EasyChair Preprint

№ 8298

Coronary Heart Disease Detection Using a Combination of Adaptive Synthetic Sampling Approach and Stacking Method on Imbalanced and Incomplete Dataset

Ahya Radiatul Kamila and Aries Subiantoro

June 18, 2022

Coronary Heart Disease Detection Using a Combination of Adaptive Synthetic Sampling Approach and Stacking Method on Imbalanced and Incomplete Dataset

Ahya Radiatul Kamila[1], Aries Subiantoro[1]

[1] Electrical Engineering Department, Universitas Indonesia, Depok 16424, Indonesia
*Corresponding author e-mail: ahya.radiatul@ui.ac.id

Abstract. *Coronary heart disease is one of the most common cardiovascular diseases that lead to death. Therefore, this study proposes an early detection system for coronary heart disease using Framingham dataset with machine learning approach. The system was developed using stacking method of two Machine Learning algorithms, such as Random Forest and Gradient Boosting. It was observed that Framingham dataset has incomplete and imbalanced data classes. Therefore, KNN algorithm and data balancing method were used to solve the problem of incomplete and imbalanced data classes. Two data balancing methods, known as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling Approach (ADASYN), were compared by evaluating the results of accuracy, precision, recall, and F1-Score. It was discovered that ADASYN with stacking method performed better with accuracy, recall, precision, and f1-score were 90.87%, 89.31%, 92.53%, and 90.89%.*

Keywords: Coronary heart disease, Machine learning, Staking method, ADASYN

I. Introduction

World Health Organization (WHO) stated that cardiovascular disease is a leading cause of 17 million premature deaths in people that are less than 70 years [1]. It is important to note that coronary heart disease is the most common cardiovascular disease, therefore its early detection which requires accuracy and speed of data by medical personnel is very essential to enable patients to take adequate measures. Meanwhile, large amounts of data or unclear data often prolong the diagnosis and increase the possibility of data reading errors. This problem is solvable through machine learning, which helps the medical personnel to provide recommendations for a patient's surgery, diagnose diseases, medical treatment, predict patient's diseases, etc.

Several studies conducted to classify coronary heart disease using machine learning methods encountered problems associated with accuracy levels, imbalanced data classes, incomplete data, or inappropriate machine learning algorithms. This is the reason Senthilkumar Mohan [2] proposed a better approach using a combination of two machine learning algorithms, known as Hybrid Random Forest with Linear Model. Jikuo Wang, *et al* [3] proposed a study using two levels of stacking, where level 1 is considered as the basic and level 2 is regarded as the meta. Kaan Uyar and Ahmet Ilhan [4] proposed a recurrent fuzzy neural networks method with a genetic algorithm, while A. Doganer and M. Kirisci [5] proposed a stacking method with random forest, naive Bayes, and sequential minimal optimization (SMO) algorithms. Another study by Tsatsral Amarbayasgalan, *et al* [6] used deep learning by combining two neural networks, while Lianke Yao, *et al* [7] employed a machine learning approach with feature extraction. Gihun Joo, *et al* [8] compared several machine learning and deep learning methods, such as logistic regression, deep neural networks, random forests, and Light GBM to determine the most suitable algorithm for the Korean National Health Insurance Service–National Health Sample Cohort (KNHSC). Durgadevi Velusamya and Karthikeyan Ramasamy [9] proposed an ensemble learning approach with KNN, Random Forest, and Support Vector Machine. Niculina Mischie and Adriana Albu[10] employed an artificial neural network approach with symptom information and patient laboratory analysis results to assist doctors in diagnosing coronary heart disease. Hossam Meshref [11] compared machine learning and deep learning algorithms, such as Artificial Neural Networks, Support Vector Machines, Naïve Bayes, Decision Trees, and Random Forests.

This present study, therefore, focuses on improving the performance of the model in order to solve the problem of incomplete and imbalanced data classes by proposing the KNN algorithm as an imputer of incomplete data and also comparing two balancing methods, such as SMOTE and ADASYN to obtain the better model performance. The better method was later retrained using a machine learning model according to the stacking method by combining two Machine Learning algorithms, such as Random Forest as a base learner and Gradient Boosting as a meta learner. The final model trained with the selected balancing method showed a significant improvement compared to those without the balancing method. This implied that the study helped to improve the model performance with a balancing method that fits the stack model.

II. Materials and Methods

This research was conducted based on machine learning stages which are broken into 4 main stages, including data entry, exploratory data analysis, data preprocessing, and model Building. The method used was divided into 3 sections with the first focused on the dataset, the second on data pre-processing, and the third on data modeling.

II.1. Dataset Description

Framingham dataset consisting of 15 independent variables and 1 dependent variable from 4238 records of male and female patients in Framingham, Massachusetts.

Table 1: Framingham Dataset

Attribute	Unit	Interpretation	Missing Value	Missing Rate	Mode/Mean
Gender	M/F	Male = 1; Female = 0	0	0%	0/2419
Age	Year	Age at the examination time	0	0%	49 years
Education	Degree	1 : High school 2 : High School or GED 3 : College or Vocational school 4 : College	105	2.40%	1/1720
currentSmoker	Yes/No	Smoker = 1; Non Smoker = 0	0	0%	0/2144
diabetes	mmol/L	Yes = 1; No = 0	29	0.68%	0/4129
totChol	mg/dL	Total cholesterol inside patient's body	53	1.25%	237mg/dL
sysBP	mmHg	Systolic blood preassure	0	0%	132.5mmHg
diasBP	mmHg	Diastolic blood preassure	0	0%	82.9mmHg
cigsPerDay	Piece	Number of cigarettes smoked/day(average)	0	0%	9 cigarettes
BPMeds	mmHg	Is the person on BP medicines	50	1.17%	0/4061
prevalentStroke	kg/m^2	if the person have any prevalent stroke	0	0%	0/4213
prevalentHyp	mmHg	Any beneath prevalent	0	0%	0/2922
BMI	kg/m^2	Body mass index	19	0.44%	25.8 kg/m^2
heartRate	BPM	The number of heartbeats per unit of time	1	0.02%	75BPM
glucose	mmol/L	Amount of glucose in the blood	388	9.15%	82mmol/L
TenYearCHD	Yes/No	Risk of developingCHD(Yes:1 ; No:0)	0	0%	0/3594

II.2 Pre-processing Data

Data pre-processing is a method of improving the performance of the model by converting raw data into an understandable format and ensures that the dataset is noise-free. Five steps proposed in the data pre-processing include data cleaning, normalization, data partitioning, feature reduction, and data balancing. Data cleaning is the initial stage of the pre-processing method consisting of identifying data that are incorrect, incomplete, or irrelevant.

KNN Imputer

It is important to note that these findings focused on solving the problem of incomplete data because there are 645 missing values scattered across several features in the Framingham dataset. The method used among several others to solve this problem is the KNN Imputer algorithm because it uses the average value of the nearest neighbor, known as K-nearest neighbor to fill in the empty values in the dataset, while the calculation of this nearest-neighbor value is based on the Euclidean distance.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

$d(p,q) : Distance$

$q_i : Training\ Data$

$p_i : Testing\ Data$

$n : Number\ of\ Training\ Data$

$i : data\ variables$

Manual input is another method of filling the missing values in the dataset. It uses the average value (mean) or median value (median) for missing value features. Therefore, Exploratory Data Analysis (EDA) is needed to identify patterns, relationships, and data distributions that determine an appropriate imputation method. KNN imputer was proposed in this study to replace the missing values due to the shortcomings of median and mean imputation methods that are likely to cause calculation errors in the standard deviation and variance estimation.

Synthetic Minority Over Sampling (SMOTE):

This is an oversampling method that creates artificial or synthetic data for minor data or fewer objects in order to have a balanced quantity with major data or more objects [13].

Adaptive Synthetic Sampling Approach (ADASYN)

This is an over-sampling approach performed on imbalanced datasets to produce synthetic data for minority data classes based on their distribution. This additional synthetic data generated from minority classes are difficult to learn than the other data [14]. The equation for making synthesis data is as follows:

$$S = x_i \times (x_u - x_i) \times \lambda \quad (2)$$

S = new synthetic data
xi = data minority class included to iteration
xu = randomly selected training data
λ = random number between 0 to 1

II.3 Data Modelling

Stacking Method

Stacking method combines multiple machine learning algorithms with the aim of improving model performance. This method considers training data as input where X_i is the independent variable and Y_i is the dependent variable consisting the number of subjects (p) and features (q). This data is trained by base classifier in order to produce an output in the form of a matrix p x N, where N is the number of base classifier $M_1, \ldots, M_N$.. afterward, the base classifier's output is retrained using the meta classifier. The stacking method of Random Forest Classifier and Gradient Boosting Classifier algorithms was proposed because the two algorithms are family of decision trees.

Table 2: Stacking Algorithm

Stacking Algorithm

Input :

D = Training Dataset $(X_i, Y_i)_{pxq}$

Output: Final Prediction of Ensemble Classifier $(\hat{Y}_{final})$

Step 1: Training data with base classifier

for t in range(T) :

$h_t = \mathcal{L}_1(D)$

end;

#New dataset generated

Step 2 : Training the prediction of base classifier with meta classifier

for i in range(T) :

$z_{it} = h_t(x_i)$

end;

$$\grave{D} = \grave{D} \cup \{\left(\left(z_{it1}, z_{it2} \dots, z_{itN}\right), Y_i\right)\}$$

end;

$$\hat{Y}_{final}) = \mathcal{L}_2(z_{it})$$

#Meta classifier prediction $(\hat{Y}_{final})$

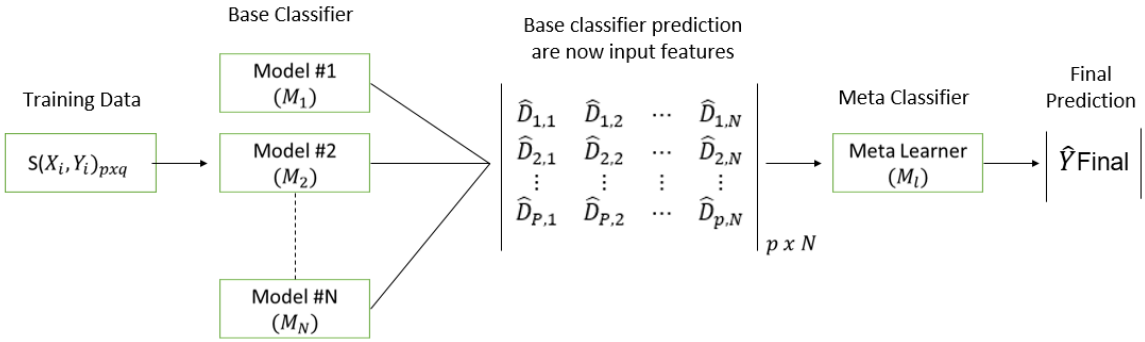

Figure 1: Stacking Model Architecture

III. Results and Discussion

Comparative Results of SMOTE and ADASYN Methods in Data Class Balancing

This method applied after the blank data has been filled using the KNN Imputer and other data preprocessing steps have been completed. This makes the processed data to be clean and correct to enable machine to perform the modeling process correctly. The comparison of these two methods aims to obtain suitable method as a balancer of data classes in coronary heart classification.

Table 3. Results of SMOTE and ADASYN

	Learning Model	Accuracy	Recall	Precision	F1 Score
BASE	Random Forest	84.66%	5.55%	57.89%	10.13%
	Gradient Boosting	83.88%	11.61%	43.39%	18.32%
	Stacking Model	85.06%	9.09%	64.28%	15.92%
SMOTE	Random Forest	78.80%	78.98%	80.13%	79.55%
	Gradient Boosting	77.53%	78.52%	77.56%	78.04%
	Stacking Model	89.63%	87.27%	91.93%	89.54%
ADASYN	Random Forest	79.00%	79.22%	81.23%	80.14%
	Gradient Boosting	77.57%	80.15%	77.62%	78.86%
	Stacking Model	90.87%	89.31%	92.53%	90.89%

The table above shows the results of model performance using three different methods, where the first experiment was conducted by training the data with a model without data balancing methods in order to discover how much the model performance changes before and after using the data balancing method. The second and third experiments utilized SMOTE and ADASYN data balancing methods to determine

which data balancing method is most appropriate for the classification problem of coronary heart disease.

The model performance results showed a better score on the accuracy values using the base method but their recall, precision, and f1 scores are worse compared to the data balancing method. Since recall and precission represent the numbers of false negatives and false positives generated by a model, it become concerns. The values of false negative and false positive are considered in this study because they are likely to be dangerous. For example, false negatives in biomedical cases can stop a patient from receiving proper medical treatment while false positives are likely to cause the patient to undergo unnecessary medical treatment.

IV. Conclusions

In conclusion, the data balanced using ADASYN was retrained using a stack of two previous basic algorithms which include random forest and xgboost to observe the performance of the model with the basic algorithm and stacking method. The heterogeneous stacking method was proposed because it can improve the model performance results compared to basic machine learning algorithms. Moreover, the data were only trained using one algorithm in the basic machine learning, but in the heterogeneous stack model, the data are trained more than once with different machine learning algorithms. Therefore, the data trained with the base classifier produces an output which was further retrained using a different machine learning algorithm in the meta classifier. This simply proved that the stacking method improves the model performance.

References

[1] S. Kaptoge *et al.*, "World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions," *The Lancet Global Health*, vol. 7, no. 10, pp. e1332–e1345, Oct. 2019, doi: 10.1016/S2214-109X(19)30318-3.

[2] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[3] J. Wang *et al.*, "A stacking-based model for non-invasive detection of coronary heart disease," *IEEE Access*, vol. 8, pp. 37124–37133, 2020, doi: 0.1109/ACCESS.2020.2975377.

[4] K. Uyar and A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," in *Procedia Computer Science*, 2017, vol. 120, pp. 588–593. doi: 10.1016/j.procs.2017.11.283.

[5] A. Doganer and M. Kirisci, "CLASSIFICATION OF CORONARY ARTERY DISEASES USING STACKING ENSEMBLE LEARNING METHOD," *THE JOURNAL OF COGNITIVE SYSTEMS*, vol. 5, no. 2, 2020, [Online]. Available: http://dergipark.gov.tr/jcs

[6] T. Amarbayasgalan, V. H. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets," *IEEE Access*, vol. 9, pp. 135210–135223, 2021, doi: 10.1109/ACCESS.2021.3116974.

[7] L. Yao *et al.*, "Enhanced Automated Diagnosis of Coronary Artery Disease Using Features Extracted from QT Interval Time Series and ST-T Waveform," *IEEE Access*, vol. 8, pp. 129510–129524, 2020, doi: 10.1109/ACCESS.2020.3008965.

[8] G. Joo, Y. Song, H. Im, and J. Park, "Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (Nationwide Cohort Data in Korea)," *IEEE Access*, vol. 8, pp. 157643–157653, 2020, doi: 10.1109/ACCESS.2020.3015757.

[9] D. Velusamy and K. Ramasamy, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset," *Computer Methods and Programs in Biomedicine*, vol. 198, Jan. 2021, doi: 10.1016/j.cmpb.2020.105770.

[10] N. Mischie and A. Albu, "Artificial neural networks for diagnosis of coronary heart disease," Oct. 2020. doi: 10.1109/EHB50910.2020.9280271.

[11] "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach Hossam Meshref," 2019. [Online]. Available: www.ijacsa.thesai.org

[12] A. Rufai, U. S., and M. Umar, "Using Artificial Neural Networks to Diagnose Heart Disease," *International Journal of Computer Applications*, vol. 182, no. 19, pp. 1–6, Oct. 2018, doi: 10.5120/ijca201891793