Non-invasive and invasive cardiovascular disease prediction with machine learning

Cara Van Uden[1†*], Srishti Bagchi[1†], Rachael Chacko[1†]

1 Department of Computer Science, Dartmouth College, Hanover, NH, USA

† These authors contributed equally to this work.
* cara.e.van.uden.19@dartmouth.edu

Abstract

Cardiovascular disease (CVD) is a term for any type of pathology of the heart which results in inefficient circulation of blood. CVD is responsible for 31% of deaths, worldwide, but 200,000 of deaths could be avoided in just the United States through early diagnosis of CVD. However, current CVD diagnostic procedures are invasive, costly, and requires a clinical setting. In this paper, we attempted to apply machine learning models on two datasets: one containing invasive data that would be collected in a clinic, and another containing non-invasive data that could be collected via survey or commercial activity-tracker. The highest accuracy obtained from training and testing on the non-invasive dataset was 60% from logistic regression. From training and testing on the invasive dataset, a logistic regression was able to achieve 87% accuracy. This suggests that invasive data may be necessary for an accurate CVD diagnosis. Future work may include using different types of non-invasive data that, hopefully, would be able to properly train a predictive model.

Introduction

Cardiovascular disease (CVD) is a broad category of heart-related diseases such as heart attack, stroke, arrhythmia, and heart valve problems, that lead to a failure to effectively circulate blood. Globally, CVD is the leading cause of morbidity and mortality. In 2012, there were 17.5 million deaths from CVD with 7.4 million deaths due to coronary heart disease (CHD) and 6.7 million deaths due to stroke [1]. In the United States alone, CVD is again the leading cause of death, with over 840,000 deaths in 2016 [2]. However, through simple lifestyle changes and screening, nearly 200,000 of these deaths per year in the United States can be avoided [3]. Established approaches to CVD risk assessment, such as that recommended by the American Heart Association/American College of Cardiology (AHA/ACC), predict future risk of CVD based on well-established risk factors such as hypertension, cholesterol, age, smoking, and diabetes. These identifying risk factors can often be evaluated easily through a standard physical exam and a blood test [4], [5]. Additionally, these risk factors are included within most CVD risk prediction tools (ACC/AHA [6], QRISK2 [7], Framingham [8], Reynolds [9]). However, there remain a large number of individuals at risk of CVD who fail to be identified by these tools. For example, approximately half of myocardial infarctions (MIs) and strokes will occur in people who are not predicted to be at risk of CVD [10]. Equivalently, some individuals not at risk are given unnecessary treatment, which may pose a financial threat or cause adverse health effects. For instance, the U.S.

Preventative Services Task Force now recommends against electrocardiograms (ECGs) for adults with low risks of heart disease for monetary reasons and because follow-up tests such as angiograms can cause inadvertent heart damage [11].

All standard CVD risk assessment models make an implicit assumption that each risk factor is related in a linear fashion to CVD outcomes [12]. Therefore, these models may oversimplify complex relationships which include many risk factors with non-linear interactions. Approaches that determine more nuanced relationships between risk factors and outcomes, such as those made possible using machine learning, need to be explored.

In this paper, we explore several machine learning approaches to detect the presence of CVD using only standard health information that would be obtained during a typical physical exam, including, but not limited to, height, weight, blood pressure (BP), and cholesterol level. We leverage both traditional machine learning and state-of-the-art deep learning techniques. The machine learning techniques include a support vector machine (SVM) with a Gaussian Radial Basis Function (RBF) kernel, a Gaussian Naive Bayes classifier (GNB), a logistic regression (LR) classifier, a k-nearest neighbors (k-NN) classifier, and a random forest (RF) classifier. For our deep learning technique, we also use a multilayer feedforward neural network for this CVD detection. By utilizing these robust machine learning and deep learning models, we hope to achieve reliable and generalizable results in detecting the presence of CVD.

In addition to using these methods to detect CVD, we will also determine feature importance from running models on non-invasive data as well as data collected from invasive procedures. With the information gained from this study, physicians could potentially alter their current case history methods to obtain more useful data from their patients. The results from this paper could also aid in streamlining the diagnostic process and improving diagnostic accuracy.

Similar studies have also generated models to predict heart disease based off of more detailed examination data, such as that obtained from coronary angiography, fluoroscopy, and myocardial scintigraphy [13], [14], [15]. Such studies have used various models including Bayesian classifiers, k-NN, and neural networks, achieving classification accuracies of up to 89% in predicting the presence of CVD. However, while the majority of these studies utilize diagnostic tests that provide very detailed information, these tests can be relatively invasive. They often require either an injection of contrast material or exposure to radiation - both of which can be uncomfortable for the patient. Additionally, such tests are often very expensive for both patients and healthcare facilities and may not always be useful toward the diagnostic process. If we can demonstrate that our model can achieve a similar level of accuracy with data that is more accessible than data from a typical medical exam, then this technology could streamline the diagnostic process and bring medicine to the home, especially if the data could be instantaneously gathered via a mobile health device such as a FitBit.

Materials and methods

Dataset description

Our non-invasive dataset [17] is an open-source Kaggle dataset, although information on how the data was collected and by whom was lacking. The dataset has health information and information on the presence/absence of CVD for 70,000 patients (34,979 presenting with CVD and 35,021 not presenting with CVD). The dataset has 11 features, as represented in Table 1.

Non-invasive dataset

Feature	Data type
age	cont.
gender	binary
height	cont.
weight	cont.
systolic bp	cont.
diastolic bp	cont.
cholesterol level	cat. (1, 2, 3)
glucose level	cat. (1, 2, 3)
alcohol usage	binary
smoker	binary
active	binary

Invasive dataset

Feature	Data type
age	cont.
gender	binary
chest pain type	cat. (0, 1, 2, 3)
max heart rate	cont.
resting bp	cont.
cholesterol level	cont.
high glucose level	binary
exercise-induced angina	binary
ST depression	cont.
ST slope	cat. (0, 1, 2)
ECG results	cat. (0, 1, 2)
fluoroscopy results	cat. (0, 1, 2, 3)
thalium stress test results	cat. (3, 6, 7)

Table 1. Non-invasive and invasive dataset features. Categorical and continuous data types are abbreviated as cat. and cont., respectively.

Features of the dataset are either continuous, assigned categorical codes, or binary. The classes are balanced, but there are more female patients included in the dataset than male patients. Further, the continuous-valued features are roughly normally distributed after normalization and removal of outliers. However, most categorical-valued features are skewed towards ”normal” as opposed to ”high” levels of potentially pathological features. A heatmap showing correlations between all data fields provides insight into potentially influential variables for predicting presence of CVD (Figure 1). From the correlations, we see that age, cholesterol, and weight may be important features in predicting the label. Likewise, other correlations between height, gender, smoking habits, glucose, and cholesterol might indicate additional influential features for predicting additional target variables other than presence of CVD.

Our invasive dataset is the open-source UCI CVD dataset [18]. It is a much smaller dataset than our non-invasive dataset, with only 303 samples. Though it does not include behavioral data like our non-invasive dataset, it includes some features that either take more time or are more invasive to obtain, such as fluoroscopy, ECG, and exercise-induced angina. Features of the dataset are either continuous, assigned categorical codes, or are binary (Table 1). When visualized with a heatmap (Figure 1), it is apparent that many of the above variables correlate strongly with one another.

Methods

Our methodology consisted of:

1. Preprocessing and cleaning the dataset. There were several outliers in the non-invasive dataset in which values for BP were abnormally high. Those outlier datapoints were removed, bringing the size of the dataset down from 70,000 to 67,007. All variables were then normalized to remove potential biases.

2. Detecting the presence or absence of CVD – a binary classification problem. As there exist many machine learning algorithms capable of accurate binary classification, we decided to implement various models and evaluate their performance (as measured by accuracy, precision, recall, and F1 score). Therefore, we chose to utilize several commonly used models, including GNB, SVM, k-NN, LR, and RF. These models implement widely different model architectures between one another, yet all are also known for their reliability and

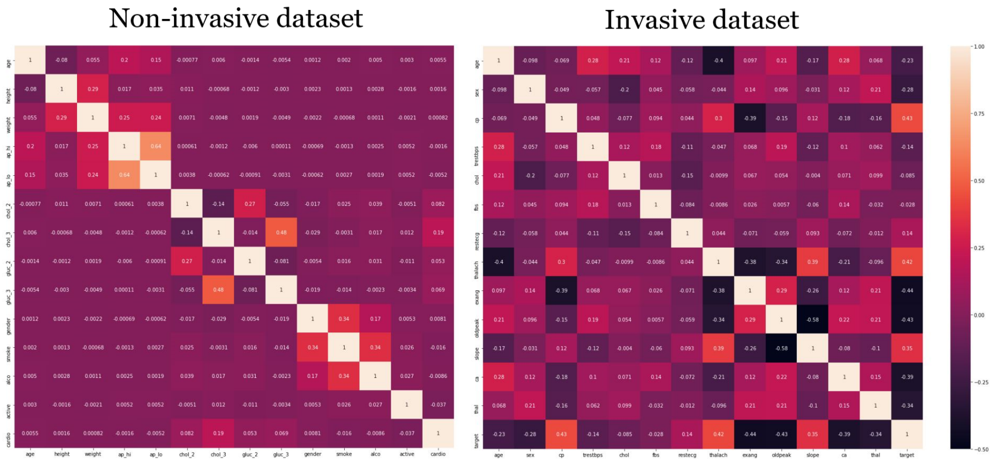

Fig 1. Heatmap of features from the non-invasive and invasive datasets. Each of the features and the label from both non-invasive and invasive datasets were plotted in heatmaps to visualize correlations in the data. In the non-invasive dataset, we note that many features are not well correlated with each other or with the label. The only notable correlation here is that between high cholesterol and CVD. However, in the invasive dataset, we see many correlations between features and the target. Specifically, chest pain and Thalium stress test results are highly correlated with CVD. This finding might suggest that the invasive dataset would be better to use for predicting CVD.

generalizability. We also implemented a MLP to explore whether a deep learning approach could result in a more successful model.

3. Comparing the performance of machine learning models trained on the non-invasive dataset versus the more invasive UCI dataset. This involved comparing performance (as measured by accuracy, precision, recall, and F1 score) of logistic regression models for CVD diagnosis after training on each dataset. We used LR for this comparison due to the promising performance of LR on the non-invasive dataset and because LR models have easily-interpretable feature importances. Feature coefficients can range between -1 (strongly pushes classification towards the negative class) and 1 (strongly pushes classification towards the positive class).

4. Predicting other target variables. We used regression models to predict BP and body mass index (BMI). We mention the performances of these models briefly in this paper but do not elaborate on them because none of the regression models demonstrated any notable performance, and because BP and BMI are already simple to measure.

Results

We achieved accuracies ranging from 0.54 to 0.60 (Table 2) in our binary classification models for CVD. The k-NN model performed the poorest, while the SVM, GNB, MLP, and LR models generated the highest accuracies. In this section, the absence of CVD is referred to as the negative class, and the presence of CVD is referred to as the positive class. The RF and k-NN models resulted in very similar precision scores for both negative and positive classes. The other models (SVM, GNB, MLP. and LR) had much higher precision for the positive class - this means that the models are more able to identify only relevant instances for the positive class. In terms of recall, all models performed better for the negative class than the positive class - this means that the model was better at identifying all relevant instances for the negative class. The LR, GNB, and MLP models resulted in especially high recall scores in this case. However, the models all had low recall for the positive class - they performed poorly on recalling relevant instances of the positive class. For medical diagnosis applications, we especially want to see high recall for the positive class. With respect to F1-scores, GNB, MLP, and LR had the highest F1-scores (0.67) for the negative class, while RF had an F1-score that was slightly better for the positive class. Notably, LR had the highest F1-score for the negative class but the lowest F1-score for the positive class.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	0.60	0.61 (0.58, 0.65)	0.60 (0.78, 0.42)	0.59 (0.66, 0.51)
RF	0.56	0.56 (0.56, 0.56)	0.56 (0.60, 0.52)	0.56 (0.58, 0.54)
GNB	0.60	0.62 (0.58, 0.66)	0.60 (0.79, 0.40)	0.58 (0.67, 0.50)
MLP	0.60	0.61 (0.58, 0.66)	0.60 (0.79, 0.41)	0.59 (0.67, 0.50)
k-NN	0.54	0.54 (0.54, 0.53)	0.54 (0.55, 0.53)	0.54 (0.54, 0.53)
LR	0.60	0.62 (0.57, 0.67)	0.60 (0.81, 0.38)	0.58 (0.67, 0.48)

Table 2. Results of the binary classification models predicting the presence/absence of CVD using the non-invasive dataset. The first value in the precision, recall, and F1-score columns refers to the weighted average for each respective score. The values inside the parentheses refer to the scores for the negative and positive classes, respectively.

Using regression models to predict other target variables, such as BP and BMI, proved to be more challenging (Table 3). Five regression models were applied to the non-invasive dataset. For predictions on BMI, we removed the height and weight variables. Likewise, for predictions of BP, we removed the systolic and diastolic BP variables. Though we used hyperparameter optimization techniques such as grid search, the regression models in this study yielded mostly negative R^2 scores. Linear regression was the only regression model tested that yielded a positive R^2 score for predicting both BP and BMI.

Regression Model	R^2 **Score**$_{\mathbf{BP}}$	R^2 **Score**$_{\mathbf{BMI}}$
Linear Regression	0.102	0.060
SVR	-0.024	n/a
MLP	-0.178	0.076
RF	-0.136	-0.076
Extra Trees	-0.260	-0.098

Table 3. Scores obtained by BP and BMI regression models on the non-invasive dataset. The n/a value for SVR predictions on BMI is due to the failure of the SVR algorithm in finding a reasonable termination point.

After implementing models to predict CVD on the non-invasive dataset, we also built models on the invasive dataset to predict CVD. The predictive ability of our non-invasive and invasive datasets for the original CVD classification task was compared. (Table 4). The invasive model outperformed the non-invasive model on all measures. While the non-invasive model had a performance around 0.60 for all of accuracy, precision, recall, and F-1 score, the invasive model achieved 0.87 or 0.88 for all of these measures. We chose to focus on the performance of logistic regression since it performed reasonably well and has a reduced computational complexity, compared to other models. This allowed for quick fine-tuning of our code. The performances of the LR models are reported in Table 4.

Dataset	Accuracy	Precision	Recall	F1-Score
Non-Invasive	0.60	0.62 (0.57, 0.67)	0.60 (0.81, 0.38)	0.58 (0.67, 0.48)
Invasive	0.87	0.88 (0.82, 0.93)	0.87 (0.93, 0.81)	0.87 (0.87, 0.87)

Table 4. Comparing performance of logistic regression models trained on the non-invasive and invasive datasets for CVD. The first value in the precision, recall, and F1-score columns refers to the weighted average for each respective score. The values inside the parentheses refer to the scores for the negative and positive classes, respectively.

Lastly, we investigated how important each of these features were in the model. We focus specifically on the logistic regression models here because, as previously mentioned, they have easily interpretable feature coefficients, and they also performed well on the classification task, as compared to the other models evaluated (Figure 5). For the non-invasive dataset, cholesterol levels were important features. Low cholesterol levels were associated with assignment to the negative class (coefficient of -0.594) while high cholesterol levels were associated with a positive class assignment (0.951). Height, gender, and systolic BP all were associated with a positive class assignment (all 0.12), while activity level (-0.204) and diastolic BP (-0.523) were associated with a negative class assignment. For the invasive dataset, chest pain was an important feature. Level 3 and 2 chest pain were strongly associated with the positive class (coefficients of 1.274 and 1.260, respectively). ECG measurements, maximum heart rate, and fluoroscopy results were other strong influential factors in the detection of the positive class, having coefficients ranging from 0.906 to 0.397. Fluoroscopy results were also strongly influential in detecting the negative class, or lack of CVD (coefficients of -1.909, -1.625, and -1.169 for varying type of fluoroscopy results). Oldpeak, which is a measurement taken from an ECG, was also strongly associated with the negative class (-1.225), as was sex (-1.039), exercise-induced angina (-0.906), blood pressure(-0.732), and thalium stress test results (-0.603).

Discussion

Regarding the results of the binary CVD classification, we attained lower accuracies with the non-invasive dataset as compared to the invasive dataset. These results may suggest that invasive data is crucial in the prediction of CVD. In terms of individual model performance, SVM, GNB, MLP, and LR achieved the most accurate classifications, which we had expected due to these models' generalizability and robustness. For example, we expected the MLP to perform well, as the depth and flexibility afforded from its neural network structure contribute to its reliability in binary classification problems. On the other hand, k-NN's poor performance could be attributed to a lack of proper hyperparameter (k) optimization. The low accuracies across the board could also be a result of the quality of the dataset itself.

One of our original objectives was to predict certain target variables, such as BMI and BP using various regression models. However, we chose not to expand upon these results, as this information can be more easily attained during routine clinical visits and therefore does not necessitate the use of machine learning models.

From our analysis of feature importance obtained from LR models, severe chest pain and high cholesterol were found to be highly associated with the presence of CVD,

Non-invasive dataset

Feature	Coefficient
chol_3	0.951
gluc_2	0.352
gluc_1	0.169
height	0.125
ap_hi	0.124
gender	0.122
age	0.075
weight	0.074
chol_2	0.026
alco	-0.089
gluc_3	-0.139
smoke	-0.199
active	-0.204
ap_lo	-0.523
chol_1	-0.594

Invasive dataset

Feature	Coefficient
cp_3	1.274
cp_2	1.260
slope_2	0.960
thalach	0.820
thal_2	0.753
cp_1	0.632
ca_4	0.500
restecg_1	0.397
fbs	0.301
thal_1	0.210
age	0.115
chol	-0.189
restecg_2	-0.200
slope_1	-0.290
thal_3	-0.603
trestbps	-0.732
exang	-0.906
sex	-1.039
ca_3	-1.169
oldpeak	-1.225
ca_1	-1.625
ca_2	-1.909

Table 5. Logistic regression models were applied on both datasets and the resultant feature coefficients for each feature are displayed here. chol_, gluc_, cp_, slope_, thal_, ca_, and restecg_ are dummy variables for cholesterol, glucose, chest pain, ECG slope, thalium stress test result, fluoroscopy test result, and ECG result, respectively. ap_hi and ap_lo refer to systolic and diastolic blood pressure, respectively. thalach, fbs, trestbps, exang, and oldpeak refer to max heart rate, fasting blood sugar, resting blood pressure, exercise-induced angina, and ST depression on an ECG, respectively.

based on both the invasive and non-invasive datasets. This concurs with pre-existing medical knowledge, as high cholesterol causes CVD in many cases and chest pain is a known symptom of CVD. Additionally, many of the fluoroscopy and ECG features were also strongly associated with the presence of CVD, which implies that data collected from invasive procedures is necessary in accurate CVD diagnoses. On the other hand, some of the feature associations were not as consistent with current medical evidence or were contrary between datasets. For instance, alcohol use and smoking were found to be more strongly associated with predicting the absence of CVD in the non-invasive dataset, while these two features are typically associated with a greater risk for CVD in the medical realm. This discrepancy may speak either to the questionable quality of the dataset or a relationship between these features and CVD presence that should be explored further in future research. Gender was also found to be associated with predicting the presence of CVD in the non-invasive dataset but very important in detecting the absence of CVD in the invasive dataset. This conflict could be better explored and clarified with further research.

Conclusion

As this paper investigates CVD as a whole, which is useful to patients in terms of a general diagnosis, future approaches could include predicting the presence of different types of CVD, such as coronary artery disease, arrhythmia, or stroke. These types of models could provide patients with more specific details about their conditions. Furthermore, though we did not find the results of the BP and BMI predictions applicable in this context, perhaps the implementation of a mobile application that uses regression models to predict these features could be beneficial to patients who do not have the financial means to attend regular clinical appointments. Our findings could also streamline CVD diagnosis. With a knowledge of which features are closely associated with the presence of heart disease, physicians could more efficiently diagnose and treat their patients.

Regarding our binary classification findings, due to the unknown origin of the non-invasive dataset, the data may not necessarily be as reliable and representative of the population. Therefore, the utility of non-invasive data cannot be completely ruled out and should be investigated further with more reliable datasets. With a more reliable dataset, this project could potentially be repeated to gain more conclusive results. If non-invasive data proves to be more useful with such valid data, then people can obtain a CVD diagnosis without needing to visit a clinic. This would aid those who may not have access to healthcare or the ability to pay for costly medical tests. As of now, however, we conclude that non-invasive data should be used in conjunction with invasive procedural data to attain accurate diagnoses.

Supporting information

Our code and further information on this project can be found at `https://sites.google.com/dartmouth.edu/cardiodiseasemodel/`.

References

1. World Health Organization. Global Status Report on Noncommunicable Diseases. Geneva, Switzerland: World Health Organization, 2014.

2. Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. AHA Journals, 2019;139(10).

3. Preventable Deaths from Heart Disease & Stroke. Centers for Disease Control and Prevention. www.cdc.gov/vitalsigns/heartdisease-stroke/index.html.

4. Bittner V. The New 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease. AHA Journals, 2019.

5. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, Himmelfarb CD, Khera A, Lloyd-Jones D, McEvoy JW, Michos ED, Miedema MD, Muñoz D, Smith SC Jr, Virani SS, Williams KA Sr, Yeboah J, Ziaeian B. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines [published online ahead of print March 17, 2019]. Circulation.

6. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2013; 135(11): 1–50.

7. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008; 336(7659): 1475–82.

8. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation 2008; 117(6): 743–53.

9. Ridker P, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The reynolds risk score. JAMA 2007; 297(6): 611–9.

10. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto AM, Kastelein JJP, et al. Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. New England Journal of Medicine 2008; 359(21): 2195–207.

11. US Preventive Services Task Force. Screening for Cardiovascular Disease Risk With Electrocardiography: US Preventive Services Task Force Recommendation Statement. JAMA. 2018;319(22):2308–2314. doi:10.1001/jama.2018.6848

12. Obermeyer Z, Emanuel EJ. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. The New England journal of medicine 2016; 375(13): 1216–9.

13. Kukar M, Kononenko I, Grošelj C, Kralj K, Fettich J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial Intel in Med, 1999;16(1).

14. Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Exp Sys with Apps, 2009;36(4).

15. Soni S, Ansari U, Sharma, D, Soni J. Predictive data mining for medical diagnosis: an overview of heart disease prediction. Intl J Comp Apps, 2011;17(8).

16. Janosi A, Steinbrunn W, Pfisterer, M, Detrano, R. Heart disease UCI dataset. Kaggle dataset, 2019. https://www.kaggle.com/ronitf/heart-disease-uci.

17. Ulianova S. Cardiovascular disease dataset. Kaggle dataset, 2019. https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

18. C. Blake, E. Keogh, and C.J. Merz, "UCI repository of machine learning databases" [http://www.ics.uci.edu/∼mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA, 1998.