Effective predictive modelling for coronary artery diseases using support vector machine

Kuncahyo Setyo Nugroho, Anantha Yullian Sukmadewa, Angga Vidianto, Wayan Firdaus Mahmudy
Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia

Article Info

Article history:

Received Jun 21, 2021
Revised Dec 24, 2021
Accepted Jan 2, 2022

Keywords:

Coronary artery disease prediction
Effective support vector machines
Good support vector machines parameters

ABSTRACT

Coronary artery disease (CAD) is a category of cardiovascular disease that causes the highest mortality rate in the world. CAD occurs due to plaque build-up on the walls of the arteries that supply blood to the heart and other organs of the body. To control the mortality rate, a practical model that is capable of predicting CAD is needed. Machine learning approaches have been used in solving various problems in various domains, including biomedicine. However, real-world data often has an unbalanced class distribution that can interfere with classifier performance. In addition, data has many features to process. This study focuses on effective modeling capable of predicting CAD using feature selection to handle high dimensional data and feature resampling to handle unbalanced data. Feature selection is very effective by eliminating irrelevant features from the training data. Hyperparameter tuning is also done to find the best combination of parameters in support vector machines (SVM). Our results show that the SVM cross-validated ten times has a more accurate training result. Furthermore, the grid search on SVM cross-validated ten times had more accurate training model results and achieved 88% accuracy on the test data.

Corresponding Author:

Kuncahyo Setyo Nugroho
Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University
Jalan Veteran, No. 8, Lowokwaru, Malang, East Java 65145, Indonesia
Email: ksnugroho26@gmail.com

1. INTRODUCTION

The heart is an essential organ in the cardiovascular system that requires a supply of blood that contains oxygen. The coronary circulation supplies blood to the heart. The aorta divides into two major coronary arteries, each of which branches off into smaller arteries and supplies blood to the entire heart muscle [1]. Cardiovascular disease (CVD) is a group of diseases. Coronary heart disease (CHD), coronary artery disease (CAD), and acute coronary syndrome (ACS) are included in CVD [2]. CAD happens because of plaque build-up in the wall of arteries that supply blood to the heart and other parts of the [3]. The condition known as atherosclerosis occurs when plaque begins to accumulate in these arteries. As the plaque hardens, the coronary arteries narrow, decreasing the blood supply to the heart. A blood clot on the plaque's surface may occur if it ruptures. In the majority of situations, a big blood clot can totally stop the coronary arteries' blood flow. A heart attack, if left untreated, can result in major health consequences and, in the worst-case scenario, death. As a result, cardiovascular disease is the leading cause of death globally [4].

There is an adequate need for the early detection of patients with CAD. A machine learning approach can solve problems in the biomedical domain [5]–[8]. Machine learning gives the computer ability to learn and improve from experience automatically. Machine learning algorithms have several major categories based on their learning approach, input and output data, and problem type: supervised, unsupervised, and reinforcement learning [9]. Support vector machines (SVMs) used in supervised learning have been shown to be extremely effective at solving classification problems in a variety of biomedical fields [6], [10]–[12].

There are many studies conducted to diagnose CAD with machine learning in recent years. The most widely used dataset in CAD diagnosis is the Z-Alizadehsani dataset [13]. Using this dataset, [14] applied data mining techniques to diagnose CAD based on the symptoms and characteristics of the patient's ECG. Their research used sequential minimal optimization (SMO) and naïve bayes (NB) classifier and a combination of both to diagnose CAD. Testing with 10-fold cross-validation shows that the combination of SMO-naïve bayes is superior by achieving more than 88.52% accuracy than SMO of 86.95% and naïve bayes of 87.22%. In another study, [15] using SMO, naïve bayes, bagging with SMO, and neural network to diagnose the same disease. Information gain is used to determine which features are most effective for diagnosing CAD. As a result, SMO with information gain obtained the best performance with an accuracy of 94.08%.

Alizadehsani *et al.* used the feature selection technique used in NB, C4.5, and SVM to diagnose CAD. Using 10-fold cross-validation, SVM has the highest accuracy of 96.40% [13]. To increase accuracy, [16] used random trees (RT), decision tree (DT), SVM, and chi-squared automatic interaction detection (CHAID) to select features based on predefined criteria for CAD diagnosis. Random trees are the best method by selecting 40 significant features and bringing out an accuracy of 91.47% [17] using hybrid PCA, DT, and firefly optimization techniques to optimize the accuracy of existing models. The PCA algorithm is used to extract features, the firefly optimization technique is used to optimize the feature selection, and DT is used to classify the data. They achieve 93% accuracy with a low classification error rate also low false positive and negative rates.

Other studies have also been conducted on the prediction of disease in the biomedical field. SVM is used to predict diabetes and pre-diabetes [10]. The SVM model is used to identify characteristics that best classify individuals into different diabetes subtypes. Their model got 83.47% for detection of diagnosed diabetes or diagnosed diabetes compared to [18] model that got 82.1%. In this research, they conclude SVM is a promising model for detecting a complex disease using common and simple variables. According to [11], [19] SVM has superior accuracy when predicting heart disease, diabetes, and parkinson’s disease. SVM was also reported to obtain better accuracy than random forest in breast cancer prediction [20].

Machine learning assuming a balanced number of instances in each class. When using unbalanced data can lead to inaccurate model prediction results. Synthetic minority oversampling technique (SMOTE) and adaptive synthetic (ADASYN) sampling are alternatives to overcome unbalanced data by creating synthetic data in the minority class [21]. SMOTE, which is integrated with the prediction model is reported to improve the prediction model's performance [22], [23]. In CAD prediction, SMOTE on artificial neural networks, DT, and SVM showed an increase in the accuracy obtained from the original data [24]. Meanwhile, [25] using ADASYN with SVM to diagnose Parkinson's disease effectively. Both studies do not employ feature selection to determine the most essential features for output prediction.

Based on previous studies, a combination of feature selection and feature resampling in CAD prediction has never been done before. Both techniques are reported to improve the performance of the resulting model. The main contribution of this study is that we propose a framework for building an effective model using feature selection and feature resampling in CAD predictions. Feature selection is used to find the most relevant features to CAD predictions. While handling imbalanced data, we reviewed several feature resampling. We use SVM with hyperparameter tuning to find the combination of parameters to make an effective CAD prediction.

2. RESEARCH METHOD

There are four main steps to complete this research, as shown in Figure 1. The first step is data exploration, followed by data preprocessing. Next, we use feature selection to determine which features have the most importance on the target variable. After identifying the relevant features, the dataset is divided into training and testing sets for the purpose of implementing multiple machine learning algorithms. The last step is model evaluation. This section discusses the processes and procedures involved in doing this research.

2.1. Dataset description

We use the Z-Alizadehsani dataset downloaded from the UCI machine learning repository. The dataset contains records of 303 patients who visited the Shaheed Rajaei Cardiovascular, Medical, and Research Center in Iran. Each patient has 54 features to diagnose CAD. These features are grouped into four categories: demographic, symptom and examination, electrocardiogram (ECG), and laboratory and echo features. Patients are categorized as having CAD if they experience stenosis in one of their coronary arteries more than or equal to 50%. A total of 216 patients in the dataset had this disease, while the rest were normal patients. This shows that the dataset has an unbalanced class distribution. The target feature on the dataset is cath with a CAD value for patients with coronary artery disease and normal for normal patients.

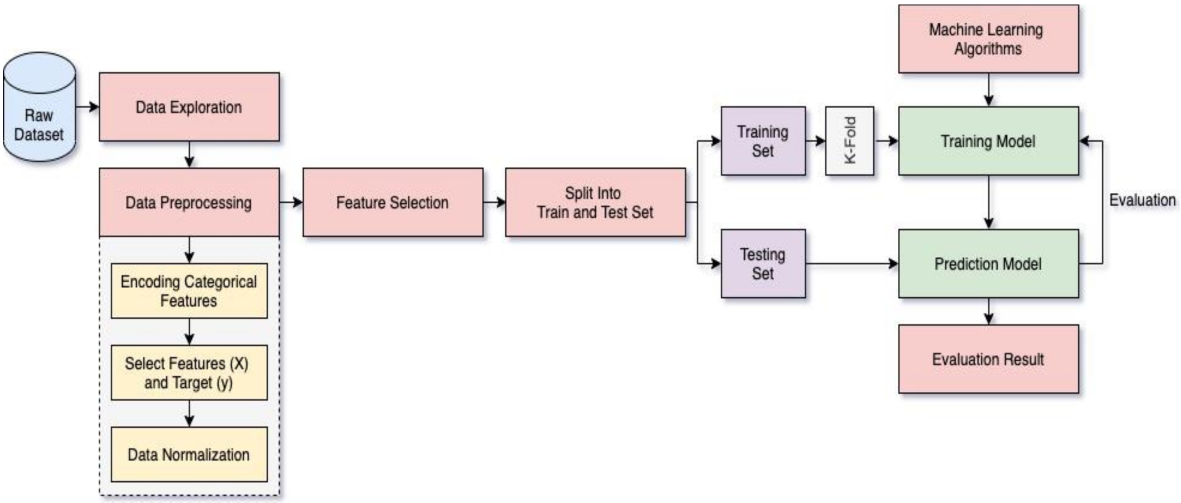


Figure 1. Research method design for CAD prediction

2.2. Data exploration

Our dataset has many diverse features, so this step is taken to explore the dataset to get useful insights through visualization and data analysis. This step also helps us find out the missing values and identify the types of numeric features and categorical features in the dataset.

2.3. Data preprocessing

Real-word datasets have incomplete, inconsistent, and even have missing value on specific features. Data preprocessing used to clean and format the raw data in the dataset so that machine learning algorithms can easily represent the feature set. In this study, we implemented several data preprocessing steps. The first step is to convert categorical features to numeric values because machine learning algorithms can only read and process numeric values. Next, we create a feature matrix that is used as the input variable and the target variable. The input feature is stored into X variable while the target feature is stored into Y variable. The final step is normalizing the data to rescale the numeric features into ranges 0 and 1 used a min-max scaler, as shown in (1).

$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{1}$$

2.4. Feature selection

The features used to train machine learning algorithms have a significant impact on the performance of the final model. Irrelevant features can have a negative impact on the resulting model [26]. To identify features that affect the target variable, feature selection can be used. Feature selection is the process of reducing the number of features in a dataset in order to improve the model's performance [27]. We used feature selection to predict which features were most important in influencing patients with CAD or not. extremely randomized trees classifier is an ensemble learning type used for feature selection. In this method, each decision tree is generated from the training sample. Then, at each test node, the decision tree is given a random sample of k features of all features, where each decision tree must choose the best feature to separate the data based on the Gini Index value. This random feature will provide several uncorrelated decision trees. This value is referred to as the feature's Gini Importance. To make the feature selection process easier, each feature is graded according to its Gini Importance.

2.5. Data separation

Data separation is used to evaluate machine learning algorithms' performance when predicting data that was not used to train the model. Divide the dataset into two subsets using the data separation process.

The first subset is utilized to train machine learning algorithms in order to generate prediction models. The second subset is the test set on which the prediction model is evaluated. We trained on 75% of the dataset and tested the model on the remaining 25%.

2.6. Stratified k-fold

When performing the data separation procedure, the main problem must be enough data to divide the dataset into training data and test data as data representations following the problem domain. Therefore, this procedure is not suitable for evaluating model performance if there are few datasets available. There will not be enough data on the training or testing subset for the model to learn the effective mapping from input to output. Prediction performance can be too optimistic (good prediction) or too pessimistic (bad).

An alternative method that can be used if do not have enough data is the K-Fold procedure by folding K as much data and repeating the process as many as K as well. One type of this procedure is a Stratified K-fold, as shown in Figure 2. Stratified K-Fold is helpful if the available dataset is few and has an unbalanced class distribution. We want to maintain the class imbalance to represent some information about what the model is trying to predict. In this study, we use a combination of the Stratified K-Fold procedure to conduct a final evaluation of the performance of the implemented model. After separated the training and testing set in the previous steps, we further divided the training set into validation set to validate the machine learning algorithms performance during the k-fold iteration process. We also performed feature resampling to balance the distribution of classes in the training set during this process. The generated prediction model is finally tested using the testing set as the final result of the predicted performance.

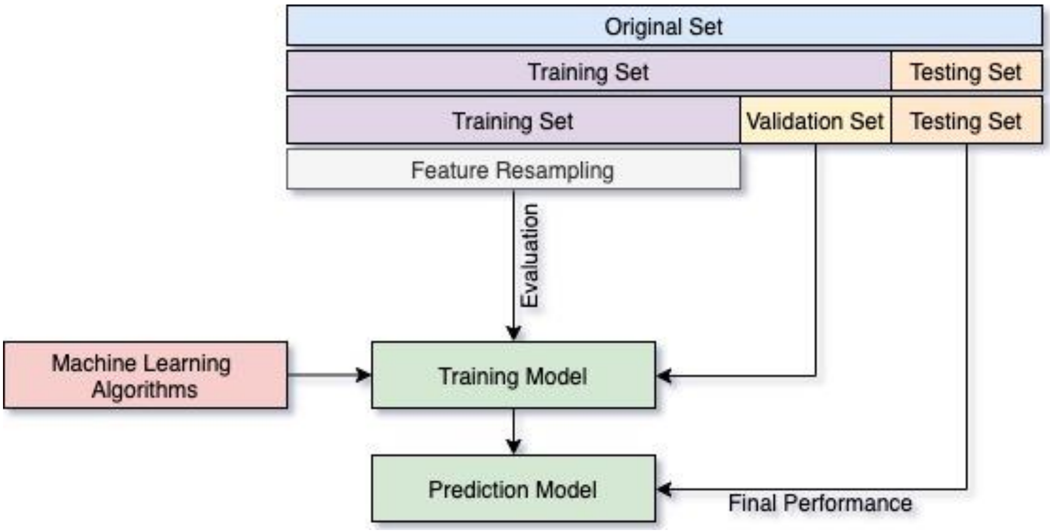


Figure 2. Stratified k-fold schema

2.7. Feature resampling

Imbalance data causes the model to be biased in choosing the majority class. There are many ways to handle a dataset with an unbalanced class distribution, including collecting more data, trying variations in machine learning algorithms, using both oversampling and undersampling techniques. Collecting more data is impossible because it requires more time and costs, while undersampling techniques can cause the loss of important information in the dataset. Therefore, in this study, we focus on oversampling techniques SMOTE and ADASYN so that we hope not to lose any information that might be valuable in the dataset. SMOTE uses a k-nearest neighbors (k-NN)-based distance approach to create synthetic data [28]. First, the data is randomly selected from the minority class, then K is the closest neighbor of the data. Synthetic data is generated between randomly selected and K-nearest data. This step is repeated until the minority class has the same proportion as the majority class. Meanwhile, ADASYN is a variation of SMOTE by creating synthetic data based on data density [29]. The synthetic data generated will be inversely proportional to the density of the minority class. That is, more synthetic data is generated in the feature space where the density of the minority class is low, and less or even less synthetic data is generated in the high density minority class [21].

2.8. Support vector machine

The SVM is a highly effective classification algorithm by finding decision boundaries known as hyperplanes. The optimal hyperplane separates the instances correctly into each class. The margins on the optimal hyperplane and instances in training are maximized to fit the data. The SVM model is not delicate to other information focuses. Its point is to track down the best division line, for example, the ideal hyperplane between the two classes of tests, to have the most significant distance conceivable to every one of the two classes of help vectors. The separator line dictates the indicator include for each prescient class. Figure 3 shows the vector machine in 2-dimensional space [16].

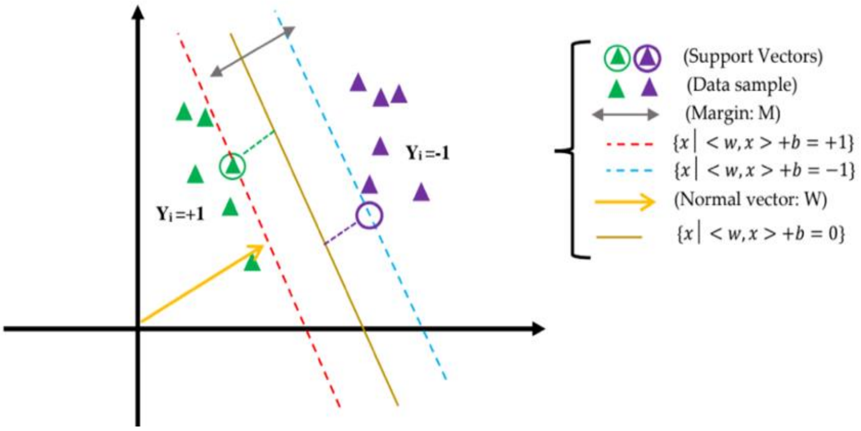


Figure 3. SVM in 2-dimensional space [16]

A hyperplane with a wider margin is projected to be more accurate than one with a smaller margin when classifying future data. Therefore, the hyperplane with the largest margin will be searched for. The function has the following in (2) [30].

$$y(x) = sgn[\textstyle\sum_{i=1}^{m} \alpha_i y_i (x_i, x) + b] \tag{2}$$

However, in (2) can be applied if the sample data used can be separated linearly. Kernel methods enable the transformation of data into huge dimensions for classification challenges. As is the case with data samples that cannot be split linearly, the kernel function converts the data to a higher-dimensional space without actually changing it to that space. In (3) can be applied when the data sample situation cannot be separated linearly.

$$y(x) = sgn[\textstyle\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b] \tag{3}$$

The kernel function $K(x_i, x)$ is equals to (x_i, x) and x is the non-linear space from the original space to high dimensional space. Where, r and d are kernel parameters, and the four basic kernels are given as follows in (4)-(7). In this study, we use all kernels to find the best SVM performance.

$$Linear: K(x_i, x_j) = x_i^T\, x_j \tag{4}$$

$$Polynomial: K(x_i, x_j) = (\gamma\, x_i^T\, x_j + r)^d, \gamma > 0 \tag{5}$$

$$Radial\ Bias\ Function\ (RBF): K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right), \gamma > 0 \tag{6}$$

$$Sigmoid: K(x_i, x_j) = tanh(\gamma x_i^T\, x_j + r) \tag{7}$$

2.9. Evaluation metrics

The model generated during the training phase is used to obtain predictive results from population data. In the classification, the confusion matrix describes the model's performance by calculating which classes are predicted correctly and incorrectly and what types of errors are made. True positive (TP) is defined as positive instances that are predicted to be true. For example, a patient with CAD is predicted to have true CAD. True negative (TN) is defined as negative instances that are predicted to be true. For example, a patient who does not have CAD is predicted not to have CAD. False positive (FP) is negative instances that are predicted as positive instances. For example, a patient who does not have CAD is predicted to have CAD. False negative (FN) is positive instances that are predicted as negative instances. For example, a patient who has CAD is predicted not to have CAD.

The most frequently used performance metric based on the confusion matrix for classification is accuracy. Accuracy is the ratio of true predictions (TP and TN) with the overall data that describes the level of closeness of the predicted value to the actual value, as shown in (8). In the training phase, the model's accuracy is obtained from the average of each fold in the cross-validation. The standard deviation was also calculated to see the variance. The problem with unbalanced data is negative instances with the majority class and positive instances with fewer classes. To interpreting the model performance with unbalanced data, receiver operating characteristic (ROC) curve are used. The ROC curve is obtained from the true positive rate (TPR) as in (9) and the false positive rate (FPR) as in (10).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{8}$$

$$TPR = \frac{TP}{TP+FN} \tag{9}$$

$$FPR = \frac{TN}{TN+FP} \tag{10}$$

3. RESULTS AND DISCUSSION

Based on the research framework in Figure 1, the first step we take is preprocessing by changing all categorical features in the dataset to numeric values and normalizing them using the minmax scaler. Next, we use feature selection to determine which features are significant and have an effect on the target variable. We choose features using the additional trees classifier. From the results of feature selection, we found that there are 16 features that are correlated with the target variables shown in Table 1. In each subsequent test, we compare the model's performance using all the features and feature selection results. We wanted to find out if feature selection could improve the performance of a given model.

Table 1. The features used are based on the selection results

Feauture name	Feature category
Age	Demographic
Weight	Demographic
BMI	Demographic
HTN	Demographic
BP	Symptom and examination
Typical chest pain	Symptom and examination
Atypical	Symptom and examination
Nonanginal	Symptom and examination
Tinversion	ECG
FBS	Laboratory and echo
TG	Laboratory and echo
ESR	Laboratory and echo
Neut	Laboratory and echo
EF-TTE	Laboratory and echo
Region RMWA	Laboratory and echo

To evaluate the predictive model, we divided the dataset into 75% for the training set and the remainder for the test set. Next, the 10-layer cross-validation procedure we applied to the training set. In this procedure, the training set is randomly divided into ten sections, 9 folds are used to train the classification model and the remaining 1 fold is divided to validate the model. This procedure is performed 10 folds. The dataset character has class imbalance, the value of cross validation is a metric that can be used to evaluate the model because the folds are made with the same number of samples for each class so that the class distribution can be optimally balanced. This procedure can ultimately provide sufficient representation of the minority and majority classes in each group.

We compared the performance of the various number of classifications in CAD predictions on different unbalanced and balanced datasets. We choose SVM as our main model. We comparing the performance of the SVM model with several other methods such as k-NN, naïve bayes, and decision tree without any cross-validation procedures. The next experiment is to implement the resampling feature in the cross-validation procedure by creating synthetic data for the minority class. We used two different oversampling techniques, SMOTE and ADASYN. Unbalanced training data shown in Figure 4, while the balanced training data results are shown in Figures 5(a) and (b).

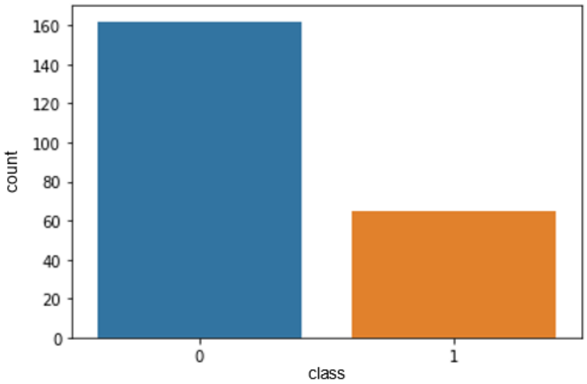
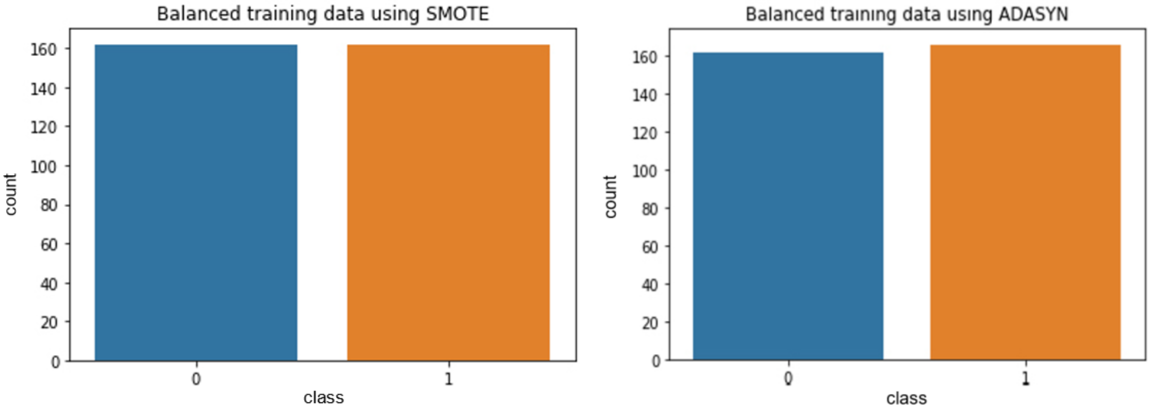


Figure 1. Unbalanced on training data



(a) (b)

Figure 2. The result of feature resampling (a) using SMOTE and (b) using ADASYN

The majority of machine learning algorithms will not produce optimal results if the parameters are not properly specified. In order to build a good classification model, it is very important to select the parameters in a machine learning algorithm. Effective parameter initialization is situation-dependent, and each situation may require unique parameters. By specifying the appropriate parameters, the model can be guided to its optimal solution [31]. Parameter optimization is time consuming if done manually, especially since it has many parameters. The biggest problem in setting up an SVM model is choosing kernel functions and their parameter values [32], [33]. Incorrect parameter settings lead to poor classification results. Therefore, the last step we took was a grid search for hyper-parameters tuning to find the optimal SVM parameter. Hyper-parameters are defined using the minimum value, maximum value, and the number of steps. The parameters we are looking for are shown in Table 2. Machine learning algorithms are trained on imbalanced data using all features and from the feature selection results, as shown in Table 3.

Based on Table 3, SVM shows that it has the highest accuracy in the training and testing phases, so it is superior to all the models we use. The SVM accuracy obtained when using the full feature is better than using feature selection. This is consistent with research [10] that SVM produces better accuracy when using all features (high dimensional data). We see that our base model does a reasonable job of modeling the data. But when viewed, the accuracy value of training and testing has quite a difference. We want the training value to be the same or close to the test value. Therefore, our next experiment uses stratified k-fold to validate the model. Based on Table 4, the average accuracy during the training phase decreases for the whole model. This shows that our model can predict the test data more accurately. Now we just focus only on our main model, the SVM. Next, we will perform feature resampling with SMOTE and ADASYN as an effort to balance the training data during the cross-validation procedure. We also perform model-based predictions on the test set. Based on Table 5, the SVM model provides good performance when using SMOTE. When using feature selection, training accuracy drops while testing accuracy goes up. This makes it possible for our model to more accurately predict test data. To improve the performance of the final model, we take hyperparameter-tuning to find the most optimal parameters in the SVM using features from the feature selection results and SMOTE in stratified k-fold. Based on grid-search, the best SVM model is obtained using the parameters C=1000, degree=1, gamma=0.001, kernel: RBF. By using cross-validation, the highest model accuracy obtained in the test set reached 88%, as shown in Table 6. Based on the confusion matrix in Figure 3, the model shows the number of TP=48 and TN=19. While the ROC curve in Figure 4 shows that the model has good performance because the curve is away from the baseline to the TPR axis. This means that the model classifies more data instances correctly.

Table 2. Grid search for hyper-parameters tuning to find optimal SVM parameters

SVM parameter	Value range
C	0.1, 1, 10, 100, 1000
Gamma	1, 0.1, 0.001, 0.001, 0.0001
Degree	1, 2, 3, 4, 5, 6
Kernel	Linear, Polynomial, RBF, Sigmoid

Table 3. Training and testing score several algorithm on the imbalanced dataset

Model	Full Feature		Feature Feature	
	Training	Testing	Training	Testing
SVM	0.942	0.855	0.925	0.815
k-NN	0.885	0.776	0.881	0.868
Naïve Bayes	0.894	0.828	0.867	0.815
Decision Tree	1.000	0.723	1.000	0.776

Table 4. Average cross-validation score trained on the imbalanced dataset

Model	Full feature	Feature feature
	Cross validation score	
SVM	0.872 ± 0.055	0.855 ± 0.072
k-NN	0.828 ± 0.037	0.842 ± 0.074
Naïve Bayes	0.854 ± 0.042	0.863 ± 0.059
Decision Tree	0.833 ± 0.041	0.797 ± 0.048

Table 5. SVM performance with feature resampling trained on the balanced dataset

	Full Feature			Feature Selection		
	Training	Cross Validation	Testing	Training	Cross Validation	Testing
SMOTE	0.942	0.859 ± 0.038	0.855	0.920	0.868 ± 0.091	0.881
ADASYN	0.942	0.849 ± 0.091	0.881	0.876	0.828 ± 0.105	0.815

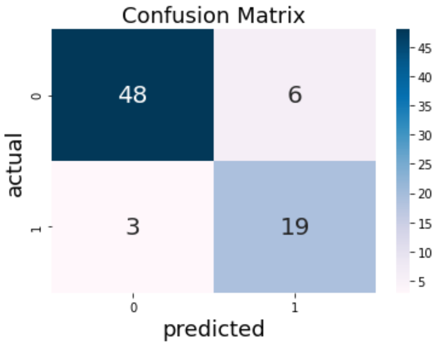


Figure 3. Confusion matrix of SVM balanced dataset with feature selection

Table 6. Hyper-parameter tuning for SVM performance in balanced dataset with feature selection

Training	Cross Validate	Testing
0.925	0.877 ± 0.05	0.881

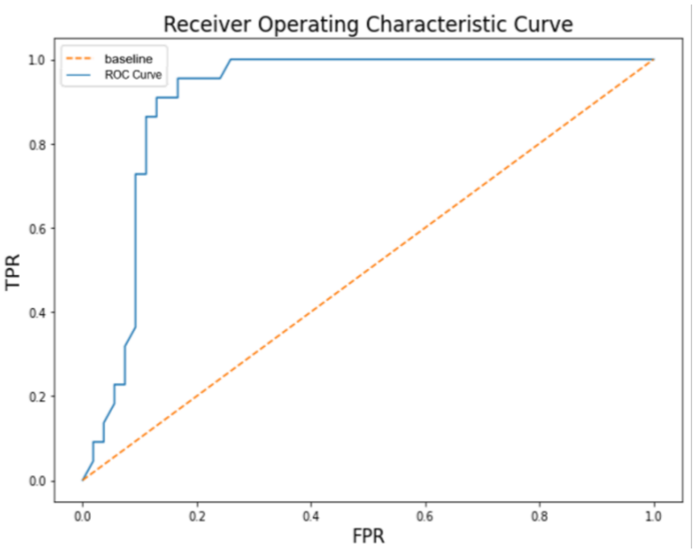


Figure 4. ROC curve of SVM balanced dataset with feature selection

4. CONCLUSION

This study has succeeded in making predictions for CAD using SVM. SVM has better performance than all tested algorithms, such as decision tree, k-NN, and naïve bayes. The main challenge of this research is to deal with unbalanced data and the many features in the dataset. Although the first testers without using feature selection and model resampling features have worked well for modeling the data, there is sufficient distance between the examiners and the training accuracy. Ideally, training and testing accuracy are relatively close. We use extra tree class-based feature selection to generate 16 updated features on the target variable. Our tests using random trees classifier for feature selection and SMOTE for feature resampling show that the best model performance is the testing accuracy of 88%.

REFERENCES

[1] "Heart disease risk factors," *Texas Heart Institute*. https://www.texasheart.org/heart-health/heart-information-center/topics/heart-disease-risk-factors (accessed May 26, 2021).

[2] F. Sanchis-Gomar, C. Perez-Quilis, R. Leischik, and A. Lucia, "Epidemiology of coronary heart disease and acute coronary syndrome," *Annals of Translational Medicine*, vol. 4, no. 13, pp. 256–256, Jul. 2016, doi: 10.21037/atm.2016.06.33.

[3] "Coronary artery disease (CAD)." https://www.cdc.gov/heartdisease/coronary_ad.htm (accessed May 26, 2021).

[4] H. Animesh, K. M. Subrata, G. Amit, M. Arkomita, and A. Mukherje, "Heart disease diagnosis and prediction using machine learning and data mining techniques: a review," *Advances in Computational Sciences and Technology*, vol. 10, no. 7, pp. 2137–2159, 2017.

[5] R. Maheshwari, K. Moudgil, H. Parekh, and R. Sawant, "A machine learning based medical data analytics and visualization research platform," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Mar. 2018, pp. 1–5, doi: 10.1109/ICCTCT.2018.8550953.

[6] L. Muflikhah, N. Widodo, W. F. Mahmudy, and Solimun, "Prediction of liver cancer based on DNA sequence using ensemble method," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Dec. 2020, pp. 37–41, doi: 10.1109/ISRITI51436.2020.9315341.

[7] A. Ridok, N. Widodo, W. F. Mahmudy, and M. Rifa'i, "A hybrid feature selection on AIRS method for identifying breast cancer diseases," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 1, pp. 728–735, Feb. 2021, doi: 10.11591/ijece.v11i1.pp728-735.

[8] S. Sumiati, H. Saragih, T. Abdul Rahman, and A. Triayudi, "Expert system for heart disease based on electrocardiogram data using certainty factor with multiple rule," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 43–50, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp43-50.

[9] S. Sah, "Machine learning: a review of learning types," *Preprints*, Jul. 2020, doi: 10.20944/preprints202007.0230.v1.

[10] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, Dec. 2010, doi: 10.1186/1472-6947-10-16.

[11] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12911-019-1004-8.

[12] A. Ridok, N. Widodo, W. F. Mahmudy, and M. Rifai, "FC-SVM: DNA binding Proteins prediction with average blocks (AB)

descriptors using SVM with FC feature selection," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, Sep. 2019, pp. 22–27, doi: 10.1109/SIET48054.2019.8986070.

[13] 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, Aug. 2018, doi: 10.1016/j.cmpb.2018.05.009.

[14] R. Alizadehsani *et al.*, "Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features," *Research in Cardiovascular Medicine*, vol. 2, no. 3, 2013, doi: 10.5812/cardiovascmed.10888.

[15] R. Alizadehsani *et al.*, "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.

[16] J. H. Joloudari *et al.*, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, Jan. 2020, doi: 10.3390/ijerph17030731.

[17] Savita, G. Sharma, G. Rani, and V. S. Dhaka, "Efficient predictive modelling for classification of coronary artery diseases using machine learning approach," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012068, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012068.

[18] K. E. Heikes, D. M. Eddy, B. Arondekar, and L. Schlessinger, "Diabetes risk calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes," *Diabetes Care*, vol. 31, no. 5, pp. 1040–1045, May 2008, doi: 10.2337/dc07-1150.

[19] K. Shankar, S. K. Lakshmanaprabu, D. Gupta, A. Maseleno, and V. H. C. de Albuquerque, "Optimal feature-based multi-kernel SVM approach for thyroid disease classification," *The Journal of Supercomputing*, vol. 76, no. 2, pp. 1128–1143, Feb. 2020, doi: 10.1007/s11227-018-2469-4.

[20] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14785.

[21] K. Davagdorj, J. S. Lee, V. H. Pham, and K. H. Ryu, "A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention," *Applied Sciences*, vol. 10, no. 9, May 2020, doi: 10.3390/app10093307.

[22] N. Santoso, W. Wibowo, and H. Hikmawati, "Integration of synthetic minority oversampling technique for imbalanced class," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 102–108, Jan. 2019, doi: 10.11591/ijeecs.v13.i1.pp102-108.

[23] K. Davagdorj, J. S. Lee, K. H. Park, P. V. Huy, and K. H. Ryu, "Synthetic oversampling based decision support framework to solve class imbalance problem in smoking cessation program," *International Journal of Applied Science and Engineering*, vol. 17, no. 3, pp. 223–235, 2020.

[24] I. D. Apostolopoulos, "Investigating the synthetic minority class oversampling technique (SMOTE) on an imbalanced cardiovascular disease (CVD) dataset," *International Journal of Engineering Applied Sciences and Technology*, vol. 4, no. 9, pp. 431–434, Jan. 2020, doi: 10.33564/IJEAST.2020.v04i09.058.

[25] C. Taleb, M. Khachab, C. Mokbel, and L. Likforman-Sulem, "A reliable method to predict parkinson's disease stage and progression based on handwriting and re-sampling approaches," in *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, Mar. 2018, pp. 7–12, doi: 10.1109/ASAR.2018.8480209.

[26] A. Ridok, W. F. Mahmudy, and M. Rifai, "An improved artificial immune recognition system with fast correlation based filter (FCBF) for feature selection," in *2017 Fourth International Conference on Image Information Processing (ICIIP)*, Dec. 2017, pp. 1–6, doi: 10.1109/ICIIP.2017.8313761.

[27] A. M. A. and P. A. Thomas, "Comparative review of feature selection and classification modeling," in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, Dec. 2019, pp. 1–9, doi: 10.1109/ICAC347590.2019.9036816.

[28] 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, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[29] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[30] I. C. Dipto, T. Islam, H. M. M. Rahman, and M. A. Rahman, "Comparison of different machine learning algorithms for the prediction of coronary artery disease," *Journal of Data Analysis and Information Processing*, vol. 8, no. 2, pp. 41–68, 2020, doi: 10.4236/jdaip.2020.82003.

[31] G. A. Fanshuri Alfarisy, W. F. Mahmudy, and M. H. Natsir, "Good parameters for PSO in optimizing laying hen diet," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 4, pp. 2419–2432, Aug. 2018, doi: 10.11591/ijece.v8i4.pp2419-2432.

[32] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, no. 4, pp. 1502–1509, Dec. 2016, doi: 10.12928/telkomnika.v14i4.3956.

[33] L. Muflikhah and D. J. Haryanto, "High performance of polynomial kernel at SVM algorithm for sentiment analysis," *Journal of Information Technology and Computer Science*, vol. 3, no. 2, pp. 194–201, Nov. 2018, doi: 10.25126/jitecs.20183260.

BIOGRAPHIES OF AUTHORS

Kuncahyo Setyo Nugroho received a bachelor of computer degree from the Department of Informatics Engineering, Faculty of Engineering, Widyagama University, Indonesia, in 2019. He is currently pursuing a master's degree at the Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Indonesia. He is a member of the Intelligent Systems Research Laboratory, with interest in affective computing. He also has research interests are in machine learning, deep learning, and natural language processing. He can be contacted at email: ksnugroho26@gmail.com or ksnugroho@student.ub.ac.id.

Anantha Yullian Sukmadewa obtained a bachelor’s degree in Educational Informatics Engineering from The Department of Informatics Engineering, Malang State University in 2019. He is currently continuing his master’s studies at The Department of Computer Science, Brawijaya University, Indonesia. He can be contacted at email: ananthayullian@gmail.com or ananthayullian@student.ub.ac.id.

Angga Vidianto completed his bachelor's degree at Department of Informatics Engineering, State Polytechnic of Malang in 2015. He started his career as a web developer and database engineer at a telecommunications company for four years. He is currently continuing his master's studies at Department of Computer Science, University of Brawijaya, Indonesia. He can be contacted at email: angga.vidianto@gmail.com or anggavidianto@student.ub.ac.id.

Wayan Firdaus Mahmudy obtained a Bachelor of Science degree from the Mathematics Department, Brawijaya University in 1995. His Master in Informatics Engineering degree was obtained from the Sepuluh Nopember Institute of Technology, Surabaya in 1999 while a Ph.D. in Manufacturing Engineering was obtained from the University of South Australia in 2014. He is a Professor at Department of Computer Science, Brawijaya University (UB), Indonesia. His research interests include optimization of combinatorial problems and machine learning. He can be contacted at email: wayanfm@ub.ac.id.