



Hybrid Feature Selection Method Based on Genetic Algorithm for the Diagnosis of Coronary Heart Disease

Wiharto^{1*} , Esti Suryani¹ , Sigit Setyawan² , and Bintang PE Putra¹ 

¹Department of Informatics, Sebelas Maret University, 57126, Indonesia

²Department of Medicine, Sebelas Maret University, 57126, Indonesia

Abstract

Coronary heart disease (CHD) is a comorbidity of COVID-19; therefore, routine early diagnosis is crucial. A large number of examination attributes in the context of diagnosing CHD is a distinct obstacle during the pandemic when the number of health service users is significant. The development of a precise machine learning model for diagnosis with a minimum number of examination attributes can allow examinations and healthcare actions to be undertaken quickly. This study proposes a CHD diagnosis model based on feature selection, data balancing, and ensemble-based classification methods. In the feature selection stage, a hybrid SVM-GA combined with fast correlation-based filter (FCBF) is used. The proposed system achieved an accuracy of 94.60% and area under the curve (AUC) of 97.5% when tested on the z-Alizadeh Sani dataset and used only 8 of 54 inspection attributes. In terms of performance, the proposed model can be placed in the very good category.

Index Terms: coronary heart disease, genetic algorithm, feature selection, ensemble learning, support vector machine

I. INTRODUCTION

The heart and lungs work together to maintain the oxygen levels in the body. When the lungs are affected by respiratory diseases, such as the novel coronavirus (COVID-19), the heart can be affected as well. The heart has to work hard to pump blood, which may be even more difficult for someone with a heart disease. Patients with coronary heart disease (CHD) are at high risk of contracting COVID-19. CHD in patients infected with COVID-19 can cause damage to the heart muscle or blood vessels [1]. The application of strict health protocols will have an impact on heart disease patients who have minimal activity. It was confirmed in a study by Hemphill et al. [2], where the number of steps during the COVID-19 pandemic was lower than before the pandemic. During the pandemic, in addition to implementing health protocols, you must also maintain a healthy lifestyle and per-

form routine health checks [3], one of which is routine heart health checks.

Cardiac health checks can be performed starting with routine examinations of risk factors, followed by electrocardiogram (ECG) examinations, laboratory examinations, and coronary angiography. Along with the development of artificial intelligence, all examination results in the diagnosis process can be used with machine learning to draw conclusions [4]. Many examinations result in several attributes that must be analyzed, and the number of attributes allows for ambiguity in drawing conclusions. There is one important process in the development of machine learning, namely, feature selection. Feature selection involves selecting the best feature that can provide a better machine learning performance [5-7]. Feature selection involves numerous methods, including filtering, wrappers, and embedding. Feature selection can be developed using a combination of several methods known as


Received 07 October 2021, Revised 16 January 2022, Accepted 04 February 2022

*Corresponding Author Wiharto (E-mail: wiharto@staff.uns.ac.id, Tel: +62-2717-20896)

Department of Informatics, Sebelas Maret University, 57126, Indonesia.

Open Access <https://doi.org/10.6109/jicce.2022.20.1.31>

print ISSN: 2234-8255 online ISSN: 2234-8883

 This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © The Korea Institute of Information and Communication Engineering

the hybrid method. The hybrid method aims to obtain better features than if only one method is used. In the development of machine learning for diagnosis, in addition to feature selection, there is an important influential component: the amount of data available for the learning process. Medical records are taken from hospitals, where people tend to check themselves when there are symptoms, therefore the diagnosis results are mostly positive. This condition causes the data obtained to obtain additional data that are diagnosed as positive than negative, which results in the availability of data being unbalanced [8, 9].

Several models for the diagnosis of CHD have been developed using feature selection methods. Numerous developments have been conducted using computational intelligence algorithms such as genetic algorithms (GAs), artificial bee colonies, and particle swarm optimization [10-13]. This study proposes a CHD diagnosis model that is preceded by feature selection using a hybrid method. The hybrid method used was a support vector machine (SVM) and a GA for the searching method. The feature selection stage ended with a filtering method that used a fast correlation-based filter (FCBF). To determine the output of the proposed system model, classification was performed using an ensemble learning algorithm, namely, the bagging-logistics model tree (bagging-LMT). At the learning stage of the bagging-LMT algorithm, to overcome unbalanced learning data, before the learning process oversampling was performed using the synthetic minority oversampling technique (SMOTE) method [14, 15]. The system testing was validated using k-folds cross-validation, as well as the z-Alizadeh Sani, Cleveland, and Statlog datasets. The performance parameters of the proposed system model include sensitivity (SSE), specificity (SPE), accuracy, area under the curve (AUC), positive prediction value (PPV), and negative prediction value (NPV).

II. LITERATURE REVIEW

CHD detection models have been developed along with machine learning [4, 16, 17]. Ghosh et al. [18] developed a diagnosis model using feature selection, namely, the Relief and LASSO techniques which can improve the performance of it. Despite the obstacles in implementing machine learning algorithms in clinical practice, the use of machine learning algorithms, such as convolutional neural networks, boosting, and SVM can provide good prospects for the development of diagnostic models [19]. Machine learning with a combination of SVM extreme gradient boosting can also be used to detect CHD, with good performance, namely, an F1 value of 91.86% and accuracy of 93.86%. However, in this study, feature selection was not optimized, and thus, the performance still requires many attribute checks [20].

Another study used the CART algorithm to detect CHD

[21]. The CART algorithm is a decision tree-based algorithm; therefore, it can perform feature selection simultaneously with the training process. The CART algorithm has the ability to use five test attributes. When tested with the z-Alizadeh Sani dataset, it was able to provide 98.61% sensitivity but low specificity, namely 77.01% and 92.41% accuracy. The ability of this model is weak when the patient is negative but detected by the system model as true negative, with a percentage of 77.01%. The CART algorithm can extract knowledge into several rules arranged in a tree diagram. The ability to compose rules, such as decision tree models, can also be achieved using hybrid binary-real particle swarm optimization (PSO) [22]. The hybrid PSO model provided an average accuracy of 84.2%. This hybrid-PSO-based rule model can produce a relatively small number of rules, that is, 10 rules. Referring to the rule, testing is also conducted for the use of 13 and 11 attributes. The result is that the use of 13 attributes is better than the use of 11 attributes.

The use of computational intelligence in addition to the PSO algorithm also uses many GA [11]. The use of Gas combined with artificial neural networks for the diagnosis of CHD can provide excellent performance, with an accuracy of 93.5%. The next development is the use of neural network-based algorithms, namely an emotional neural network (ENN) combined with PSO (ENN+PSO). The use of ENN+PSO can reduce attributes from 55 to 22 with an accuracy of 88.34% [12]. The use of particle swarm can also provide good performance for the diagnosis of CHD, as shown by the test results with an accuracy parameter value of 87.097% [13]. The subsequent development uses a combination of PSO and GA. The use of the combination of the two algorithms can reduce attributes from 55 to 22, and the resulting performance is 93.08% for accuracy and 91.51% F-Score [23]. PSO+GA is capable of producing the same number of reduced attributes as PSO+ENN for the z-Alizadeh Sani dataset; however, its performance is better when using PSO+GA.

The ability of GAs in the feature selection process can also be demonstrated by Karegowda et al. [24]. This study proposed a combination of GA+correlation feature selection (CFS) with a radial basis function (RBF) classification algorithm. This model provided better CHD diagnosis system performance than the combination of a decision tree and RBF. The GA combined with SVM was also able to provide good performance compared to PSO, using the objective function in the form of accuracy [25]. A similar study was also conducted by Ephzibah [26], which showed that the ability of SVM+GA was better than that of SVM. This capability was indicated by the minimum number of features. Subsequent research using the genetic fuzzy system-logit-boost (GFS-LB) for the diagnosis of CHD provided better performance than without using GA [27]. Subsequent research was conducted using a feature selection model that combined

filtering and wrapper methods [28]. The filtering algorithm used was conditional mutual information maximization (CMIM) combined with a binary GA (BGA), with the fitness function in the GA as an accuracy function. The resulting accuracy performance parameter was better than that of several other feature selections.

With the use of GAs in the feature selection process in machine learning, the majority of studies only focus on accuracy performance parameters, so that the fitness function used is also accurate [11, 23, 27]. The accuracy fitness function is also used in the PSO algorithm [12, 13, 22, 23]. This is certainly inappropriate in the medical field, particularly for the screening process or early diagnosis of the disease. In the screening or diagnosis of CHD, the sensitivity performance parameter is crucial. These parameters are used to measure when a patient is positive for CHD. Machine learning will also detect CHD, as indicated by the sensitivity parameter, so that in the screening or diagnosis system, it must suppress the incidence of positive patients, and using machine learning it is also diagnosed as positive [29]. Referring to this, the sensitivity performance parameter is crucial to be included in determining feature selection in the development of a CHD diagnosis system model.

The selection of the right feature selection method is important, particularly for high-dimensional data. FCBF is an effective feature selection method for high-dimensional data [30]. Sánchez-Marzoño et al. [31] confirmed that the performance of accuracy of FCBF is better than the Relief when the number of attributes is more than 40 attributes. The FCBF feature selection algorithm combined with the SVM classification algorithm was able to provide better performance than the k-nearest neighbor, random forest, and naive Bayesian algorithms [32]. The CHD diagnosis system model with the z-Alizadeh Sani dataset has more than 40 attributes; therefore, the FCBF is very precise. The ability of feature selection was also demonstrated by Djellali et al. [25], where FCBF was able to reduce the number of large features but was still able to maintain performance in a good category. The ability of FCBF when combined with GA with a fitness function in the form of accuracy results in better performance than when using only FCBF.

The problem of developing machine learning is not limited in feature selection. The availability of good data for the learning process will make the learning process successful in building a system model for CHD diagnosis. The challenge in processing medical data is imbalanced data; this condition can cause the machine learning model built to be poor [8, 33]. The problem of imbalanced data can be overcome using the oversampling method. The development of a prognostic model for patients with heart failure also utilizes the oversampling method. Kim et al. [34] used oversampling algorithms such as SMOTE, borderline-SMOTE, and adaptive synthetic sampling (ADASYN). In addition, a study con-

ducted by Brandt and Lanzén [35] analyzed the performance of SMOTE and ADASYN. The results of the two studies show that SMOTE is better than the others. SMOTE is also used in the classification of hypertension, where the data are unbalanced, and the results show an increase in accuracy from 91 to 98% [36]. The use of SMOTE is also effective for high-dimensional datasets [37], such as the z-Alizadeh Sani CHD dataset. The ability of SMOTE to predict compound-protein interactions has also been demonstrated [38]. In this study, SMOTE is better than random under-sampling (RUS), a combination of over-undersampling (COUS), and Tomek link (T-Link) algorithms, with reference to the AUC performance parameter.

III. METHODS

In this study, the z-Alizadeh Sani dataset [39-42] was used, along with the Cleveland and Statlog datasets [43] as support. The z-Alizadeh Sani dataset has relatively complete types of examinations, such as demographic, symptom and examination, ECG, and laboratory and echo features. Another advantage of this dataset is that the data used are relatively new compared to their predecessors, such as the Cleveland and Statlog datasets. These datasets can be accessed online, and their distribution is presented in Table 1. This study used the research methods shown in Fig. 1. The stages are divided as follows: the first stage is pre-processing in the form of data normalization. The second stage performs feature selection using hybrid SVM-GA, which classifies data, and then measures the accuracy and sensitivity performance. The two performance parameters are then used as fitness functions in the GA, as shown in Equation (1). The SVM algorithm used is a binary SVM with a kernel using a radial basis function (RBF) [44]. The SVM algorithm works using nonlinear transformations, one of which uses the RBF kernel. The transformation is performed to map the input data to a higher-dimensional space and then perform a linear classification of the input data in the dimensional feature space to build the optimal hyperplane. Finally, the mapping returns to the original space and becomes a nonlinear classification in the input space [44, 45]. The GA serves to select a subset of features using the benchmark as the fitness function. Chromosomal representation in the form of feature subsets or attributes of CHD examination. The GA works by performs several steps, as shown in Algorithm-1 [28, 46].

Table 1. Coronary heart disease dataset

Dataset	#Feature	#Instance	Ratio Normal/CHD
z-Alizadeh Sani	54	303	1:2.50
Cleveland	13	303	1:0.85
Statlog	13	270	1:0.80

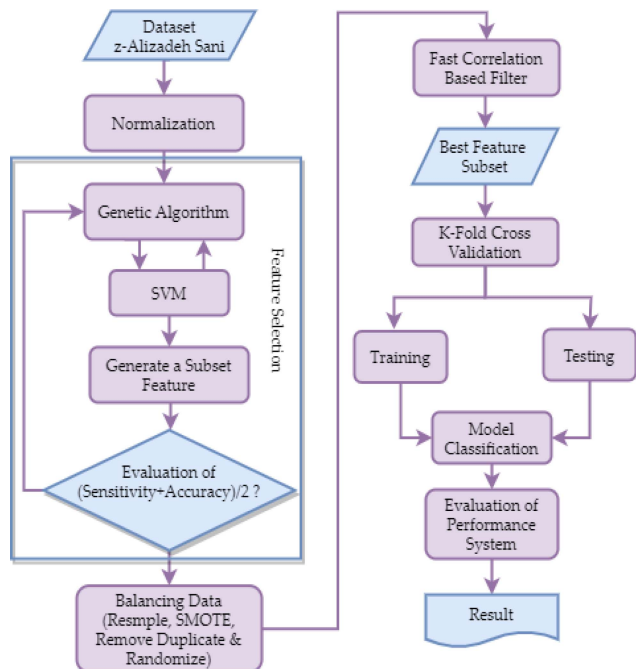


Fig. 1. Research methods.

$$f(\text{ACC}, \text{SEN}) = (\text{ACC} + \text{SEN})/2. \tag{1}$$

The third stage is the oversampling process used to balance the data. Oversampling was performed using the SMOTE algorithm [14, 36]. The percentage of oversampling with reference to the ratio of positive and negative data for CHD is shown in Table 1. The SMOTE process is preceded by resampling. The next stage is the final stage of the feature selection process, namely, filtering using the FCBF algorithm [30, 47]. The FCBF algorithm generates a weight that indicates the weight of the attribute to the output. The next stage is the classification process using the bagging algorithm, with each bootstrap process using the LMT algorithm [48], which is shown in Algorithm-2 [49, 50]. The system model was also tested using the random forest algorithm, forest by penalizing attributes (ForestPA) [51], C4.5, multilayer perceptron (MLP), and bagging-forestPA.

Algorithm-1: Genetic-Algorithm

- 1: Set generation=0 (The first generation)
- 2: Initialized initial population, P(generator), randomly //P(generator) is the population of one generation
- 3: Evaluate the fitness value for each individual
- 4: //equation (1)
- 5: **For** generation **to** generation=Maximum
- 6: Generation = generation +1
Population selection to get the parent candidate
- 8: (P'(generation))

- 9: Crossover on P'(generation)
 - 10: Mutation on P'(generation)
 - 11: Evaluate the fitness value for each individual
New population form = {P(generator) that survive,
 - 12: P'(generation)}
- endFor**

Algorithm-2: Bagging-LMT

- 1: The training set $S = (X_i, y_i), i = 1, 2, 3, 4, \dots, m$
- Input** Machine learning L (LMT algorithm)
- 2: The number of base classifier T
 - 3: For $t = 1, 2, 3, \dots, T$
 - 4: Create m -th samples randomly from S , designated bootstrap samples S_t
 - 5: Using S_t to learning $L: N_t=L(S_t)$
 - 6: Combining using majority voting: $N^*(x) = \arg \max_{y \in I} \sum_{t: N_t(x)=y} 1$
 - 7: **endFor**
- Output** ensemble N^*

The last stage is the evaluation of the system performance. The diagnosis system model was developed with feature selection using a hybrid SVM-GA and FCBF, and the resulting performance was measured by referring to the confusion matrix, as shown in Table 2. The performance parameters used included accuracy (ACC), sensitivity (SEN), specificity (SPE), AUC, positive prediction value (PPV), and negative prediction value (NPV). Referring to Table 2, the performance parameters can be calculated using Equations (2-4).

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{3}$$

$$\text{Accuracy} = \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}. \tag{4}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{5}$$

$$\text{NPV} = \frac{\text{TN}}{\text{FN} + \text{TN}}. \tag{6}$$

Table 2. Confusion Matrix

Actual Class	Predictive Class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

IV. RESULTS

Testing of the CHD diagnosis system model using the SMV-GA hybrid feature selection can produce 25 attributes, as shown in Fig. 2. The number of attributes is obtained when using the GA parameters with population 1000, generation 100, probability crossover 0.55, and probability mutation 0.3. Fig. 2 shows the feature weights of the 25 attributes resulting from the FCBF process. Referring to the process, it was shown that there are eight attributes that have a high weight in influencing the diagnosis of CHD. The attributes that have a high weight are typical chest pain, hypertension (HTN), age, diabetes mellitus (DM), regional wall motion abnormality (RWMA) region, T inversion, Q wave, and triglyceride (TG).

The z-Alizadeh Sani dataset was used for the test. Subsequent testing of the same model using the Cleveland dataset yielded 10 attributes. These results are shown in Fig. 3, which also shows the weights for each attribute using the FCBF algorithm. The attributes that had significant weights in the diagnosis process were thal, cp, slope, ca, age, and restecg. The final test used the StatLog dataset. The test results are shown in Fig. 4 with 10 attributes. The filtering results with the FCBF show that only six attributes had a

significant weight in the diagnosis process. The attributes were thal, cp, oldpeak, thalach, ca, and restecg.

System testing after the feature selection and data balancing process using SMOTE was used to test the classification results. The classification process used the bagging-LMT algorithm. The performance parameters used for the analysis are shown in Equations (2-6). The results of testing the performance of the bagging-LMT algorithm using eight attributes, which were the result of feature selection in the z-Alizadeh Sani dataset, are shown in Fig. 5. Fig. 5 shows that the AUC performance of the bagging-LMT algorithm was better than that of the other ensemble algorithms. The highest AUC value was 97.5%, and the results for all performance parameters are listed in Table 2. In addition to the results for eight attributes in the z-Alizadeh Sani dataset, performance was also shown to determine changes in attribute reduction from 25 to eight attributes. These results are shown in Fig. 6, where changes in the number of attributes do not indicate a significant change in performance.

The result of subsequent system testing using Cleveland datasets are shown in Figs. 7, while using Statlog datasets in Fig. 8. The test results using the Cleveland and Statlog datasets show that the ability of the bagging-LMT algorithm is

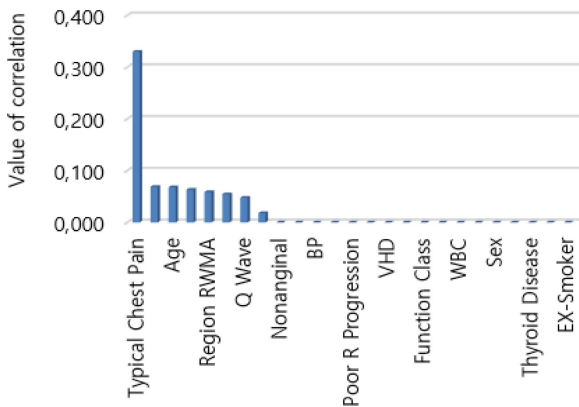


Fig. 2. Feature of result SVM-GA & FCBF: z-Alizadeh Sani.

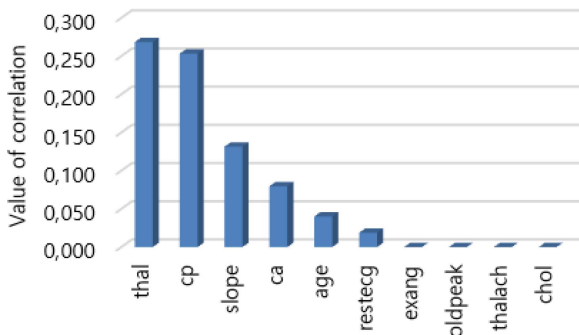


Fig. 3. Feature of result SVM-GA & FCBF: Cleveland.

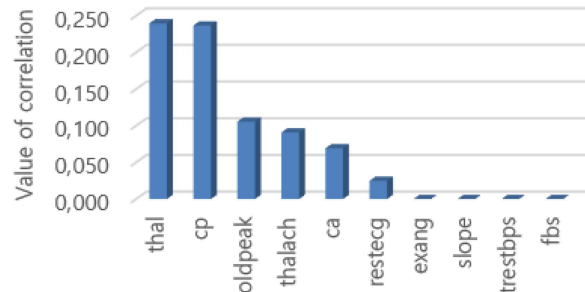


Fig. 4. Feature of result SVM-GA & FCBF: Statlog.

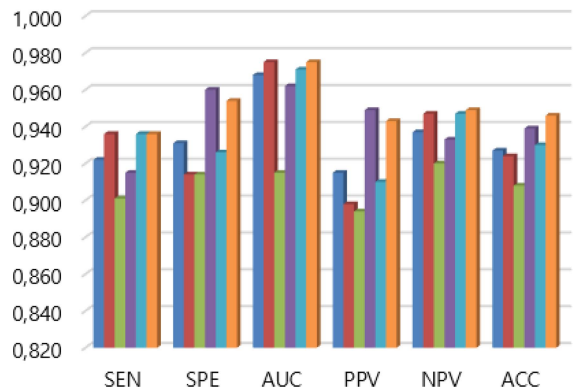
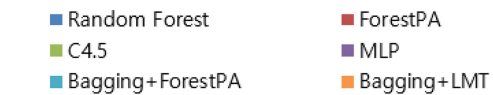


Fig. 5. Performance using 8 attribute dataset z-Alizadeh Sani.

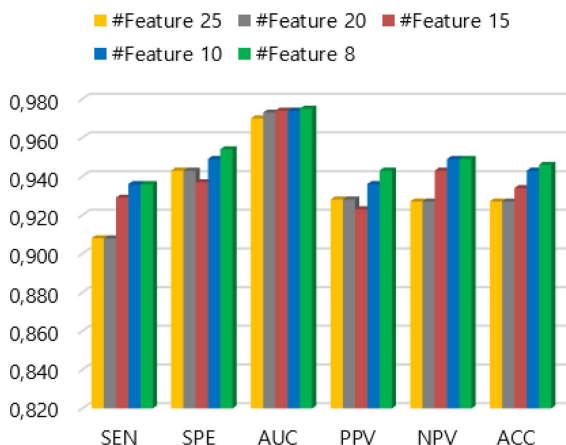


Fig. 6. Performance of 8-25 attribute using Bagging+LMT and dataset z-Alizadeh Sani

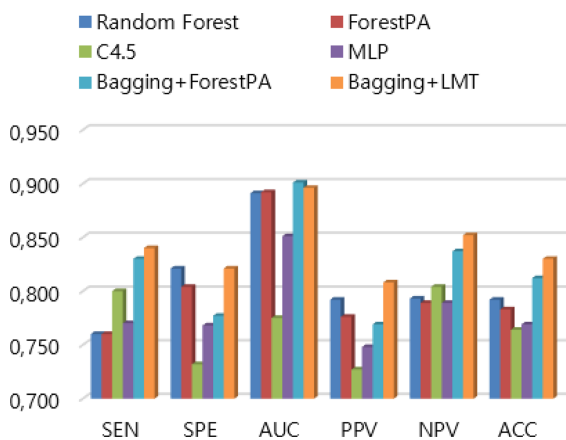


Fig. 7. Performance using 6 attribute dataset Cleveland

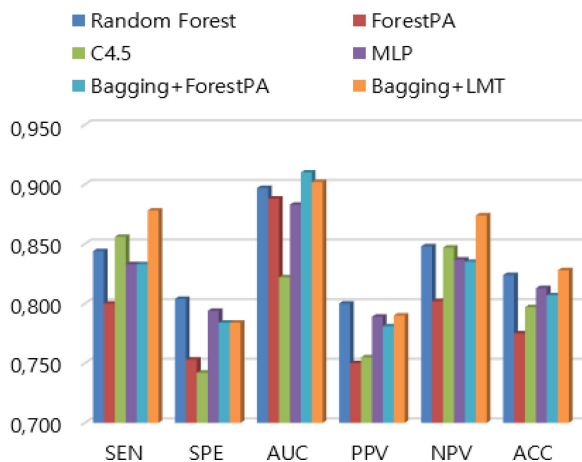


Fig. 8. Performance using 6 attribute dataset Statlog

better than that of the random forest, forestPA, C4.5, MLP, and bagging-forestPA algorithms. The performance of the bagging-LMT algorithm for the AUC performance parameter

Table 3. Performance of Bagging-LMT algorithm

Dataset	SEN	SPE	AUC	ACC
z-Alizadeh Sani	0.936	0.954	0.975	0.946
Cleveland	0.840	0.821	0.896	0.830
Statlog	0.878	0.784	0.902	0.828

was still lower than that of the z-Alizadeh-Sani dataset. This is because the z-Alizadeh Sani dataset has a high level of data imbalance, as shown in Table 3; therefore, the use of SMOTE is very effective when compared to the Cleveland and Statlog datasets. The Cleveland and Statlog datasets had the same attributes, however the feature selection results had two different attributes.

V. DISCUSSION AND CONCLUSIONS

A. Discussion

The hybrid SVM-GA model combined with FCBF can significantly reduce the number of attributes, particularly for the z-Alizadeh Sani dataset. In the z-Alizadeh Sani dataset, there was a decrease in the number of attributes from 54 to eight. Attribute reduction occurs in two stages, from 54 to 25 attributes, using a hybrid SVM-GA with accuracy and sensitivity fitness functions, while 25 to eight attributes refer to the weights generated by the FCBF algorithm. Changes in performance based on the number of attributes (25 to 8) are shown in Fig. 7. Performance changes that occur from the use of 25 to eight attributes with the bagging-LMT classification algorithm, as well as 10-fold cross-validation validation, are not significant. Tests using the Cleveland and Statlog datasets were also able to reduce the number of attributes from 13 to 10 when using hybrid SVM-GA and after being combined with FCBF to six attributes. The resulting performance is not as good as that when tested with the z-Alizadeh Sani dataset because the z-Alizadeh Sani dataset has a high level of imbalanced data, as shown in Table 1. When tested using the z-Alizadeh Sani dataset, the performance of the proposed system can provide an AUC performance of 97.5%, which is included in the very good category [52]. Meanwhile, in testing with the Cleveland dataset, the AUC value of 89.6% was still in the good category, whereas the statlog dataset AUC value of 90.2% was in the very good category [52].

Changes in the number of features when using features 25 to 8, as shown in Fig. 7. It can be observed that the AUC performance parameter decreases the number of attributes, and there is an increase in performance. This shows that when using 25 attributes, there is ambiguity between attributes that causes performance to decrease, which also confirms previous studies that states feature selection can

improve performance [5-7]. In the Cleveland and Statlog datasets, when compared using 10 attributes with 6 attributes 6, the resulting AUC performance parameters are relatively the same. The Cleveland and Statlog datasets have the same number and type of attributes, however the result of the feature selection process produced has two different attributes. The difference is in the Cleveland dataset which includes slope and age, whereas the Statlog dataset includes old-peak and thalach. The difference is, of course, caused by the data in each dataset; therefore, the results of the weight ranking of the FCBF feature selection process are different.

In the proposed system model, by adding a data-balancing process, the sensitivity and specificity performance parameters were both high. Many studies have reported high accuracy performance parameters, high sensitivity, and low specificity. A significant difference between sensitivity and specificity was shown in the model of CHD diagnosis using the CART algorithm [21], bagging-SMO, naive Bayesian, SMO, and neural networks [53]. A significant difference between the sensitivity and specificity will also have an impact on the low AUC performance value if the AUC calculation is the average between the sensitivity and specificity [15]. In the biomedical field, if there is an AUC value of 97.5%, it indicates that if 100 patients are positive for CHD disease, then there are 97 people correctly diagnosed with CHD by the system, while three patients are wrong. The capability of the proposed system with reference to the AUC parameter, is better than that of some previous studies, such as the research conducted by Joloudari et al. [54], which uses a random tree algorithm.

Using SVM-GA hybrid feature selection with fitness, which is a function of accuracy and sensitivity, causes the best feature subset to be determined by the sensitivity and accuracy parameters. Using the fitness function, as shown in Equation (1), results in better accuracy performance and a smaller number of attributes compared to using only the fitness function with accuracy [26, 55]. The same is true for the PSO algorithm [12]. A complete comparison with previous studies using either GA, PSO, or other algorithms, with

testing using the z-Alizadeh Sani dataset, is shown in Table 4.

Using the bagging-LMT algorithm in the proposed diagnosis system model can provide better performance than several other algorithms, such as RF and C4.5. The LMT algorithm is a combination of logistic regression and C4.5. This combination can be used to prune and prevent overfitting [56]. The ability of the LMT algorithm is even better when bagging is used. The application of bagging resolves the problem of unstable classifications. This makes the bagging ability of the LMT better than that of C4.5, as well as the MLP, because the MLP has a serious problem of overfitting. The random forest algorithm is a decision-tree-based ensemble algorithm that does not prune the resulting decision tree [57]. Failure to perform this pruning can lead to high prediction errors in new cases, and another weakness is the slow classification process, therefore it is not suitable for real-time cases. The forestPA algorithm is almost the same as a random forest, except that it is built using the CART algorithm [58]. The weakness of this algorithm is that the determination of the wrong weight on the attribute affects its performance.

B. Conclusion

The diagnosis system model using the hybrid feature selection method SVM-GA and FCBF, as well as the bagging-LMT algorithm, is able to provide good performance. The best performance occurs when using the z-Alizadeh Sani dataset because this dataset has a high level of imbalanced data compared to the Cleveland and Statlog datasets. This makes the addition of the SMOTE algorithm highly effective. The best performance was achieved with a sensitivity of 93.6%, specificity of 95.4%, AUC of 97.5%, PPV of 94.3%, NPV of 94.9%, and accuracy of 94.6% for the z-Alizadeh Sani dataset. This performance also shows that the LMT bagging ability is better than that of the random forest, MLP, C4.5, forestPA, and forestPA bagging algorithms. The resulting attribute reduction also showed a significant decrease from 54 to eight attributes. Referring to the resulting performance, the proposed diagnostic system model performed better than several previous studies. This makes the proposed model an alternative for the diagnosis of CHD with minimal examination attributes.

ACKNOWLEDGEMENTS

We would like to thank the National Research and Innovation Agency of the Republic of Indonesia for providing research funding under the Basic Research Grant scheme (Contract Number: 221.1/UN27.22/HK.07.00/2021).

Table 4. Comparison with previous research (z-Alizadeh Sani)

Ref	Method	#Feature	SEN	SPE	AUC
[59]	Var-IBLMM	54	85.6	73.7	-
[53]	Bagging-SMO	33	95.8	87.4	-
[42]	Combined IG for all arteries- SVM	27	86.0	-	-
[11]	ANN+GA	22	97.0	92.0	-
[10]	PSO-based FS	27	-	-	98.7
[12]	Hybrid PSO+ENN	22	-	-	-
[54]	Random Tree	40	-	-	96.7
[60]	EHBM-DNN	54	95.8	96.5	-
[40]	SVM along with Feature engineering	32	1.00	88.0	92.0
Proposed SVM-GA&FCBF		8	93.6	95.4	97.5

REFERENCES

- [1] S. Bae, S. R. Kim, N. Kim, W. J. Shim, and M. Park, "Impact of cardiovascular disease and risk factors on fatal outcomes in patients with COVID-19 according to age: a systematic review and meta-analysis," vol. 107, no. 5, pp. 373-380, 2021.
- [2] N. M. Hemphill, M. T. Y. Kuan, and K. C. Harris, "Reduced physical activity during COVID-19 pandemic in children with congenital heart disease," *Canadian Journal of Cardiology*, vol. 36, no. 2020, pp. 1130-1134, 2020.
- [3] J. Kim, J. Lee, and Y. Lee, "Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree," *Healthcare Informatics Research*, vol. 21, no. 3, pp. 167-174, 2015, DOI: 10.4258/hir.2015.21.3.167.
- [4] W. Wiharto, H. Kusnanto, and H. Herianto, "System diagnosis of coronary heart disease using a combination of dimensional reduction and data mining techniques: A review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 2, pp. 514-523, 2017, DOI: 10.11591/ijeecs.v7.i2.pp514-523.
- [5] N. M. Khan, N. Madhav C, A. Negi, and I. S. Thaseen, "Analysis on improving the performance of machine learning models using feature selection technique," in *Intelligent Systems Design and Applications*, vol. 941, A. Abraham, A. K. Cherukuri, P. Melin, and N. Gandhi, Eds. Cham: Springer International Publishing, 2020, pp. 69-77. DOI: 10.1007/978-3-030-16660-1_7.
- [6] K. P. Shroff and H. H. Maheta, "A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy," in *2015 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, Jan. 2015, pp. 1-6. DOI: 10.1109/ICCCI.2015.7218098.
- [7] E. M. Karabulut, S. A. Özel, and T. İbrikçi, "A comparative study on the effect of feature selection on classification accuracy," *Procedia Technology*, vol. 1, pp. 323-327, 2012, DOI: 10.1016/j.protcy.2012.02.068.
- [8] Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui, "a framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection," *Journal of Healthcare Engineering*, vol. 2018, pp. 1-11, 2018, DOI: 10.1155/2018/6275435.
- [9] S. Belarouci and M. A. Chikh, "Medical imbalanced data classification," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 2, no. 3, pp. 116-124, Apr. 2017, DOI: 10.25046/aj020316.
- [10] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *BioMed Research International*, vol. 2020, pp. 1-10, Apr. 2020, DOI: 10.1155/2020/9816142.
- [11] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," *Computer Methods and Programs in Biomedicine*, vol. 141, no. 2017, pp. 19-26, 2017, DOI: 10.1016/j.cmpb.2017.01.004.
- [12] A. H. Shahid and M. P. Singh, "A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1568-1585, Oct. 2020, DOI: 10.1016/j.bbe.2020.09.005.
- [13] R. P. Cherian, N. Thomas, and S. Venkitachalam, "Weight optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm," *Journal of Biomedical Informatics*, vol. 110, p. 103543, Oct. 2020, DOI: 10.1016/j.jbi.2020.103543.
- [14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002, DOI: 10.1613/jair.953.
- [15] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "SMOTE-RSB *: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory," *Knowledge and Information Systems*, vol. 33, no. 2, pp. 245-265, 2012, DOI: 10.1007/s10115-011-0465-6.
- [16] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, and P. A. Patel, "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure," *American Heart Journal*, vol. 229, pp. 1-17, Nov. 2020, DOI: 10.1016/j.ahj.2020.07.009.
- [17] N. Kumar, N. N. Das, D. Gupta, K. Gupta, and J. Bindra, "Efficient Automated Disease Diagnosis Using Machine Learning Models," *Journal of Healthcare Engineering*, vol. 2021, pp. 1-13, 2021.
- [18] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304-19326, 2021, DOI: 10.1109/ACCESS.2021.3053759.
- [19] C. Krittanawong, "Machine learning prediction in cardiovascular diseases: A meta-analysis," *Scientific Reports*, vol. 2020, no. 10, pp. 1-11, 2020.
- [20] L. Ashish, S. K. V, and S. Yeligi, "Ischemic heart disease detection using support vector machine and extreme gradient boosting method," *Materials Today: Proceedings*, p. S2214785321008129, Feb. 2021, DOI: 10.1016/j.matpr.2021.01.715.
- [21] M. M. Ghiasi, S. Zendejboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Computer Methods and Programs in Biomedicine*, vol. 192, pp. 1-14, Aug. 2020, DOI: 10.1016/j.cmpb.2020.105400.
- [22] M. Zomorodi-moghadam, M. Abdar, Z. Davarzani, X. Zhou, P. Plawiak, and U. R. Acharya, "Hybrid particle swarm optimization for rule discovery in the diagnosis of coronary artery disease," *Expert Systems*, vol. 38, no. 1, Jan. 2021, DOI: 10.1111/exsy.12485.
- [23] M. Abdar, W. Książek, U. R. Acharya, R.-S. Tan, V. Makarenkov, and P. Plawiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 179, p. 104992, Oct. 2019, DOI: 10.1016/j.cmpb.2019.104992.
- [24] A. G. Karegowda, A. S. Manjunath, G. Ratio, and C. F. Evaluation, "Comparative study of attribute selection using gain ratio," *International Journal of Information Technology and Knowledge and Knowledge Management*, vol. 2, no. 2, pp. 271-277, 2010.
- [25] H. Djellali, S. Guessoum, N. Ghoulmi-Zine, and S. Layachi, "Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection," in *2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B)*, Boumerdes, pp. 1-6, Oct. 2017. DOI: 10.1109/ICEE-B.2017.8192090.
- [26] E. P. Ephzibah, "Cost effective approach on feature selection using genetic algorithms and LS-SVM classifier," *IJCA*, vol. ecot, no. 1, pp. 16-20, Dec. 2010. DOI: 10.5120/1532-135.
- [27] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A hybrid approach for heart disease diagnosis and prediction using machine learning techniques," in *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*, vol. 12157, M. Jmaiel, M. Mokhtari, B. Abdulrazak, H. Aloulou, and S. Kallel, Eds. Cham: Springer International Publishing, pp. 299-306, 2020. DOI: 10.1007/978-3-030-51517-1_26.
- [28] A. K. Shukla, P. Singh, and M. Vardhan, "A new hybrid feature subset selection framework based on binary genetic algorithm and information theory," *Int. J. Comp. Intel. Appl.*, vol. 18, no. 03, p. 1950020, Sep. 2019, DOI: 10.1142/S1469026819500202.

- [29] W. Wiharto, H. Herianto, and H. Kusnanto, "A tiered approach on dimensional reduction process for prediction of coronary heart disease," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 2, pp. 487-495, 2018, DOI: 10.11591/ijeecs.v11.i2.
- [30] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings, Twentieth International Conference on Machine Learning*, Washington, DC, United States, pp. 856-863, Aug. 2003.
- [31] N. Sánchez-Marroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection - A comparative study," in *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, vol. 4881, H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 178-187. DOI: 10.1007/978-3-540-77226-2_19.
- [32] Y. Khouridfi and M. Bahaj, "Feature selection with fast correlation-based filter for breast cancer prediction and classification using machine learning algorithms," in *2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, Rabat, Morocco, pp. 1-6, Nov. 2018. DOI: 10.1109/ISAECT.2018.8618688.
- [33] W. Xie, G. Liang, Z. Dong, B. Tan, and B. Zhang, "An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data," *Mathematical Problems in Engineering*, vol. 2019, pp. 1-13, May 2019, DOI: 10.1155/2019/3526539.
- [34] Y. -T. Kim, D. -K. Kim, H. Kim, and D. -J. Kim, "A comparison of oversampling methods for constructing a prognostic model in the patient with heart failure," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea (South), pp. 379-383, Oct. 2020. DOI: 10.1109/ICTC49870.2020.9289522.
- [35] J. Brandt and E. Lanzen, "A comparative review of SMOTE and ADASYN in imbalanced data classification," Dissertation, Uppsala University, Sweden, 2021.
- [36] N. Matondang and N. Surantha, "Effects of oversampling SMOTE in the classification of hypertensive dataset," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 5, no. 4, pp. 432-437, Aug. 2020, DOI: 10.25046/aj050451.
- [37] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 106, pp. 1-6, Dec. 2013, DOI: 10.1186/1471-2105-14-106.
- [38] A. R. Purnajaya, W. A. Kusuma, and M. K. D. Hardhienata, "Performance comparison of data sampling techniques to handle imbalanced class on prediction of compound-protein interaction," *Bio*, vol. 8, no. 1, pp. 41-48, Jun. 2020, DOI: 10.24252/bio.v8i1.12002.
- [39] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, "Diagnosis of coronary artery disease using cost-sensitive algorithms," in *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, pp. 9-16, 2012, DOI: 10.1109/ICDMW.2012.29.
- [40] R. Alizadehsani *et al.*, "Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries," *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 119-127, 2018, DOI: 10.1016/j.cmpb.2018.05.009.
- [41] R. Alizadehsani *et al.*, "Diagnosis of coronary artery disease using data mining based on lab data and echo features," *Journal of Medical and Bioengineering*, vol. 1, no. 1, pp. 26-29, 2013, DOI: 10.12720/jomb.1.1.26-29.
- [42] R. Alizadehsani *et al.*, "Coronary artery disease detection using computational intelligence methods," *Knowledge-Based Systems*, vol. 109, pp. 187-197, Oct. 2016, DOI: 10.1016/j.knsys.2016.07.004.
- [43] R. Detrano, A. Janosi, W. Steinbrunn, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304-310, 1989, DOI: 10.1016/0002-9149(89)90524-9.
- [44] Y. Zhang, F. Liu, Z. Zhao, D. Li, X. Zhou, and J. Wang, "Studies on application of support vector machine in diagnose of coronary heart disease," *2012 6th International Conference on Electromagnetic Field Problems and Applications, ICEF'2012*, 2012, DOI: 10.1109/ICEF.2012.6310380.
- [45] R. Jing and Y. Zhang, "A view of support vector machines algorithm on classification problems," in *2010 International Conference on Multimedia Communications*, TBD, TBD, Hong Kong, pp. 13-16, Aug. 2010. DOI: 10.1109/MEDIACOM.2010.21.
- [46] K. Uyar and A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Computer Science*, vol. 120, pp. 588-593, 2017, DOI: 10.1016/j.procs.2017.11.283.
- [47] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast Correlation Based Filter (FCBF) with a different search strategy," in *2008 23rd International Symposium on Computer and Information Sciences*, Istanbul, Turkey, pp. 1-4, Oct. 2008. DOI: 10.1109/ISCIS.2008.4717949.
- [48] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, pp. 161-205, 2005, DOI: 10.1007/s10994-005-0466-3.
- [49] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier," *Artificial Intelligence in Medicine*, vol. 98, pp. 35-47, Jul. 2019, DOI: 10.1016/j.artmed.2019.07.005.
- [50] M. C. Tu, D. Shin, and D. Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms," in *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, Chengdu, China, pp. 183-187, Dec. 2009. DOI: 10.1109/DASC.2009.40.
- [51] M. N. Adnan and M. Z. Islam, "Forest PA : Constructing a decision forest by penalizing attributes used in previous trees," *Expert Systems with Applications*, vol. 89, pp. 389-403, Dec. 2017, DOI: 10.1016/j.eswa.2017.08.002.
- [52] F. Gorunescu, "Data mining: Concepts, models, and techniques," Berlin, Heidelberg: Springer, 2011.
- [53] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 52-61, Jul. 2013, DOI: 10.1016/j.cmpb.2013.03.004.
- [54] J. H. Joloudari, E. H. Joloudari, H. Saadatfar, M. Ghasemigol, S. M. Razavi, A. Mosavi, N. Nabipour, S. Shamshirband, and L. Nadai, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *IJERPH*, vol. 17, no. 3, p. 731, Jan. 2020, DOI: 10.3390/ijerph17030731.
- [55] N. Jothi, W. Husain, N. Abdul Rashid, and S. M. Syed-Mohamad, "Feature selection method using genetic algorithm for medical dataset," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 6, p. 1907, Dec. 2019, DOI: 10.18517/ijaseit.9.6.10226.
- [56] X. Luo, F. Lin, Y. Chen, S. Zhu, Z. Xu, Z. Huo, M. Yu, and J. Peng, "Coupling logistic model tree and random subspace to predict the landslide susceptibility areas with considering the uncertainty of environmental features," *Sci Rep*, vol. 9, no. 1, p. 15369, Dec. 2019, DOI: 10.1038/s41598-019-51941-z.

- [57] J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck, "A comparison of random forests, boosting and support vector machines for genomic selection," *BMC Proc.*, vol. 5, no. S3, p. S11, Dec. 2011, DOI: 10.1186/1753-6561-5-S3-S11.
- [58] M. Pramanik, R. Pradhan, P. Nandy, A. K. Bhoi, and P. Barsocchi, "Machine learning methods with decision forests for parkinson's detection," *Applied Sciences*, vol. 11, no. 2, p. 581, Jan. 2021, DOI: 10.3390/app11020581.
- [59] C. Hu, W. Fan, J. -X. Du, and N. Bouguila, "A novel statistical approach for clustering positive data based on finite inverted Beta-Liouville mixture models," *Neurocomputing*, vol. 333, pp. 110-123, Mar. 2019, DOI: 10.1016/j.neucom.2018.12.066.
- [60] P. Verma, V. K. Awasthi, and S. K. Sahu, "A novel design of classification of coronary artery disease using deep learning and data mining algorithms," *Revue d'Intelligence Artificielle*, vol. 35, no. 3, pp. 209-215, 2021.



Wiharto

Wiharto is an Associate professor of Computer Science at Department of Informatics, Sebelas Maret University, Surakarta, Indonesia. He received his Ph.D. degree from Gadjah Mada University, Indonesia in 2017. He is conducting research activities in the areas of artificial intelligence, computational intelligence, expert system, Machine learning, and data mining.



Esti Suryani

Esti Suryani received obtained a Bachelor of Science (B.S.) from Gadjah Mada University, Yogyakarta, Indonesia, 2002 and master's degree in Computer Science (M.Cs.) from Gadjah Mada University, Yogyakarta, Indonesia, 2006. He is presently working as an Assistant professor in the Department of Informatics, Faculty of mathematics and natural sciences, Sebelas Maret University, Surakarta, Indonesia. His experience and areas of interest focus on image processing and fuzzy logic.



Sigit Setyawan

Sigit Setyawan received obtained a Bachelor of Medicine from Sebelas Maret University, Surakarta, Indonesia, 2005 and master's degree in Medicine (M.Sc.) from Gadjah Mada University, Yogyakarta, Indonesia, 2015. He is presently working as an Assistant professor in the Department of Medicine, Faculty of Medicine, Sebelas Maret University, Surakarta, Indonesia. His experience and areas of interest focus on Biologi molecular, Genomic, and health informatics.



Bintang PE Putra

student of undergraduate program in informatics, 2018, Faculty of Mathematics and Natural Sciences, Sebelas Maret University, Surakarta, Indonesia. The area of research being carried out is the image processing, data mining, artificial intelligence, machine learning, and computational intelligence.