shown in Table 4, the most stable performance for almost all machine learning classifiers is FS_5. Therefore, the most significant features for dengue prediction are the combination of features of FS_5. The random forest classifier yields the highest accuracy for FS_5 with the accuracy of 0.93 and precision of 0.90.

Table 4 Performance comparison of features sets generated with the standard deviation values

Feature set (FS)	Feature	LR		KNN		XGBoost		RF		NB		SVM	
		Acc (SD)	Pre (SD)	Acc (SD)	Pre (SD)	Acc (SD)	Pre (SD)	Acc (SD)	Pre (SD)	Acc (SD)	Pre (SD)	Acc (SD)	Pre (SD)
FS_{FI}	F_1, S_1, S_2, S_3, S_4, S_5, S_{13}, S_{14}, S_{16}, S_{17}, S_{19}, S_{20}, S_{24}, S_{25}	0.91 (0.04)	0.89 (0.17)	0.90 (0.04)	0.70 (0.17)	0.90 (0.03)	0.71 (0.14)	0.91 (0.04)	0.81 (0.21)	0.82 (0.10)	0.50 (0.15)	0.89 (0.05)	0.66 (0.23)
FS_{RFE}	F_2, S_1, S_3, S_4, S_5, S_7, S_9, S_{12}, S_{13}, S_{14}, S_{16}, S_{17}, S_{18}, S_{19}, S_{20}, S_{21}, S_{22}, S_{23}, S_{24}	0.89 (0.06)	0.78 (0.32)	0.86 (0.06)	0.57 (0.25)	0.83 (0.07)	0.47 (0.23)	0.83 (0.06)	0.47 (0.23)	0.42 (0.07)	0.21 (0.05)	0.87 (0.05)	0.80 (0.40)
FS_{PCC}	F_1, S_1, S_2, S_4, S_5, S_7, S_8, S_{14}, S_{15}, S_{16}, S_{17}, S_{18}, S_{19}	0.91 (0.04)	0.84 (0.19)	0.90 (0.03)	0.69 (0.14)	0.90 (0.04)	0.70 (0.20)	0.92 (0.04)	0.83 (0.16)	0.89 (0.06)	0.70 (0.18)	0.90 (0.04)	0.70 (0.17)
FS_{KBest}	F_2, S_3, S_4, S_6, S_7, S_8, S_{10}, S_{11}, S_{12}, S_{20}, S_{21}, S_{23}, S_{24}, S_{25}	0.85 (0.04)	0.10 (0.30)	0.82 (0.03)	0.15 (0.32)	0.79 (0.05)	0.23 (0.21)	0.80 (0.03)	0.22 (0.30)	0.30 (0.04)	0.18 (0.04)	0.84 (0.04)	0.00 (0.00)
FS_1	S_1, S_3, S_4, S_5, S_7, S_{14}, S_{16}, S_{17}, S_{19}, S_{20}	0.88 (0.05)	0.73 (0.33)	0.88 (0.06)	0.69 (0.24)	0.89 (0.07)	0.76 (0.24)	0.89 (0.07)	0.71 (0.23)	0.82 (0.16)	0.54 (0.19)	0.88 (0.05)	0.87 (0.30)
FS_2	FS_1, S_2, S_{13}, S_{18}, S_8, S_{12}, S_{21}, S_{23}, S_{24}, S_{25}, F_1, F_2	0.90 (0.04)	0.72 (0.30)	0.90 (0.04)	0.70 (0.17)	0.89 (0.01)	0.70 (0.16)	0.92 (0.03)	0.88 (0.13)	0.66 (0.10)	0.31 (0.09)	0.88 (0.05)	0.59 (0.24)
FS_3	FS_2, S_6, S_{10}, S_9, S_{11}, S_{15}, S_{22}	0.90 (0.05)	0.74 (0.33)	0.89 (0.03)	0.66 (0.14)	0.92 (0.02)	0.77 (0.18)	0.93 (0.03)	0.89 (0.13)	0.56 (0.08)	0.25 (0.06)	0.88 (0.04)	0.61 (0.22)
FS_4	FS_1, S_6, S_{10}, S_9, S_{11}, S_{15}, S_{22}	0.90 (0.05)	0.77 (0.35)	0.88 (0.05)	0.65 (0.22)	0.87 (0.05)	0.59 (0.17)	0.88 (0.04)	0.58 (0.18)	0.57 (0.09)	0.25 (0.07)	0.88 (0.05)	0.87 (0.31)

FS_5	S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_{13}, S_{14}, S_{15}, S_{16}, S_{17}, S_{18}, S_{19}, S_{20}, S_{25}, F_1	0.90 (0.04)	0.86 (0.22)	0.89 (0.04)	0.67 (0.14)	0.91 (0.03)	0.77 (0.13)	0.93 (0.04)	0.90 (0.13)	0.88 (0.09)	0.66 (0.20)	0.89 (0.04)	0.66 (0.18)

Acc: Accuracy; Pre: Precision, SD: Standard deviation

DISCUSSION

Table 4 shows that significant features for dengue prediction are fever (S_1), fever duration (S_2), headache (S_3), muscle joint pain (S_4), nausea (S_5), vomiting (S_6), abdominal pain (S_7), shivering (S_8), malaise (S_{13}), loss of appetite (S_{14}), sneezing (S_{15}), coughing (S_{16}), shortness of breath (S_{17}), rash (S_{18}), bleeding nose (S_{19}), bitter mouth (S_{20}), temperature (S_{25}) and age (F_1). However, not all these features are dengue symptoms. It is important to note that the dataset consists of dengue records and other medical records including malaria, COVID-19, dyspepsia, gastritis, typhoid fever and pneumonia. We will discuss which symptoms and risk factors that are important for dengue predictions or dengue diagnosis with the confirmation of medical doctors knowledge.

Fever, fever duration and high temperature are three important dengue symptoms. For fever, three out of four feature selection methods select this symptom as an important feature. All fifteen medical doctors interviewed also agree that one of the most important dengue features is fever. Even though only two feature selection methods including FI and PCC chose fever duration as an important feature, fever normally starts 4-10 days after infection and last for 2-7 days [32]. Based on the medical doctors interviewed and medical records collected, temperature also has a significant contribution for dengue prediction that can reach 39-40°C. Eleven medical doctors interviewed agree that high temperature of fever is important to distinguish dengue from other diseases such as malaria and typhoid fever. Therefore, it is important to include fever, fever duration and high temperature of fever as three important features for dengue diagnosis.

Arthralgia/joint pain and myalgia/muscle pain are two symptoms that are considered as the most significant features for the dengue prediction and dengue diagnosis [33]. All the four feature selection methods indicate these two symptoms are important for distinguishing dengue from other diseases

including malaria, typhoid fever, COVID-19, dyspepsia and pneumonia. Nine medical doctors also consider these two symptoms as significant symptoms for dengue diagnosis.

Headache is one of the most important symptoms in diagnosing and predicting dengue [33]). The three feature selection methods other than PCC consider this symptom essential for dengue diagnosis. In addition, it is also confirmed by nine medical doctors interviewed.

Nausea is considered as one of the most significant symptoms for dengue diagnosis [33]. That also applies for vomiting [34]. However, if persistent vomiting occurs then the individual might progress to the severe state [33]. Two medical doctors agree that nausea is part of dengue symptoms whereas three medical doctors agree that vomiting is an important symptom for dengue diagnosis. In the prediction perspective, nausea is more considered significant because it is selected by three feature selection methods. Whereas vomiting is least significant as only KBest selects this symptom. However, these two symptoms are highly correlated, thus it is important to consider both symptoms as dengue symptoms.
Loss of appetite is considered a symptom that can indicate individuals suffer from dengue. Eight medical doctors interviewed confirm that this symptom is also considered as a dengue symptom. This symptom is also selected by three feature selection methods other than KBest.

Even though shivering is associated with malaria [2,33], shivering is also important for dengue diagnosis and prediction. Eight medical doctors interviewed also agree that shivering is also a dengue symptom. In the dengue prediction perspective, shivering is also an important feature for dengue prediction as it is selected by two feature selection methods including PCC and KBest as part of significant features.
Malaise is an important symptom for dengue diagnosis and it normally happens when individuals are in severe condition [33]. In addition, ten medical doctors also confirm that this symptom is essential in dengue diagnosis. It is also selected by two feature selection methods including FI and RFE.

Bleeding nose is one of the most important symptoms in dengue diagnosis as part of bleeding manifestations [33,34]. This symptom with other bleeding manifestations indicate that individuals

progress is in severe condition. Fourteen medical doctors interviewed agree that to determine an individual suffers from dengue is to check the presence of the bleeding nose. Moreover, three feature selection methods selected this symptom as a significant feature for dengue prediction.

Similar to the bleeding nose, the presence of rashes in skin is also pivotal in distinguishing dengue from other similar diseases such as malaria and typhoid fever [33]. Fourteen medical doctors confirm that a rash in an individual's body is a distinguishing symptom that led their initial diagnosis to dengue. This symptom is also selected in two feature selection methods including RFE and PCC.

Abdominal pain is considered as one of the dengue symptoms especially when someone in the severe state [33,34]. Thirteen medical doctors also confirm that this symptom is essential to determine dengue from other diseases. This symptom is also selected by three feature selection methods other than FI as a significant symptom for dengue prediction.

Shortness of breath or fast breathing is one of dengue symptoms that indicates the severe state of dengue [33]. This is also confirmed by one medical doctor interviewed. This symptom is also selected by three feature selection methods other than KBest as the important feature for dengue prediction.

Age can be considered as one of the important risk factors for dengue diagnosis [35]. Even though six medical doctors do not consider this factor as an important feature for dengue diagnosis, nine medical doctors include this factor as feature that should not be overlooked when diagnosing potential dengue patients. Two feature selection methods including FI and PCC also consider this factor important for dengue prediction. Normally, individuals younger than 15 years old are prone to dengue infection [36]. Bitter mouth is associated with malaria as this symptom is considered as one of malaria symptoms [37,38]. However, interestingly eight medical doctors interviewed agree that this symptom also can be found in individuals who suffer from dengue. This symptom also appears in three feature selection methods other than PCC. Thus, this symptom should not be ignored when diagnosing potential dengue patients.

From the dengue prediction perspective, sneezing and coughing are important features. Three feature selection methods select this symptom as significant features for dengue prediction. However, no medical doctors confirm that sneezing and coughing are part of dengue symptoms. Sneezing and coughing might be the distinguished symptom to determine COVID-19 from dengue. It is important to know that the dataset consists of medical records from COVID-19 patients. Besides, sneezing and coughing are known as COVID-19 symptom [39,40]. Therefore, sneezing and coughing are important for dengue prediction but not necessarily are dengue symptoms.

This study does not include other features such as orbital pain, history of previous suffering from dengue and history of visiting endemic dengue areas. In this study, all this information were not found in the medical records collected. This opens the room for the future studies. The significant features as results from this study can be used to develop reliable and powerful machine learning techniques, which later can be used to develop early-stage dengue prediction tools.

In conclusion, there are four findings of this study. First, there are 17 symptom features including fever, fever duration, headache, muscle and joint pain, nausea, vomiting, abdominal pain, shivering, malaise, loss of appetite, sneezing, coughing, shortness of breath, rash, bleeding nose, bitter mouth, temperature and one risk factor feature including age that are important for dengue prediction. However, sneezing and coughing are not necessarily important for dengue diagnosis. Second, arthralgia/joint pain and myalgia/muscle pain are the most significant features for the dengue prediction. Third, even though a bitter mouth symptom is highly related to malaria diagnosis, this study suggests that the medical doctors should not ignore the bitter mouth symptom in diagnosing dengue as this symptom is also important for dengue prediction. Fourth, random forest classifier yields the most stable performance for dengue prediction. Knowledge of these features are essential to educate society about significant symptoms and risk factors for dengue to avoid progression to severe conditions, which can lead to death. The findings of this study can also be used as a reference for medical doctors in differentiating dengue from non-dengue diseases including malaria, COVID-19 and typhoid fever.

ACKNOWLEDGMENTS

We would like to thank Widya Mandira Catholic University for providing the research grant for this research project. We would also like to express our gratitude to the Director of Kewapante Hospital Sikka, and the Director of Soe Hospital South Central Timor, and to the Department of Permission Affair in Sikka, South Central Timor, and in East Nusa Tenggara Province — Indonesia. We are grateful to the medical records staff and the fifteen medical doctors for their cooperation.

AUTHOR CONTRIBUTIONS

Conceptualization: Bria YP, Nani PA, Siki YCH, Meolbatak EM, Guntur RD, Mamulak NMR. Data curation: Bria YP, Nani PA, Mamulak NMR. Formal analysis: Bria YP, Siki YCH, Meolbatak EM, Guntur RD. Funding acquisition: Bria YP. Methodology: Bria YP, Nani PA, Siki YCH, Meolbatak EM, Guntur RD, Mamulak NMR. Project administration: Mamulak NMR. Writing – original draft: Bria YP, Nani PA, Siki YCH, Meolbatak EM, Guntur RD, Mamulak NMR. Writing – review & editing: Bria YP, Nani PA, Siki YCH, Meolbatak EM, Guntur RD, Mamulak NMR.

CONFLICT OF INTEREST

The authors have no conflicts of interest associated with the material presented in this paper.

FUNDING

This research was supported by Widya Mandira Catholic University Kupang East Nusa Tenggara Province Indonesia (044/WM.H9/SKP/IX/2023).

REFERENCES

1. Ministry of Health of Indonesia. Information on DHF Cases in 2023 Week 19. [cited 2023 Aug 19]. Available from: https://p2pm.kemkes.go.id/publikasi/infografis/info-kasus-dbd-2023-minggu-ke-19 (Indonesian)

2. Bria YP, Yeh CH, Bedingfield S. Significant symptoms and nonsymptom-related factors for malaria diagnosis in endemic regions of Indonesia. Int J Infect Dis 2021;103:194 – 200. https://doi.org/10.1016/j.ijid.2020.11.177
3. World Health Organization. Update on the Dengue situation in the Western Pacific Region. 2015;2014(482):5. [cited 2024 March 20]. Availabel from : Dengue-20151229.pdf (who.int)
4. WHO. Dengue Guidelines for Diagnosis, Treatment, Prevention and Control. 2009. [cited 2024 Feb 15]. Availabel from : https://iris.who.int/handle/10665/44188
5. WHO. Comprehensive Guidelines for Prevention and Control of Dengue and Dengue Haemorrhagic Fever. 2011. [cited 2023 Dec 25]. Available from : https://iris.who.int/handle/10665/204894
6. Ministry of Health of Indonesia. National Guidelines for Medical Services for the Management of Dengue Infection in Children and Adolescents. 2021 p. 1–67. [cited 2024 Jan 10]. Available from: https://yankes.kemkes.go.id/unduhan/fileunduhan_1660187378_126303.pdf (Indonesian)
7. Central Bureau of Statistics East Nusa Tenggara Province. Number of Disease Cases According to Regency/City and Type of Disease (Inhabitant) in 2022. [cited 2024 Jan 10]. Available from: https://ntt.bps.go.id/indicator/30/1485/1/jumlah-kasus-penyakit-menurut-kabupaten-kota-dan-jenis-penyakit.html (Indonesian)
8. Rakhmani AN, Zuhriyah L. Knowledge, Attitudes, and Practices Regarding Dengue Prevention Among Health Volunteers in an Urban Area – Malang, Indonesia. J Prev Med Public Health 2024;57:176-184. https://doi.org/10.3961/jpmph.23.484
9. Noroozi Z, Orooji A, Erfannia L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Sci Rep 2023;13(1):1–15. https://doi.org/10.1038/s41598-023-49962-w
10. Álvarez JD, Matias-Guiu JA, Cabrera-Martín MN, Risco-Martín JL, Ayala JL. An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders. BMC Bioinformatics 2019;20(1):1–12. https://doi.org/10.1186/s12859-019-3027-7
11. Song J, Li Z, Yao G, Wei S, Li L, Wu H. Framework for feature selection of predicting the

diagnosis and prognosis of necrotizing enterocolitis. PLoS One 2022;17(8 August). http://dx.doi.org/10.1371/journal.pone.0273383

12. Gu F, Ma S, Wang X, Zhao J, Yu Y, Song X. Evaluation of Feature Selection for Alzheimer's Disease Diagnosis. Front Aging Neurosci 2022;14(June):1–7. https://doi.org/10.3389/fnagi.2022.924113
13. Spencer R, Thabtah F, Abdelhamid N, Thompson M. Exploring feature selection and classification methods for predicting heart disease. Digit Heal 2002;6:2–12. https://doi.org/10.1177/2055207620914777
14. Ramasamy V, Vadivel S, Kothandapani S, Mahilraj J, Sivaram P, Sharma B. An Optimal Feature Selection with Neural Network-Based Classification Model for Dengue Fever Prediction. 2023 6th Int Conf Inf Syst Comput Networks, ISCON 2023. 2023;1–5. https://doi.org/10.1109/ISCON57294.2023.10112011
15. Ramadona AL, Tozan Y, Wallin J, Lazuardi L, Utarini A, Rocklöv J. Predicting the dengue cluster outbreak dynamics in Yogyakarta, Indonesia: a modelling study. Lancet Reg Heal - Southeast Asia 2023;15:100209. https://doi.org/10.1016/j.lansea.2023.100209
16. Tanawi IN, Vito V, Sarwinda D, Tasman H, Hertono GF. Support Vector Regression for Predicting the Number of Dengue Incidents in DKI Jakarta. Procedia Comput Sci 2021;179(2020):747–53. https://doi.org/10.1016/j.procs.2021.01.063
17. Nuraini N, Fauzi IS, Fakhruddin M, Sopaheluwakan A, Soewono E. Climate-based dengue model in Semarang, Indonesia: Predictions and descriptive analysis. Infect Dis Model 2021;6:598–611. https://doi.org/10.1016/j.idm.2021.03.005
18. Anggraeni W, Nurmasari R, Riksakomara E, Samopa F, Wibowo RP, Condro LT, et al. Modified Regression Approach for Predicting Number of Dengue Fever Incidents in Malang Indonesia. Procedia Comput Sci 2017;124:142–50. https://doi.org/10.1016/j.procs.2017.12.140
19. Lestari NA, Tyasnurita R, Vinarti RA, Anggraeni W. Long Short-Term Memory forecasting model for dengue fever cases in Malang regency, Indonesia. Procedia Comput Sci 2021;197(2021):180–8. https://doi.org/10.1016/j.procs.2021.12.131
20. Thamrin SA, Aswi, Ansariadi, Jaya AK, Mengersen K. Bayesian spatial survival modelling for

dengue fever in Makassar, Indonesia. Gac Sanit 2021;35:S59–63. https://doi.org/10.1016/j.gaceta.2020.12.017

21. Aswi A, Cramb S, Duncan E, Hu W, White G, Mengersen K. Climate variability and dengue fever in Makassar, Indonesia: Bayesian spatio-temporal modelling. Spat Spatiotemporal Epidemiol. 2020;33. https://doi.org/10.1016/j.sste.2020.100335
22. Fauzi IS, Nuraini N, Ayu RWS, Lestari BW. Temporal trend and spatial clustering of the dengue fever prevalence in West Java, Indonesia. Heliyon 2022;8(8):e10350. https://doi.org/10.1016/j.heliyon.2022.e10350
23. Joshi A, Miller C. Review of machine learning techniques for mosquito control in urban environments. Ecol Inform 2021;61:101241. https://doi.org/10.1016/j.ecoinf.2021.101241
24. Hoyos W, Aguilar J, Toro M. Dengue models based on machine learning techniques: A systematic literature review. Artif Intell Med 2021;119(August):102157. https://doi.org/10.1016/j.artmed.2021.102157
25. Shaikh MSG, SureshKumar DB, Narang DG. Development of optimized ensemble classifier for dengue fever prediction and recommendation system. Biomed Signal Process Control 2023;85(March):104809. https://doi.org/10.1016/j.bspc.2023.104809
26. Bria YP, Yeh CH, Bedingfield S. Machine Learning Classifiers for Symptom-Based Malaria Prediction. Proc Int Jt Conf Neural Networks. 2022;2022-July. 10.1109/IJCNN55064.2022.9891945
27. Qiu P, Niu Z. TCIC_FS: Total correlation information coefficient-based feature selection method for high-dimensional data. Knowledge-Based Syst 2021;231:107418. https://doi.org/10.1016/j.knosys.2021.107418
28. Senan EM, Abunadi I, Jadhav ME, Fati SM. Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms. Comput Math Methods Med. 2021;2021. https://doi.org/10.1155/2021/8500314
29. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. J Big Data 2020;7(1). https://doi.org/10.1186/s40537-020-00327-4

30. Alshanbari HM, Mehmood T, Sami W, Alturaiki W, Hamza MA, Alosaimi B. Prediction and Classification of COVID-19 Admissions to Intensive Care Units (ICU) Using Weighted Radial Kernel SVM Coupled with Recursive Feature Elimination (RFE). Life. 2022;12(7). https://doi.org/10.3390/life12071100
31. Mathew TE. A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. 2019;10(3):55–63.
32. Gupta G, Khan S, Guleria V, Almjally A, Alabduallah BI, Siddiqui T, et al. DDPM: A Dengue Disease Prediction and Diagnosis Model Using Sentiment Analysis and Machine Learning Algorithms. Diagnostics 2023;13(6). https://doi.org/10.3390/diagnostics13061093
33. World Health Organization. Dengue and severe dengue [Internet]. 2023 [cited 2023 Aug 6]. Available from: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#:~:text=The incidence of dengue has,dengue cases are under-reported.
34. Tatura SNN, Denis D, Santoso MS, Hayati RF, Kepel BJ, Yohan B, et al. Outbreak of severe dengue associated with DENV-3 in the city of Manado, North Sulawesi, Indonesia. Int J Infect Dis 2021;106:185–96. https://doi.org/10.1016/j.ijid.2021.03.065
35. Santoso MS, Yohan B, Denis D, Hayati RF, Haryanto S, Trianty L, et al. Diagnostic accuracy of 5 different brands of dengue virus non-structural protein 1 (NS1) antigen rapid diagnostic tests (RDT) in Indonesia. Diagn Microbiol Infect Dis 2020;98(2):115116. https://doi.org/10.1016/j.diagmicrobio.2020.115116
36. Ministry of Health of Indonesia. Dengue Hemorrhagic Fever [Internet]. 2022 [cited 2024 Feb 5]. Available from: https://ayosehat.kemkes.go.id/topik/demam-berdarah-dengue (Indonesian)
37. Nwokolo E, Ujuju C, Anyanti J, Isiguzo C, Udoye I, Bongos-Ikwue E, et al. Misuse of artemisinin combination therapies by clients of medicine retailers suspected to have malaria without prior parasitological confirmation in Nigeria. Int J Heal Policy Manag 2018;7(6):542–8. https://doi.org/10.15171/ijhpm.2017.122
38. Liu DT, Besser G, Oeller F, Mueller CA, Renner B. Bitter Taste Perception of the Human Tongue Mediated by Quinine and Caffeine Impregnated Taste Strips. Ann Otol Rhinol Laryngol. 2020;129(8):813–20. https://doi.org/10.1177/0003489420906187

39. Romero-Castro NS, Colín-Hernández I, Godoy-Reyes ME, Hernández-Hernández M, García-Verónica A, Paredes-Solis S, et al. Clinical Signs and Symptoms Associated with COVID-19: A Cross Sectional Study. Int J Odontostomatol 2022;16(1):112–9. http://dx.doi.org/10.4067/S0718-381X2022000100112

40. Costeira R, Lee KA, Murray B, Christiansen C, Castillo-Fernandez J, Lochlainn MN, et al. Estrogen and COVID-19 symptoms: Associations in women from the COVID Symptom Study. PLoS One 2021;16(9 September):1–14. http://dx.doi.org/10.1371/journal.pone.0257051

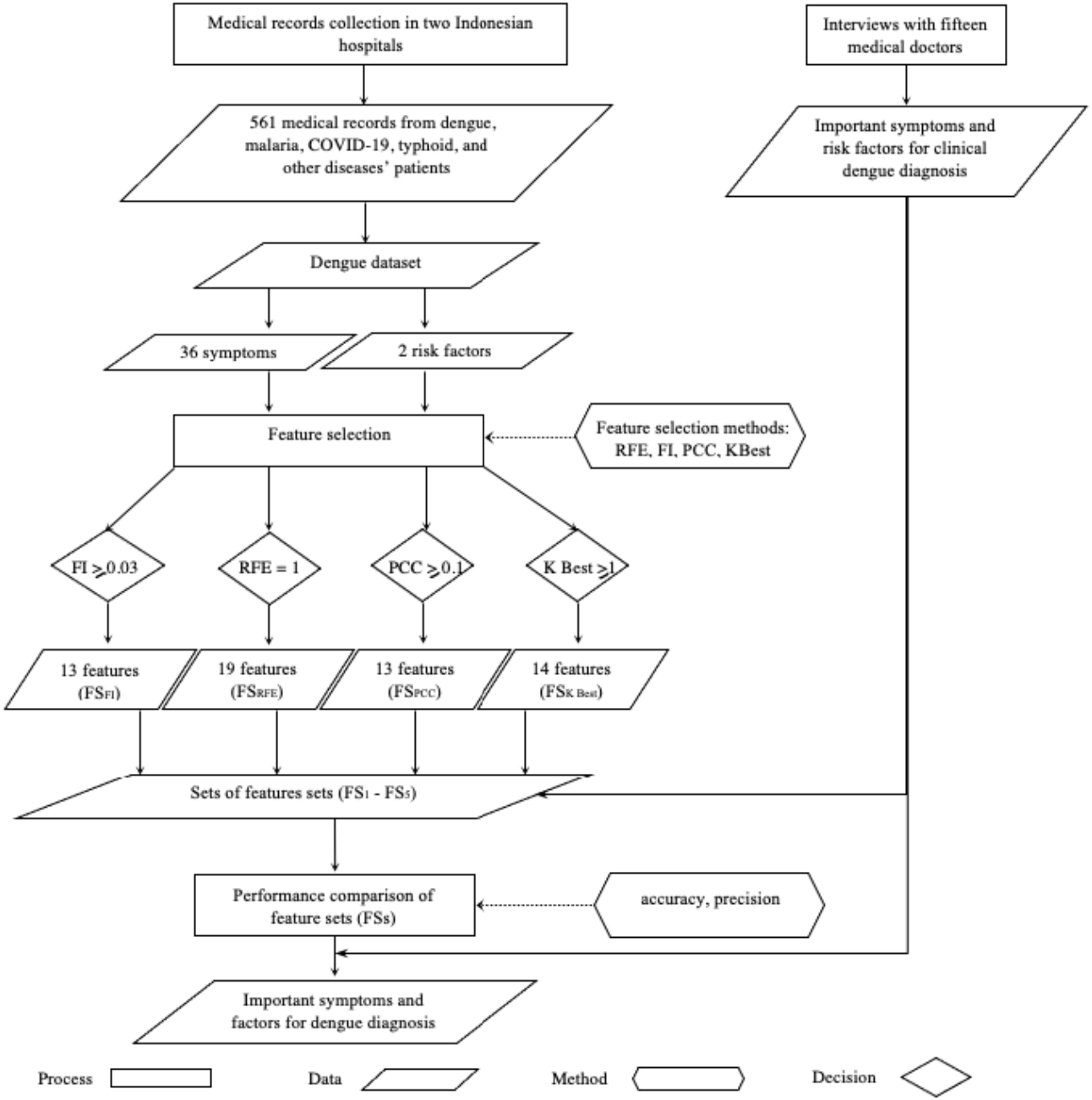

Figure 1. The approach for determining important features for dengue diagnosis