Project report
Using machine learning to predict heart disease

An Phuc Huynh
University of Canberra
u3093019@uni.canberra.edu.au

Quang Loc Hoang
University of Canberra
u3209067 @uni.canberra.edu.au

Tuan Anh Pham
University of Canberra
u3211107 @uni.canberra.edu.au

***Abstract*— Heart disease has put the burden to the economy and people deaths because of this disease are increasing significantly all over the world. This report compares the performance of three models which use Regression logistic, Support Vector Machine and Random forest. The result illustrates that Regression logistic outperforms than the others with accuracy, F1 and Area under the Receiver Operating Characteristic scores.**

I. Introduction

Heart disease has become more popular and contributed to majority cause of death for people in the world. Predicting heart disease diagnoses is crucial when it can provide better treatment and save cost at early stage. Machine learning which is a combination between mathematics and computer science has played an important role in predicting healthcare outcome as data has been increasing dramatically and becomes a rich source for analysis [1].

The challenge of heart disease is in detection and medical devices which used for detecting are not affordable for people. We realise that there are not many self-assessments for heart disease so that people can use to predict if they are diagnosed with it or not. Although these assessments cannot replace medical professional advices, it helps people to determine to seek medical treatment at early stage. Moreover, several algorithms have been used in machine learning and each of them has their own advantages and disadvantages. There is a need to find an appropriate algorithm to be used in prediction of these self-assessments [1].

This report will compare three models which use Logistic regression, Support Vector Machine (SVM) and Random forest to choose the one which outperforms than the others.

II. Literature review

A. Logistic Regression

Logistic regression is a type of regression analysis which is most used in machine learning. It is implemented to measure the probability of a binary event occurring and to solve problems relating to classification. A binary event is defined as there are only two possible outcomes which are event happens or not happen.

The sigmoid function is S shaped curve to fit any practical value of binary outcome between 0 and 1. This function will provide a probability score based on value of probability score and Threshold value. In case probability score above the threshold value, the function returns as True and as False for opposite [2].

B. Support Vector Machine

Support Vector Machine (SVM) has been applied for based diagnostic system in recent years especially for heart valve disease identification. This disease classification is divided into two stage. First of all, sound heart signal is used to classify as if it has a systolic murmur or a diastolic murmur using a two-class SVM classifier. Secondly, once again different two-class SVM classifiers are used to classify the heart sound signals into systolic murmur or diastolic murmur.

SVM was used in previous study for classification of structural brain Magnetic Resonance images. The result of this study showed accuracy score is above 91% after conducting feature selection for cross-validation procedure [3].

C. Random Forest

Random forest is an algorithm which is used in machine learning because of its flexibility, simplicity and better result most of time. Moreover, a random forest is made of several decision trees which are defined that it is suitable for the studies relating to genome-wide, gene-by-environment and gene-by- gene interaction [4].

The random forest operates by choosing the best decision tree prediction and form groups including growing the trees and adding randomness to the model. Random forest shows a good result via searching best feature among a random subset of features to create a better model and prevent of over fitting [5].

D. Model evaluation

1. Accuracy

Accuracy is one of the metrics for evaluating classification models. It is calculated by divide number of correct predictions for total number of predictions [6].

2. F1-score

F1-score is usually used in every binary classification problem where you care more about the positive class. F1 is usually more useful than accuracy in case of an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it is better to look at both Precision and Recall [7].

3. Area under the ROC (AUC)

ROC stands for Receiver Operating Characteristics. A ROC plot is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It shows how many true positive classifications can be gained as allow for more false positives. It is commonly used to visualize the performance of a binary classifier.

AUC stands for areas under curve and it measures the areas under the ROC curve. It quantifies the performance of a classifier to a single number and bounds it between 0 and 1. The closer the AUC is to 1, the better performance the classifier has [6].

III. IMPLEMENTATION STRATEGIES

A. *Reasons for choosing models*

This project aims to find out the most accurate model to predict the heart disease using machine learning. The data is well-labelled and the outcome of the model should be yes/no or positive/negative. Therefore, it is clearly supervised learning as the outcome already categorized. In this project, we will use the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. Therefore, in this project, we will use three classification algorithms:

The first one is Logistic Regressions because of its efficient and straightforward nature. It doesn't require high computation power, easy to implement, easily interpretable. It performs well when the dataset is linearly separable. It not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).

The second algorithm is Support Vector Machines. SVMs offers good accuracy and perform faster prediction compared to Naïve Bayes algorithm. It also uses less memory because it uses a subset of training points in the decision phase. It works well with a clear margin of separation and with high dimensional space.

The final classification algorithm is Random forests. It is considered as a highly accurate and robust method because of the number of decision trees participating in the process. It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases. We can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

B. *Research methodology*

Cleveland database of heart disease was chosen for dataset. We choose the heart disease dataset from UCI Machine Learning Repository. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used for this project. Firstly, this dataset will be cleaning by checking for missing values then conduct Exploratory Data Analysis (EDA).

For data processing: we need to convert some categorical variables into dummy variables and scale all the values before training the models.

With Regression logistic, Support Vector Machine and Random Forest, we will evaluate model with default parameters. Beside that, we use GridSearchCV function to find the best parameters for each model to check if we can get better evaluation metrics.

Model evaluations after applied parameters tunning including accuracy, F1 and AUC scores are used to compare.

IV. EVALUATION AND RESULTS

A. *Exploratory Data Analysis*

1. *Target variable*

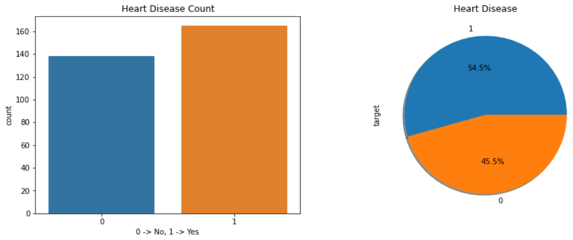

Fig. 1. Heart Disease plot.

The graph shows that more than half of the population suffering from heart disease with percentage of 54.5%. The target data is nearly balanced.

2. *Sex vs. Target*

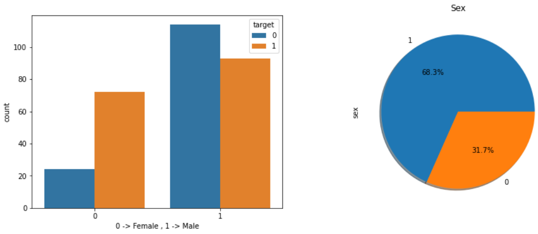

Fig. 2. Sex vs. Target plot.

The graph shows that number of women suffering from heart disease are more than those having no heart disease and vice versa for men. The men population is more than women with percentage of 68.3%.

3. *Chest Pain (cp) vs. Target*

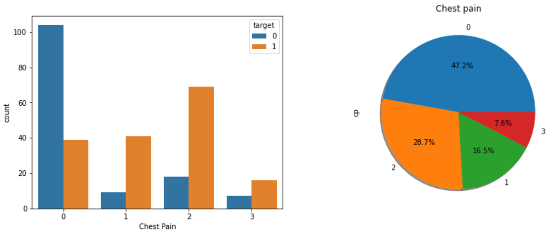

Fig. 3. Chest Pain (cp) vs. Target plot.

There are 4 levels of chest pain with people at level 2 suffering heart disease more than others while people at level 0 having no heart disease more than others. People at level 0 accounts for the biggest proportion with percentage of 47.2%.

4. *Fasting blood sugar (fbs) vs. Target*

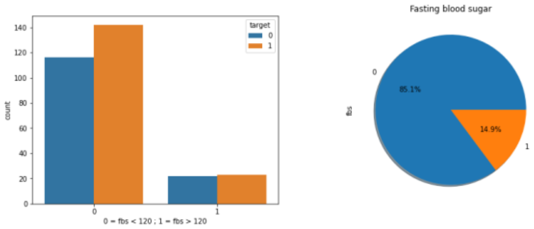

Fig. 4. Fasting blood sugar (fbs) vs. Target plot.

People with fbs lower than 120 more likely to suffer from heart disease. People having fbs lower than 120 accounts for the biggest proportion with percentage of 85.1%.

5. *Resting electrocardiographic results (restecg) vs. Target*

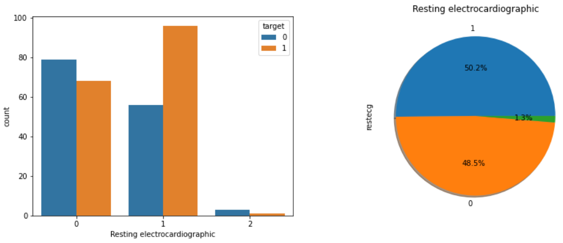

Fig. 5. Resting electrocardiographic results (restecg) vs. Target plot.

People with resting electrocardiographic result 1 is suffering from heart disease more than people with result 0. Most of the people have resting electrocardiographic results as 0 and 1.

6. *Exercise induced angina (exang) vs. Target*

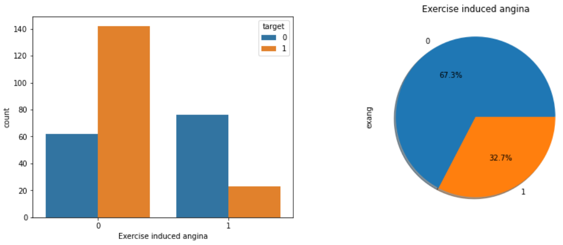

Fig. 6. Exercise induced angina (exang) vs. Target plot.

People without exercise induced angina is suffering from heart disease more than other. People without exercise induced angina accounts for the biggest proportion with percentage of 67.3%.

7. *The slope of the peak exercise ST segment (slope) vs. Target*

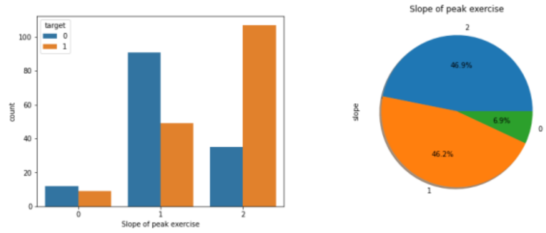

Fig. 7. The slope of the peak exercise ST segment (slope) vs. Target plot.

People with down slope is suffering heart disease more than others. Most people have flat slope and down slope on the peak exercise ST segment.

8. *Number of major vessels colored by flouroscopy (CA) vs. Target*

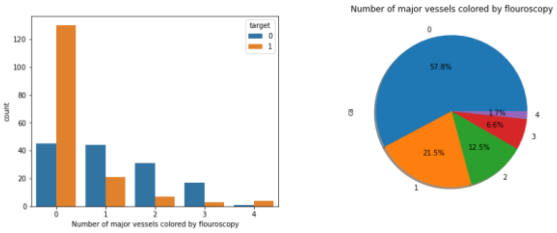

Fig. 8. Number of major vessels colored by flouroscopy (CA) vs. Target plot.

People having 0 major vessels colored by flouroscopy are suffering heart disease more than others. More than half of the population have 0 major vessels colored by flouroscopy with percentage of 57.8%.

9. *Thallium stress result (thal) vs. Target*

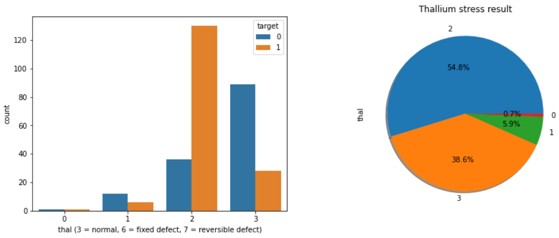

Fig. 9. Thallium stress result (thal) vs. Target plot.

People with thal value 2 is suffering heart disease more than others. Most of the population have thal value 2 with percentage of 54.8%.

10. *Age vs. Target*

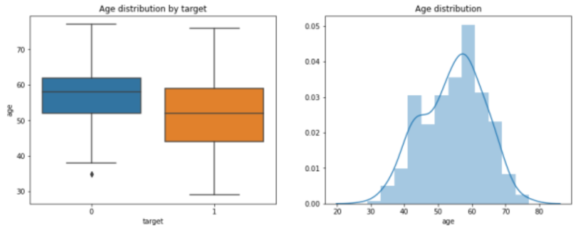

Fig. 10. Age vs. Target plot.

The average age of people suffering heart disease is lower than those of people having no heart disease. The age distribution is normally with most of population having age between 50 and 70.

11. *Resting blood pressure (trestbps) vs. Target*

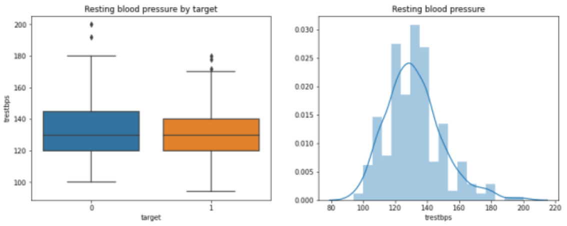

Fig. 11. Resting blood pressure (trestbps) vs. Target plot.

There is no big difference between the trestbps of heart disease and no heart disease. There are several outliers and the resting blood pressure distribution is normal with most of population having trestbps between 120 and 140.

12. *Serum cholestoral (chol) vs. Target*

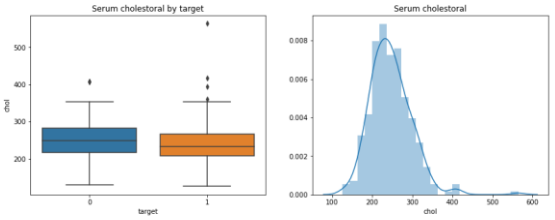

Fig. 12. Serum cholestoral (chol) vs. Target plot.

There is no big difference in the distribution of serum cholestoral between heart disease and no heart disease. There are several outliers. The distribution of serum cholestoral is normal with most of population having chol between 200 and 300.

13. *Maximum heart rate achieved (thalach) vs. Target*

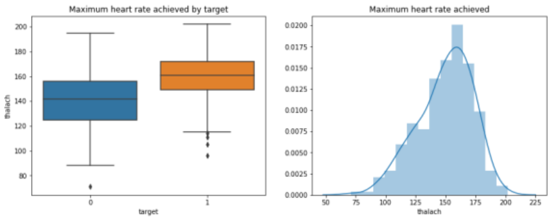

Fig. 13. Maximum heart rate achieved (thalach) vs. Target plot.

The average maximum heart rate achieved of people having heart disease is higher than people having no heart disease. There are several outliers. The distribution of maximum heart rate achieved is normal with most of population having thalach between 150 and 175.

14. *ST depression induced by exercise relative to rest (oldpeak) vs. Target*

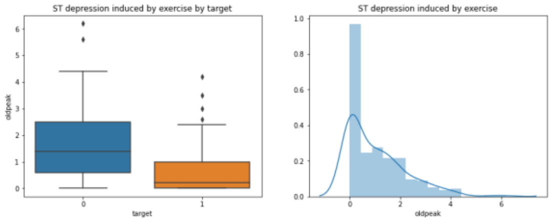

Fig. 14. ST depression induced by exercise relative to rest (oldpeak) vs. Target plot.

The average oldpeak of people having heart disease is lower than those of people having no heart disease. There are several outliers. The distribution of oldpeak is right skewed with most of the population have value 0.

15. *Correlation matrix*

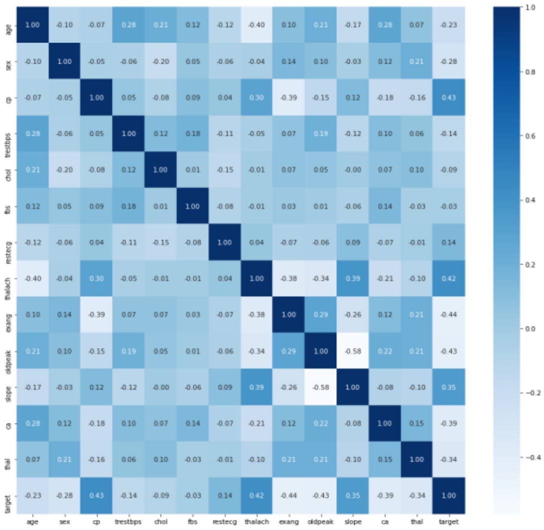

Fig. 15. Correlation matrix plot.

fbs and chol are the lowest correlated with the target variable. Most of the variables are correlated with each other and the target variable.

B. *Logistic Regression*

1. *Default parameters result*

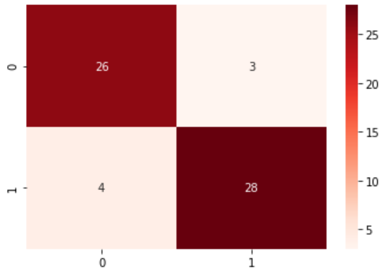

Fig. 16. Confusion matrix plot.

The correlation matrix shows there are problems in only 7 values, otherwise all the values were predicted right.

Accuracy: 0.8852

Precision: 0.8858

Recall: 0.8852

F1-score: 0.8853

AUC-score: 0.8858

2. *Parameters tunning result*

With the best parameter C of 10, the model will be built with the best parameter and calculate evaluation metrics for the best model.

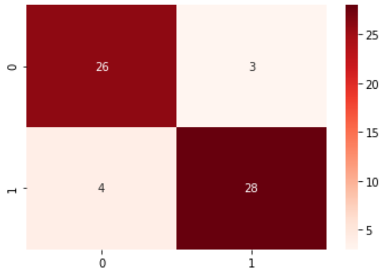

Fig. 17. Confusion matrix for best model plot.

We have the same confusion matrix with the logistic regression model before tuning with 7 problems

Accuracy: 0.885

Precision: 0.8858

Recall: 0.8852

F1-score: 0.8853

AUC-score: 0.8858

We have the same evaluation metrics with the logistic regression model before tuning. The parameter tuning did not improve the logistic regression model.

C. *Support Vector Machine*

1. *Default parameters result*

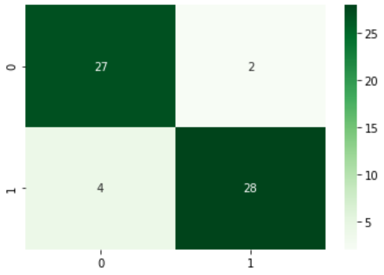

Fig. 18. Confusion matrix plot.

The confusion matrix shows there are problems in only 6 values, otherwise all the values were predicted right.

Accuracy: 0.9016

Precision: 0.9037

Recall: 0.9016

F1-score: 0.9017

AUC-score: 0.9030

2. *Parameters tunning result*

With the best parameters {'C': 2, 'gamma': 0.1, 'kernel': 'rbf'}, the model will be built with the best parameter and calculate evaluation metrics for the best model.

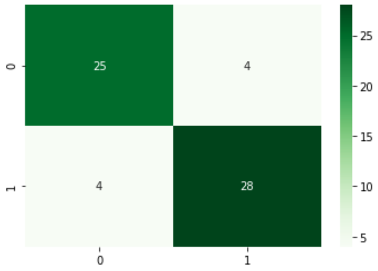

Fig. 19. Confusion matrix for best model plot.

The result shows that there are more problem values in confusion matrix after tuning with 8 problems.

Accuracy: 0.8688

Precision: 0.8688

Recall: 0.8688

F1-score: 0.8688

AUC-score: 0.8685

The evaluation metrics of the support vector machine model after tuning decrease compare to the model before tuning. Thus, the parameter tuning did not improve the support vector machine model.

D. *Random Forests*

1. *Default parameters result*

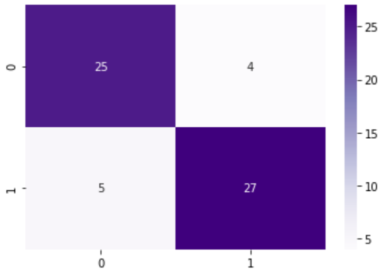

Fig. 19. Confusion matrix plot.

The confusion matrix shows there are problems in 9 values, otherwise all the values were predicted right.

Accuracy: 0.8524

Precision: 0.8530

Recall: 0.8524

F1-score: 0.8525

AUC-score: 0.8529

2. Parameters tunning result

With the best parameter {'criterion': 'gini', 'max_features': 'log2', 'n_estimators': 1000}, the model will be built with the best parameter and calculate evaluation metrics for the best model.

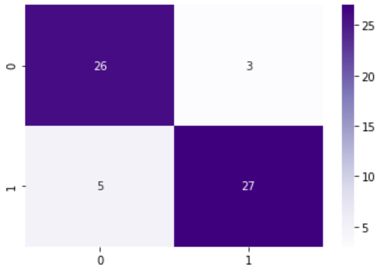

Fig. 20. Confusion matrix for best model plot.

The result shows that there are fewer problem values in the confusion matrix after tuning with 8 problems

Accuracy: 0.8688

Precision: 0.8688

Recall: 0.8688

F1-score: 0.8688

AUC-score: 0.8685

The evaluation metrics of the random forest model after tuning increase compare to the model before tuning. The parameter tuning helped to improve the random forest model.

E. Final result

Based on data when models were built with three algorithms, we chose parameters tunning results to compare evaluation metrics.

TABLE 1. TABLE OF COMPARATION EVALUATION METRICS

Models	**Evaluation metrics**		
	Accuracy	***F1 score***	***AUC score***
Logistic regression	0.885	0.8853	0.8858
Support Vector Machine	0.8688	0.8688	0.8685
Random forest	0.8688	0.8688	0.8685

The data shows that Logistic regression model is the best model for predicting.

V. CONTRIBUTION AND SUGGESTION

A. Contributions

This study compared three different models which are Regression logistic, Support Vector Machine and Random forest and as a result, Regression logistic outperforms than other two for prediction heart disease. This provides a useful tool for people to use easily to do self-assessment so that they can find out if they are diagnosed with disease or not at early stage. Thus, it can save more lives.

This study support guidelines for further research to use same method and models for prediction of other disease.

This study once again emphasizes the importance of applying machine learning in healthcare sector especially in prediction disease.

B. Suggestions

With models in machine learning, dataset is a crucial factor contributing to the success of models especially in accuracy prediction. We suggest in future, there is a need to increase size of dataset like adding more records or combine with other data sources.

Because the limitation of time and resources, we could not compare more algorithms. We suggest in the future, researchers can add more algorithms to compare each other so that they can have a clear picture of which algorithms performs better for heart disease predictions.

REFERENCES

[1] M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning", *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020. Available: 10.1109/iccci48352.2020.9104210 [Accessed 20 October 2020].

[2] S. Ambesange, V. A, S. S, Venkateswaran and Y. B S, "Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques", *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 827-832, 2020. Available: https://ieeexplore-ieee-org.ezproxy.canberra.edu.au/stamp/stamp.jsp?tp=&arnumber=9210404&tag=1. [Accessed 20 October 2020].

[3] I. Maglogiannis, E. Loukis, E. Zafiropoulos and A. Stasis, "Support Vectors Machine-based identification of heart valve diseases using heart sounds", *Computer Methods and Programs in Biomedicine*, vol. 95, no. 1, pp. 47-61, 2009. Available: https://www-sciencedirect-com.ezproxy.canberra.edu.au/science/article/pii/S0169260709000339?via%3Dihub. [Accessed 20 October 2020].

[4] M. Maenner, L. Denlinger, A. Langton, K. Meyers, C. Engelman and H. Skinner, "Detecting gene-by-smoking interactions in a genome-wide association study of early-onset coronary heart disease using random forests", *BMC Proceedings*, vol. 3, no. 7, p. S88, 2009.

Available: https://search-proquest-com.ezproxy.canberra.edu.au/docview/1030083605?rfr_id=info%3Axri%2Fsid%3Aprimo. [Accessed 20 October 2020].

[5] Y. Prawira Putra, D. Khrisne and I. Suyadnya, "Expert System for Early Diagnosis of Heart Disease Using the Random Forest Method", *Journal of Electrical, Electronics and Informatics*, vol. 3, no. 1, p. 15, 2019. Available: https://ojs.unud.ac.id/index.php/JEEI/article/view/46590/29855. [Accessed 20 October 2020].

[6] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997. Available: https://www-sciencedirect-com.ezproxy.canberra.edu.au/science/article/pii/S0031320396001422?via%3Dihub. [Accessed 20 October 2020].

[7] S. Sakr et al., "Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project", *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, pp. 1-15, 2017. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0566-6#citeas. [Accessed 20 October 2020].