Constructing Multiple Layers of Machine Learning for the Early Detection of Cardiovascular Diseases

Hao-Yun Hsieh, Chang-Fu Su and Shu-I Chiu

February 15, 2022

Constructing Multiple Layers of Machine Learning for The Early Detection of Cardiovascular Diseases

Hao-Yun Hsieh
Department of Sociology
National Chengchi University
Taipei, Taiwan
107204028@g.nccu.edu.tw

Chang-Fu Su
Department of Anesthesia,
Division of Medical Quality
En-Chu-Kong Hospital
New Taipei City, Taiwan
colus.su@gmail.com

Shu-I Chiu
Department of Computer Science
National Chengchi University
Taipei, Taiwan
sichiu@g.nccu.edu.tw

***Abstract*—Healthcare is an inevitable task to be done in human life. Cardiovascular diseases are one of the most common diseases among elders, as treatments at early stage has been proven to be able to reduce greatly the death rate, early diagnosis of cardiovascular diseases and the detection of high-risk patients are therefore extremely important. In response to this need in future society, we developed some machine learning models that can provide cardiovascular diseases diagnosis with lower budgets. We hope our work can reduce death risks for elder patients and preserve medical resources at the same time. Our proposed models are imported to the graphical user interface we developed. It helps users detect cardiovascular diseases at the early stage.**

Keywords— Cardiovascular disease, classification, graphical user interface, multi-layered machine learning, prediction

I. INTRODUCTION

Demographic researchers of the United Nations have revealed that the percentages of elders in the total population are increasing rapidly in 21th century. Globally, there were 703 million older persons aged 65 or over in 2019. Eastern and South-Eastern Asia was home to the largest number of the world's older population (260 million), followed by Europe and Northern America (over 200 million). Over the next three decades, the global number of older persons is projected to more than double, reaching over 1.5 billion persons in 2050 [1]. With the growth and the aging of the population in recent years, the number of people who need to take health examinations regularly are also increasing considerably. It has resulted in the huge consumption of medical resources. Medical resources in many regions are barely sufficient for such large numbers of elder patients. And among all common illnesses, cardiovascular diseases (CVDs), including chronic ischemic heart disease, congestive heart failure, and arrhythmia remains the most common cause of death of older adults [2]. Generally, diagnosis of CVDs requires either blood test or electrocardiographic (ECG) test, which are both needy of expensive equipment and few hours to know the results; therefore, it is logical to assume that the diagnosis of CVDs is or will soon become one of the major burdens on hospitals and clinics around the world as the population aging continuing. In response to this issue, we developed some machine learning models that can provide early diagnosis of CVD in order to help medical staves to identify those who need treatment or further examination in a simpler way, and then we built these machine learning models under the framework of a graphical user interface (GUI). With the GUI, users can receive cardiovascular diagnosis by inputting their information (e.g., age, sex, habit, and blood pressures).

As our goal is to reduce the consumption of medical resources, the False-Positive cases (people who actually do not have CVDs but are diagnosed as CVDs patients by algorithms) are what should be avoided as much as possible. We therefore built another set of machine learning models that can detect possible False-Positive cases. These models for False-Positive cases detection would be Layer 2 of our models. The result of the Layer 2 would be combined with predictions of Layer 1 and be inputted into Layer 3 of our models, by this way we can eliminate certain numbers of False-Negative cases. The workflow of our proposed models are shown in Fig. 1. Finally, we constructed GUI for users to input their basic information and then our GUI system response the probability with CVDs and without CVDs.

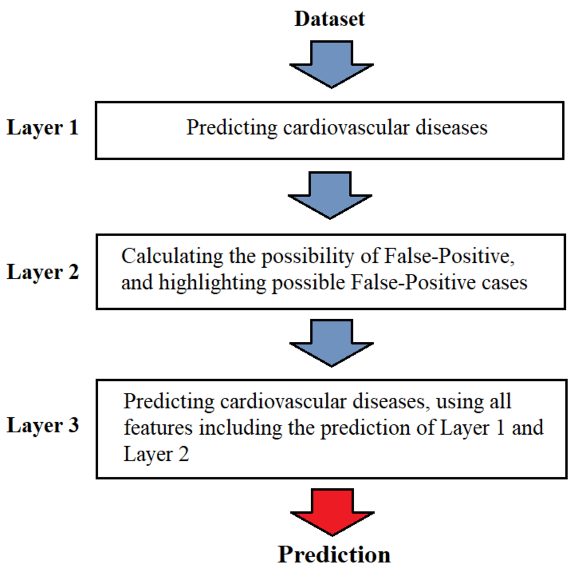

Fig. 1. The workflow of our GUI and 3 layers of models.

Major contributions of our work are:

- Providing a technical solution to the exacerbating burdens on medical workers for early diagnosis of cardiovascular diseases, reducing the consumptions of medical resources in the aging societies. The user interface we constructed only requires data that is easy to obtain (e.g. blood pressure, BMI), while previous researchers generally adopted complex data like ECG, which requires high level professional equipment to know. Our models are therefore able to reduce the burdens on the healthcare system.
- Preventing elders from dying caused by severe cardiovascular diseases, because our models can theoretically detect cardiovascular disorders in the early stage; and medical treatments and habit changing in the early stage can significantly prevent worsening of related conditions.

- Using multiple layers of models allows us to eliminate False-Positive cases as much as possible. Note that we use different training data in different layers to avoid the occurrence of model over-fitting.

The remainder of this paper is organized as follows: We will briefly review the related works in Section 2, describe the proposed methods in Section 3, depict experiments in Section 4, report results from experiments in Section 5, and finally conclude this paper in Section 6.

II. Related Work

CVDs are the leading cause of death globally. Most CVDs can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol. The early methods of forecasting the CVDs helped in making decisions about the changes to have occurred in high-risk patients which resulted in the reduction of their risks [3]. Some studies show that Naive Bayes, Random Forest, Gradient boosting, Logical Regression and Support Vector Machine for predicting CVD [3, 4]. In this paper, we use these classifiers to predict CVD and then we construct the website based on our predictive models.

The behavioral risk factor surveillance system (BRFSS) is a state-based telephone survey coordinated by the Centers for Disease Control and Prevention (CDC) [5]. More than 400,000 adults complete the survey annually, making the BRFSS the largest telephone survey in the world [6]. It focuses on self-reported information regarding chronic conditions and health risk behaviors [5]. A number of scholars and researchers have conducted studies of the reliability and validity of the BRFSS estimates in the context of these changes [5]. They studied a review of reliability and validity studies of the BRFSS. New analyses and comparisons of BRFSS dataset includes the new methodologies and cell phone data [5]. In this paper, we used BRFSS dataset to predict CVDs and then built the graphical user interface based on our proposed models for general users.

In recent years, artificial intelligence (AI) plays a crucial role in “earlier medicine” [7]. Machine learning is an extension of a century-long quest for AI [8]. In [8], they concluded by reviewing the limitations associated with contemporary application of machine learning algorithms within the CVD field. Machine learning could provide a powerful platform for integration of clinical and imaging data, which would be useful for multifactorial and complex CVDs [8]. Many studies use echocardiography, electrocardiography, or imaging data to predict CVDs [8-10]. In [9], they proposed to develop an artificial neural network (ANN)-based machine learning technique, combining both individualized medical information and clinical ECG data, to train the cell phone to learn to adapt to its user's physiological conditions to achieve better ECG feature extraction and more accurate CVD classification results. For general users, it is hard to get such data. Therefore, we use general data like BRFSS or surveys to predict CVDs.

III. Methods

Data Descriptions.

Data used in this paper is the BRFSS dataset of the year 2015. BRFSS is a health-related survey collected by the United States government. This dataset contains over 400,000 responses from 50 states, the District of Columbia, Guam, and Puerto Rico, in which there are 251,450 valid responses that are related to CVDs. The definition of CVDs in this dataset is coronary heart disease and/or myocardial infarction. Features we used are shown in Table 1.

We took socioeconomic factors into our consideration because there have been many sociological researches indicated that socioeconomic classes influenced health conditions. For example, income reflects living quality and type of working (white-collar job or blue-collar job), and these two factors affect health conditions.

Table 1. Features in the raw data

Feature Type	Features
Target	Cardiovascular Disease (binary)
Common Health Related Factors	High Blood Pressure, Smoking, Diabetes, Sex, Age-Group (per 5 years), BMI, Regular Exercise, Diet Habit (fruits and vegetables), High Cholesterol Level, Regular Alcohol Drinking, Self-Report Health Condition (numerical score)
Socioeconomic Factors	Household Income, Education Attainment, Insurance

We normalized numeric features such as income, education, BMI and self-reported health conditions in order to avoid possible bias (normalization is done by calculating z-score). In the original dataset, the ratio of CVDs patients and people without CVDs is roughly 10.3 percent (23,659/227,791). Due to the imbalanced ratio, cases without CVDs are under-sampled by random sampling. As for test dataset, we directly extracted it from the original dataset, so the distribution and ratio of test dataset is exactly the same as the original dataset. The workflow for the processing is illustrated in Fig. 2.

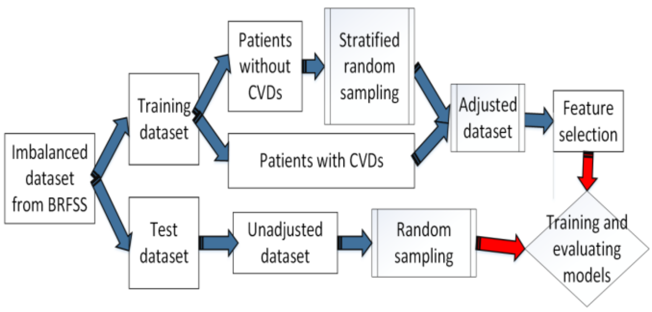

Fig. 2. The data flow

After the adjustment, the training dataset now has an approximately 1:1 ratio between CVDs patients and people without CVDs. For model training workflow, as explained in the introduction section, this paper used three layers of models. In Layer 1, we trained models using CVDs as target; and then we imported the best Layer 1 model into another set of data, and then we mark the False-Positive cases in the dataset by inspecting whether the prediction of Layer 1 model corresponds to the real cardiovascular observation;

then we dropped real cardiovascular observation and the prediction of Layer 1 of our model from the Layer 2 training dataset. And then we trained Layer 2 of our model using the False-Positive marks as target. After that, we imported the best Layer 2 model into Layer 3 training dataset. Layer 3 models are trained afterwards using CVDs as target, predictions of Layer 1 and Layer 2 as new factors. We want to test whether it is possible to reduce the occurrence of False-Positive cases significantly by adding the possibility of False-Positive into training features. Eventually, we constructed the 3 layers model into the GUI we developed. Workflow of different layers is shown in Fig. 1 in the introduction section.

Feature Selection In machine learning and statistics, feature selection is the process of selecting a subset of relevant features for use in model construction. Feature selection methods can be distinguished into three categories: filters, wrappers, and embedded/hybrid method [11]. Wrappers methods perform better than filter methods because the feature selection process is optimized for the classifier to be used [11]. We use the sequential forward selection (SFS) manner by Whitney [12] based on wrapper method. SFS is one of the commonly used heuristic methods for feature selection. We use random forest (RF) [13] and the 10-fold cross-validation test for the area under the receiver operating characteristic curve (AUROC) value estimate.

IV. EXPERIMENTS

Machine learning in this paper is implemented with the "caret" package in R. For machine learning, regarding classification algorithms, we use top ones [14]: Naïve Bayes [15, 16], Support Vector Machines (SVM) [17], Logistic Regression [18], *k*-Nearest Neighbor (*k*NN) [19], Classification and Regression Tree (CART) [20], Multilayer Perceptron (MLP) [21] and C5.0 Decision Tree [22]. In addition, we also use the RF algorithm. RFs are an ensemble learning for classification, regression, and other tasks by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

For our GUI development, we used the "gradio" package on Python. This package allows us to import our machine learning model into an user interface. However, since Python has not yet supported the C5.0 decision tree, we replaced the C5.0 decision tree with the C4.5 decision tree [23] when constructing our GUI in the Python environment.

A. *Multiple Layers of Modeling*

There are three layers of modeling, layer 1 would provide the first prediction of cardiovascular diseases, layer 2 would provide the prediction of False-Positive cases possibility, and layer 3 would provide final prediction of CVDs using all selected features and results of layer 1 and layer 2 models. Features each layer uses are summarized in Table. 2.

Table. 2. Features Different Layers Used for Modeling.

Layer	Features
Layer 1	**Target**: Cardiovascular Disease. **Features**: Age, Sex, High Blood Pressure, High Cholesterol Level, Self-Reported Health Conditions.
Layer 2	**Target**: False-Positive. **Features**: Age, Sex, High Blood Pressure, High Cholesterol Level, Self-Reported Health Conditions, Insurance, Diet Habit (Vegetable).
Layer 3	**Target**: Cardiovascular Disease. **Features**: Age, Sex, High Blood Pressure, High Cholesterol Level, Self-Reported Health Conditions, Prediction of Layer 1, Prediction of Layer 2.

In terms of training data of each layer, we adjusted the imbalance ratio of its target features. Feature selection has been done respectively for each layer, some features that were not used in the first layer were therefore adopted in the second layer. In order to avoid the overfitting situation, we used different training dataset and different test dataset for different layers. The description is in Table. 3 and Table. 4. The reason why we used the same test dataset for layer 1 and layer 3 was to make a comparison of these two layers. There is no overlapped data among these dataset.

Table. 3. Training dataset and test dataset for each layer.

Layer	Training Dataset	Test Dataset
Layer 1	A	*Alpha*
Layer 2	B	*Beta*
Layer 3	C	*Alpha*

Table. 4. Description of different dataset.

Dataset	Target	Target Ratio	Note
Training Dataset A	CVDs	Adjusted	Under-sampling
Training Dataset B	False Positive	Unadjusted	CVDs ratio remains the original ratio
Training Dataset C	CVDs	Adjusted	Under-sampling
Test Dataset *alpha*	CVDs	Unadjusted	Random Sampling
Test Dataset *beta*	False Positive	Unadjusted	Random Sampling

B. *Evaluation Metrics*

In this paper, we use the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and F-Score (see Equation 1). F-Score is the harmonic mean of precision and recall and gives a good combination of the two. Generally, F-Score with β=3 can emphasize recall.

$$F_\beta = (1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{(\beta^2\cdot\text{precision})+\text{recall}}. \quad (1)$$

Precision and recall are calculated by Equation 2 and Equation 3 separately, in which TP represents the number of True-Positive cases; FP represents the number of False-Positive cases, and FN represents False-Negative cases.

$$precision = \frac{TP}{TP+FP} \quad (2)$$

$$recall = \frac{TP}{TP+FN} \quad (3)$$

Furthermore, we use 15-fold cross-validation. It divides the dataset into 15 disjoint subsets. It uses 14 subsets to create a new dataset, and use the new dataset to train a classifier. Then, it uses the remaining 1 subset to test the classifier. It repeats the above two steps 15 times, and each time it uses a different subset. The final result is an aggregate of the 15 test results. Cross-validation is almost the standard way to evaluate classifiers and compare classification algorithms (and find an optimal set of parameters for a classification algorithm) in data mining.

V. Results

In this section, we would present the results for our experimental settings.

A. *The Layer 1*

For models in the first layer, results of AUROC and AUPRC are shown in Fig. 3. As what can be seen in Fig. 3, SVM and logistic regression appeared to have the best performance.

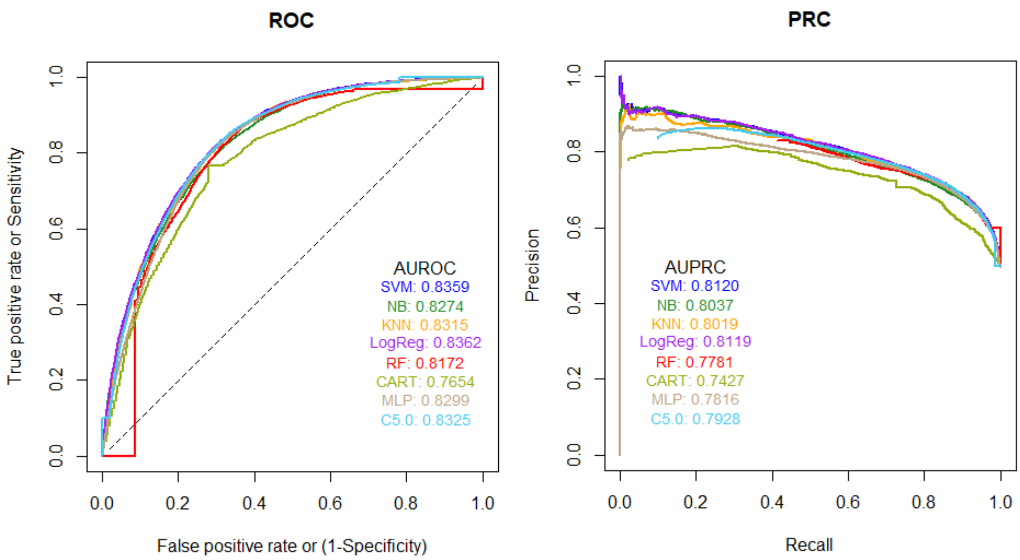

Fig. 3. AUROC and AUPRC for layer 1 Models.

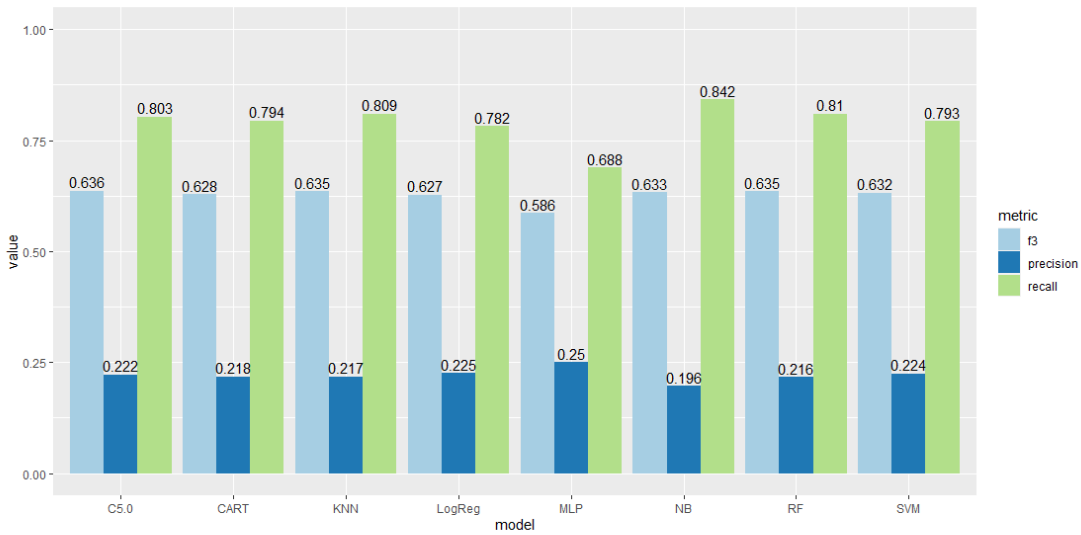

Fig. 4. Comparison of different models in layer 1 (Test Data).

As for the F_3-score, recall and precision of different models in layer 1, the comparison are shown in Fig. 4. Note that metrics in Fig. 4 were all calculated within the test dataset alpha. After comprehensively inspecting the aforementioned evaluation metrics including AUPRC, *k*NN model is adopted as our used model for layer 1, because we value F_3-score in test data more than just AUROC and AUPRC. And *k*NN model shows both good performance on its AUROC and F_3-score, whereas SVM and logistic regression do not have such performance in test data.

B. *The Layer 2.*

For models in the second layer, results of AUROC and AUPRC are shown in Fig. 5 and Fig. 6, which indicates that C5.0 decision tree and *k*NN have the best performance.

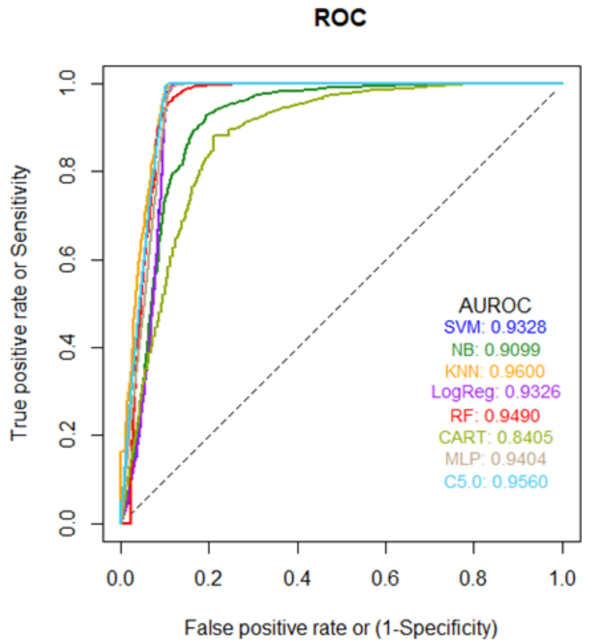

Fig. 5. AUROC for layer 2 models.

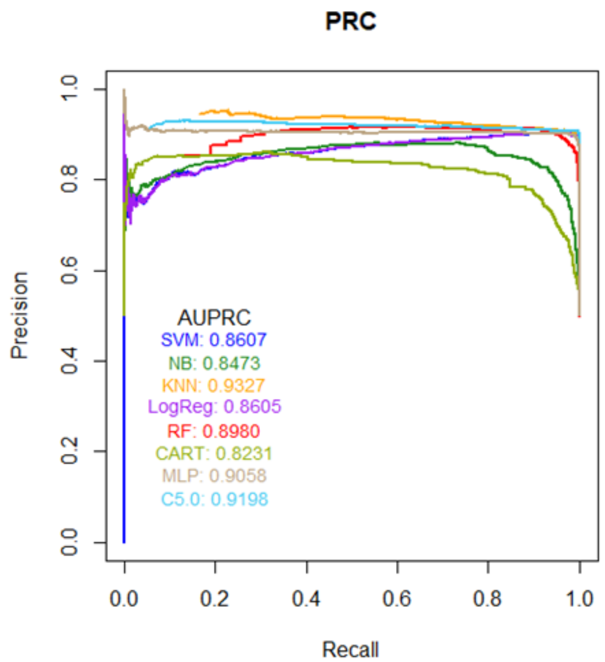

Fig. 6. AUPRC for the layer 2 models.

It is fair to say that the second layer received a very good result of modeling. Almost every model has AUROC and AUPRC above 0.9, this indicates that marking False-Positive cases for CVDs is totally practical.

As for the F_3-score, recall and precision of different models in layer 2, the comparison is shown in Fig. 7. Metrics in Fig. 7 were all calculated within the test dataset beta. After comprehensively inspecting aforementioned evaluation metrics including AUPRC, C5.0 decision tree is adopted as our used model for layer 2, because in spite of the better AUROC value of *k*NN model in layer 2, C5.0 has a better performance in test dataset.

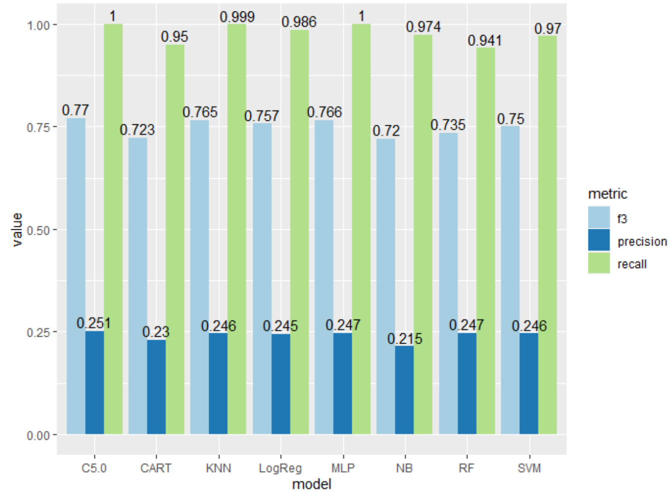

Fig. 7. Comparison of different models in layer 2 (Test Data)

C. *Layer 3.*

For models in the third layer, results of AUROC and AUPRC are shown in Fig. 8 and Fig. 9. According to AUPRC, it can be said that *k*NN and logistic regression have the best performance.

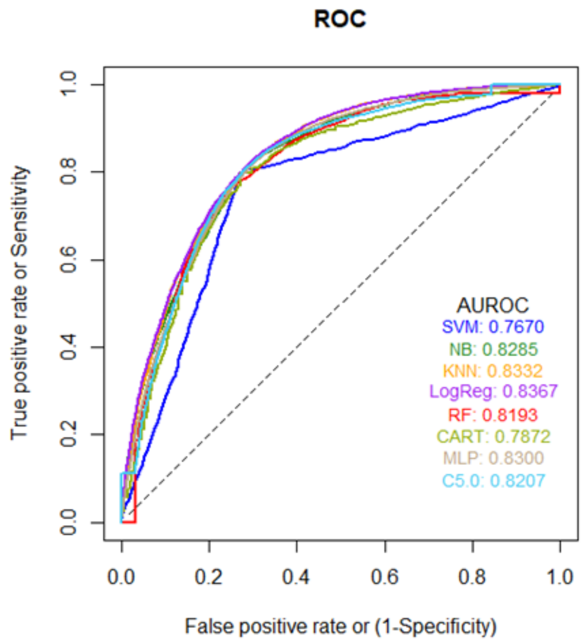

Fig. 8. AUROC for models in layer 3.

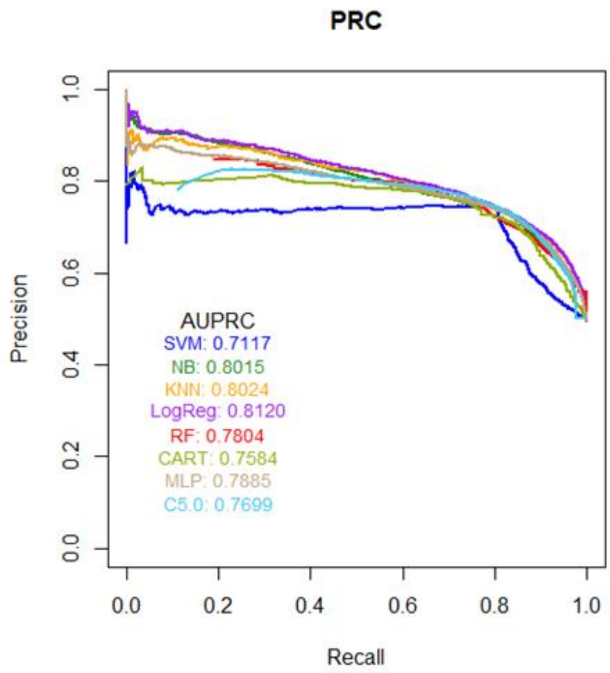

Fig. 9. AUPRC for models in layer 3.

For the F_3-score, recall and precision of different models in layer 3, the comparison are shown in Fig. 10. Note that metrics in Fig. 10 were all calculated within the test dataset alpha. After inspecting aforementioned evaluation metrics, *k*NN is adopted as our used model for layer 3.

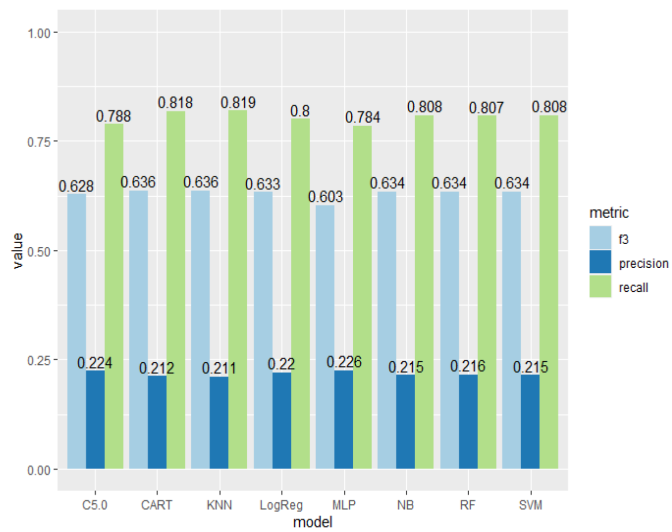

Fig. 10. Comparison of different models in layer 3(Test Data)

In Fig. 10, it can be noticed that F_3-score of certain models have indeed increased compared to models in layer 1. However, the effects of enhancement are not as good as what was expected. Despite the increased F_3-score, the addition of the second layer does not consolidate the precision, it boosts the recall on the contrary; which means our original goal of eliminating False-Positive cases from layer 1 failed. But overall, the third layer indeed has a better performance compared to Layer 1, our method of modeling, which uses multiple layers of models, still shows some potential. For example, the high AUROC and AUPRC of Layer 2 in Fig. 5 and Fig. 6 indicated that highlighting False-Positive cases is fully possible. It can be expected that in other cases of disease detection, by using deep learning algorithms in the second layer, better accuracy can still be provided by the third layer

As for our GUI, even though in terms of the cost of computation, the larger cost of computation of our multi-layered model is not directly proportional to the degree of improvement in its accuracy; it is undeniably still a better option due to its better performance. We therefore select the combined model that has the best performance (with *k*NN model in layer 1, C5.0 model in layer 2 and *k*NN model in layer 3) for our GUI.

The output value of our GUI would be possibility instead of simply a binomial classification, so users can still decide whether to go for further examination or not based on possibility. This would be useful for those who received close possibilities like 51% or 49%. Our GUI would operate like what are shown in Fig. 11 and Fig. 12.

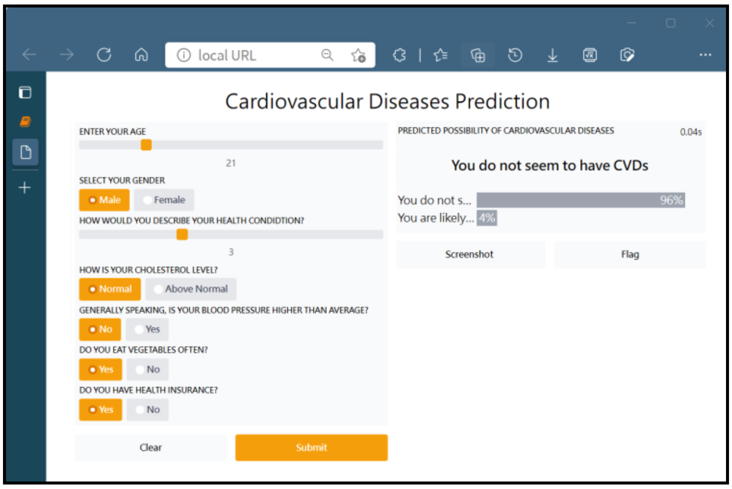

Fig. 11. Our GUI operating on a local URL. (negative result)

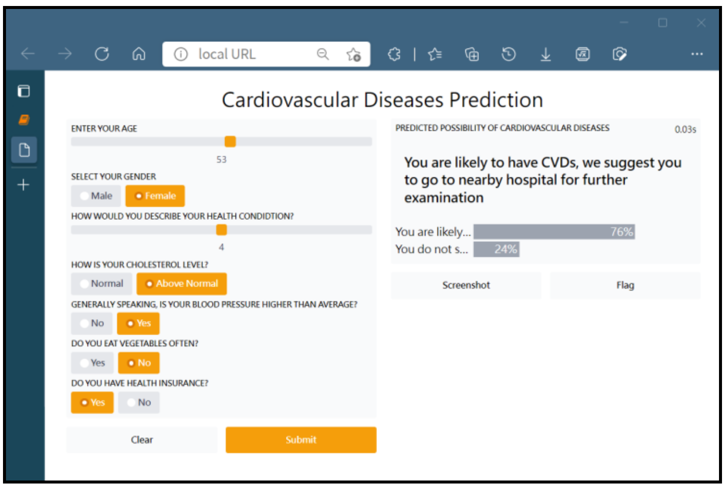

Fig. 12. Our GUI operating on a local URL. (positive result)

Note that the recall value of our models in the test dataset are around 0.8, which means our GUI can detect over 80% of CVDs patients. Despite the F_3-score around only 0.63 and 78% of False-Positive rate, our GUI can still make its contribution to early diagnosis of CVDs, in order to save elder's lives and preserve medical resources at the same time.

VI. Conclusion

Machine learning is one of the most exciting technologies that one would have ever come across. It can be explained as automating and improving the learning process of computers based on their experiences without any human assistance. We applied classifiers to detect CVDs with features that are easier to obtain, which is different from previous similar researches that used ECG data as features. For single-layered model, the best classifier is *k*NN, the AUROC value is 0.83.and its F_3-score in test dataset is 0.635. For multi- layered model, *k*NN + C5.0 + *k*NN is the best; the AUROC value is closed to 0.84, with the F_3-score around 0.64 in test dataset.

In the future, we will add more features to our models. And as what was mentioned, we might try using deep learning model for Layer 2 as results presented in Fig. 5 and Fig. 6 have revealed the high possibility of highlighting False-Positive cases. Once we can improve the accuracy of Layer 2 on False-Positive cases marking, we would be able to perform multi-layered models with better accuracy in Layer 3.

Acknowledgment

The authors would like to express their sincere gratitude to the anonymous reviewers for their constructive comments on the manuscript of this paper.

References

1. Division of Population, U.N., *World Population Ageing 2019: Highlights*. 2019.

2. Jaul, E. and J. Barron, *Age-related diseases and clinical and public health implications for the 85 years old and over population*. Frontiers in public health, 2017. 5: p. 335.

3. Dinesh, K.G., et al. *Prediction of cardiovascular disease using machine learning algorithms. in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. 2018. IEEE.

4. Patel, J., D. Tejal Upadhyay, and S. Patel, *Heart disease prediction using machine learning and data mining technique.* Heart Disease, 2015. 7(1): p. 129-137.

5. Pierannunzi, C., S.S. Hu, and L. Balluz, *A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS)*, 2004–2011. BMC medical research methodology, 2013. **13**(1): p. 1-14.

6. Control, C.f.D. and Prevention, *Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System*. 2010, Washington, DC: Center for Disease Control and Prevention.

7. Iqbal, U., L.A. Celi, and Y.-C.J. Li, *How can artificial intelligence make medicine more preemptive?* Journal of Medical Internet Research, 2020. **22**(8): p. e17211.

8. Al'Aref, S.J., et al., *Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging*. European heart journal, 2019. 40(24): p. 1975-1986.

9. Jin, Z., et al. *HeartToGo: a personalized medicine technology for cardiovascular disease prevention and detection. in 2009 IEEE/NIH Life Science Systems and Applications Workshop*. 2009. IEEE.

10. Jin, Z., Y. Sun, and A.C. Cheng. *Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone. in 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2009. IEEE.

11. Khalid, S., T. Khalil, and S. Nasreen. *A survey of feature selection and feature extraction techniques in machine learning. in 2014 science and information conference.* 2014. IEEE.

12. Whitney, A.W., *A direct method of nonparametric measurement selection.* IEEE Transactions on Computers, 1971. **100**(9): p. 1100-1103.

13. Ho, T.K. *Random decision forests. in Proceedings of 3rd international conference on document analysis and recognition*. 1995. IEEE.

14. Wu, X., et al., *Top 10 algorithms in data mining. Knowledge and information systems*, 2008. **14**(1): p. 1-37.

15. Hand, D.J. and K. Yu, *Idiot's Bayes—not so stupid after all?* International statistical review, 2001. **69**(3): p. 385-398.

16. Zhang, H., *The optimality of naive Bayes*. AA, 2004. 1(2): p. 3.

17. Cortes, C. and V. Vapnik, *Support-vector networks*. Machine learning, 1995. 20(3): p. 273-297.

18. Wright, R.E., *Logistic regression*. 1995.

19. Laaksonen, J. and E. Oja. *Classification with learning k-nearest neighbors. in Proceedings of International Conference on Neural Networks (ICNN'96)*. 1996. IEEE.

20. Lewis, R.J. *An introduction to classification and regression tree (CART) analysis. in Annual meeting of the society for academic emergency medicine in San Francisco, California.* 2000. Citeseer.

21. Attali, J.-G. and G. Pagés, *Approximations of functions by a multilayer perceptron: a new approach.* Neural networks, 1997. **10**(6): p. 1069- 1081.

22. Rajeswari, S. and K. Suthendran, *C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud.* Computers and Electronics in Agriculture, 2019. **156**: p. 530-539.

23. Chauhan, H. and A. Chauhan, *Implementation of decision tree algorithm c4. 5.* International Journal of Scientific and Research Publications, 2013. **3**(10): p. 1-3.