Relevant Features Independence of Heuristic Selection and Important Features of Decision Tree in the Medical Data Classification

Yusi Tyroni Mursityo [1], Irfany Rupiwardani [2], Widhy H. N. Putra [1], Dewi Sri Susanti [3], Titis Handayani [4], and Samingun Handoyo [5,6,*]

[1] Information System Department, Brawijaya University, Malang, Indonesia
[2] Environmental Health Department, Widyagama Husada School of Health Science, Malang, Indonesia
[3] Statistics Study Program, Lambung Mangkurat University, Banjarbaru, Indonesia
[4] Information System Study Program, Semarang University, Semarang, Indonesia
[5] Statistics Department, Brawijaya University, Malang, Indonesia
[6] EECS–IGP Department, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
Email: yusi_tyro@ub.ac.id (Y.T.M.); irfany@widyagamahusada.ac.id (I.R.); widhy@ub.ac.id (W.H.N.P.); ds_susanti@ulm.ac.id (D.S.S.); titis@usm.ac.id (T.H.); samistat@ub.ac.id (S.H.)
*Corresponding author

***Abstract*—The input of predictive models has an important role in directing the classification model to have a satisfactory performance in predicting an unknown label of the instance class. The predictor features should be not only relevant to the target feature but also should be independent of each other. The research objective is to obtain the predictor features that are relevant and independent through feature selection using the Chi-square test, one-way Analysis of Variance (ANOVA), and Pearson's correlation test, and to show important features of the decision tree are different from the features of relevant and independent. The evaluation of irrelevant features yields 44 of 67 relevant features which as many as 18 discrete types with two labels are dropped. The dataset with 44 relevant features is used to train the first decision tree. The relevant features mean that the target feature depends on them. The best predictor features should not only be relevant but also independent of each other. The evaluation of independent among features yields 11 of 44 independent features where the features of numeric and discrete with 2 labels are represented by 1 and three features respectively. The dataset with 11 relevant and independent features is used to train the second decision tree. The important features of both models are very different and the second model has better performance than the first one for the metrics of accuracy, recall, and F1-Score.**

***Keywords*—data mining, decision tree, feature selection, important feature, relevant feature, Pearson's correlation**

I. Introduction

Development models in medical datasets are challenging in the data science field because many datasets have been collected without experimental design or planned sample frame. The data collection has only the goal of recording data for documentation [1]. Medical datasets possibly contain useful information such as patterns underlying the data, or a relationship causality among features [2]. However, Medical datasets tend to not only have a large number of feature dimensions but also have a large number of instances. The datasets consist of both categorical and numerical features. The categorical feature can be a qualitative or discrete data type, and also it has various label numbers [3]. While numerical features have various measurement units [4]. Furthermore, some features may be completely unrelated to the target feature that will be predicted, and some features may be redundant in the sense that two or more such features contain the same predictive information [5].

A simple model is a better one which just involves only relevant features [6]. A good model can be developed when among predictor features are independent of each other, and the target feature is dependent on the predictor features [7]. Selection of features must be done to get a good model that involves only relevant features as the predictor features [8]. The feature selection involving a structure model called the filter approach is not an easy task because the selection process has stages as many as the factorial of d features dimension [9]. Forward selection and backward elimination are popular filter methods applied in linear models both the regression [10] and classification models [11]. However, the application of the filter method in non-linear or assembled models such as a decision tree based on the concept of divide and conquer steps is not a precise decision choice [12]. Zhang and Yang [13] even use the features importance of the decision tree model as the relevant features to build another classification model.

The selection of subset features on the medical dataset with a large number of instances and various scale features should be tackled by effective approaches. One such approach based on some statistical tests in the process of

Manuscript received October 27, 2023; revised December 7, 2023; accepted January 22, 2024; published May 16, 2024.

feature selection is known as the heuristic method [14]. The heuristic method is based on intuition which employs a practical method that is not guaranteed to be optimal or perfect approach but it is nevertheless sufficient to yield an approximation well [15]. The evaluation of dependency between the target and predictor features, and the evaluation of independence among predictor features are key concepts in running the heuristic method [16]. There are 3 statistical tests usually used for the dependency or independency test namely the Chi-square test for evaluating dependency among 2 categorical features [17], Pearson's correlation test for evaluating dependency among two numerical features [18], and one-way Analysis of Variance (ANOVA) for evaluating dependency between the categorical and numerical features [19]. Each of the statistical tests needs to meet a particular condition called the assumption underlying inferences in the tests to produce a valid decision. Fortunately, the assumptions in the inferential statistics are automatically fulfilled when the dataset has a large number of samples or instances. The process of feature selection employed in the heuristic method does not involve any optimization algorithms, and the resulting subset feature is not the optimal one

The research has goals to apply the heuristic feature selection method in the dataset with a large number of instances and various scales of predictor features and to show that important features of the decision tree are different from relevant and independent features yielded by the proposed method. The relevant features obtained by evaluating the dependency between the target and predictor features produce the first dataset employed in training the first decision tree model. The second dataset is produced by evaluating the independence among predictor features of the first dataset and it is employed in training the second decision tree model. Both models are explored in their performances and important features to acquire the insight differences between selected subset features and subset features importance.

The remaining parts of the paper are organized as reviewing the related works presented in Section II, the material and proposed methods presented in Section III, detailing the results in the selection process of relevant and independent features, and exploring both model performance and features importance presented in Section IV, and the conclusion presented in Section V.

II. Literature Review

Finding a good model is an important task in data mining where machine learning algorithms are the most methods implemented there. In general, machine learning methods are divided into unsupervised and supervised learning depending on the existence of the target feature. Unsupervised learning includes the method to make instances rank [20] and to make instances groups by minimizing the variance within groups and maximizing the variance between groups [21]. Supervised learning models are also known as predictive models which are regression models if the target feature has a numerical scale [22] and classification models if the target feature has a categorical scale [23]. The decision tree model is built by using the repeating of divide and conquer steps until the stopping criteria are fulfilled. Widodo *et al.* [24] showed that the C4.5 tree performed similarly to the Convolution Neural Networks of 1 Dimension (CNNID) where the building CNN1D is a very complex task. Marji and Handoyo [25] explored the comparison performance between ridge logistic and decision tree models where the decision tree is moderately better than the ridge regression. Zaini and Awang [26] show the decision tree performance compared to 10 other machine learning methods is better excluding the Random Forest is the best one. A decision tree model is not only easily developed directly from the original features of a dataset but also needs a low computational resource. The model interpretation is also simple that is by the traveling tree from the root node to leaf nodes which are directly presenting the instance labels. There are enough reasons to explore the deeper and better implementation of the decision tree model.

The garbage in and garbage out are trademarks in system development which means the quality of system inputs is a determinant factor in the system output quality [27]. The selection of relevant features is an effort to acquire a high-quality subset of predictor features as the model inputs [28]. Feature selection is absolutely a needed task that has to be conducted before building models. As the development of a tree model directly operates feature by feature, the original features should be employed in the process of model building as the input features. A heuristic feature selection is one approach that is adequate or suitable for the feature selection of the decision tree [29]. Some intensive researches are conducted to acquire the most relevant subset features, subsequently, the classification model performance on the medical datasets can be increased. Tran and Tran [30] developed a model for heart disease prediction based on the rank and weights assigned by the Infinite latent feature selection method. Nguyen *et al.* [31] explored swarm intelligence techniques to acquire the subset relevant features in classification models focusing on the representation and searching mechanisms. Sabeena and Sarojini [32] carried out the selection large set of features by using statistical analysis and applied the Ant Colony algorithm on the cancer dataset. Furthermore, Grissa *et al.* [33] conducted feature selection based on evaluating a combination of numeric and symbolic features of metabolisms to acquire the best one for yielding an effective and accurate classification model.

Even though feature selection methods employed in machine learning models have involved sophisticated algorithms, some researchers still rely on important features of the decision tree model as the relevant features subset for the predictor features of classification models. The selection of the best relevant features subset by using the decision tree to train multilayer perceptron networks was carried out by Ahmed and Jameel [34]. The random forest tree model for handling the feature selection issue in the number of higher features in three popular datasets namely Bank Marketing, Car Evaluation, and Human Activity Recognition using Smartphones employed in training the machine learning models was published in Chen's study [35]. The extraction of useful service quality

features by using a hybrid model of the Information System and the decision tree to enhance customer satisfaction and loyalty was conducted by Romalt and Kumar [36]. The feature importance of the Decision tree to acquire symbolized sets of damaged buildings as the relevant features only using post-earthquake information was conducted by Wang *et al.* [37]. Finding the feature relevant by identifying and ranking the scores of feature importance generated by three techniques namely impurity-based, permutation-based, and Shap values for building a model to predict breast cancer was done by Mathew [38]. Another subset features selection approach published in Le's research [39] employed the hybrid between Bio-Inspired Optimization (GA and PSO) and 11 popular classifier models of machine learning employed in classifying Parkinson's disease patients. Pasha and Latha [40] predict the probability of the patient of wart skin disease belonging to each possible label rather than predicting a label value directly in the multivariate healthcare dataset.

Based on the literature review above, it can be perceived that much research conducted in acquiring the optimal subset of predictor features involving machine learning models ranging from simple models to sophisticated ones Even some published works reported the hybrid between the optimization methods and machine learning models. Those approaches are based on the model accuracy as the selection criteria and the optimal subset features can be acquired. Nevertheless, the trade-off has to be paid including the expensive computation resources required in conducting the approaches. Some works also directly employed the important features of decision tree models as the relevant subset features subsequently employed in training the machine learning models where the important features are very possibly not the relevant features. The heuristic approach is one of the feature selection methods without involving machine learning models and solely employing statistical tests. The method produces the subset features retaining the original form although the subset features acquired are not assured the optimal ones. There are still open problems in acquiring the relevant subset features as the input of machine learning models. Hence, the research offers the implementation of the heuristic approach to acquire the relevant and independent subset features and implicitly shows that the important features of the decision tree are different from the relevant features acquired by the proposed method.

III. Materials and Methods

The dataset consists of 52,159 examples (instances) with 68 features (columns) which are 23 numerical features and 45 categorical features including the "outcome" feature as the target feature with binary classes namely a surgery treatment was not conducted on the patient (Class 0) or the surgery treatment was conducted (Class 1). The categorical features consist of the qualitative and discrete types where there are only 2 qualitative types namely the "SEX" and "joint" features. The dataset was obtained from a medical recording part of a Taiwanese hospital. The work is a part of the data mining project.

Table I presents a brief description of all features in the medical dataset employed in the research.

TABLE I. The Description of Features in the Medical Record Dataset

Feature Name	Number of columns and their description
ID	1, Col_1 (Decimal)
Outcome	1, Col_2 (Binary)
AGE	1, Col_3 (Numeric)
SEX	1, Col_4 (Binary)
Length of stays ("LOS")	1, Col_5 (Numeric)
Routine blood test ("OP_time_minute", "OP_time_hour", "ASA", "CBC_WBC", "CBC_RBC", "CBC_HG", "CBC_HT", "CBC_MCV", "CBC_MCH", "CBC_MCHC", "CBC_RDW", "CBC_Platelet", "CBC_RDW", "BUN", "Crea", 'GOT', 'GPT', "ALB", "Na", "K")	20, Col_6toCol_25 (Numeric)
Uric Acid ("UA")	1, Col_26 (Numeric)
History disease category A ("Drain", "Cemented", "Commercial_ALBC", "Non_commercial_ALBC", "Blood_trans", "Congestive Heart Failure", "Cardiac Arrhythmia", "Valvular Disease", "Pulmonary Circulation Disorders", "Peripheral Vascular Disorders", "Hypertension Uncomplicated", "Paralysis", "Other Neurological Disorders", "Chronic Pulmonary Disease", "Diabetes", "Hypothy roidism", "Renal Failure", "Liver Disease", "Peptic Ulcer Disease excluding bleeding", "AIDS/HIV", "Lymphoma", "Metastatic Cancer", "Solid Tumor without Metastasis", "Rheumatoid Arthritis/collagen" "Coagulopathy", "Obesity", "Weight Loss", "Fluid and Electrolyte Disorders", "Blood Loss Anemia", "Deficiency Anemia", "Alcohol Abuse", "Drug Abuse", "Psychoses", "Depression")	35, Col_27toCol_61 (Binary)
History disease category B ("Lung disease", "Anemia", "Psyciatric disorder")	3, Col_62toCol_64 (3_labels)
History disease category C ("Heart disease", "Cancer history")	2, Col_65toCol_66 (4_labels)
Diagnosis	1, Col_67 (6_labels)
elx_index	1, Col_68 (13_labels)
cci_index	1, Col_69 (16_labels)

The categorical features consist of various label numbers ranging from 2 to 16. The numerical features have various measurement units, missing values dominating them, and also some outliers suffered by many instances. Table II shows the occurrence of missing value numbers on the numerical features. On the other side, there is no missing value on the categorical features.

The numerical feature consists of 21 features with missing value numbers ranging from 2106 to 48,564 instances with an exception on 2 features of the LOS and AGE only 94 instances with missing values. The data

preprocessing concerns the missing value imputation, commensurate measures unit scale, and treatment to outlier observations that have to get the attention seriously. Some limitations employed in the data preprocessing are determined subjectively namely the missing values imputation is tackled by using 0 or mean value, the commensurate measure is conducted by transforming numerical features into the standardized normal score, the observation is justified as an outlier when it has a value greater than absolute of three and the associated instance is dropped out dataset. The data preprocessing concerns the missing value imputation, commensurating measures unit scale, and treatment to outlier observations that have to get the attention seriously.

TABLE II. THE NUMBER OF MISSING VALUES IN THE DATASET

No.	Missing Value		No.	Missing Value	
	Feature	Count		Feature	Count
1	CBC RDWCV	48564	13	GPT	34312
2	UA	39442	14	Crea	30030
3	ALB	36891	15	GOT	29282
4	Na	36265	16	BUN	28057
5	K	35993	17	CBC_HT	27542
6	CBC MCV	35800	18	CBC_HG	27425
7	CBC_RDW	35795	19	ASA	20158
8	CBC MCHC	35790	20	OP_time hour	20106
9	CBC MCH	35787	21	OP_time minute	20106
10	CBC RBC	35779	22	LOS	94
11	CBC WBC	35751	23	AGE	94
12	CBC Platelet	35605			

There are four main processes involved in researching to acquire the important features of both decision tree models based on the relevant and independent subset features. The stages are summarized in the following: a). Evaluating the dependency of the target features on the predictor features by employing the Chi-Square and One-way ANOVA tests on the dataset produced by the preprocessing data, and ultimately acquiring the first dataset; b). Evaluating the independence among the predictor features by employing the Chi-Square, One-way ANOVA, and Pearson Correlation tests on the first dataset, and ultimately acquiring the second dataset; c). Training both decision tree models employing the first and second acquired datasets respectively to build Model 1 and Model 2; d). Assessing the performance metrics and exploring the important features of both decision tree models.

In stage (a), there are employed the chi-square test for evaluating the categorical features, and the One-way ANOVA for evaluating numerical features. In stage (b), there are some limitations in evaluating independence among predictor features including the Chi-square test only conducted among binary predictor features, and the One-way ANOVA conducted only between numerical and binary features. The categorical features with the class of more than 2 labels are automatic as elements of subset selected features. All statistical tests employ a significant level of 5% in making inferences. The building of the decision tree models tunes adequately the tree depth and the minimum number of instances in the leaf node as the pruning criteria. The description of methods employed in this research is briefly presented in the following.

A. Chi-Square Test for Dependency between 2 Categorical Features

The test proves the hypothesis that two categorical features are independent of each other. The statistics Chi-square is calculated based on a contingency table which has the row and column numbers associated with the class number of both categorical features. The formula of Chi-square statistics is given in Eq. (1) as follows [41]:

$$\chi^2 = \sum \frac{(Expected - Observed)^2}{Expected} \tag{1}$$

The steps to conduct the Chi-square test are:

- Create a contingency table where the cell value is associated with the number of instances that came from the cross-level of both features.
- Calculate the expected frequencies of each cell.
- Calculate the Chi-square statistic and the associated *p*-value.
- Make a decision which is to reject H_0 if *p*-value < 0.05.

B. One Way ANOVA for Evaluating Dependency between the Categorical and Numerical Features

The one-way ANOVA is the extended T-test for differences between 2 population means on unpaired data. The dependency between the numerical and categorical features can be evaluated by using the one-way ANOVA in general. The categories or levels of a categorical feature as groups or blocks, while the numerical feature data is the response values. The F-statistic test has the main role that can be calculated by using Eq. (2) as the following [42]:

$$F = \frac{Variance_{\ between\ groups}}{Variance_{\ within\ groups}} \tag{2}$$

$$SS_T = SS_B + SS_W \tag{3}$$

The sum square between blocks of SS_B and the sum square within blocks of SS_W must satisfy Eq. (3). Both variances in Eq. (2) can be calculated by dividing both sum squares by the associated degree of freedom. The steps in conducting a one-way ANOVA test are given as follows:

- Format data of both features into 2 dimensions where the column represents the categorical feature levels and the row represents instances of the numerical features.
- Calculate the SS_T.
- Calculate the SS_B.
- Calculate the $SS_W = SS_T - SS_B$.
- Calculate the *Variance between groups.*
- Calculate the *Variance within groups.*
- Calculate F statistic and the associated *p*-value.
- Make a decision which is to reject H_0 if *p*-value < 0.05.

C. Pearson's Correlation Test for Evaluating Dependency between 2 Numerical Features

A level of relationship between 2 numerical features is measured using the correlation value which is a popular metric called Pearson's correlation coefficient [43]. The

coefficient is calculated by standardized the $cov(x, y)$ covariance value that is divided by the multiplied standard deviations both features of S_x and S_y [44]. Let considering 2 features of the *X* and *Y*, the Pearson's correlation is calculated by using Eq. (4) given as follows [45]:

$$r_{xy} = \frac{cov(x,y)}{S_x \times S_y} \quad (4)$$

Because $F_{\left(df_{\tilde{y}}, df_{er}\right)} = \frac{r^2 \times (n-2)^2}{1-r^2}$ and $F = t^2$ then the T-statistic of $t_{(n-2)}$ can be calculated through the Person's correlation with the following formula:

$$t_{(n-2)} = \frac{r \times (n-2)}{\sqrt{1-r^2}} \quad (5)$$

The T statistic in Eq. (5) has the *T* distribution with *n*-2 degrees of freedom that is used for the evaluation of dependency between 2 numerical features where the steps are the following:

- Calculate the covariance of both features.
- Calculate each feature's standard deviation.
- Calculate the Person's correlation coefficient.
- Calculate the T-statistic and the associated *p*-value.
- Make a decision which is to reject H_0 if $p < 0.05$.

D. Feature Importance of Decision Tree Classification

Considering a dataset (x, y) which is the predictor feature of *X* with d dimension, and the response feature of *Y* with one dimension for developing a prediction function f as the predictive model. A feature importance method can be loosely understood as a function that maps each feature into a score. The scores rank features by how much they "contribute" to the predictive model. In general, feature importance is not consistently or rigorously defined. Feature importance is not equal to the dependence of the response feature on the predictor feature but it is the contribution of a predictor feature to the predictive model [46].

Consider node *t* in a decision tree built on *N* training data instances and let node *t* have N_t node samples. The formula to calculate the score of feature importance is given in Eq. (6) as the following [47]:

$$imp(X_m) = \sum_{u(s_t)=X_m} P(t)\Delta i(s,t) \quad (6)$$

where $\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R)$, $P(t) = N_t/N$ is the proportion of samples reaching node *t* and $u(s_t)$ is the feature used in split s_t .with P_L and P_R are the probabilities an instance splits left and right, respectively, and $i(t)$ is an impurity measure [48].

E. Performance Metrics of Classification Model

In binary classification model, confusion matrix elements namely TN, FN, FP, TP which stand for True Negative, Fall Negative, Fall Positive, and True Positive have main role as the raw material in calculating of the popular performance metrics given in Eq. (7) to Eq. (10) as the following [49].

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

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

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

$$F_1\ Score = \frac{2 \times Recall \times Precision}{Recall+Precision} \quad (10)$$

These metrics except the Accuracy metric give more stress on True Positive occurrence which is the true prediction of the concerning class. If a large portion of instances come from Class 1 and are truly predicted then those metrics will have large values. On the other hand, the Accuracy metric considers all of the instances predicted true over to the number of instances in the dataset which means the proportion of instances with true predicted by the model [50].

IV. Result and Discussion

The section covers the preprocessing data, feature selection with an evaluation of the dependency between predictor features and the target feature, and the evaluation of independence among predictor features. The predictor features that are independent of the target feature are dropped from the dataset, and among predictor features are supposed independent from each other. The set of dependent predictor features is represented by one of them. The last part of the section discusses the comparison of the important features and the both of model's performance.

A. Preprocessing Data

Feature engineering was carried out including handling missing values, commensurate measures, and dropping the instances with outlier values. The missing values imputation was conducted differently concerning both groups of numerical features in Table II. The first group is Feature 1 to Feature 21, and the second group is the 2 remaining features. The missing value imputation of features on the first group is based on the assumption that they occurred when the recording process related to data with no observation yielded. It is supposed to have a zero value, then the imputation filled in them is 0 value. The missing values imputation on the second group is filled in with each associated of the feature mean value. A commensurate measure of features in the dataset has to be fulfilled before further analyses are conducted on the numerical features. They are normalized by transforming into the Z score to have the same measurement unit. The box plot diagram is used to display the data distribution on each numerical feature. The outlier or anomaly observation can be known as the value which is far away from its central tendency. The observation value is categorized as an outlier if its value on the feature lies outside the range of −3 and 3. The instances with outlier values were dropped from the dataset. After the dropping of outliers, the number of instances in the dataset decreases to 45,958 instances. Furthermore, the dataset is divided into training and testing data. The testing data are the 5% of the dataset that is picked up randomly and the 95% remaining as the training data.

B. Dependency Test between Predictor Features and the Target Feature

A feature named "outcome" is the target feature with 0 or 1 observation values. The evaluation of dependency between a categorical predictor feature and the target feature can be done by the Chi-square test while the evaluation of dependency between a numerical predictor feature and the target feature is employed one-way ANOVA. The Chi-square test that evaluates the dependency of categorical features is presented in Table III as follows:

TABLE III. THE CHI-SQUARE STATISTIC AND THE *P*-VALUE ON THE DEPENDENCY TEST OF DISCRETE PREDICTORS TO THE TARGET

Categorical Features	No.	Feature Name	Chi-Square	*P*-Value	Test
Categorical Features with 2 Labels	1	Drain	32.38	0.000	Sig.
	2	Cemented	1.18	0.278	-
	3	Commercial_ALBC	7.96	0.005	Sig.
	4	Non_commercial_ALBC	7.29	0.007	Sig.
	5	Blood_trans	25.7	0.000	Sig.
	6	Congestive Heart Failure	2.25	0.133	-
	7	Cardiac Arrhythmia	0.84	0.360	-
	8	Valvular Disease	1.48	0.224	-
	9	Pulmonary Circulation Disorders	0.35	0.553	-
	10	Peripheral Vascular Disorders	2.99	0.084	-
	11	Hypertension Uncomplicated	0.52	0.470	-
	12	Paralysis	4.64	0.031	Sig.
	13	Other Neurological Disorders	0.59	0.442	-
	14	Chronic Pulmonary Disease	1.03	0.310	-
	15	Diabetes	3.28	0.070	-
	16	Hypothyroidism	1.06	0.302	-
	17	Renal Failure	3.58	0.059	-
	18	Liver Disease	7.81	0.005	Sig.
	19	Peptic Ulcer Disease excluding bleeding	5.15	0.023	Sig.
	20	AIDS/HIV	0.32	0.574	-
	21	Lymphoma	1.49	0.223	-
	22	Metastatic Cancer	0.53	0.469	-
	23	Solid Tumor without Metastasis	9.71	0.002	Sig.
	24	Rheumatoid Arthritis/collagen	39.08	0.000	Sig.
	25	Coagulopathy	9.75	0.002	Sig.
	26	Obesity	0.01	0.925	-
	27	Weight Loss	1.64	0.200	-
	28	Fluid and Electrolyte Disorders	26	0.000	Sig.
	29	Blood Loss Anemia	1.7	0.192	-.
	30	Deficiency Anemia	11.41	0.001	Sig.
	31	Alcohol Abuse	15.49	0.000	Sig.
	32	Drug Abuse	14.34	0.000	Sig.
	33	Psychoses	5.9	0.015	Sig.
	34	Depression	7.03	0.008	Sig.
	35	SEX	53.42	0.000	Sig.
	36	Joint	0.85	0.355	-
Categorical Features with 3 Labels	1	Lung disease	1.18	0.554	-
	2	Anemia	17.64	0.000	Sig.
	3	Psyciatric disorder	11.88	0.003	Sig.
Categorical Features with 4 Labels	1	Heart disease	3.12	0.373	-
	2	Cancer history	7.99	0.046	Sig.
Categorical Features with 6 Label	1	Diagnosis	96.6	0.000	Sig.
Categorical Features with 13 Labels	1	elx_index	51.55	0.000	Sig.
Categorical Features with 16 Labels	1	cci_index	45.19	0.000	Sig.

There are 18 of 34 categorical predictor features with 2 labels that do not have a significant dependency from the target feature, and they must be dropped from the dataset. It is also found that 1 of 3 and 1 of 2 categorical predictor features with 3 and 4 labels respectively which are not significant dependencies, and also are dropped from the dataset. The categorical features with labels number 6, 13, and 16 consisted of one feature respectively have a significant dependency. The dependency evaluation of qualitative predictors acquires the "SEX" feature has significant dependency but the "joint" feature does not. In total, there are as many as 23 categorical features where the target feature has a significant dependency on them.

The dependent test between numerical predictor and target feature is evaluated by using one-way ANOVA. The F-statistic and the *p*-value are given in the Table IV as the following:

TABLE IV. THE F STATISTIC AND THE *P*-VALUE ON THE DEPENDENCY TEST OF NUMERICAL PREDICTORS TO THE TARGET

No.	Numerical Feature			
	Feature name	F-statistic	*P*-Value	Test
1	AGE	29.75	0.000	Sig.
2	LOS	77.63	0.000	Sig.
3	OP_time_minute	4.252	0.039	Sig.
4	OP_time_hour	4.252	0.039	Sig.
5	ASA	1.909	0.167	-
6	CBC_WBC	122.6	0.000	Sig.
7	CBC_RBC	135	0.000	Sig.
8	CBC_HG	228.7	0.000	Sig.
9	CBC_HT	232.1	0.000	Sig.
10	CBC_MCV	130.9	0.000	Sig.
11	CBC_MCH	134.7	0.000	Sig.
12	CBC_MCHC	134.8	0.000	Sig.
13	CBC_RDW	129.1	0.000	Sig.
14	CBC_Platelet	125.9	0.000	Sig.
15	CBC_RDWCV	1.609	0.217	-
16	BUN	64.69	0.000	Sig.
17	Crea	115.4	0.000	Sig.
18	GOT	79.39	0.000	Sig.
19	GPT	35.28	0.000	Sig.
20	ALB	46.36	0.000	Sig.
21	Na	73.33	0.000	Sig.
22	K	86.66	0.000	Sig.
23	UA	24.44	0.000	Sig.

Table IV shows that the target feature does not have a significant dependency on Both "ASA" and "CBC_RDWCV" features which must be dropped from the dataset. There are 21 of 23 numerical features maintained in the dataset. The final result of the remaining categorical and numerical features maintained in the dataset is 44 features in total given in the following list: ("outcome", "SEX", "AGE", "LOS", "OP_time_minute", "OP_time_hour", "CBC_WBC", "CBC_RBC", "CBC_HG", "CBC_HT", "CBC_MCV", "CBC_MCH", "CBC_MCHC", "CBC_RDW", "CBC_Platelet", "BUN", "Crea", "GOT", "GPT", "ALB", "Na", "K", "UA", "Drain", "Commercial_ALBC", "Non_commercial_ALBC", "Blood_trans", "Paralysis", "Liver Disease", "Peptic Ulcer Disease excluding bleeding", "Solid Tumor without Metastasis", "Rheumatoid Arthritis/collagen", "Coagulopathy", "Fluid and Electrolyte Disorders", "Deficiency Anemia", "Alcohol Abuse", "Drug Abuse", "Psychoses", "Depression", "Anemia", "Psyciatric disorder", "Cancer history", "Diagnosis", "elx_index", "cci_index"). The acquired dataset with the list of features above is treated as the first dataset that will be divided into the training and testing data to train and evaluate the first decision tree model. The predictor features in the first dataset are called the relevant subset features.

C. Independency Test among Predictor Features

The evaluation of independence among predictor features is conducted on the first dataset consisting of 23 categorical and 21 numerical features. For a simplification purpose, the statistical tests are done only on two groups of features, and all of the features exclude both groups automatically as a part of selected independent features. The groups are the 16 categorical features with 2 labels and the 21 numerical features. The acquired independent features from both groups are automatically the elements of the selected independent features which means the evaluation test between the numerical and categorical predictor features is not conducted.

Figs. 1 and 2 present the Chi-square and *p*-values of the independency test among categorical features with 2 labels respectively. Some parts both 2 figures are not displayed because of the limited space. Both the row and column in the both features have the same name as the Cf1 to Cf16 standing for the categorical Feature 1 to categorical Feature 16. Two features are said independent when the associated cell in Fig. 2 is greater than 0.05. As an example, the "Cf5" and "Cf1" features have the cell with a *p*-value of 0.85 which is greater than 0.05 and implies that both features are independent of each other.

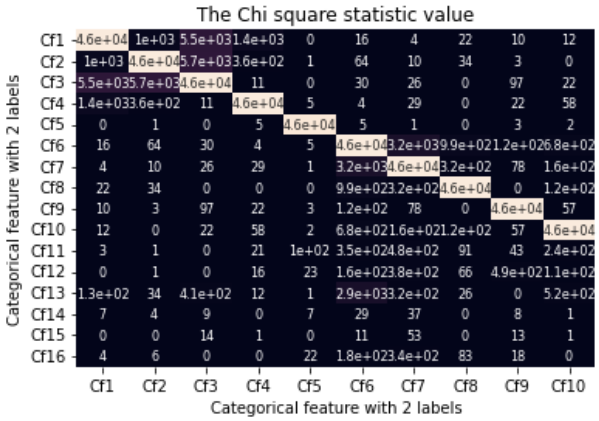

Fig. 1. The Chi-square values of the independent test among 16 categorical features with 2 labels.

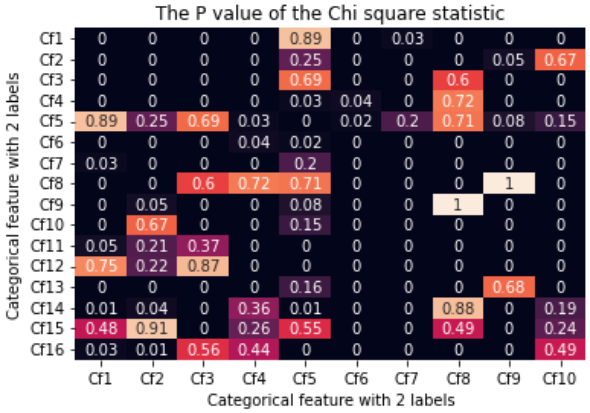

Fig. 2. The *p*-value of Chi-square statistic on the independent test among 16 categorical features with 2 labels.

The search for 2 features that are independent of each other is started by using the Cf5 row in the Fig. 2 as a decision basis row across all columns. The rows and columns in which the cell value is greater than 0.05 are maintained and otherwise are dropped. The step leading to the Cf5 feature is independent of all of the remaining features namely the Cf1, Cf2, Cf3, Cf7, Cf8, Cf9, Ct10, Cf13, and Cf15. It is a notice the Cf5 is the first selected feature which is not only the relevant feature but also the independent one. The second step of searching for independent features is to pick up the Cf1 row as the decision basis row. The independency evaluation of the

Cf1 and Ct15 features has a *p*-value of 0.48 which means both features are independent of each other but all of the other remaining features are dependent on the Cf1 feature. Subsequently, it is acquired that the categorical features with 2 labels that have independence of each other consist of 3 features namely the Cf5 ("Paralysis"), Cf1 ("Drain"), and Cf15 ("Psychoses").

Fig. 3 presents the Pearson correlation value among 21 features where the main diagonal has a value of 1 stating the self-correlation of a feature. The row and column have feature names of Nf1 to Nf21 standing for the numerical Feature 1 to numerical Feature 21. Some parts of Fig. 3 are not presented because of the limited space. The associated *p*-values of all Pearson correlation in Fig. 3 are dominated by 0 and some values of 0.01 that are less than 0.05 which means all 21 features are dependent on each other. All numerical features can be represented by one of them and it is picked up the Nf1 ("AGE") as the representation feature.

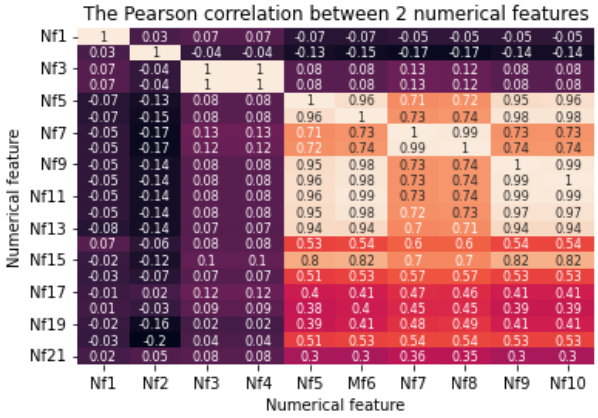

Fig. 3. The *p*-value of the Pearson correlation statistic on the independent test among 21 numerical features.

Finally, the evaluation of independence among relevant predictor features in the first dataset leads to a reduction in the number of features to 11 including the "output" feature as the target. The relevant and independent subset features acquired are given in the list of ["outcome", "SEX", "AGE", "Paralysis", "Drain", "Psychoses", "Anemia", "Psyciatric disorder", "Cancer history", "Diagnosis", "elx_index", "cci_index"]. Where the "SEX" feature comes from the qualitative type, the "AGE" feature comes from the numerical type, the "Paralysis", "Drain", and "Psychoses" come from the categorical type with 2 labels, and the remaining features in the list come from the categorical features having greater than 2 labels.

D. Features Importance and Model's Performance

Two datasets which are the first one yielded by the selection feature based on the dependency between predictor features and the target feature, and the second one is yielded not only by the first method but also is evaluated by satisfying the independence among predictor features are employed to build and evaluate decision tree models. Both datasets are divided into the training and testing data where it is decided as many as 5% part selected randomly as the testing data and the 95% remaining part as the training data. Each of the training data is used to train the decision tree model. The feature importance of both models is explored in the form of the bar chart presented in Figs. 4 and 5, and the table of feature scores presented in Tables V and VI. The performance of both models is evaluated by using the associated testing data in the 4 popular metrics given in Fig. 6.

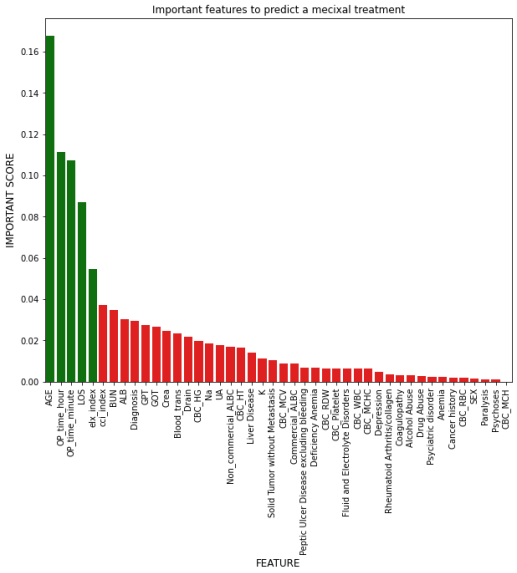

Fig. 4. The bar chart of the first model features importance.

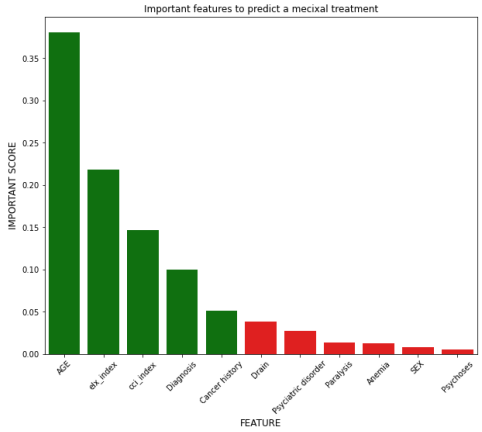

Fig. 5. The bar chart of the second model features importance.

TABLE V. THE SCORE OF FEATURE IMPORTANCE IN THE FIRST DECISION TREE MODEL

No.	Feature Importance of the Model 1	
	Feature Name	Score
1	AGE	0.16781
2	OP_time_hour	0.11149
3	OP_time_minute	0.10733
4	LOS	0.08685
5	elx_index	0.05453
6	cci_index	0.03695
7	BUN	0.03479
8	ALB	0.03025
9	Diagnosis	0.02958
10	GPT	0.02729
11	GOT	0.02673

TABLE VI. THE SCORE OF FEATURE IMPORTANCE IN THE SECOND DECISION TREE MODEL

No.	Feature Importance of the Model 2	
	Feature Name	Score
1	AGE	0.38001
2	elx_index	0.21781
3	cci_index	0.14641
4	Diagnosis	0.09978
5	Cancer history	0.05125
6	Drain	0.03845
7	Psyciatric disorder	0.02740
8	Paralysis	0.01328
9	Anemia	0.01218
10	SEX	0.00795
11	Psychoses	0.00550

Fig. 4 presents the bar chart of 44 features with the normalized importance scores on the X-axis. The red color bars state the features whose importance scores are lower than 5%. While Table V presents the features with the 11 of first scores rank. The numerical features dominate in the importance scores. Only 3 of the 11 features are categorical namely the elx_index, cci_index, and Diagnosis features with the rank order of 5, 6, and 9 which the total score is around 11%. In the first model, there are still many redundant or overlapping features because independence among predictors has not been evaluated yet. The yielded tree model needs 44 inputs whose are characteristics of the patient.

Fig. 5 is a bar chart of feature importance in the second tree model. The 'AGE' feature is only one of the numerical features in the dataset where it has the highest score of feature importance around 38%. Table VI presents the scores of feature importance of the second tree model. The categorical features with a large number of labels have a greater score than the categorical features with a few number of labels. The 3 categorical features namely the elx_index, cci_index, and Diagnosis have the total scores of around 46% which increases around 35% compared to the total score in the previous tree model. The reduction of predictor features number from 44 to 11 features has a meaningful use of the model to predict an instance with an unknown where it comes from the class label.

The above results confirm that the feature selection using the feature's importance of the tree model which is done by researchers in [34, 36, 37] is not a good choice. The subset of important features in Table V are the input features when building the decision tree based on the approaches in [34, 36, 37]. The subset of relevant and independent features in Table VI has confirmed that both subset features have very different elements. Employing the important features of the decision tree as the input of machine learning models is not a precise decision and it should be avoided. The relevant and independent features are similar to the Principal Component Analysis (PCA) which explains how much it represents the variability of the dataset while the important features describe how much the features have contributed to the model for predicting of unknown label of an instance [28]. The selection feature through the sequential evaluation of dependency between the predictor and target feature and continued evaluation of independence among predictor features will yield the dataset with the precise predictor features which is similar to the works conducted by [29, 30, 32]. However, the approach leaves some subjective decisions when evaluating the independence among predictor features which consist of mixing large numerical and categorical features. Even the heuristic approach is impossible to employ when the dataset has very large predictor features. The work published in [39] offered a good approach that combined the optimization methods (GA and PSO) and machine learning models. Nevertheless, in the case of the medical record dataset, the heuristic approach is very supportive in building an efficient classification system because it only employs the relevant and independent features as the input model

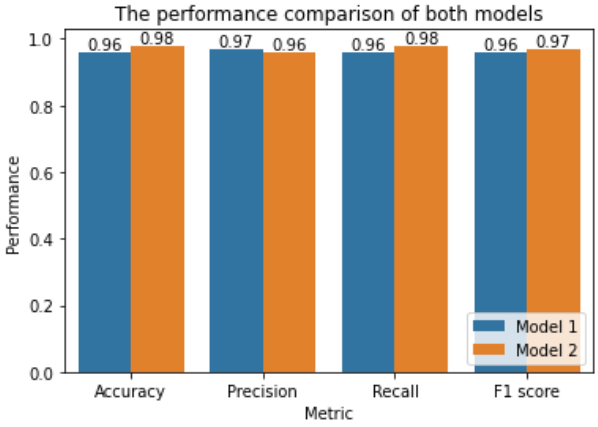

Fig. 6. The performance of both decision tree models on the 4 popular metrics.

Finally, the comparison of both model's performance in the testing data is presented in Fig. 6 Both decision tree models are evaluated for their performances with the associated testing data which is 5% of the dataset. The second decision tree model has a higher performance than the first one. The second model has the values of accuracy, recall, and F1-Score of 98%, 98%, and 97%, respectively, while the first model has the same performance value of 96% on those 3 metrics. While the first model has better performance in the precision metric of 97% compared to the second model of 96%. The performance gap of both models is only slightly different, nevertheless, the number of selected subset features is very different which is 43 predictors compared to 10 predictor features. The second decision tree model is very effective in the implementation of a system of real-life applications. It means a class label of a new instance can be known by using the second model by only observing the 10 features.

V. CONCLUSION

Medical dataset usually consists of many features some of which are irrelevant and redundant features. They must be selected before they are as the input of predictive models. The irrelevant features are evaluated by a dependent test between the target and predictor features, while the redundant features are evaluated by an independence test among predictor features where a group of dependent predictors is represented by one of them. The evaluation result of irrelevant features degrades the

number of predictor features from 67 to 44 where the dropped feature numbers are respectively 18 discrete types with 2 labels, 2 numeric types, 1 qualitative type, and 2 discrete types with more than 2 labels. The evaluation result of redundant features causes the degradation of predictor features from 44 to 11 features where 13 discrete types with 2 labels and 20 numeric types are dropped. 2 datasets will be divided into the training and testing data. The first dataset contains 44 relevant features, and the second dataset contains 11 relevant and independent features. The decision tree models yielded have very different important features on the 11 first scores rank because the important features of the first model are dominated by the numerical features while the important features of the second model consist only of 1 numeric feature. The second model also has better performance than the first model where the performance metrics of the accuracy, recall, and F1-Score are 98%, 98%, and 97%, respectively. A challenge of future work is to make deeper confirmation that picking up the important features of the decision tree model as the input of machine learning models is not a wise choice. It is done through a direct comparison of 2 group models one is trained by important features of the decision tree and the other one is trained by relevant and independent features.

Conflict of Interest

The authors declare no conflict of interest related to this research project.

Author Contributions

Conceptualization was done by Mursityo and Handoyo, the methodology was done by Mursityo and Rupiwardani; software was done by Mursityo and Putra, validation was done by Susanti, Handayani, and Mursityo, formal analysis was done by Putra, investigation, and resources were done by Handayani, data curation was done by Susanti and Rupiwardani, writing—original draft preparation was done by Mursityo and Susanti, writing—review and editing was done by Handoyo and Putra; visualization was done by Handayani, supervision was done by Handoyo, project administration was done by Putra, funding acquisition was done by Mursityo. All authors had approved the final version.

References

[1] M. H. Avizenna, R. A. Widyanto, D. K. Wirawan, T. A. Pratama, and A. S. Nabila, "Implementation of a priori data mining algorithm on medical device inventory system," *Journal of Applied Data Sciences*, vol. 2, no. 3, pp. 55–63, 2021.

[2] P. Wang and J. Li, "Implementation of real-time medical and health data mining system based on machine learning," *J. Healthc Eng.* vol. 2021, pp. 1–5, 2021.

[3] J. Podani, D. Schmera, and S. Bagella, "Correlating variables with different scale types: A new framework based on matrix comparisons," *Methods Ecol. Evol.*, vol. 14, no. 4, pp.1049–1060, 2023

[4] Marji, S. Handoyo, I. N. Purwanto, and M. Y. Anizar, "The effect of attribute diversity in the covariance matrix on the magnitude of the radius parameter in fuzzy subtractive clustering," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 12, pp. 3717–3728, 2018.

[5] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, "Feature selection: A review and comparative study," in *Proc. E3S Web of Conferences, 10th International Conference on Innovation, Modern Applied Science & Environmental Studies (ICIES'2022)*, 2022, vol. 351, 01046.

[6] Z. Zhang and Y. Liu, "Parsimony-enhanced sparse Bayesian learning for robust discovery of partial differential equations," *Mech. Syst. Signal Process*, vol. 171, 108833, 2022.

[7] I. Kavakiotis, P. Samaras, A. Triantafyllidis, and I. Vlahavas, "FIFS: A data mining method for informative marker selection in high dimensional population genomic data," *Comput. Biol. Med.*, vol. 90, pp. 146–154, 2017.

[8] S. K. Nayak, P. K. Rout, A. K. Jagadev, and T. Swarnkar, "Elitism based multi-objective differential evolution for feature selection: A filter approach with an efficient redundancy measure," *Journal of King Saud University—Computer and Information Sciences*, vol. 32, no. 2, pp. 174–187, 2020.

[9] P. Michel, N. Ngo, J. F. Pons, S. Delliaux, and R. Giorgi, "A filter approach for feature selection in classification: Application to automatic atrial fibrillation detection in electrocardiogram recordings," *BMC Med. Inform Decis. Mak.*, vol. 21, pp. 1–17, 2021.

[10] H. Kusdarwati and S. Handoyo, "Modeling treshold liner in transfer function to overcome non normality of the errors," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 546, no. 5, 052039, 2019.

[11] S. Handoyo, Y. P. Chen, G. Irianto, and A. Widodo, "The varying threshold values of logistic regression and linear discriminant for classifying fraudulent firm," *Mathematics and Statistics*, vol. 9, no. 2, pp. 135–143, 2021.

[12] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J. M. Lagniez, and P. Marquis, "On the explanatory power of Boolean decision trees," *Data Knowl. Eng*, vol. 142, 102088, 2022.

[13] F. Zhang and X. Yang, "Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection," *Remote Sens Environ*, vol. 251, 112105, 2020.

[14] S. Handoyo, N. Pradianti, W. H. Nugroho, and Y. J. Akri, "A heuristic feature selection in logistic regression modeling with Newton Raphson and gradient descent algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 119–126, 2022.

[15] L. Fu, T. Zhu, G. Pan, S. Chen, Q. Zhong, and Y. Wei, "Power quality disturbance recognition using VMD-based feature extraction and heuristic feature selection," *Applied Sciences (Switzerland)*, vol. 9, no. 22, 4901, 2019.

[16] H. Wang, Y. Ou, Y. Wang, T. Xing, and L. Tan, "Semi-supervised bacterial heuristic feature selection algorithm for high-dimensional classification with missing labels," *International Journal of Intelligent Systems*, vol. 2023, Feb. 2023.

[17] K. F. Lalonde and W. Cotten, "Use of contingency tables for determining statistical dependence of attribute data from aluminum reduction cell processes," *TMS Light Metals*, vol. 555, 2007.

[18] A. G. Dufera, T. Liu, and J. Xu, "Regression models of Pearson correlation coefficient," *Stat. Theory Relat. Fields*, vol. 7, no. 2, pp. 1–10, 2023.

[19] N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way ANOVA F-test for e-mail spam classification," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 7, no. 3, pp. 625–638, 2014.

[20] I. N. Purwanto, A. Widodo, and S. Handoyo, "System for selection starting lineup of a football players by using Analytical Hierarchy Process (AHP)," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 1, pp. 19–31, 2018.

[21] S. Handoyo, A. Widodo, W. H. Nugroho, and I. N. Purwanto, "The implementation of a hybrid fuzzy clustering on the public health facility data," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 6, pp. 3549–3554, 2019.

[22] H. N. Utami, S. Handoyo, and Sandra, "The effect of self efficacy and hope on occupational health behavior in east java of Indonesia," *International Journal of Scientific and Technology Research*, vol. 9, no. 2, pp. 3571–3575, 2020.

[23] W. H. Nugroho, S. Handoyo, and Y. J. Akri, "An influence of measurement scale of predictor variable on logistic regression modeling and learning vector quantization modeling for object classification," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 1, pp. 333–343, 2018.

[24] A. W. Widodo, S. Handoyo, I. Rupiwardani, Y. T. Mursityo, I. N. Purwanto, and H. Kusdarwati, "The performance comparison between C4.5 Tree and One-Dimensional Convolutional Neural Networks (CNN1D) with tuning hyperparameters for the classification of imbalanced medical data," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 5, pp. 748–759, 2023.
[25] Marji and S. Handoyo, "Performance of ridge logistic regression and decision tree in the binary classification," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 15, pp. 4698–4709, 2022.
[26] N. A. M. Zaini and M. K. Awang, "Performance comparison between meta-classifier algorithms for heart disease classification," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, pp. 323–328, 2022.
[27] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *PLoS ONE*, vol. 15, no. 12, 0243300, 2021.
[28] S. Zhao, M. Wang, S. Ma, and Q. Cui, "A feature selection method via relevant-redundant weight," *Expert. Syst. Appl.*, vol. 207, 117923, 2022.
[29] M. Mera-Gaona, D. M. López, R. Vargas-Canas, and U. Neumann, "Framework for the ensemble of feature selection methods," *Applied Sciences (Switzerland)*, vol. 11, no. 17, 8122, 2021.
[30] H. M. Le, T. D. Tran, and L. V. Tran, "Automatic heart disease prediction using feature selection and data mining technique," *Journal of Computer Science and Cybernetics*, vol. 34, no. 1, pp. 33–48, 2018.
[31] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm Evol. Comput.*, vol. 54, 100663, 2020.
[32] S. Sabeena and B. Sarojini, "Optimal feature subset selection using ant colony optimization," *Indian J. SCI Technol.*, vol. 8, no. 35, pp. 1–5, 2015.
[33] D. Grissa, M. Pétéra, M. Brandolini, A. Napoli, B. Comte, and E. Pujos-Guillot, "Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data," *Front Mol. Biosci.*, vol. 3, no. 7, 30, 2016.
[34] W. Ahmed and N. G. M. Jameel, "Malicious URL detection using decision tree-based lexical features selection and multilayer perceptron model," *UHD Journal of Science and Technology*, vol. 6, no. 2, pp. 105–116, 2022.
[35] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 52, 2020.
[36] A. A. Romalt and R. M. S. Kumar, "An analysis on feature selection methods, clustering and classification used in heart disease prediction—A machine learning approach," *Journal of Critical Reviews*, vol. 7, no. 6, pp. 138–142, 2020.
[37] C. Wang, X. Qiu, H. Liu, D. Li, K. Zhao, and L. Wang, "Damaged buildings recognition of post-earthquake high-resolution remote sensing images based on feature space and decision tree optimization," *Computer Science and Information Systems*, vol. 17, no. 2, pp. 619–646, 2020.
[38] T. E. Mathew, "An optimized extremely randomized tree model for breast cancer classification," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 16, pp. 5234–5246, 2022.
[39] A. Pasha and P. H. Latha, "Bio-inspired dimensionality reduction for Parkinson's Disease (PD) classification," *Health Information Science and Systems*, vol. 8, pp. 1–22, 2020.
[40] A. Pasha and P. H. Latha, "Well-calibrated probabilistic machine learning classifiers for multivariate healthcare data," *International Journal of Advanced Research in Computer Science*, vol. 12, no. 2, pp. 39–45l, 2021.
[41] T. Yamamoto, K. Sakurai, M. Watanabe, I. Sakuma, N. Kanahara, A. Shiina, T. Hasegawa, H. Watanabe, M. Iyo, and R. Ishibashi, "Cyclothymic temperament is associated with poor medication adherence and disordered eating in type 2 Diabetes patients: A case-control study," *Diabetes Therapy*, vol. 12, no. 9, pp. 2611–2624, 2021.
[42] M. Alassaf and A. M. Qamar, "Improving sentiment analysis of arabic tweets by one-way ANOVA," *Journal of King Saud University—Computer and Information Sciences*, vol. 34, no. 6, pp. 2849–2859, 2022.
[43] S. Sreedevi, "Study of test for significance of Pearson's correlation coefficient," *Peer Reviewed and Refereed Journal*, no. 2, pp. 1–4, 2022.
[44] P. Oranpattanachai, "Relationship between the reading strategy, reading self-efficacy, and reading comprehension of Thai EFL students," *LEARN Journal: Language Education and Acquisition Research Network*, vol. 16, no. 1, pp. 194–220, 2023.
[45] E. I. Obilor and E. C. Amadi, "Test for significance of Pearson's correlation coefficient," *International Journal of Innovative Mathematics, Statistics & Energy Policies*, vol. 6, no. 1, pp. 11–23, 2018.
[46] K. Gajowniczek and T. Ząbkowski, "Interactive decision tree learning and decision rule extraction based on the ImbTreeEntropy and ImbTreeAUC packages," *Processes*, vol. 9, no. 7, 1107, 2021.
[47] L. Xu, L. Wang, Y. Li, and A. Du, "Big model and small model: Remote modeling and local information extraction module for medical image segmentation," *Appl. Soft. Comput.*, vol. 136, 110128, 2023.
[48] F. Bollwein and S. Westphal, "A branch & bound algorithm to determine optimal bivariate splits for oblique decision tree induction," *Applied Intelligence*, vol. 51, no. 10, pp. 7552–7572, 2021.
[49] M. F. Amin, "Confusion matrix in binary classification problems: A step-by-step tutorial," *Journal of Engineering Research*, vol. 6, no. 5, 2022.
[50] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019.