Machine Learning Classification with Logistic Regression Feature Selection Approach on Health Datasets

J. Pongthao, A. Na-udom, J. Rungrattanaubol

Abstract— **The study of health data has been of interest for decades. Early detection of illnesses and diseases is crucial for improving healthcare and quality of life. In the era of big data, numerous health datasets have been accumulated, explored, and analyzed. Various classification models have been developed to enhance the accuracy of illness and disease prediction. The purpose of this research was to analyze and compare the performance of four classification techniques on three health datasets containing both quantitative and qualitative data. The classification methods examined in this study included Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. The three datasets analyzed focused on cardiovascular disease, diabetes, and smoking data. The research also emphasized the feature selection process to extract the most significant features. This paper proposes a feature selection process based on the Logistic Regression analysis, discussing three methods: forward selection, backward elimination, and stepwise methods. Additionally, Spearman's rank correlation coefficient was used for feature selection. The classification models were constructed and evaluated using the 10-fold cross-validation. Accuracy and F1(weight) were used to measure and compare model performance. The results showed that the Support Vector Machine outperformed other models on the three datasets, achieving accuracies of 73.45% for cardiovascular disease, 74.87% for diabetes, and 75.78% for smoking. Additionally, the number of features was reduced for diabetes. Feature selection methods based on Logistic Regression primarily improved the performance of Logistic Regression and Decision Tree classification models, with minimal impact on the performance of Random Forest and Support Vector Machine models. The number of features of many models was reduced with comparative performance. The selected features of health datasets from the proposed feature selection methods were summarized and discussed.**

Index Terms— **Decision Tree, Logistic Regression, Random Forest, Support Vector Machine**

Manuscript received December 9, 2024; revised April 8, 2025.

This work was supported in part by Naresuan University, and National Science, Research and Innovation Fund (NSRF) under Grant R2566B032.

J. Pongthao is a postgraduate student of Naresuan University, Phitsanulok, 65000, Thailand. (e-mail: jirananp64@nu.ac.th)

A. Na-udom is an associate professor of Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, 65000, Thailand. (corresponding author phone: 66-55963201; e-mail: anamain@nu.ac.th).

J. Rungrattanaubol is an assistant professor of Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, 65000, Thailand. (e-mail: jaratsrir@nu.ac.th).

I. Introduction

Medical studies investigating and explaining diverse aspects of human health, diagnosis of illnesses and diseases, and treatments have long been a subject of significant interest and research. In the era of big data, large collections of health datasets can be used to improve healthcare delivery, detect early diseases, predict patient outcomes, and provide answers to health and disease-related questions. These data can come from a variety of sources, including electronic health records, clinical trials, and genomic databases.

A study of health datasets to assess the risk of developing various illnesses and diseases has been widely considered a crucial step. The earlier this health data is identified and understood, the more effective treatment can be applied. This is important for both public health policies and private healthcare practices.

Machine learning plays a crucial role in predictive health modelling using health datasets in classifying diseases and illnesses. A predictive classification model is a type of supervised machine learning that predicts a discrete value, called 'class'. Effective models are then used to predict or classify new or upcoming data [1]. This approach is becoming more popular in a variety of applications such as image analysis, natural language processing, finance, marketing, education, and healthcare [2].

Numerous studies have explored and developed classification models using healthcare and health datasets [3-6]. The data and datasets are from machine learning repositories such as UCI [7-8], and Kaggle [9], as well as datasets collected by specific organizations [10-11]. Most of the studies identified in the literature review focused on modelling the classifiers on one health dataset, for example, the diabetes datasets [10, 12-17], the cardiovascular and heart diseases datasets [3, 7, 9, 11, 18-19] and breast cancer datasets [8, 20].

There are many popular classification techniques, including, inter alia, Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbors (K-NN), Decision Tree (DT), Random Forest (RF), Artificial Neural Networks (ANN), Support Vector Machine (SVM). Each technique is developed using distinct algorithms and modeling concepts. For example, LR, NB, and SVM are rooted in statistical and mathematical principles, while DT and RF utilize tree-like structures. ANN are inspired by the organization and functioning of the human brain, whereas K-NN operates as an instance-based learning algorithm.

Feature selection is a crucial step in building predictive models, as it can reduce the number of features, improve

model performance, and, most importantly, enhance the interpretability of the model. Various methods exist for feature selection, including Correlation Analysis, Chi-Square Test, Relief, Forward Selection, and Backward Elimination. These techniques help identify significant features while discarding less important ones, making the model more interpretable and comprehensible. However, it is important to recognize that feature selection does not always yield better model performance; some classification models can be effectively developed without it. Table I provides a summary of reviewed papers on health data classification models, indicating that LR, SVM, DT, and RF have demonstrated superior performance, with and without the application of feature selection.

TABLE I
SUMMARY OF CLASSIFICATION TECHNIQUES USED ON HEALTH DATA

	Classification Techniques	Best	Feature selection
[7]	LR, RF, SVM, ANN, K-NN	DT	Yes
[8]	RF, SVM(RBF)	SVM	Yes
[11]	LR, DT, RF, ANN, K-NN	LR	No
[12]	LR, RF, ANN, NB, K-NN	ANN and LR	No
[13]	LR, SVM(RBF)	LR	Yes
[16]	LR, DT, RF, SVM, ANN	RF	No
[19]	DT, SVM, NB, K-NN	SVM	No
[20]	LR, SVM	LR	Yes

Most related studies have concentrated on developing accurate classification models for a single health dataset, utilizing various classification techniques and employing different feature selection methods, as shown in Table I. In contrast, our research explores the construction of classification models across multiple health datasets. The machine learning classification techniques we selected include LR, DT, RF, and SVM. Utilizing Logistic Regression analysis, which inherently selects relevant features during the modeling process, we propose three feature selection approaches: forward selection, backward elimination, and stepwise selection. Additionally, we employ Spearman's Rank Correlation Coefficient as a foundational filter for feature selection.

We selected three different mixed-type health datasets that varied in the ratio of qualitative (discrete) to quantitative (numerical) input features: equal, larger, and smaller. The four classification techniques chosen for this study were LR, DT, RF, and SVM. These techniques were selected based on insights gained from the literature review and the various objectives of the modeling concepts. LR uses statistical regression, DT relies on information gain to construct decision trees, RF is an ensemble method utilizing multiple trees, and SVM is grounded in mathematical principles. We explored the performance of these four machine-learning classification techniques using the selected health datasets. Feature selection processes, which included LR analysis and Spearman's rank correlation, were applied to identify relevant features. The main purpose of this study was to analyze and compare the performance of the classification models while examining the impact of feature selection on each technique.

The process of constructing the model and selecting features is described in Section III. We evaluated the performance of the models using a 10-fold cross-validation approach, measuring average accuracy and F1 weighted scores from the test sets. Each classification technique was repeated ten times for consistency. Section IV presents a comparison of the performance of the classification models for each technique, both with and without feature selection methods. Furthermore, the features selected by each method are presented and analyzed in Section IV.

II. Methodology

A. Logistic Regression (LR)

Logistic Regression (LR) is rooted in statistical concepts and aims to estimate or predict the probability of an event's success or failure based on various influencing factors or input features. The output is qualitative, specifically categorized into 'classes.' In this research, a binary logistic regression model, which provides a two-class output, was employed. The possible outcomes are defined as failure (Y=0) or success (Y=1). The probability model for logistic regression (π_i) is formulated as illustrated in equation (1).

$$\pi_i = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)} \quad (1)$$

Given that the relationship between the input features and the output is non-linear, the relationship must be adjusted into a linear function, called the logit response function, as shown in equation (2).

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \quad (2)$$

where X_i represents an input feature, n denotes the total number of input features, and β_i stands for the regression coefficient. During the regression analysis conducted in the logistic regression modeling process, the most relevant input features can be identified and selected for modeling purposes. This makes logistic regression analysis a valuable tool for feature selection [10]. Additionally, when constructing LR models, there is an option to include or exclude feature selection.

B. Decision Tree (DT)

A decision tree (DT) is a tree-like structure used to classify data into discrete outputs or classes. It is one of the most popular supervised machine learning techniques due to its interpretable structure, which facilitates understanding how the model derives its predictions.

The construction of a decision tree begins at the top and proceeds downward. It starts by calculating the information gain of all input features, which measures how effectively the data can be separated. The input feature with the highest information gain, regarded as the most effective separator, is chosen as the 'root' node. The data is then split based on the value of this selected feature, which creates 'branches.' This process of expanding the tree continues recursively until a stopping criterion is reached.

There are variations of decision trees that differ in how they measure information. For example, the ID3 algorithm uses the Gini Index, while C4.5 and J48 use Entropy.

C. Random Forest (RF)

Random Forest (RF) is an ensemble method that combines multiple decision trees (DTs) to make predictions [21]. Unlike a single decision tree, a random forest constructs a large collection of decision trees, each utilizing a random subset of data. Moreover, during the construction of each tree, a random

selection of input features is considered. Once the forest is built, the final output is determined by taking the majority vote from the predictions of all individual decision trees. Therefore, when developing RF models, it is crucial to specify both the number of trees to generate and the number of input features to select for each tree [22]. While RF may not be as interpretable as a single decision tree, it typically achieves greater accuracy.

D. Support Vector Machine (SVM)

Support Vector Machine (SVM) were introduced by Cortes and Vapnik in 1995 [23] to address both classification and regression problems. The SVM classifier defines a hyperplane to classify data, aiming to minimize classification errors and maximize margins as much as possible. It can handle both linear and non-linear data. However, since most data tend to be non-linear, the kernel trick is employed to simplify data complexity and prevent overfitting issues. The hyperplane function is detailed in equation (3).

$$f(x) = sgn\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b\right) \quad (3)$$

Various functions can be utilized for a kernel (K), including polynomial, radial basis function, and sigmoid. The application of the radial basis function is demonstrated in equation (4).

$$K(x_i, x_j) = exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \quad (4)$$

where the value α_i ranges from 0 to Cost. In this context, when constructing SVM models, the value of Gamma (γ) and Cost need to be predefined to use as parameters of the model.

E. Feature Selection

Feature selection is a crucial process in machine learning that involves selecting a relevant subset of input features from a dataset. It is considered one of the most important factors for improving the accuracy of predictive models while discarding irrelevant features. By reducing the number of features, this process can decrease processing time and make predictive models easier to interpret and deploy.

There are three general approaches to feature selection: filter, wrapper, and embedded methods. The filter approach selects features based on statistical properties that define the relationship between input features and the output. Techniques used in this approach include the Correlation Coefficient, Chi-squared test, and Information Gain. The wrapper method evaluates subsets of features by training and testing the model on each subset. In contrast, the embedded method incorporates feature selection as part of the model training process [5].

In this paper, two key methods of feature selection are applied, which are described as follows:

A statistical Spearman's Rank Correlation Coefficient is a non-parametric measure of the strength and direction of association between two ranked variables. It evaluates how well the relationship between these variables can be described using a monotonic function.

Logistic Regression Analysis is a statistical technique used to model the relationship between a discrete outcome variable and one or more input variables. The feature selection process in Logistic Regression (LR) analysis begins with identifying input variables that are statistically significant and contribute to developing accurate predictive models. The main methods for feature selection in LR analysis are Forward Selection, Backward Elimination, and Stepwise Regression.

Forward Selection (*f*) process involves gradually adding one statistically significant input variable at a time to the model, as demonstrated in the equation (2), and repeating this process until a suitable model is achieved.

Backward Elimination (*b*) is the reverse of Forward selection process. It involves removing one variable at a time from the model instead of adding them.

Stepwise Regression (*s*) combines both Forward Selection and Backward Elimination approaches. In this method, variables are added to the model one by one, testing their correlation and goodness of fit with the previously added variables using the partial F (or t) test statistic. Similarly, when determining which variables to remove, the same testing process is applied.

F. Model Evaluation

Developing a classification model involves dividing the dataset into two parts: a training set and a test set. The training set is used to build the classification model, while the test set is utilized to evaluate the model's performance. The accuracy of the classification model depends on its ability to correctly classify instances from the test set. To prevent overfitting, we apply the concept of 10-fold cross-validation. The effectiveness of the model is assessed using two key metrics: Accuracy (Acc) and Weighted F1 Score (F1w). Accuracy represents the percentage of correct predictions made by the model. In contrast, Weighted F1 score ($F1_w$), is calculated as the harmonic mean of precision (Pre) and recall (Re), incorporating both 'Yes' and 'No' classification. These metrics are derived from formulas (5) and (6), respectively, and are based on a confusion matrix, as shown in Table I.

TABLE I
CONFUSION MATRIX

Actual Class	Predicted Class	
	Yes	**No**
Yes	TP	FN
No	FP	TN

True Positive (TP): The number of correct predictions classified as **Yes**.
True Negative (TN): The number of correct predictions classified as **No**.
False Positive (FP): The number of incorrect predictions classified as **Yes** (actual: **No**).
False Negative (FN): The number of incorrect predictions classified as **No** (actual: **Yes**).

$$\text{Accuracy (Acc)} = (TP+TN) / (TP+FN+FP+TN) \quad (5)$$

$$F1_w = \frac{N_{Yes}}{N}\left(\frac{2\times Re(Yes)\times Pre(Yes)}{Re(Yes)+Pre(Yes)}\right) + \frac{N_{No}}{N}\left(\frac{2\times Re(No)\times Pre(No)}{Re(No)+Pre(No)}\right) \quad (6)$$

where N is the total number of records, N_{yes} is the number of 'Yes', N_{no} is the number of 'No' in the dataset and

$$Re(Yes) = TP/(TP+FN),\ Re(No) = TN/(TN+FP)$$
$$Pre(Yes) = TP/(TP+FP),\ Pre(No) = TN/(TN+FN)$$

III. Experimental setup

A. The datasets

Three health datasets were chosen based on the proportion of discrete and numeric input features, as well as a similar number of classes to prevent imbalanced issues. As shown in Table II, the counts of the 'Yes' and 'No' classes are relatively close. The selected datasets include Cardiovascular Disease, Diabetes, and Smoking Cessation, all sourced from the Kaggle database. The datasets underwent preprocessing, which involved checking for missing data and duplicates. After eliminating records with missing information and duplicates, the preprocessed datasets intended for classification modeling are summarized below and presented in Table II.

1. Cardiovascular Disease (Cardio) – The dataset consists of 68,433 records of patients diagnosed with cardiovascular disease. It includes 6 discrete features and 5 numeric input types.

2. Diabetes (Diabetes) – The dataset consists of 68,134 records of diabetes patients, including 18 discrete features and 3 numeric features.

3. Smoking Cessation (Smoke) – The dataset includes 27,285 records of smokers who have quit smoking, featuring 4 discrete and 18 numeric input characteristics.

TABLE II
CHARACTERISTICS OF EACH DATA SETS

Dataset	(Number of input features) Discrete: Numeric	No of Classes Yes : No	Ratio of Yes and No
1. Cardio	(11) 6:5	33,830 : 34,603	1:1.023
2. Diabetes	(21) 18:3	34,394 : 33,740	1.019:1
3. Smoke	(22) 4:18	9,173 : 18,112	1:1.975

An essential step in constructing a machine learning model is understanding the data. Tables III to V provide information on the input features, including their brief descriptions and feature types for each dataset. Additionally, the tables display the Spearman's Rank Correlation (r) between each feature and the class. A symbol (*) indicates a correlation that is significant at the 0.05 level. Features that are in bold appear in more than one dataset. Notably, the feature 'age' is present in all three datasets and is classified as both numeric and discrete.

B. A process for constructing classification models

The study first focused on constructing classification models with LR, DT, RF, and SVM on the three health datasets without feature selection. To ensure a fair evaluation and comparison of these four techniques, 10-fold cross-validation principles were applied. The process is described step by step as follows:

Step 1: Stratified random sampling was applied to divide the dataset into 10 sets. As illustrated in Fig 1, each set is intended to maintain an equal representation of the 'Yes' and 'No' classes, ensuring balanced representation within the dataset.

Step 2: The construction of the RT and SVM models was carried out with the following parameters:

RF: Number of features (n): 3, 4, and 5 (calculated using the square root of the number of features in a dataset [23]).

Number of trees (m): 250 and 500

SVM: A kernel with a radial basis function has gamma values of 0.001, 0.01, 0.1, and 2. The cost values are 1 and 10.

TABLE III
INPUT FEATURES, MEANING, SPEARMAN'S RANK CORRELATION OF CARDIO

Features	Meanings	Type	r
age	**Age**	**numeric**	**0.236***
sex	**Gender**	**discrete**	**0.007**
height	**Height**	**numeric**	**-0.013***
weight	**Weight**	**numeric**	**0.180***
sys	**Systolic**	**numeric**	**0.451***
dia	**Diastolic**	**numeric**	**0.356***
chol	**Cholesterol level**	**discrete**	**0.215***
glu	Glucose level	discrete	0.091*
smoke	**Smoker or non-smoker**	**discrete**	**-0.017***
alc	Drink alcohol or not	discrete	-0.009*
exer	**Take exercise regularly**	**discrete**	**-0.038***

TABLE IV
INPUT FEATURES, MEANING, SPEARMAN'S RANK CORRELATION OF DIABETES

Features	Meanings	Type	r
hiBP	Has high blood pressure	discrete	0.371*
hiChol	Has high cholesterol	discrete	0.282*
cholCk	Regular cholesterol checkups	discrete	0.118*
bmi	BMI	numeric	0.316*
smoke	**Smoke**	**discrete**	**0.077***
stroke	Has a history of stroke	discrete	0.124*
heartAtt	Has coronary heart disease	discrete	0.208*
exer	**Take exercise regularly**	**discrete**	**-0.146***
fruits	Regularly eating fruits	discrete	-0.044*
veg	Regularly eating vegetable	discrete	-0.072*
alcoAdd	Alcohol addiction	discrete	-0.098*
healthcare	Has health care protection	discrete	0.028*
noDr	Cannot pay for the doctor	discrete	0.035*
physH	Physical health score	discrete	0.400*
dayOfMent	Mental health sick days in a month	numeric	0.077*
dayOfInj	Sick and injured days in a month	numeric	0.202*
diffWalk	Difficulty in walking	discrete	0.262*
sex	**Gender**	**discrete**	**0.045***
age	**Age**	**discrete**	**0.260***
educ	Education level	discrete	-0.156*
income	Income level	discrete	-0.217*

TABLE V
INPUT FEATURES, MEANING, SPEARMAN'S RANK CORRELATION OF SMOKE

Features	Meanings	Type	r
age	**Age**	**numeric**	**-0.178***
height	**Height**	**numeric**	**0.403***
weight	**Weight**	**numeric**	**0.311***
waist	Waist circumference	numeric	0.212*
eye_left	Measure of left-eye eyesight	numeric	0.093*
eye_right	Measure of right-eye eyesight	numeric	0.105*
hear_left	Measure of left-ear hearing	discrete	-0.024*
hear_right	A measure of right-ear hearing	discrete	-0.018*
sys	**Systolic**	**numeric**	**0.057***
dia	**Diastolic**	**numeric**	**0.086***
fbs	Fasting blood sugar	numeric	0.071*
chol	**Cholesterol**	**numeric**	**-0.047***
trig	Triglyceride	numeric	0.222*
hdl	High-density lipoprotein	numeric	-0.199*
ldl	Low-density lipoprotein	numeric	-0.051*
hemo	Hemoglobin value	numeric	0.412*
urineP	Urine protein level	discrete	-0.004
serum	Serum creatinine value	numeric	0.273*
AST	Aspartate aminotransferase enzyme	numeric	0.052*
ALT	Alanine aminotransferase enzyme	numeric	0.173*
GP	Guanine nucleotide-binding proteins	numeric	0.350*
dental	Dental caries level	discrete	0.113*

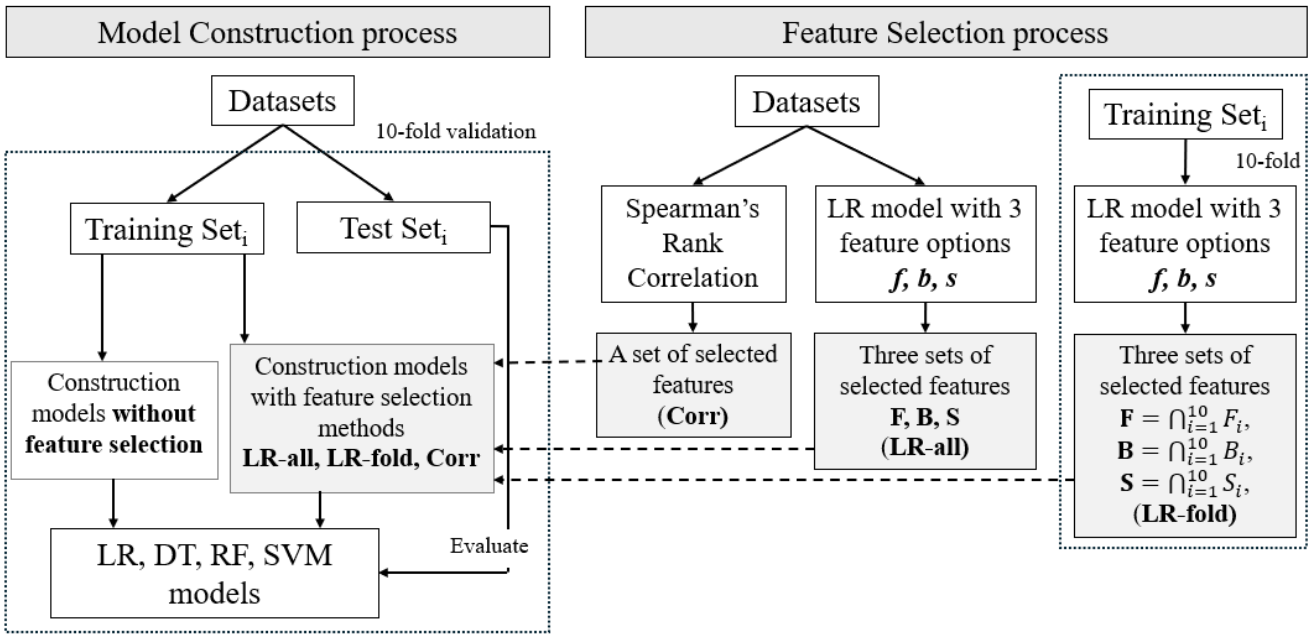

Fig. 2. The implementation process for the model construction and feature selection process for each method.

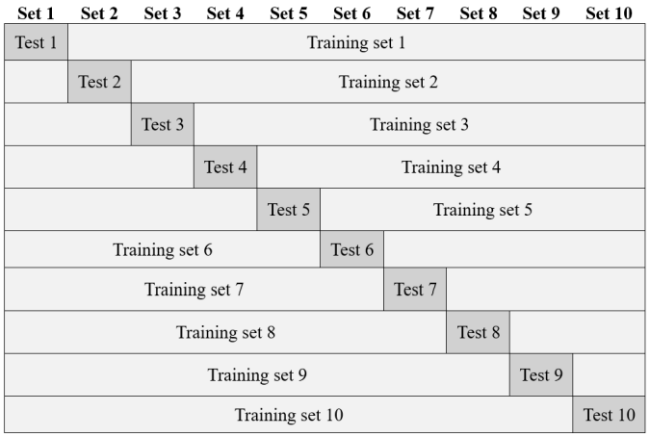

Fig. 1. Stratified random sampling on 10-fold cross-validation.

Step 3: Classification models were constructed using nine datasets as the training set, with one dataset reserved as the test set. For instance, as illustrated in Fig. 2, data from Sets 2 to 9 were combined to form the training set, with Set 1 designated as the test set. The model was developed using the training set and evaluated with the test set over 10 repetitions. The model's performance was assessed by calculating the average results from the test sets, as shown in Fig. 2.

C. Feature selection processes

This research utilizes three methods for feature selection, as illustrated in Fig 2. The first method, referred to as **Method 1 (Corr)**, employs basic statistical analysis using the correlation coefficient. The other two methods are based on logistic regression (LR) analysis and are detailed as **Method 2 (LR-all)** and **Method 3 (LR-fold)**.

Method 1 (Corr): The basic statistics process involved calculating the correlation coefficients for the datasets. Specifically, the Spearman's Rank Correlation Coefficient was computed for each feature in relation to the class (output) across all datasets. Features were selected based on their correlation coefficient values, as presented in Tables III to V. The top features, identified by the highest absolute correlation values, included 10 features for the Cardio dataset, 16 for the Diabetes dataset, and 14 for the Smoke dataset. The Spearman's rank correlation coefficient was chosen because the class values were qualitative.

Method 2 (LR-all): Using the complete datasets for the LR analysis, with three options (*f*, *b*, and *s*) at a significance level of 0.05, this method ultimately resulted in three sets of selected features: **F**, **B**, and **S**.

Method 3 (LR-fold): Our proposed feature selection process, which builds on the 10-fold concept, utilizes each training set in the LR analysis. As a result, we obtained ten sets of selected features for each option (F_i, B_i, and S_i, where *i* ranges from 1 to 10). The final sets of selected features (**F**, **B**, and **S**) for each option were created by taking the intersection of these ten sets. Consequently, the number of selected features was always less than or equal to the total number of features considered in **Method 2 (LR-all)**.

D. Implementing Design and Tools

The study was conducted using the R programming language along with various packages, including blorr, RWeka, randomForest, and e1071. The implementation steps are outlined as follows:

1. Construct classification models using four different techniques, applying the specified parameter values for each dataset without using feature selection methods. Each model is repeated 10 times and evaluated using the test set.
2. Select input features using three feature selection methods as described in Section III - C.
3. Construct classification models again using the same four techniques with the specified parameter values, this time utilizing the set of selected features from step 2. Each model is repeated 10 times and evaluated on the test set.

IV. Results and Discussion

The classification models of each dataset were developed using LR, DT, RF, and SVM techniques, without any feature selection. These models were constructed based on the 10-fold cross-validation approach. The average accuracy of the predictive models on both the training and test sets is presented in Table VI. As outlined in Section III – B, the RF

and SVM techniques were models and tests with various parameter values.

TABLE VI
THE AVERAGE ACCURACY OF TRAINING AND TEST SETS FOR CLASSIFICATION MODELS ACROSS EACH DATASET BY TECHNIQUES

	Cardio		Diabetes		Smoke	
	Accuracy		Accuracy		Accuracy	
	Train	Test	Train	Test	Train	Test
LR	**72.73**	**72.71**	**74.48**	**74.43**	**73.78**	**73.64**
DT	**75.23**	**73.02**	**83.72**	**71.75**	**90.32**	**70.80**
RF (4, 250)	88.18	72.14	**94.31**	**74.01**	100	75.24
RF (4, 500)	**88.22**	**73.16**	**94.34**	**74.01**	**100**	**75.32**
RF (5, 250)	96.23	72.49	97.26	73.55	100	75.20
RF (5, 500)	96.34	72.55	97.33	73.60	100	75.19
SVM (0.001,1)	72.73	72.72	74.47	74.41	74.12	74.00
SVM (0.01,1)	73.14	73.09	74.85	74.74	75.95	75.38
SVM (0.1,1)	**73.70**	**73.45**	75.79	74.70	83.26	75.16
SVM (0.2,1)	74.08	73.41	77.59	74.53	90.80	74.26
SVM (0.001,10)	72.96	72.95	74.76	74.67	74.91	74.65
SVM (0.01,10)	73.27	73.19	**75.13**	**74.87**	**77.06**	**75.78**
SVM (0.1,10)	74.19	73.34	78.53	74.08	94.07	72.45
SVM (0.2,10)	75.61	72.94	84.71	72.38	99.68	71.33

The accuracy of the models on training sets is consistently higher than on test sets. Some models, particularly the Random Forest (RF) models, show significantly higher accuracy. For instance, the RF model achieved 100 percent accuracy on the training set but only 75.32 percent on the test set for the smoking dataset. Additionally, as illustrated in Table VI, high accuracy on the training set does not guarantee high accuracy on the test set.

In this research, we evaluated performance by calculating the average accuracy across ten test sets. Fig. 3 shows the top-performing models from the LR, DT, RF, and SVM techniques applied to Cardio, Diabetes, and Smoke datasets. These results will serve as a baseline for comparison when using feature selection methods in modeling.

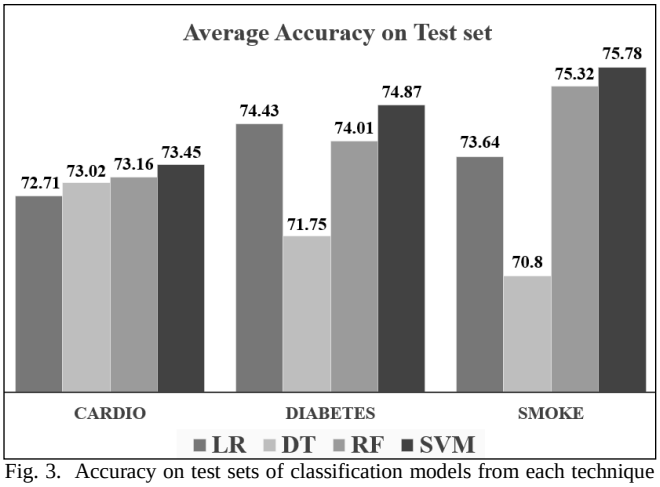

Fig. 3. Accuracy on test sets of classification models from each technique on each dataset based on the 10-fold concept (without feature selection)

As shown in Fig. 3, SVM outperformed other models on the three datasets: Cardio (73.45% accuracy), Diabetes (74.74%), and Smoke (75.78%). RF performed the second best on Cardio and Smoke datasets, and LR was the second best on Diabetes, while DT performed the worst on Diabetes and Smoke.

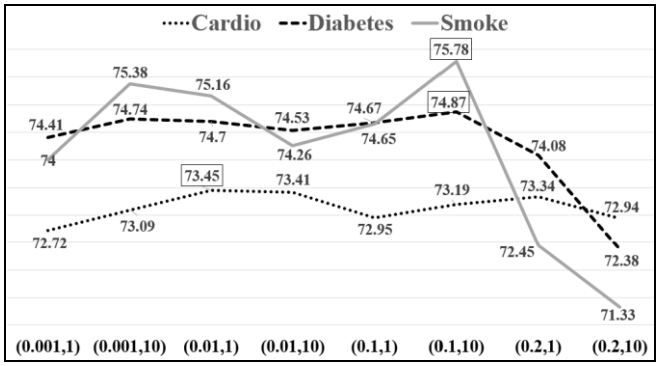

Fig. 4. The accuracy of SVM models on test sets with designed parameter values of grammar and cost: (grammar, cost).

The SVM models constructed from various parameter settings are shown in Fig. 4. The minimum and maximum SVM accuracy values, determined by the designed parameter settings are 72.72% and 73.45% for Cardio, 72.38% and 74.87% for Diabetes, and 71.33% and 75.78% for Smoke, as shown in Fig. 5. The best SVM models for Cardio were obtained with gamma = 0.1 and cost = 1, while for both Diabetes and Smoke were gamma = 0.01 and cost = 10. The parameter setting values with gamma = 0.001, cost = 1 and gamma = 0.2, cost = 10 gave the lowest accuracy values for SVM models on three datasets. Hence, these setting values are not recommended.

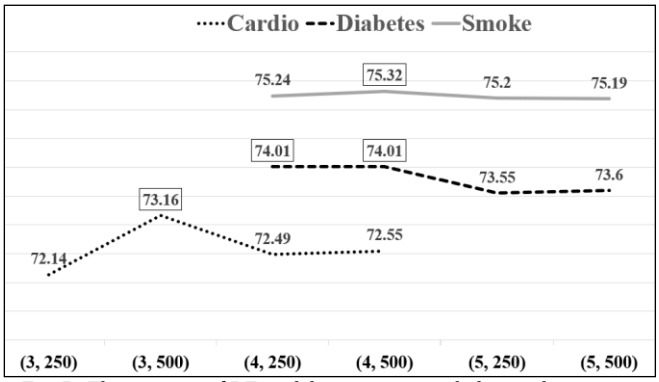

Fig. 5. The accuracy of RF models on test sets with designed parameter values of number of features (n) and number of trees (m): (n, m).

As explained previously in Section III-B, RF has two key parameters: the number of features (n) and the number of trees (m). To determine the number of features, we calculated n as the square root of the total number of features in the dataset. For example, the Cardio dataset has 11 features, so n is the square root of 11, which is approximately 3.312. Consequently, we set n for the RF models for Cardio to 3 and 4. Similarly, we calculated n for the Diabetes and Smoke datasets, obtaining values of 4 and 5, respectively. We set the number of trees (m) to 250 and 500. The accuracy of the RF models with these specified parameters is shown in Fig. 5. Among these, setting the number of trees to 500 yielded the best accuracy for the RF models across all three datasets.

TABLE VII
SELECTED FEATURES FROM EACH METHOD AND OPTION

Methods	Selected Features	No.
	1. Cardio	
LR-all (F, B, S) LR-fold (F, B, S) Corr	age, height, weight, sys, dia, chol, glu, smoke, alc, exer	10
	2. Diabetes	
LR-all (F, B, S)	hiBP, hiChol, cholCk, bmi, stroke, heartAtt, veg, alcoAdd, physH, dayOfMent, dayOfInj, diffWalk, sex, age, educ, income	16
LR-fold (F, B, S)	hiBP, hiChol, cholCk, bmi, stroke, heartAtt, alcoAdd, physH, dayOfMent, dayOfInj, diffWalk, sex, age, income	14
Corr	hiBP, hiChol, cholCk, bmi, smoke, stroke, heartAtt, exer, alcoAdd, physH, dayOfMent, dayOfInj, diffWalk, age, educ, income	16
	3. Smoke	
LR-all (F, S)	height, weight, waist, sys, fbs, chol, trig, hdl, hemo, serum, AST, ALT, GP, dental	14
(B)	height, weight, waist, sys, fbs, trig, hdl, ldl, hemo, serum, AST, ALT, GP, dental	14
LR-fold (F, S)	height, weight, waist, sys, fbs, chol, trig, hdl, hemo, ALT, GP, dental	12
(B)	height, weight, waist, sys, fbs, trig, hdl, hemo, ALT, GP, dental	11
Corr	age, height, weight, waist, eye_left, eye_right, dia, trig, hdl, hemo, serum, ALT, GP, dental	14

The second part of the implementation focused on the feature selection process. At the conclusion of this process, we identified a set of selected features for each dataset based on logistic regression (LR) feature selection methods, including forward selection (*f*), backward elimination (*b*), and stepwise selection (*s*). The selected features for the Cardio dataset were consistent across all three methods. For the Diabetes dataset, the selected features varied among the three methods for the LR-all and LR-fold approaches; however, the three options (*f, b*, and *s*) ultimately yielded the same set of selected features. In the case of the Smoke dataset, the features selected using forward selection (*f*) and stepwise selection (*s*) remained the same, while the backward elimination (*b*) method produced a different set, as detailed in Table VII.

Because the parameter values for the most accurate RF and SVM models were unstable as depicted in Fig. 4 and Fig. 5, we constructed the models using all the designed parameters along with a selection of features from three methods, as illustrated in Table VIII.

The performance of the models was evaluated using 10-fold cross-validation, measuring accuracy (Acc) and weighted F1 score ($F1_w$). The results are presented in Table VIII, which shows the performance of classification models for each technique. The highlighted row indicates the performance of the models without feature selection, as depicted in Fig. 3. This served as the baseline for assessing improvements.

The symbol "%" represents the percentage difference in accuracy from the baseline models of each technique and dataset, calculated as follows: $\% = (Acc - Acc_{base}) / Acc_{base}$. In comparing the performance, we first consider the accuracy (Acc), followed by the weighted F1 score ($F1_w$), and finally, the number of features used in the models. A smaller number of features indicates a better and more straightforward model.

In the case of Logistic Regression (LR), the models that included feature selection outperformed those that did not across three datasets. For the Cardio and Diabetes datasets, the accuracy remained the same; however, the weighted F1 score ($F1_w$) improved, and the number of features was reduced to 10 and 16, respectively. In the Smoke dataset, the accuracy increased to 73.76%, with the number of features reduced to 12.

In the Decision Tree (DT) technique, no improvement was observed in predicting cardiovascular disease. However, when feature selection was applied using all three methods, there was a moderate increase in the models' accuracy for predicting diabetes and smoking. This suggests that feature selection can enhance the performance of DT models.

Similarly, for the Random Forest (RF) method, there was also no improvement in the predictions for cardiovascular disease. Nevertheless, the accuracy of the RF models improved for diabetes and smoking when specific feature selection methods, such as LR-fold and correlation, were utilized.

The accuracy of the SVM models with feature selection methods mostly decreased, suggesting that the feature selection methods applied in this study could not improve the performance of the SVM models. However, there was a case in the Diabetes dataset where the accuracy of the model with LR-all was unchanged (74.87), but the number of features was reduced to 16. Therefore, it could be considered as an improvement.

Table IX presents the accuracy, number of features, and feature selection methods of the best classification models for each dataset, comparing the results with and without feature selection methods across each technique.

In the case of the Cardio dataset, the feature selection methods improved the LR model by reducing the number of features to 10. However, there was no improvement in the models constructed using DT, RF, and SVM. This lack of improvement may be due to the small number of features in the dataset, which consists of only 11 features. In contrast, for the Diabetes and Smoke datasets, the feature selection methods enhanced the accuracy of the models across almost all techniques, except for the SVM model on the Smoke dataset. Among the feature selection methods, the LR-fold method generally yielded the best performance, followed by the LR-all and correlation (Corr) methods.

The feature selection methods applied to the Diabetes and Smoke datasets increased the accuracy of DT models from 71.75% to 73.13%, representing an improvement of approximately 1.93%. For the Smoke dataset, the accuracy of RF models improved from 74.01% to 74.42% (about 0.55%) and from 75.32% to 75.34% (about 0.03%), respectively. Additionally, for the Smoke dataset, the most accurate models were achieved using the logistic regression feature selection option, specifically LR-fold (*f, s*) for LR models and LR-fold (*b*) for DT models.

When considering the number of feature reductions, Cardio was possibly reduced from 11 features to 10, Diabetes from 16 features to 14, and Smoke from 22 features to 11.

In conclusion, the SVM model outperformed other techniques across three datasets. It achieved an accuracy of 73.45% using 11 features for the Cardio dataset, 74.79% with 16 features for the Diabetes dataset (utilizing the LR-all

TABLE VIII
THE PERFORMANCE OF FOUR CLASSIFICATION MODELS WITH FEATURE SELECTION METHODS ON THREE DATASETS

Dataset	Feature selection (number of selected features)	LR			DT			RF			SVM		
		$F1_w$	**Acc**	**%**	**$F1_w$**	**Acc**	**%**	**$F1_w$**	**Acc**	**%**	**$F1_w$**	**Acc**	**%**
Cardio	without selection (11)	72.56	72.71		**72.97**	**73.02**		**73.12**	**73.16**		**73.34**	**73.45**	
	LR-all (10)	**72.60**	**72.71**	0.000	72.94	73.01	-0.014	73.01	73.06	-0.137	73.29	73.41	-0.054
	LR-fold (10)	72.60	72.71	0.000	72.94	73.01	-0.014	73.01	73.06	-0.137	73.29	73.41	-0.054
	Corr. (10)	72.60	72.71	0.000	72.94	73.01	-0.014	73.01	73.06	-0.137	73.29	73.41	-0.054
Diabetes	without selection (21)	74.40	74.43		71.73	71.75		73.96	74.01		74.71	74.87	
	LR-all (16)	**74.41**	**74.43**	0.000	72.40	72.44	*0.962*	73.96	74.01	0.000	**74.71**	**74.87**$^{\#}$	0.000
	LR-fold (14)	74.40	74.42	-0.013	**73.07**	**73.13**	*1.923*	**74.35**	**74.42***	*0.554*	74.58	74.76	-0.147
	Corr. (16)	74.25	74.27	-0.215	72.26	72.30	*0.767*	73.67	73.74	-0.365	74.45	74.61	-0.347
Smoke	without selection (22)	73.00	73.64		70.59	70.80		75.10	75.32		**75.47**	**75.78**	
	LR-all (14)	73.02	73.66	*0.027*	71.37	71.58$_{f,s}$	*1.102*	74.51	74.72$^{+}_{b}$	-0.797	75.04	75.41$^{\wedge}_{b}$	-0.488
	LR-fold (11/12)	**73.10**	**73.76**$_{f,s}$	*0.163*	**72.52**	**72.70**$_{b}$	*2.684*	74.26	74.44$_{f,s}$	-1.168	74.56	75.03$^{\wedge}_{f,s}$	-0.990
	Corr. (14)	72.71	73.40	-0.326	71.47	71.57	*1.088*	**75.12**	**75.34***	*0.027*	75.03	75.54$^{\wedge}$	-0.581

$_{b, f, s}$ delivers the best accurate models on Smoke dataset. *RF(4, 500) $^{+}$RF(4, 250) $^{\#}$SVM(0.01, 1) $^{\wedge}$SVM(0.01, 10)

TABLE IX
THE BEST CLASSIFICATION MODELS FOR EACH DATASET BY TECHNIQUES

	Cardio	**Diabetes**	**Smoke**
	Acc (No. of features) / Method	Acc (No. of features) / Method	Acc (No. of features) / Method
LR	72.71 (10) / LR-fold, LR-all, Corr	74.43 (16) / LR-all	73.76 (12) / LR-fold (f, s)
DT	73.02 (11) / without selection	73.13 (14) / LR-fold	72.70 (11) / LR-fold (b)
RF	73.16 (11) / without selection	74.42 (14) / LR-fold	75.34 (14) / Corr
SVM	**73.45 (11) / without selection**	**74.87 (16) / LR-all**	**75.78 (22) / without selection**

feature selection), and 75.78% with 22 features for the Smoke dataset, as presented in Table IX.

Feature selection methods were evaluated on the three datasets. For the Cardio dataset, models without feature selection achieved the best results. In the case of the Diabetes dataset, the feature selection methods LR-all and LR-fold improved the accuracy of all models while also reducing the number of features.

In the case of the Smoke dataset, the feature selection methods LR-fold and Corr enhanced accuracy and minimized the number of features. However, it was observed that the SVM model without feature selection achieved the highest accuracy. Notably, the RF model using the Corr method significantly reduced the number of features by nearly half while maintaining accuracy comparable to that of the SVM.

V. CONCLUSION

This paper focuses on developing classification models using three health datasets—cardio, diabetes, and smoke—and employing four different techniques: Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), and Support Vector Machine (SVM). Additionally, it examines the impact of various feature selection methods, including Spearman's Rank Correlation Coefficient and LR analysis, which includes three approaches: forward selection, backward elimination, and stepwise selection. Two methods were used to derive sets of features selected by LR: LR-all and LR-fold. The results indicated that the LR-fold approach performed better, as it increased the accuracy of many models while also reducing the number of selected features compared to other methods. Furthermore, the LR analysis feature selection method led to a more significant improvement in accuracy than the correlation feature selection method. The important features identified from these feature selection methods for each dataset are presented in Table VII.

The results indicated that the Support Vector Machine (SVM) outperformed the other techniques across all datasets. However, constructing these models was quite complex, as varying parameter values were necessary to achieve the most accurate model. The study revealed that feature selection methods could reduce the number of features while enhancing the accuracy of the models, particularly for the Diabetes and Smoke datasets, which contain more features than the Cardiovascular dataset. These feature selection techniques primarily improved the accuracy of the Logistic Regression (LR) and Decision Tree (DT) models, with only slight improvements seen in the Random Forest (RF) and Support Vector Machine (SVM) models.

REFERENCES

[1] J. Han, M. Kamber and J. Pei, *Data Mining Concepts and Techniques, 3rd Edition*, Morgan Kaufmann Publishers, Burlington, 2011.

[2] S. H. Liao, P. H. Chu and P. Y. Hsiao, "Data Mining Techniques and Applications – A decade review from 2000 to 2011," *Expert Systems with Applications*, vol.39, no.12, pp11303-11311, 2012.

[3] N. A. Fridausanti, R. A. Ningrum and S. Oomariyah, "Comparisons of Logistic Regression and Support Vector Machine in Classification of Echocardiogram Dataset," *INFERENSI*, vol.5, no.2, pp85-90, 2022.

[4] S.M. Birjandi and S.H. Khasteh, "A survey on data mining techniques used in medicine," *Journal of Diabetes & Metabolic Disorders*, vol.20, no.2, pp2055-2071, 2021.

[5] Y. Saeys, I. Inza, P. Larrañaga, "A review of feature selection methods in bioinformatics," *Bioinformatics*, vol.23, no.9, pp2507–2517, 2007.

[6] R. Sangeetha and T.N. Ravi, "Random Probit Regressive Decision Forest Classification based IoT aware Content Caching with Healthcare Data," *IAENG International Journal of Computer Science*, vol.51, no.6, pp582-593, 2024.

[7] K. Dissanayake and M. G. Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, 5581806, 17 pages, 2021. Available: https://doi.org/10.1155/2021/5581806

[8] C. Aroef, Y. Rivan and Z. Rustam, "Comparing random forest and

support vector machines for breast cancer classification," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol.18, no.2, pp815-821, 2020.

[9] K. Slime, A. Maizate, L. Hassouni, and N. Mouine, "Toward a Model to Predict Cardiovascular Disease Risk Using a Machine Learning Approach," *IAENG International Journal of Computer Science*, vol.51, no.5, pp519-527, 2024.

[10] R. Paisanwarakiat, A. Na-udom, J. Rungrattanaubol, "Combining Logistic Regression Analysis with Data Mining Techniques to Predict Diabetes," *Lecture Notes in Networks and Systems: Proceedings of the 18th International Conference on Computing and Information Technology (IC2IT 2022)*, pp88-98, 2022.

[11] N. Nai-arun and R. Moungmai, "Diagnostic Prediction Models for Cardiovascular Disease Risk using Data Mining Techniques," *ECTI Transactions on Computer and Information Technology*, vol.14, no.2, pp113-121, 2020.

[12] A. R. Olivera, V. Roesler, C. Lochpe, M. I. Schmidt, A. Vigo, S. M. Barreto and B. B. Duncan, "Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes – ELSA-Brasil: accuracy study," *Sao Paulo Medical Journal*, vol.135, no.3, pp234-246, 2017.

[13] B. S. F. Astuti, N. A. Fridausanti and S. W. Purnami, "Model Evaluation for Logistic Regression and Support Vector Machines in Diabetes Problem," *INFERENSI*, vol.1, no.2, pp77-82, 2018.

[14] S. Raghavendra and K. J. Santosh, "Performance evaluation of random forest with feature selection methods in prediction of diabetes," *International Journal of Electrical and Computer Engineering*, vol.10, no.1, pp353-359, 2020

[15] W. Chansra and S. M. Isa, "Diabetes prediction using ensemble stacking with LASSO and Genetic Algorithm for feature selection," *ICIC Express Letters*, vol.16, no.12, pp1341-1349, 2022.

[16] E. Dritsas and M. Trigka, "Data-driven machine-learning methods for diabetes risk prediction," *Sensors*, vol.22, no.14, pp5304, 2022.

[17] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," *24th International Conference on Automation and Computing (ICAC)*, Newcastle upon Tyne, UK, pp1-6, 2018.

[18] I. B. K. Manuaba, I. Sutedja and R. Bahana, "The evaluation of supervised classifier models to develop a machine learning API for predicting cardiovascular disease risk," *ICIC Express Letters*, vol.14, no.3, pp219-226, 2020.

[19] R. Assari, P. Azimi and M. R. Taghva, "Heart Disease Diagnosis Using Data Mining Techniques," *International Journal of Economics & Management Sciences*, vol.6, no.3, pp72-79, 2017.

[20] N. M. Ali, N. A. Aziz and R. Besar, "Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol.20, no.2, pp712-719, 2020.

[21] L. Breiman. Random Forests. *Machine Learning, 45*(1), 5-32, 2001.

[22] G. Rebala, A. Ravi and S. Churiwala, *An Introduction to Machine Learning*, Cham: Springer, 2019.

[23] V. N. Vapnik, *The Nature of Statistical Learning Theory, 2nd Edition*, Springer New York, NY, 2000.

[24] R. Gholami and N. Fakhari, "Support Vector Machine: Principle, Parameters and Applications," *Handbook of Neural Computation*, pp515-535, 2017.