

Comparing Classifiers for the Prediction of the Stenosis of Coronary Artery

Hataichanok Sakkara*, Utumporn Maisueb[†], Teerapong Leelanupab[‡]

Faculty of Information Technology

King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand 10520

Email: *hataichanok.skra@gmail.com, [†]aal.maisueb@gmail.com, [‡]teerapong@it.kmitl.ac.th

Abstract—Myocardial Ischemia is the main cause of mortality in patients with Coronary Artery Disease (CAD). One of the methods used in screening patients with this disease is the diagnosis of radionuclide myocardial perfusion imaging (rMPI). In this paper, we conducted a comparative study by experimenting on several machine learning models, such as Logistic Regression, Random Forest, XGBoost, etc., to classify the stenosis of coronary artery. High-level features from rMPI computed by 4D-MSPECT polar map were used to train/test the models. rMPI features of the risk group of CAD patients were obtained from a public hospital. With the hypothesis that patient characteristics (e.g., Diabetes Mellitus, Hypertension, Dyslipidemia) could improve the prediction performance of the models, this study also included patient characteristics in our experimentation as important parts of feature selection. All other processes (i.e., data cleaning, feature selection, feature engineering and feature transformation) in machine learning pipeline were also deliberately experimented in this study. For model selection, two-level validation regarding generalization and hyperparameter tuning were also performed.

Keywords—Radionuclide Myocardial Perfusion Imaging; Stenosis Classification; Coronary Artery Disease, Machine Learning

I. INTRODUCTION

Myocardial Ischemia is a condition when myocardium (heart muscle) does not receive sufficient oxygen and nutrients from blood supply. Basically, Myocardial Ischemia happens because blood flow to heart is obstructed by a partial or complete blockage of coronary arteries by plaques being built up over time. Due to extended ischemia, the death of cardiac myocytes brings about a heart attack also known as a Myocardial Infarction (MI). Myocardial Ischemia and MI are closely related, with the former being an early condition of the latter. MI is a major Coronary Artery Diseases (CAD) and causes mortality globally from the statistics of WHO 2016.

This research paper presents a comparative study by experimenting on seven machine learning models, i.e., Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Adaptive Boosting (AdaBoost), XG Boosting (XGBoost) and Gradient Boosting, for classifying the stenosis of coronary artery. High-level features derived from radionuclide myocardial perfusion imaging (rMPI) were analyzed to train/test the predictive models. In addition to rMPI features, patient characteristics, such as Age, Sex, Body Mass Index (BMI), are also included in our experimentation with an aim to enhance the performance of prediction. All models are finely tuned using cross validation and subsequently tested by unseen samples for performance generalization.

II. RELATED WORKS

A. Myocardial Perfusion Imaging

Methods for screening CAD patients can be categorized into two groups, invasive (e.g., Invasive Coronary Angiography) and non-invasive (e.g., Exercise Electrocardiography, Stress Echocardiography, SPECT and PET.) One of the non-invasive methods is a diagnosis of radionuclide myocardial perfusion imaging (rMPI). rMPI can show the data of three coronary arteries, including left anterior descending coronary artery (LAD), left circumflex coronary artery (LCX) and right coronary artery (RCA). Normal coronary blood flow is about 0.6-0.8 mL/min/g of myocardium. When the narrowing of a coronary artery diameter is less than 50% of the diameter of the vessel, the effect on blood flow generally is diagnosed as being clinically insignificant. The patients with suspected CAD will receive cardiac catheterization with coronary angiography (CAG) to confirm the degree of stenosis.

Commonly, the interpretation of rMPI requires experience of cardiologists based on visual assessment (see the polar map in Fig. 1 (Left)), aka. qualitative analysis [1]. Distinct cardiologists may interpret the same perfusion image differently with respect to myocardial activity as well as cardiac wall motion and thickening. Therefore, qualitative analysis is subjective and suffers from the problem of suboptimal reproducibility. Quantitative analysis [2] is an alternative approach to diagnosing CAD from the numbers or scores calculated by rMPI software, e.g. 4D-MSPECT. Fig. 1 (Right) illustrates, in a red square, the average perfusion severity scores indicating coronary flow reserve of three myocardial regions (LAD, LCX and RCA) as well as their mean average score. Slomka *et al.* [2] reported that quantitative analysis had superior reproducibility with comparable diagnostic and prognostic performance to qualitative analysis. Besides, quantitative analysis enables automated quantification of myocardial perfusion for cardiology practitioners. Nevertheless, there has been no prior research studying quantitative analysis by using 4D-MSPECT polar map severity scores. Therefore, it would be interesting to see whether the perfusion severity scores, as high-level features of rMPI, can be used to classify the stenosis of coronary artery or not, and in particular when they are used to train machine learning model for stenosis prediction.

B. Classification Models

1) *Logistic Regression*: Performing well in linearly separable classes, logistic regression [3] is a classification model that addresses the problem of estimating a probability due to a *logit* function to be outside the range of [0, 1]. The logistic

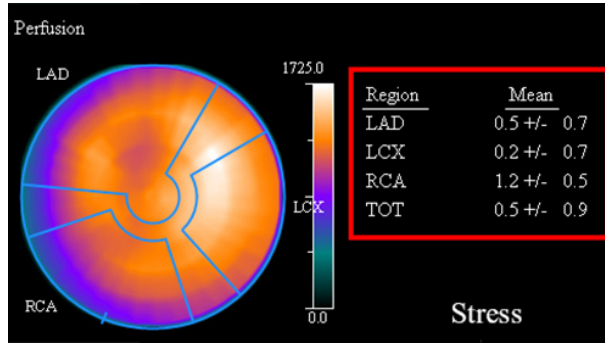


Fig. 1. Post-stress radionuclide myocardial perfusion imaging (rMPI), displayed in polar map (Left) with a gradient color scale indicating a perfusion deficit (purple to black) and severity scores (Right) computed by the software, 4D-MSPECT, with respect to perfusion areas (i.e., basal, mid, and apical), vertical and horizontal axes, and perfusion percentage. The rMPI is segmented by 3-dimensional perfusion distribution according to the regions related to three coronary arteries, i.e., LAD, LCX and RCA.

regression uses a sigmoid function as an inverse form of the logit function to predict the probability that a certain example belongs to a particular class. With its characteristic S-shaped curve, the sigmoid function takes real-number as an input and transforms it into the range [0,1]. This estimate is then mapped to the prediction of the model over a categorical class. The sigmoid function can be computed as follows:

$$h(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (1)$$

where $h(x)$ is the hypothesis function of logistic regression for an example x . β_0 is a constant to shift the S-curve right or left, and β_1 is a slope coefficient.

2) *Support Vector Machine*: Based on geometric principles, Support Vector Machines (SVM) [3] treats certain examples as points in some n -dimensional space. SVM aims to find a hyperplane (decision boundary) that maximizes the margin to separate positive examples from negative ones.

$$\text{Margin}(w) = \frac{|w \cdot x^-| + |w \cdot x^+|}{\|w\|} \quad (2)$$

where $\|w\|$ is the length of the vector w , and x^- and x^+ are the closest negative and positive examples to the hyperplane.

The training examples that are closest to this hyperplane are so-called support vectors. Whereas SVM models with large margins tend to have a lower generalization, models with small margins are more prone to overfitting.

3) *Decision tree*: With an aim to minimize the depth of the tree, a tree induction model [3] finds informative features that can split examples at the parent node into two or more child nodes, which together have minimal impurity of classes in a target variable. By computing information gain, decision tree can create a tree-like structure for prediction. Information gain can be derived from two common criteria, i.e., Entropy and Gini¹.

¹Due to a space limit, we formally define only a Gini score as an example of Information Gain.

$$\text{Gini} = 1 - \sum_{i=1}^n p^2(c_i) \quad (3)$$

where $p(c_i)$ is the probability or percentage of the i -th class c_i in a node, and n is the total number of categorical classes.

4) *Ensemble Learning*: A family of ensemble methods create a meta classifier by aggregating predictions made by multiple classifiers with the goal of improving generalization performance. An ensemble of classifiers can be built from a mix of different types of classification algorithms, such as logistic regression, SVM, decision tree and so on. Alternatively, we can create the ensemble by using the same base classification algorithm, fit by different subsets of a training dataset. These subsets can be generated by randomly selecting either samples or features. This type of ensemble methods is so-called Bagging, one of which that applies the decision tree algorithm on random feature subsets is the Random Forest [4].

The next generation of ensemble is a Boosting approach, which aims to train a series of classifiers by examples that are difficult to classify. This approach is performed in several iterations, by which classifiers in later iterations are learned from error examples misclassified by earlier ones to improve the performance of the ensemble. In this study, we examine three state-of-the-art Boosting algorithms, i.e., AdaBoost [5], Gradient Boosting [6], XGBoost² [7]. Note that these three methods differ mainly in terms of how weights are updated and classifiers are aggregated.

C. Evaluation Metrics

In the domain of machine learning, five common evaluation measures are employed to evaluate predictions in a classification problem. These include Accuracy, Precision, Recall, F1 and AUC of ROC. They all are computed from a confusion matrix when comparing the classifier predictions against real classes. Shortly called AUC, the ROC AUC (ROC area under the curve) summarizes, into a single number, classifier performance based on Receiver Operating Characteristic (ROC) plotted with respect to True Positive Rate (TPR) and False Positive Rate (FPR).

In medical domain, Sensitivity (aka. Recall or TPR) and Specificity (aka. True Negative Rate - TNR) are the two most significant metrics for the clinical test performance. This is because Sensitivity measures the ability of, in our case, the predictive models to correctly identify patients *with* a disease, and Specificity to correctly determine those *without* a disease. Lower Sensitivity means that our tests or the models are not good for catching actual cases of the disease, leaving the patients at risk for their lives. Similarly but slightly less significant regarding life-threatening, lower Specificity means that wrong cases have to receive invasive screening (e.g., CAG) unnecessarily. The trade off between TPR and TNR is explored in ROC as a trade off between TPR and FPR. Accordingly, AUC is the right balance between Sensitivity and Specificity, becoming the most interesting measure, followed by Accuracy, in this study. Formally, Accuracy, Sensitivity and Specificity can be defined as follows³:

²A more advanced algorithm of Gradient Boosting

³TP denotes true-positive; TN stands for true-negative; FP is false-positive; and FN denotes false-negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity = \frac{TN}{FP + TN} \quad (6)$$

III. EXPERIMENTATION

A. Experimental Data

All researchers in this work have been certified⁴ for conducting human research. Access to patient data has been approved by Institutional Review Boards⁵ (IRBs). All data attributes that could identify the patients were anonymized. The data was collected by the Institute Of Cardiovascular Diseases of Rajavithi public hospital between 2018 and 2019. The data was from only rMPI patients, who was suspected by physicians to have CAD and sent to receive cardiac catheterization with CAG. Only rMPI samples with CAG data were used in this study because CAG can ensure whether patients have the narrowing of any coronary arteries or not.

Only 208 examples (male 29.8 % and female 70.2 %) are left for our analysis and to scrap high-level features from rMPI images computed by 4D-MSPECT polar map. Dataset has three subsets of features, i.e., high-level rMPI features, patient characteristics and the three original target variables CAGs. Ten high-level rMPI features⁶ consist of one Level of Ejection Fraction (LVEF) and, belonging to each of the three coronary arteries, three features, i.e., the average perfusion severity scores of 4D-MSPECT, wall thickness and wall motion. Each CAG is a per-vessel target variable, i.e., LADCAG, LCXCAG and RCACAG, associated with each coronary artery as well. As a result from Invasive Coronary Angiography, the CAG value of 1 indicates a certain patient has the stenosis in a particular coronary artery or 0 otherwise. The ratios of positive to negative examples are 50:50 for LADCAG, 34.15:65.85 for LCXCAG, and 39.63:60.37 for RCACAG.

For patient characteristics, there are totally seven features, including Age, Gender, BMI, Diabetes Mellitus (DM), Hypertension (HT), Dyslipidemia (DLP) and Chronic Kidney Disease (CKD).

B. Data Preparation

1) *Data Cleaning*: Raw data contained duplicate cases (i.e., putting records of the same patients) and false indexes (i.e., input errors, non-sequential numbers, null values), manually filled in by a medical technician for anonymization. Accordingly, we removed duplicate cases and re-generate the index of all records.

⁴The certificate of completion or attendance for Good Clinical Practice and Basic Research Ethics.

⁵Please see the Acknowledgment section.

⁶The mean average score of perfusion severity (indicated by TOT in Fig. 1) is reserved in this work.

2) *Feature Transformation*: For the three original target variables, CAGs indicate the actual cases of the stenosis in each coronary artery. We transform the three original CAGs (i.e., LADCAG, LCXCAG and RCACAG) into a new target variable, namely aggCAG (aggregated CAG). That is, when the value of any one of the three original CAGs is 1, we set the value of aggCAG to 1 or 0 otherwise, meaning that a certain patient has the stenosis regardless of which coronary artery has. This aggCAG is then used as a target variable as a per-patient prediction in our experiment. The ratio of positive to negative examples then becomes 59.15:40.85 for the aggCAG.

For a better understanding of data, we computed and visualized the Pearson correlation, using the correlation matrix with heatmap, to see how features are related to each other or to the target variable aggCAG. The visualization illustrates that aggCAG has a little correlation with other features. Thus, making predictions with existing features might lead to poor performance. We therefore assumed that feature engineering could help in reducing noise in features and resulting in more accurate predictions.

3) *Feature Engineering*: For feature engineering, we applied principal component analysis (PCA) [8] on only three features of high-level rMPI feature (i.e., perfusion severity, wall thickness and wall motion), grouped by each coronary artery. That is, the three rMPI features of one coronary artery were transformed by PCA into a single feature, resulting in three PCA features in total. After applying the PCA, we evaluate the correlation of the given features again. The result showed that their correlation with the aggCAG has increased.

4) *Feature Selection*: We were suggested by a cardiologist that, together with high-level rMPI features, patient characteristics (PC) could help in aggCAG predictions. We therefore divided the experimental dataset into four sub-datasets regarding subsets and types of features. The four sub-datasets are:

- D1: original high-level rMPI features plus PC
- D2: PCA features and LVEF plus PC
- D3: only original high-level rMPI features
- D4: only PCA features and LVEF

When comparing datasets D1 against D3, and D2 against D4, one can answer whether PC is a set of good features for predicting aggCAG⁷, or not.

C. Experimental Settings

In this research, we aim to find the best predictive models for classifying the stenosis of coronary artery from high-level features of rMPI. Unfortunately, our dataset is rather small due to a few number of rMPI patients only with CAGs, the small number of available scanners in a public hospital, the limited number of cases that the institute can accept daily for scanning, and the high cost of rMPI and CAG scanning. Consequently, classifiers based on deep learning techniques and neural network are not suitable for the small number of training examples and features in a structured dataset. Note

⁷We also experimented different classifier-independent (or Filter) feature selection strategies. However, the results from those were weaker than those presented here. We believe that our experimenting models, like decision tree and ensemble models, have already included a process of feature selection within their model training.

TABLE I. COMPARE AUC, SENSITIVITY, SPECIFICITY AND ACCURACY OF ALL MODELS, TRAINED AND TESTED BY FOUR SUB-DATASETS FOR THE PREDICTIONS OF CATEGORICAL CLASSES OF AGGCAG.

	AUC				Sensitivity				Specificity				Accuracy			
	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
Logistic Regression	0.76	0.71	0.76	0.74	0.53	0.60	0.60	0.67	0.83	0.83	0.75	0.79	0.72	0.74	0.69	0.74
Decision Tree	0.72	0.74	0.75	0.55	0.60	0.73	0.67	0.60	0.83	0.75	0.83	0.50	0.74	0.74	0.77	0.54
SVM	0.76	0.75	0.72	0.69	0.60	0.73	0.67	0.60	0.92	0.83	0.96	0.71	0.79	0.79	0.85	0.67
Random Forest	0.88	0.85	0.85	0.81	0.67	0.67	0.67	0.73	0.88	0.75	0.83	0.75	0.79	0.72	0.77	0.74
AdaBoost	0.68	0.76	0.68	0.69	0.87	0.73	0.60	0.67	0.50	0.79	0.75	0.71	0.64	0.77	0.69	0.69
XGBoost	0.83	<u>0.80</u>	<u>0.82*</u>	0.74	0.60	<u>0.73</u>	0.80*	0.53	0.88	0.83	<u>0.71*</u>	0.79	0.77	0.79	<u>0.74*</u>	0.69
Gradient Boosting	0.78	0.79	0.85	0.78	0.73	0.80	<u>0.73</u>	0.73	0.79	0.50	0.79	0.75	0.77	0.62	0.77	0.74
Mean	0.74	0.75	0.75	0.69	0.62	0.70	0.62	0.59	0.76	0.75	0.81	0.74	0.71	0.73	0.74	0.68

that this assumption cannot be applied to some tasks of image datasets, in which every pixel is considered training features.

Therefore, we experimented on seven predictive models, i.e., Logistic Regression, Decision Tree, SVM, Random Forest, AdaBoost, XGBoost and Gradient Boosting. For model selection, our study aims to closely control the behavior of predictive models when optimizing for performance and finding the right balance between bias and variance. We thus performed a robust two-level validation for, at the first-level, estimating the generalization performance and for, at the second level, finding the best hyperparameter settings of models.

We opted for a classic strategy, called “Model selection via k -fold cross validation” [9]. That is, in an outer process, we used a two-way holdout sampling by randomly splitting dataset into a training set and a test set with a ratio of 80 to 20, respectively. In an inner process, we used 10-fold cross-validation (i.e., $k=10$) for performance optimization and hyperparameter tuning. Recall that as the most interesting metric of our task was AUC, we defined AUC as a key scorer when optimizing our predictive models. After the inner process was complete, the best hyperparameters of each model were selected to train a specific classifier in the outer process. For model selection and comparison, the predictions of classifiers was validated against unseen test set for an unbiased estimate of the generalization performance.

IV. EVALUATION RESULTS

Table I shown the experimental results in terms of four metrics, i.e., AUC, Sensitivity, Specificity and Accuracy. The results highlighted in bold indicated the best performing models in each column or sub-dataset (D1, D2, D3, or D4.) Italic means the model obtained the highest score in each metric. As we can see, the best performing model in terms of AUC was Random Forest. However, it performed relatively worse than other models regarding the Sensitivity, which is a significant metric indicating that the model lacked an ability to correctly classify positive examples (i.e., actual patients with CAD.)

Therefore, with due care we inspected the results by considering Sensitivity more important than Specificity (but still acceptably high) while regarding AUC and Accuracy as equally important. We accordingly highlighted such four models by underlining them as candidates for possible deployment in production. As indicated by asterisk (*), we eventually selected the XGBoost model that was trained by the sub-dataset D3 as the most appropriate classifier for the prediction of the stenosis. Although the XGBoost model with D3 did not perform the best in terms of AUC, it still performed quite well with AUC 0.82; meanwhile, relatively high Sensitivity 0.80 and acceptable Specificity 0.71. This was because XGBoost usually performs a cross-validation at each iteration of its

boosting process and also has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization that prevents the model from overfitting. Moreover, the results suggested that although including patient characteristics (D1 and D2) did not harm the results of stenosis predictions, it did not improve the performance noticeably either.

V. CONCLUSION

In this paper, we have evaluated the prediction performance of seven machine learning models for the stenosis classification. High-level rMPI features from 4D-MSPECT and patient characteristics have been employed to train and test classifiers, with a purpose to find whether patient characteristics could enhance the predictions or not. PCA has also been applied to reduce noise in the rMPI features, grouped by coronary arteries. We have set up an empirical study using two-level validation for model selection and performance optimization. Results of our study have suggested that the XGBoost trained by only original high-level rMPI features is the most suitable model for predicting the stenosis of coronary arteries.

ACKNOWLEDGMENT

This work was a collaboration between KMITL and Rajavithi hospital. Two IRBs have ethically approved this research with the certificate no. “EC-KMITL_63_035” and “Rajavithi 075/2563”. Many special thanks should also go to Miss Taratip Narawong and Dr. Tarit Taerakul (MD.) for their patient guidance and data collection.

REFERENCES

- [1] F. A. Mettler and M. J. Guiberteau, “Cardiovascular system,” in *Essentials of Nuclear Medicine Imaging*, 6th ed. Elsevier Health Sciences, 2012.
- [2] P. Slomka, Y. Xu, D. Berman, and G. Germano, “Quantitative analysis of perfusion studies: strengths and pitfalls,” *Journal of nuclear cardiology: official publication of the American Society of Nuclear Cardiology*, vol. 19, no. 2, pp. 338–346, April 2012.
- [3] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.
- [4] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [5] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.
- [6] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.
- [7] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [8] I. Jolliffe, “Principal component analysis,” in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Springer Berlin Heidelberg, 2011, pp. 1094–1096.
- [9] S. Raschka, “Model evaluation, model selection, and algorithm selection in machine learning,” *arXiv preprint arXiv:1811.12808*, 2018.