SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Semi-Supervised Learning: Assisted Cardiovascular Disease Forecasting using Self-Learning Approaches

Ekramul Haque Tusher[1], Mohd Arfian Ismail[1,2,*], Ferose Khan[3], Anis Farihan Mat Raffei[1], Jalal Uddin Md Akbar[1]

[1] Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia
[2] Center of Excellence for Artificial Intelligence & Data Science, Universiti Malaysia Pahang Al-Sultan Abdullah, 26300 Gambang, Kuantan, Pahang, Malaysia
[3] Department of Computer Science and Engineering, International Islamic University Chittagong, Kumira 4318, Bangladesh

ABSTRACT

Cardiovascular diseases (CVDs) are characteristics that affect both the heart and the blood vessels. This disease is the main factor contributing to the greatest number of deaths globally. In the present global context, it is very difficult to detect cardiovascular diseases by early-stage symptoms. If this isn't diagnosed early, it could lead to death. In order to improve the accuracy of the CVD prediction system, a wide variety of supervised and unsupervised learning approaches from the fields of machine learning were used. Only labelled data is used in supervised learning systems to create a classification model but acquiring sufficient amounts of labelled data takes time and typically requires the cooperation of field experts. However, unlabelled samples are readily available in a variety of real-world situations. More effectively than any other machine learning approach, semi-supervised learning (SSL) addresses this problem by integrating quantities of labelled and unlabelled data to improve the classification model. In this work, we propose semi-supervised learning approaches based on self-learning with Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest (RB). According to the comparison's findings, SVM has a high classification accuracy rate of 94.68%, a recall rate of 94.41%, a sensitivity rate of 94.49%, a F1 score rate of 92.99%, a precision rate of 91.59% a low, a balanced accuracy rate 94%, a G-mean rate of 94.45 and low Error-rate 5.32%. The model may be used to forecast cardiovascular disorders in the medical profession.

Keywords:

Cardiovascular disease; Naïve Bayes; Semi-supervised learning; Random Forest; Support vector machine

1. Introduction

The human heart, as it is the organ responsible for pumping blood to the rest of our body, is considered to be the most vital component of the human body. The heart, the blood vessels that carry blood throughout the human body make up the circulatory system. The collective term for conditions that arise from problems with the cardiovascular system is cardiovascular disease. Approximately, 18 million people all over the world have passed away as a result of cardiac disorders

* *Corresponding author.*
E-mail address: arfian@ump.edu.my

https://doi.org/10.37934/araset.56.1.136150

reported by the World Health Organization (WHO). This accounts for an estimated 32 % of all deaths that have occurred around the world [1]. Currently, cardiovascular diseases (CVDs) are the leading cause of mortality worldwide. The term "cardiovascular diseases" relates to a collection of diseases that can impair a person's heart & blood vessels, eventually lead to organ failure. Heart disease represents one of the major causes of morbidity and mortality around the world. Because of this, people often consider cardiovascular diseases to be a major source of concern for their health. It is characterized by a wide range of symptoms, some of which include chest pain, stroke, a rapid heart rate, trouble breathing, and disorientation [2]. Despite this, cardiovascular diseases are quite prevalent among individuals of all ages, including the younger generation, mostly as a result of their unhealthy lifestyles. The prediction of cardiovascular disease is often regarded as one of the most important and contentious issues in the field of healthcare informatics. Therefore, in order to avoid being affected by cardiovascular disease need to examine the root factors and symptoms.

Although semi-supervised learning techniques have showed potential in classification problems, the question of how to efficiently use the objectifying encoded in labelled and unlabelled training data remains undetermined. Additionally, semi-supervised learning is still something that has to be improved on for CVD. In this research, provide a novel model with semi-supervised learning that utilizes self-training using Random Forest, Support Vector Machine, Naïve Bayes. In order to solve, even in part, the difficulty of accurately predicting the breakdown of CVD. Due to a lack of expertise resources, it is hard and expensive to get labelled training to predict CVD. On the other side, there are a ton of unlabelled images that are simple to get from public picture repositories. When dealing with issues with a small sample size, supervised learning performs poorly since there isn't much single or multivariate variation, and the differences in unlabelled data can't be used well [3]. Research on semi-supervised learning has also been carried out in great depth to make full use of unlabelled examples. This research has been carried out to propose a variety of novice semi-supervised learning models with self-learning to forecast CVD with SVM, NB and RF. Improving the security of patient data and increasing the classification accuracy.

The work that is relevant to this topic in section 2. The technique that was used in our research is outlined in Section 3. Also, Section 4 contains the analysis and discussion. In Section 5, conclusion and suggestions are made for future research.

2. Related Works

2.1 Semi-Supervised Learning

The approach of SSL has been widely used in machine learning. It is viewed as a middle ground between the approaches to supervised learning and unsupervised learning, which enables it to make use of the benefits provided by both types of learning methods. H. J. Scuder was the one who first presented the idea of SSL in 1965 [4]. J.A. Patwary *et al.,* [5] SSL is a technology that consists of two steps. A limited amount of labelled data is used to build an initial classifier in the first step. And this classifier is used in the second step to annotate a large quantity of unlabelled data. The original starting classifier is then updated with the newly labelled data, and the model is retrained so far as the first accuracy is increased, or the model satisfies some predetermined stopping criterion. Ashfaq *et al.,* SSL approaches give labels by taking into consideration unlabelled instances in addition to labelled data, which allows for the construction of a more accurate classifier [6]. Researchers have shown a great deal of interest in a variety of SSL approaches, including self-training, co-training, expectation maximization (EM), transudative support vector machines (TSVMs) using generated mixture models and graph-based methods.

2.2 Existing Works on CVD Forecast

Rachael *et al.,* [7] examined the use of three different machine learning approaches to diagnose cardiovascular diseases, paying particular attention to the range of hyper-parameters associated with each methodology. The approaches used in this study were SVM, decision trees, and multi-layer perceptron neural networks. They emphasize how important it is to conduct 10 folds cross validation in order to assess the ambiguity of trained models that are created from two datasets that have different features, demonstrating an accuracy that varies from 0.21 to 0.92 depending on the settings that were applied to the SVM model.

Barbara *et al.,* [8] in order to prevent premature deaths, those who are more likely to acquire CVD must be identified early. To be more precise, it intends to design a system that is capable of predicting the presence or absence of cardiovascular diseases via the use of data mining in order to fulfil the urgent demand of extracting essential information buried in clinical data. As a result, the period between admission to hospital and diagnosis will be minimized, and the patient will be able to get quick and sufficient treatment.

In Shilaskar *et al.,* [9] work, the existence of CVD should be able to be predicted with greater precision using a smaller set of characteristics, which is the purpose of this research. They investigated developing an intelligent system to build a feature subset that would increase diagnostic performance. In order to discover a subset of features that provides a better classification result, features are rated using a distance measure and then searched using forward selection, forward inclusion, backward elimination search strategies. For the diagnosis of cardiovascular disease, they suggested using a hybrid forward selection method. Comparing this method to back-elimination and forward inclusion methods, their experiment shows that it identifies smaller subsets and improves diagnostic accuracy.

MD Samiul Islam *et al.,* [10] developed a method to detect risk factors of CVD by using the attention module-based Long Short-Term Memory, which has around the accuracy of 94% with 0.90 Matthews Correlation Coefficient (MCC) scores, in contrast to other previously published approaches. Additionally, they recommended creating a one-of-a-kind smart health care platform that would enable ongoing patient monitoring and data collecting. Initially, the proposed platform is utilized to collect data, and they determine the dataset's best-suited characteristics for using different machine learning techniques. The experimented result demonstrated that the attention module-based LSTM performs better than the existing statistical machine learning algorithms when it comes to the prediction and identification of CVD risk factors. This may encourage patients with CVD to make changes to their lifestyle.

Study by Najmul *et al.,* [11] find out how well different feature selection techniques can predict cardiovascular disease. They used a brand-new multi-stage method for selection of features and comparison analysis in classification problems to predict CVD. As mentioned earlier, predicting CVD is an essential part of medical informatics; despite this, only a tiny number of studies depend on traditional prediction techniques as opposed to innovative data mining methodologies. According to the findings of this research, there is a fresh option available to medical data scientists for the forecasting of CVD. Their study presents a novel method for CVD prediction by medical data scientists.

Patro *et al.,* [12] study proposed a framework that system helps to predict the heart disease based on key risk indicators and different classifier designs, including KNN, NB,SVN, Lasso and ridge regression methods. Principal component analysis (PCA) and linear discriminant analysis were also used in addition to these data classifications. The rate of F1 accuracy is 85%, whereas the accuracy

of the SVM classifier is 92%. Using precision, sensitivity and accuracy, the achievement of the suggested research task is evaluated.

Maiga *et al.,* [13] the authors used machine learning models to predict CVD then compare using risk factors such body mass index (BMI), cholesterol levels, blood sugar levels, systolic and diastolic blood pressure, in addition to a number of other risk factors. The algorithms utilized were KNN, NB, LR, RF. According to the results of the research, the RF model, is the best choice since it has a high level of accuracy (73%) as well as a sensitivity (65%) and a specificity (80%).

Alfaidi *et al.,* [14] Use machine learning methods, a model was created for predicting the likelihood that a person will suffer from CVD. Seven distinct algorithms were used in the execution of the tests, and a dataset of cardiovascular diseases that is freely accessible to the public was used to train the models. A Chi-square test was utilized to figure out which factors were most important for predicting cardiovascular disease. The experimented result demonstrated that Multi-Layer Perceptron is the best at a disease predicting, with 87.23% accuracy.

In Pooja Rani *et al.,* [15] work, a decision system was developed and used machine learning techniques to predict CVD based on the medical data of a patient. According to their results, the RF model's accuracy was the highest at 86.60%.

Alalawi *et al.,* [16] in their research, heart disease was diagnosed using CVD and heart disease datasets and a variety of machine learning algorithms, including SVM, ANN, LR, DT, RF, KNN, VC, NB and GB. The performance of each of the models was experimented based on their respective levels of accuracy, recall, precision, and f-score. As a consequence of this, the RF model had the best performance in the dataset for cardiac disease, achieving an accuracy of 94%, while the GB model had the best result in the dataset for cardiovascular diseases, obtaining an accuracy of 73%, a 73% F1-score, a 73% Recall, and a 74% Precision. Both of these models had the highest performance in the dataset.

Sabab *et al.,* [17] proposed a method that applies feature selection strategies to increase the precision of the classification techniques and techniques for data mining to maximize the detection of cardiac diseases. The accuracy of the SVM, NB classifiers and C4.5 DT used by the authors to identify cardiovascular disease was 87.8%, 86.80%, and 79.9%, respectively.

According to the findings of such research, a prediction model that is powered by data mining and machine learning techniques will unquestionably be able to handle the worrying rise in the prevalence of cardiovascular diseases. The majority of emphasis has been focused on machine learning that operates independently. Within the scope of this study, a comparison is made between the performance of semi-supervised learning models depending on the accuracy of the model's parameters.

3. Methodology

3.1 Support Vector Machine

SVM is applied for regression and classification issues, and it generates remarkable outcomes across a variety of fields. The goal of the SVM classifier is to achieve the greatest feasible margin of separation between the target classes and the other classes. Different kernel functions deal with various data types of linear kernel, polynomial kernel, Radial Basis Function (RBF) [18]. SVM is used to differentiate each, and every piece of data that is part of the n-dimensional feature set. Support vectors have been made up of the class data points that are closest to the hyperplane [19]. It is possible to construct a hyperplane using the equation that is presented below:

$$P_0 : mT\, n + c = 0 \quad (1)$$

As shown in the following equations, two more hyperplanes, P_1 and P_2, are formed in parallel with the existing hyperplane:

$$P_1 : mT\, n + c = -1, \quad (2)$$

$$P_2 : mT\, n + c = 1 \quad (3)$$

The equations that follow should be satisfied by the hyperplane in order to fulfil the requirements for each input vector Ij.

$$mIj + c \geq +1 \text{ for } Ij \text{ having class 1}, \quad (4)$$

$$mIj + c \geq -1 \text{ for } Ij \text{ having class 0.} \quad (5)$$

3.2 Naïve Bayes

The NB method has become one of the most well-known classifiers, and it is used in the process of classifying this disease. The NB classifier is one example of a statistical tool that may provide probabilistic predictions and forecasts about class membership [20].The Bayes Theorem provides the foundation for the NB classifier, which is a statistical classifier that makes independent assumptions about the relationships between the predictors [21]. Bayes' Theorem states that:

$$F(h \setminus x) = (F(x \setminus h) * F(h))/(F(x)) \quad (6)$$

where,
F(x) = Prior probability of x.
F(h) = Prior probability of h.
F(h\ x) = Posterior probability of h condition on x.
P(x\h) = Posterior probability of x condition on h.
Working procedure of Naive Bayes classifier:

Let C represent the training set linked to the class labels. An element vector with n dimensions is used to represent each tuple,

$$Y = (y_1, y_2, y_3....., y_n) \quad (7)$$

Suppose there are m classes A_1, A_2, A_3....A_m. Let's assume we wish to categorize tuple Y, which is unknown. Thereafter, conditioned on Y, the classifier will forecast that Y corresponds to the class with a greater posterior probability. i.e., the Naive Bayesian classifier allocates an unidentified tuple Y to the class Ai, if and only if $Q(Ai|Y) > Q(Aj|Y)$ for $1 \leq j \leq m$, and $I \neq j$. Bayes' Theorem is used to figure out the above posterior probabilities.

3.3 Random Forest

The RF classifier is one that takes the average of a number of decision trees which have been applied to various subsets of a given dataset. The goal of this classifier is to improve the accuracy of predictions made using the given dataset. The increasing number of trees within the forest increases

precision and eliminates the issue of overfitting. It is able to handle big datasets with lots of dimensions. A RF is a collection of n different decision trees. Every decision tree within the forest is trained using various subset of the training set, which were created using bagging to create the original labelled data. While the tree is developing, Random Forest utilizes randomized feature selection. This property is absolutely essential when it comes to multidimensional datasets [22]. This is due to the fact that when there are hundreds or thousands of features, such as in medical diagnosis as well as documents, many features that are only marginally relevant might not appear in such a single decision tree at all. The majority voting procedure among the trees is used in RF to obtain the final hypothesis [23]. Demonstrate of Random Forest procedure:

The RF's Generalization Error (TE*) is described as,

$$E^* = T_{a,b}\,(ng(A,B)) < 0 \tag{8}$$

Margin function is present when ng (A, B). The Margin function calculates how much the average vote at (A, B) for the correct class is more than the average vote for each other class. A is the predictor vector in this instance, while B is the classification.

The formula for the Margin function is,

$$ng\,(A,B) = avk\,I(bk\,(A) = B) - \max j \neq B\; avk\,I(bk\,(A) = j) \tag{9}$$

'I' stands for the indicator function here. Confidence in the categorization is inversely correlated with margin. The predicted value of the Margin function is used as a measure of Random Forest's strength as,

$$Z = E\,A,\,B\,(ng\,(A,\,B)) \tag{10}$$

An ensemble classifier's generalization error is constrained above with a function of the mean correlation between its base classifiers with their average strength (z). where 't' is the average correlation value, an upper limit on the generalization error is provided by:

$$TE^* \leq t\,(1 - z_2) / z_2 \tag{11}$$

3.4 Self-Learning

Self-learning is probably the earliest type of semi-supervised learning approach. Self-learning is a popular approach of semi-supervised learning in several domains, including object detection, Natural Language Processing and identification. Self- learning is an excellent learning algorithm to use when there is only a small amount of labelled data and most of the data is unlabelled [24]. A given supervised classifier may use this class to serve as a semi-supervised classifier, enabling it to learn from data that is unlabelled [25]. In order to do this, it predicts pseudo-labels for the data that are unlabelled and then adds those predictions to the training set. The classifier may continue to iterate until either the maximum number of allowed iterations has been achieved or until the most recent iteration did not add any pseudo-labels to the training set. It may also determine the rankings of the instances based on the confidence of their predictions and adds most confident examples to the labelled dataset [26]. The initial classifier is then retrained using the larger training dataset. This process is done a predefined number of times or until a certain predetermined heuristic criterion is achieved. When every classification is able to correctly identify the data which is given unlabelled

data, the accuracy of the classification would only improve [27]. In real-world applications, higher accurate and confident criteria as well as more accuracy-confidence assessments are employed to avoid mislabelling of the presented data [28]. Following is a summary of the self-training algorithm:

i. Develop the predictor, g using the labelled training data (xl, yl).
ii. Make a prediction using data that has not been labelled, x > xu.
iii. Include (x, g(x)) in your labelled data.
iv. Continue iterating until the first predictor's accuracy of predictions is satisfied.

3.5 Proposed Model

Expanded the model and proposed a novel approach in favour of predicting CVD by using semi-supervised learning by employing self-learning based SVM, RF, and NB. This autonomous self-learning allows the models to take advantage of abundant unlabelled data to uncover hidden relationships and patterns associated with increased cardiovascular risk. The self-training iterates in an automated, unsupervised manner until specified stopping criteria are met - ultimately yielding enhanced risk stratification capabilities compared to supervised methods. In Figure 1, presents an overview of the strategy that has suggested together with its corresponding diagram.

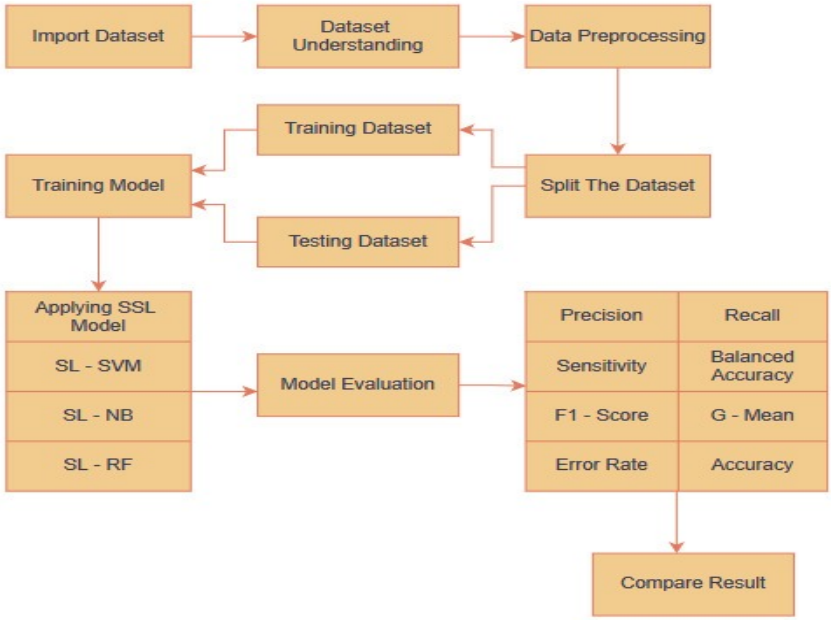

Fig. 1. Proposed Methodology

4. Results Analysis

4.1 Environment

Python 3 language was used to develop the suggested method, and the tests were carried out on a personal computer, running Windows 10 (64-bit operating system) and equipped with an Intel Core i5-4590 central processor unit operating at 3.30 gigahertz and 12 gigabytes of random-access memory (RAM).

4.2 Details of Dataset

Results The CVD dataset from the Kaggle. The cardiovascular disease dataset on Kaggle provides a binary classification task which records cardiovascular disease either being present or not. This dataset contains 70000 records; however, each record only contains 11 features [29]. Additionally, these characteristics can be separated into three categories:

i. Subjective: Patient self-reported information, such as whether or not they smoke.
ii. Objective: information that is factual, like gender.
iii. Examination: values are obtained from clinical examinations, like the weight.

For the interest of clarification, we would like to note out that the Kaggle repository does, in fact, include a variety of datasets that are associated with heart disease and that specifically selected a single data set with thousands of measurements. The fields that are included in each record are stated in Table 1, along with each data type.

Table 1
Explanation of the Features of the Cardiovascular Dataset on Kaggle

Feature	Type	Detail
Age	Objective	Age in days (int)
Gender	Objective	Categorical code
Weight	Objective	Weight in kg (float)
Height	Objective	Height in cm (int)
Diastolic blood pressure	Examination	int
Systolic blood pressure	Examination	int
Blood glucose	Examination	1: normal, 2: above normal, 3: well above normal
Cholesterol	Examination	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective	Binary
Physical activity	Subjective	Binary
Alcohol intake	Subjective	Binary
Cardiovascular disease	Target variable	Binary

The three groups for blood glucose and cholesterol readings were swapped out for values in range [0, 1] and the binary data, including the categorization itself, were all subjected to a single hot encoding. The range [0, 1] was scaled for the remaining numerical variables. Class distribution CVDs are 49% and those without CVDs are 51%. CVDs as a class distribution 49% and not having CVDs as a class distribution 51%. This is a balanced dataset. Due to the small difference between the 49% of people without CVDs and the 51% of people with CVDs, this dataset looks to have a balanced class distribution.

4.3 Data Preprocessing

Firstly, data was required to be integrated, and then the process of data cleaning had to be applied. In more depth, the presence of missing values, duplicated data, outliers, and inconsistencies were investigated. This was discovered that there weren't any missing values, duplicated data, or inconsistencies.

After that, this was the time to move on to the process of transforming the data. Due to the fact that the value of the attribute age has been determined in terms of days, it was transformed into years using the equation "age/365." However, active, alcohol use, cardiovascular activity, and

smoking are examples of the binominal type of characteristics, but cholesterol is an example of the polynomial type. By default, the majority of attributes had been imported as Integers. Therefore, these properties were changed in accordance. Before doing further analysis, implemented data pre-processing methods that accounted for missing values and data outliers. In the beginning, the categorical data were verified and then transformed into numerical form. In target class, the number "1" denotes the "presence" of cardiovascular disease, while the number "0" represents its "absence". In a ratio of 80:20, the training samples make up 80% of the samples, while the testing samples make up the remaining 20%.

4.4 Performance Measurement

A confusion matrix, often known as a contingency table, is a visualization technique used to show how well a classification system performs. The confusion matrix is comprised of the following elements: True Positives (TP), True Negatives (TN), False Positives (FP), and False Positives (FN). The term "true positive" refers to the proportion of accurate forecasts that an instance is true, while the term "false positive" refers to the proportion of inaccurate predictions that an instance is true. The term "true negative" refers to the proportion of accurate guesses that an instance is false, whereas the term "false negative" refers to the proportion of inaccurate guesses:

i. Accuracy: Accuracy is one factor to consider when rating categorization models. Accuracy is the proportion of forecasts that model predicted successfully. For binary classification, accuracy can also be assessed in terms of positives and negatives, as shown below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

ii. Precision: A classification system's performance can also be evaluated using precision. It is calculated as the ratio of true positives to the sum of true and false positives for each class. It represents the actual positive cases among all the optimistic predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{13}$$

iii. Recall: Recall is a metric used to show the percentage of examples that the model was able to properly recognise out of all the potential positive labels. It is the proportion of genuine positives to the total of true positives and false negatives.

$$\text{Recall} = \text{TPR} = \frac{TP}{TP+FN} \tag{14}$$

iv. Selectivity: Selectivity is a measurement of the percentage of the negative group that was accurately predicted to be negative, as shown below:

$$\text{Selectivity} = \text{TNR} = \frac{TN}{TN+FP} \tag{15}$$

v. F1-score: The accuracy statistic counts the number of times a model correctly predicted the entire dataset, as shown below:

$$\text{F1–score} = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (16)$$

vi. G-mean: The Geometric Mean (G-Mean) calculates the balance of categorization results for both majority and minority classes, as shown below:

$$\text{G-Mean} = \sqrt{(TPR \times TNR)} \quad (17)$$

vii. Balanced Accuracy: Balanced accuracy is one measure of a binary classifier's effectiveness. It is quite useful when the classes are not evenly distributed, as shown below:

$$\text{Balanced Accuracy} = \frac{1}{2}(TPR + TNR) \quad (18)$$

viii. Error Rate: The "error rate" describes a calculation of the size of a model's prediction error in relation to the real model, as shown below:

$$\text{Error Rate} = 1 - Accuracy \quad (19)$$

Class 0 patients are those without cardiovascular diseases, and Class 1 patients are those with cardiovascular diseases. This is a balanced dataset because of the small difference between the 49% of people without CVDs and the 51% of people with CVDs.

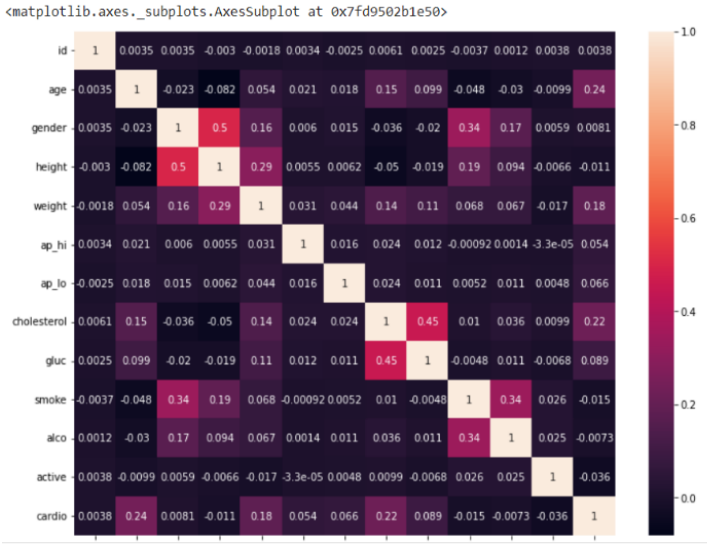

Fig. 2. Correlation of the feature

By providing information about the relationship between variables, correlation analysis serves a crucial function in statistics. It assesses how closely changes in one variable are related to those in another. A correlation is stronger when its value is closer to 1, while a correlation is weaker when its value is closer to 0. The correlation coefficient's magnitude shows the relationship's strength.

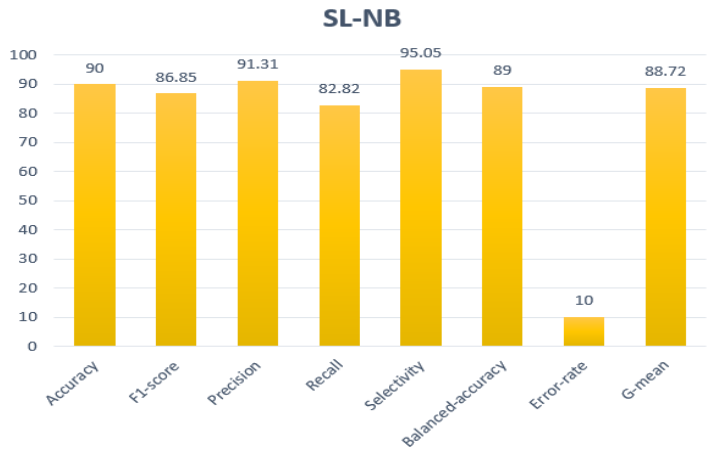

Fig. 3. Performance of SL-NB

The SL-NB model performs well in terms of precision, recall, selectivity, balanced accuracy, and the G-mean and displays a high accuracy. These metrics show how well the model performs in terms of making accurate predictions and reducing false positives and false negatives.

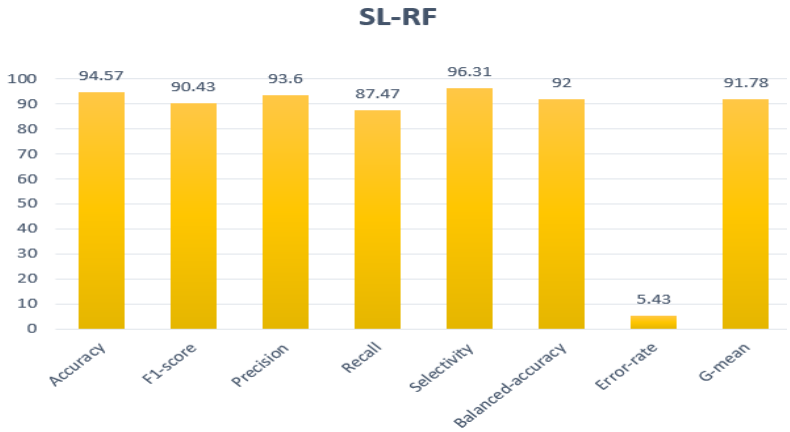

Fig. 4. Performance of SL-RF

The SL-RF model performs excellently overall, excelling in the G-mean, recall, selectivity, balanced accuracy, and accuracy across all categories. It's accuracy, G-mean, recall, selectivity, balanced accuracy, and precision better than SL-NB and error-rate is less than SL-NB.

Fig. 5. Performance of SL-SVM

Based on the results from the figures show that SL-SVM is the best model. The accuracy, f1-score, precision, recall, selectivity, balanced accuracy, error-rate and G-mean of the SL-SVM are 94.68%, 92.99%, 91.59%, 94.41%, 94.49%, 94%, 5.32% and 94.45% respectively. The model has a high accuracy of 94.68%, which means it is reliable. It also has a high F1-score of 92.99%, which means it is balanced in terms of both precision and recall. The model has a high precision of 91.59%, and a high recall of 94.41%. This means it is good at avoiding false positives and catching true positives. In addition to being quite accurate for binary classification issues, the model has a high selectivity of 94.49%. Because of its excellent balanced accuracy of 94%, the model does not favour one class over the other. With an error rate of only 5.32%, the model is very precise. A high G-mean of 94.45% indicates that the model is consistent and robust across both classes.

Table 2 provides a comparison of cardiovascular disease prediction using several classification models. SL-SVM has the highest accuracy at 94.68%, followed by SL-RF at 94.57%, and SVM at 94%. Self-learning techniques have the potential to enhance the effectiveness of base classifiers, particularly in the case of SVM and NB.

Table 2
Comparison of Testing Outcomes for SSL Based Self-learning using SVM, NB AND RF With Various Models

Classifier	Accuracy (%)
LR	85.54
SVM	94
KNN	84.56
RF	86.03
DT	85.93
MLP	87.23
NB	83.38
Gradient Boosting	74
Bagging	73
SL-NB	90
SL-RF	94.57
SL-SVM	94.68

5. Conclusions and Future Works

A large number of people have been suffering with cardio problems all over the world. At this point in time, cardiovascular disease (CVD), is one of the major common causes of mortality on a global scale. But there is a huge need to give people the best and most affordable health care. In this research, cardiovascular disease dataset used for experiment which acquired from the Kaggle repository. This dataset contains a sample size of 70000 records of patients and 11 features and proposed semi-supervised learning approaches based on self-learning with SVM, NB and RB. According to the comparison's findings, Support Vector Machine has a high classification accuracy rate. This study showed a better improvement in the accuracy for detecting cardiovascular disease. This study also has a greater effect as a result of the suggested system's clearly superior performance in terms of sensitivity, specificity, precision, and accuracy when compared to the vast majority of the existing methods.

In future, increase the size of the dataset and create innovative algorithms based on deep learning that can identify CVD abnormalities. In addition to this, create an approach to deep learning that is capable of rapidly detecting and categorizing CVD data. There is a hope that machine learning will be able to handle semi-supervised learning: it may be utilized to discuss future study on

cardiovascular diseases (CVD). We firmly believe that the self-learning based semi-supervised machine learning technique described in this study may aid future researchers in the development of unique and advantageous CVD prediction schema and encourage additional research in the field of recognizing CVD cataclysms.

Acknowledgement
This study was supported by Post Graduate Research Scheme (PGRS) with PGRS2303110 from the University Malaysia Pahang Al-Sultan Abdullah.

References

[1] Jinjri, Wada Mohammed, Pantea Keikhosrokiani, and Nasuha Lee Abdullah. "Machine learning algorithms for the classification of cardiovascular disease-A comparative study." In *2021 International Conference on Information Technology (ICIT)*, pp. 132-138. IEEE, 2021. https://doi.org/10.1109/ICIT52682.2021.9491677

[2] Alam, Md Sakib Bin, Muhammed JA Patwary, and Maruf Hassan. "Birth mode prediction using bagging ensemble classifier: A case study of bangladesh." In *2021 International conference on information and communication technology for sustainable development (ICICT4SD)*, pp. 95-99. IEEE, 2021.

[3] Nordin, Noor Syahirah, and Mohd Arfian Ismail. "A hybridization of butterfly optimization algorithm and harmony search for fuzzy modelling in phishing attack detection." *Neural Computing and Applications* 35, no. 7 (2023): 5501-5512. https://doi.org/10.1007/s00521-022-07957-0

[4] Patwary, Muhammed JA, Xi-Zhao Wang, and Dasen Yan. "Impact of fuzziness measures on the performance of semi-supervised learning." *International Journal of Fuzzy Systems* 21 (2019): 1430-1442. https://doi.org/10.1007/s40815-019-00666-2

[5] Patwary, Muhammed JA, Weipeng Cao, Xi-Zhao Wang, and Mohammad Ahsanul Haque. "Fuzziness based semi-supervised multimodal learning for patient's activity recognition using RGBDT videos." *Applied Soft Computing* 120 (2022): 108655. https://doi.org/10.1016/j.asoc.2022.108655

[6] Ashfaq, Rana Aamir Raza, Xi-Zhao Wang, Joshua Zhexue Huang, Haider Abbas, and Yu-Lin He. "Fuzziness based semi-supervised learning approach for intrusion detection system." *Information sciences* 378 (2017): 484-497. https://doi.org/10.1016/j.ins.2016.04.019

[7] Hagan, Rachael, Charles J. Gillan, and Fiona Mallett. "Comparison of machine learning methods for the classification of cardiovascular disease." *Informatics in Medicine Unlocked* 24 (2021): 100606. https://doi.org/10.1016/j.imu.2021.100606

[8] Martins, Bárbara, Diana Ferreira, Cristiana Neto, António Abelha, and José Machado. "Data mining for cardiovascular disease prediction." *Journal of medical systems* 45 (2021): 1-8. https://doi.org/10.1007/s10916-020-01682-8

[9] Shilaskar, Swati, and Ashok Ghatol. "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases." *Expert systems with applications* 40, no. 10 (2013): 4146-4153. https://doi.org/10.1016/j.eswa.2013.01.032

[10] Islam, MD Samiul, Haider Muhamed Umran, Samir M. Umran, and Mohammed Karim. "Intelligent healthcare platform: cardiovascular disease risk factors prediction using attention module based LSTM." In *2019 2nd international conference on artificial intelligence and big data (ICAIBD)*, pp. 167-175. IEEE, 2019. https://doi.org/10.1109/ICAIBD.2019.8836998

[11] Hasan, Najmul, and Yukun Bao. "Comparing different feature selection algorithms for cardiovascular disease prediction." *Health and Technology* 11, no. 1 (2021): 49-62. https://doi.org/10.1007/s12553-020-00499-2

[12] Patro, Sibo Prasad, Neelamadhab Padhy, and Dukuru Chiranjevi. "Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning." *Evolutionary intelligence* 14, no. 2 (2021): 941-969. https://doi.org/10.1007/s12065-020-00484-8

[13] Maiga, Jaouja, and Gilbert Gutabaga Hungilo. "Comparison of machine learning models in prediction of cardiovascular disease using health record data." In *2019 international conference on informatics, multimedia, cyber and information system (ICIMCIS)*, pp. 45-48. IEEE, 2019. https://doi.org/10.1109/ICIMCIS48181.2019.8985205

[14] Alfaidi, Aseel, Reem Aljuhani, Bushra Alshehri, Hajer Alwadei, and Sahar Sabbeh. "Machine learning: assisted cardiovascular diseases diagnosis." *International Journal of Advanced Computer Science and Applications* 13, no. 2 (2022). https://doi.org/10.14569/IJACSA.2022.0130216

[15] Rani, Pooja, Rajneesh Kumar, Nada MO Sid Ahmed, and Anurag Jain. "A decision support system for heart disease prediction based upon machine learning." *Journal of Reliable Intelligent Environments* 7, no. 3 (2021): 263-275. https://doi.org/10.1007/s40860-021-00133-6

[16] Alalawi, Hana H., and Manal S. Alsuwat. "Detection of cardiovascular disease using machine learning classification models." *International Journal of Engineering Research & Technology* 10, no. 7 (2021): 151-7.

[17] Sabab, Shahed Anzarus, Md Ahadur Rahman Munshi, and Ahmed Iqbal Pritom. "Cardiovascular disease prognosis using effective classification and feature selection technique." In *2016 international conference on medical engineering, health informatics and technology (MediTec)*, pp. 1-6. IEEE, 2016. https://doi.org/10.1109/MEDITEC.2016.7835374

[18] Idris, Nur Farahaina, and Mohd Arfian Ismail. "Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition." *PeerJ Computer Science* 7 (2021): e427. https://doi.org/10.7717/peerj-cs.427

[19] Majid, Hanafi, Syahid Anuar, and Noor Hafizah Hassan. "TPOT-MTR: A Multiple Target Regression Based on Genetic Algorithm of Automated Machine Learning Systems." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 30, no. 3 (2023): 104-126. https://doi.org/10.37934/araset.30.3.104126

[20] Phoenix, Peter, Richard Sudaryono, and Derwin Suhartono. "Classifying promotion images using optical character recognition and Naïve Bayes classifier." *Procedia Computer Science* 179 (2021): 498-506. https://doi.org/10.1016/j.procs.2021.01.033

[21] Yang, Feng-Jen. "An implementation of naive bayes classifier." In *2018 International conference on computational science and computational intelligence (CSCI)*, pp. 301-306. IEEE, 2018. https://doi.org/10.1109/CSCI46756.2018.00065

[22] Speiser, Jaime Lynn, Michael E. Miller, Janet Tooze, and Edward Ip. "A comparison of random forest variable selection methods for classification prediction modeling." *Expert systems with applications* 134 (2019): 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

[23] Mohapatra, Niva, K. Shreya, and Ayes Chinmay. "Optimization of the random forest algorithm." In *Advances in Data Science and Management: Proceedings of ICDSM 2019*, pp. 201-208. Springer Singapore, 2020. https://doi.org/10.1007/978-981-15-0978-0_19

[24] Perez-Siguas, Rosa, Hernan Matta-Solis, Eduardo Matta-Solis, Luis Perez-Siguas, Hernan Matta-Perez, and Alejandro Cruzata-Martinez. "Emotion Analysis for Online Patient Care using Machine Learning." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 30, no. 2 (2023): 314-320. https://doi.org/10.37934/araset.30.2.314320

[25] Cao, Weipeng, Yuhao Wu, Chengchao Huang, Muhammed JA Patwary, and Xizhao Wang. "MFF: Multi-modal feature fusion for zero-shot learning." *Neurocomputing* 510 (2022): 172-180. https://doi.org/10.1016/j.neucom.2022.09.070

[26] Zhang, Yong, Dun-wei Gong, Xiao-zhi Gao, Tian Tian, and Xiao-yan Sun. "Binary differential evolution with self-learning for multi-objective feature selection." *Information Sciences* 507 (2020): 67-85. https://doi.org/10.1016/j.ins.2019.08.040

[27] Meyer, Anneke, Suhita Ghosh, Daniel Schindele, Martin Schostak, Sebastian Stober, Christian Hansen, and Marko Rak. "Uncertainty-aware temporal self-learning (UATS): Semi-supervised learning for segmentation of prostate zones and beyond." *Artificial Intelligence in Medicine* 116 (2021): 102073. https://doi.org/10.1016/j.artmed.2021.102073

[28] Jinjri, Wada Mohammed, Pantea Keikhosrokiani, and Nasuha Lee Abdullah. "Machine learning algorithms for the classification of cardiovascular disease-A comparative study." In *2021 International Conference on Information Technology (ICIT)*, pp. 132-138. IEEE, 2021. https://doi.org/10.1109/ICIT52682.2021.9491677

[29] Ulianova, Svetlana. "Cardiovascular disease dataset." *Kaggle. com* (2019).