Journal of Data Analysis and Information Processing, 2023, 11, 1-10
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Heart Disease Prediction Using Machine Learning Algorithms with Self-Measurable Physical Condition Indicators

Huating Sun[1], Jianan Pan[2]

[1]Department of Geography, University of Washington, Seattle, USA
[2]School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China
Email: paulsun8b8@gmail.com, Pjn18260178293@163.com

How to cite this paper: Sun, H.T. and Pan, J.N. (2023) Heart Disease Prediction Using Machine Learning Algorithms with Self-Measurable Physical Condition Indicators. *Journal of Data Analysis and Information Processing*, **11**, 1-10.
https://doi.org/10.4236/jdaip.2023.111001

Received: October 10, 2022
Accepted: January 15, 2023
Published: January 18, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

In recent years, the number of cases of heart disease has been greatly increasing, and heart disease is associated with a high mortality rate. Moreover, with the development of technologies, some advanced types of equipment were invented to help patients measure health conditions at home and predict the risks of having heart disease. The research aims to find the accuracy of self-measurable physical health indicators compared to all indicators measured by healthcare providers in predicting heart disease using five machine learning models. Five models were used to predict heart disease, including Logistics Regression, K Nearest Neighbors, Support Vector Model, Decision tree, and Random Forest. The database used for the research contains 13 types of health test results and the risks of having heart disease for 303 patients. All matrices consisted of all 13 test results, while the home matrices included 6 results that could test at home. After constructing five models for both the home matrices and all matrices, the accuracy score and false negative rate were computed for every five models. The results showed all matrices had higher accuracy scores than home matrices in all five models. The false negative rates were lower or equal for all matrices than home matrices for five machine learning models. The conclusion was drawn from the results that home-measured physical health indicators were less accurate than all physical indicators in predicting patients' risk for heart disease. Therefore, without the future development of home-testable indicators, all physical health indicators are preferred in measuring the risk for heart diseases.

Keywords

Machine Learning, Data Visualization, Feature Engineering, Health, Heart Disease

1. Introduction

Heart disease, caused by abnormal heart and blood vessel conditions, is widely considered a direct threat to human life and health. It is one of the significant diseases exerting irreversible effects on many middle-aged and older people, in which fatal complications are highly likely to result [1]. Makino states that the absolute risk of cardiovascular heart disease is associated with disability and death among people 65 years or older [2]. The World Health Organization (WHO) declared an estimated 17.7 million people died from cardiovascular disorders in 2015, accounting for one-third of all deaths that year [3]. According to the Australian Bureau of Statistics, heart ailment was one of Australia's two highest causes of mortality [4]. As its extremely negative influence on human health, a great deal of effort has been devoted to the study of the onset of heart disease, trying to prevent and reduce the incidence of heart disease with a timely and efficacious method. Moreover, purpose to prevent the adverse effects of heart disease, it is recommendable to use sophisticated equipment to detect potential heart risks in advance. Currently, qualified health organizations can conduct many tests, including blood tests, echocardiography, chest X-Rays, magnetic resonance imaging (MRI), electrocardiogram, physical examination, and exercise stress test that provide medical doctors with valuable information in their diagnosis and their views on the patient's heart failure risk level [5].

There are several risk factors for heart failure, corresponding to different test indexes. A significant amount of relevant research has been carried out to reveal the potential attributes of a heart attack. Sex, age, smoking, hypertension, and diabetes depend on heart disease [6]. Peter *et al.* [7] suggest that indexes including blood pressure, total cholesterol, and age are essential in predicting coronary heart disease. The effects of sex differences on traditional cardiovascular risk factors are considered to be notable [8]. Heart rate is also a powerful indicator of a patient's potential heart attack risk [9]. The attributes of heart disease could be approximately divided into two types according to whether the indicators could be measured at home. It is considered worthwhile to compare the accuracy of indicators measured at home with those measured in hospitals, which is useful for future tests of heart disease.

Computational technology and statistical approach have been popular in discovering the relationship between heart diseases and patients' health conditions [10] [11]. They can help predict the potential risk of heart disease based on the patient's underlying physical condition in advance, thereby reducing the probability of dying from a heart attack. Many statistical methods based on computer calculation have been applied to predict heart attacks [12]. Due to its high accuracy, SVM has been prevalently applied as a classification method to predict heart attacks [13]. Akkaya used logistic regression and the k-NN algorithm to estimate heart failure and accomplished compromising outcomes [13]. With the adoption of Random Forest, the best accuracy of 82.18% has been achieved by modification of feature selection [14]. These algorithms have been proved to predict the risk of heart disease effectively, which helps researchers and doctors

make better judgments about heart disease.

Although these machine learning technique has been acknowledged and refined continuously to increase the performance of prediction, few investigators has examined and compare the accuracy of home-tested versus in-hospital measures for predicting heart disease risk. Few investigators have examined the relative accuracy of home-tested versus in-hospital measures for predicting heart disease risk. If the indicators measured at home can well predict the patient's risk of heart disease, then the patient can be tested by themselves or their families instead of having to go to the hospital for testing. Therefore, the innovation of this article lies in that not only did it use five machine learning algorithms to regress data on heart patients, but it also compared the contribution of these algorithms to the prediction of heart disease measured at home and measured in the hospital.

This study aims to compare the patient's physical condition indicators measured at home and in the hospital, using 5 different prediction methods to explore their accuracy of heart disease prediction. Moreover, the research question "How is machine learning algorithms' performance with only self-measurable physical condition indicators compared to algorithms with all physical condition indicators?" would be answered accordingly.

2. Data Description

We used the data from the Cleveland heart data set from the UCI machine learning repository. The data we selected is made up of 14 variables and 303 instances. Overall speaking, there are 13 variables and 1 categorical response variables (target). Among these variables, numerical variables are age, trtbps, chol, thalach, old peak; Categorical variables are sex, exang, cp, fbs, rest_ecg, slp, thall, target. The table below illuminates the meaning of each variable. Detailed information could be seen in Table 1.

From Figure 1 we can see that in the data set, most patients with heart attack are aging between 50 and 60, while only few people have heart failure aged under 30 or above 70. The range of this attribute is 29 - 77, illustrating the wide span of age.

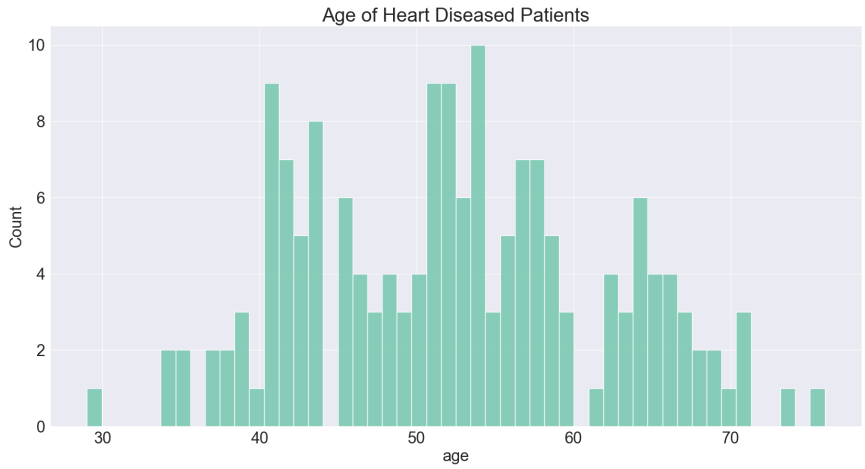

Figure 1. Age of heart diseased patients.

The Chol means cholestoral of patients, fetched via BMI sensor. According to Figure 2, it seems that most patients' cholestoral is around 230 mg/dl and the whole distribution shows a slightly right skewness.

According to Figure 3, most maximum heart rates of patients gathers between 140 to 180. Some particular patients have extremely low and high heart rate, specifically lower than 100 and surpassing 200.

When it comes to resting blood pressure (Figure 4), a great number of patients have resting blood pressure around 100 to 140. Only a few have abnormal values of around 160 mm/Hg and below 100 mm/Hg.

Table 1. Variable description.

Variable Name	Descriptions	Rage of value
Age	Age of the patient	29 - 77
Sex	Sex of the patient	0, 1
exang	Exercise induced angina: 1 = yes; 0 = no	0, 1
cp	Chest Pain type : 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic;	1, 2, 3, 4
trtbps	Resting blood pressure (in mm/Hg)	94 - 200
chol	Cholestoral in mg/dl fetched via BMI sensor	126 - 564
fbs	Fasting blood sugar: 1 = fasting blood sugar > 120 mg/dl; 0 = fasting blood sugar ≤ 120 mg/dl	0, 1
restecg	Resting electrocardiographic results: 0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria;	0, 1, 2
thalach	Maximum heart rate achieved	71 - 202
old peak	ST depression induced by exercise relative to rest	0 - 6.2
slp	The slope of the peak exercise ST segment	0, 1, 2
thall	Thalassemia: 0 = null; 1 = fixed defect; 2 = normal; 3 = reversable defect	0, 1, 2, 3
target	Output: 0 = less chance of heart attack; 1 = more chance of heart attack	0, 1

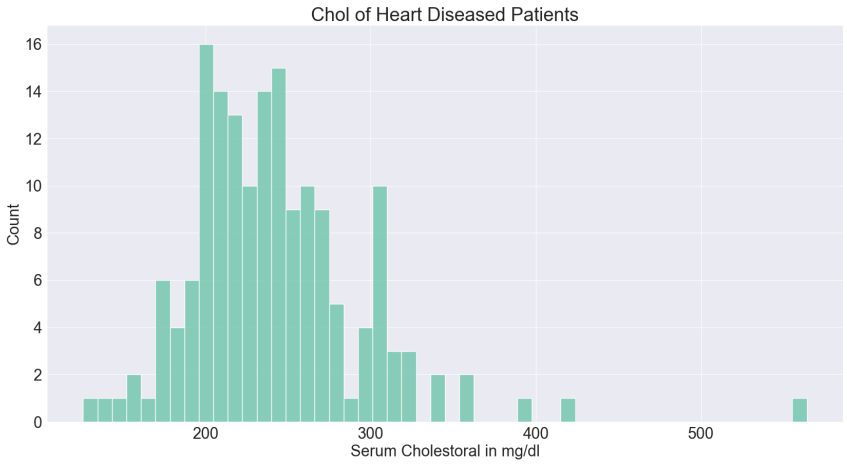

Figure 2. Chol of heart diseased patients.

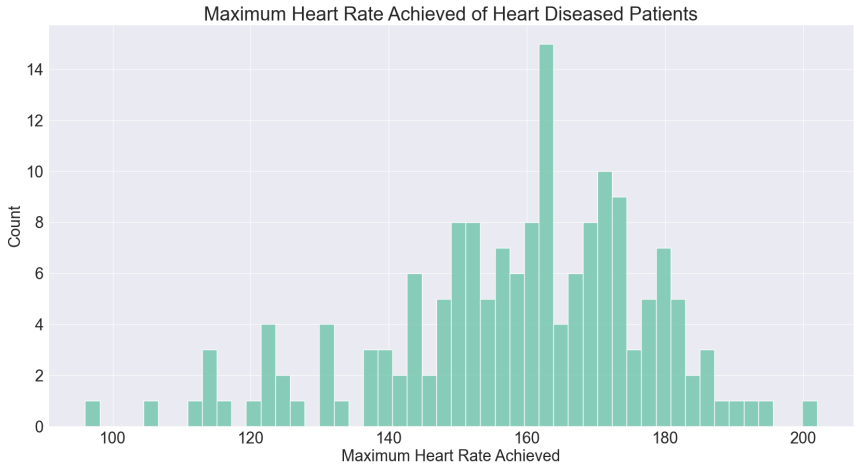

Figure 3. Maximum heart rate achieved of heart diseased patients.

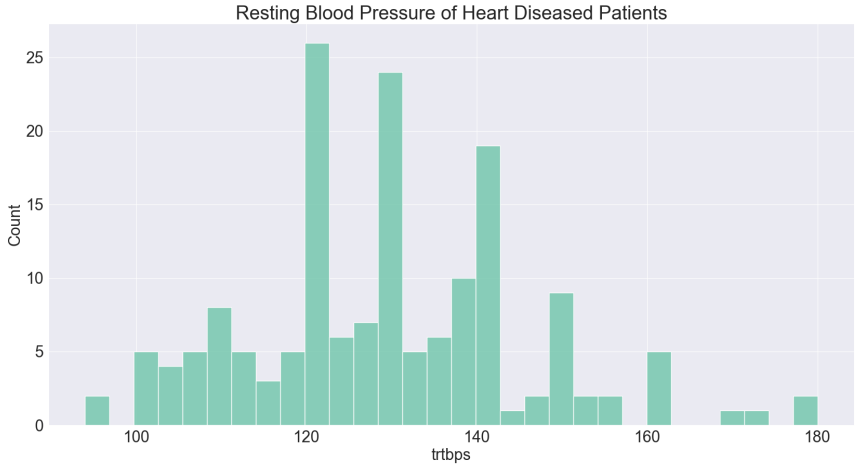

Figure 4. Resting blood pressure of heart diseased patients.

3. Methodology

3.1. Data Processing

For data description, the research utilized the describe function and pandas profiling in Python to summarize the dataset. The raw data contained 14 variables for 303 patients. Chi-square values, extra-tree classifiers, and correlation matrices were measured to conduct data analysis. The Chi-square values and correlation matrices showed that no variables were highly correlated, and all variables were selected for model building. Moreover, all numerical variables were scaled to normal using *Standard Scaler*.

The 13 independent variables were divided into home matrices and all matrices. Home matrices consisted of 6 variables—age, sex, resting blood pressure, cholesterol, fasting blood sugar, and thalassemia. All matrices included all 13 independent variables. The research created the training set and test sets with 80% training data and 20% testing data.

The helper function was used in Python to show each model's accuracy score,

false negative rate, and confusion matrix. The accuracy score was used to measure the percentage of correctly predicted patients who had or did not have a risk for heart disease. The score showed the accuracy of each model in predicting the correct heart disease risks for patients. The false negative rate measured the percentage of patients with a high risk for heart disease but was mispredicted as having a low risk for heart disease. The false negative rate was significant because misprediction may lead to late treatment for the patients. Those values were used in the final model comparison to conclude the accuracy of self-measured home matrices compared to all matrices.

3.2. Machine Learning Algorithms

The research built five models for both the home matrices and all matrices.

3.2.1. Logistics Regression

Logistics Regression is a model for predicting a binary outcome utilizing the observations of a data set. The research selected this model because the output variable is a binary outcome taking either the high risk or no risk for heart disease. The *Logistic Regression* from the *sklearn* package in Python was used to build the model. Library for large linear classification was chosen for logistics models because the dataset size was relatively small.

3.2.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a classification algorithm that tests the likelihood of a data point belonging to a group according to the distance to the nearest point. The research chose 1 to 20 as the number of neighbors. The K Neighbors Classifier Scores were calculated for each number of neighbors. The line chart using the number of neighbors as x and the K Neighbors Classifier Scores as y was created. The research chose K equal 8 since it had the highest K Neighbors Classifier Score.

3.2.3. Support Vector Machine

Support Vector Machine was chosen as one of the models because it is an algorithm for classification and regression. The research used *svm* from *sklearn.svm* package in Python. The Radial basis function kernel was selected, gamma equaled 0.01, and the ragularization parameter equaled 1 for the two machine learning models.

3.2.4. Decision Tree

Decision tree was chosen because it is a nonparametric machine learning model for classification and regression. The research drew the line graph using the number of maximum depth from 1 to 30 as x and Decision Tree Classifier Score as y. Maximum depth equal to 10 was picked for the model building because it has the highest scores.

3.2.5. Random Forest

Random Forest is an algorithm consisting of decision trees. *Random Forest Clas-*

sifier from the *sklearn. ensumble* package was used to build the home and all matrices models. The number of estimators equaled 1000 in both the home and all matrices models.

4. Result

Raw data, after some preprocessing, are fed into machine learning algorithms. Afterward, the accuracy score and the false negative rate are obtained.

4.1. Accuracy

According to Table 2, the Logistic Regression and Support Vector Model have the highest accuracy score at 88.52% within the machine learning algorithms with all physical condition indicators. In comparison, the Decision Tree has the lowest accuracy score with only 85.25%. Within the machine learning algorithms with only physical condition indicators measured at home, Logistic Regression has the highest accuracy score at 73.77%, while the Support Vector Model has the lowest accuracy at only 68.85%.

After comparing the accuracy between machine learning algorithms with only physical condition indicators measured at home and algorithms with all physical condition indicators, it is concluded that algorithms with only physical condition indicators measured at home do not perform as accurately as algorithms with all physical condition indicators. The difference in accuracy ranges from 14.75% to 19.67%.

4.2. False Negative Rate

From the false negative rate perspective (Table 3), it is observed that the Decision Tree has the highest false negative rate within the algorithms with all physical condition indicators. In contrast, Logistic Regression has the lowest false negative rate. Within the algorithms with only physical condition indicators measured at home, K Nearest Neighbors and Random Forest have the highest false negative rate, while Decision Tree has the lowest false negative rate.

Table 2. The table shows the accuracy score of machine learning algorithms with all physical condition indicators and only self-measurable indicators. Orange represents the algorithm with the highest accuracy score. Green represents the algorithm with the lowest accuracy score.

Model	All Indicators Accuracy	Self-Measurable Indicators Accuracy
Logistic Regression	88.52%	73.77%
K Nearest Neighbors	86.89%	70.49%
Support Vector Model	88.52%	68.85%
Decision Tree	85.25%	70.49%
Random Forest	86.89%	72.13%

Table 3. The table shows the false negative rate of machine learning algorithms with all physical condition indicators and only self-measurable indicators. Orange represents the algorithm with the highest false negative rate. Green represents the algorithm with the lowest false negative rate.

Model	False Negative Rate—All Indicators	False Negative Rate—Self-Measurable Indicators
Logistic Regression	8.82%	26.47%
K Nearest Neighbors	11.76%	29.41%
Support Vector Model	11.76%	23.53%
Decision Tree	17.65%	17.65%
Random Forest	11.76%	29.41%

After comparison, it is concluded that machine learning algorithms with all physical condition indicators have a much lower false negative rate than algorithms with only physical condition indicators measured at home. Note that the false negative rate for the Decision Tree is the same for both groups. This is probably due to the randomness of the data splitting process, as the test set is only 20% of the entire data set, which is about 60 data samples. The difference between the algorithms ranges from 0% to 17.65%.

5. Conclusions and Discussion

5.1. Conclusion

To answer the research question of this study, it is concluded that the machine learning algorithms with only self-measurable physical condition indicators do not predict as accurately as machine learning algorithms with all physical condition indicators. Not only do algorithms with self-measurable physical condition indicators not predict the heart disease outcome as accurately as algorithms with all physical condition indicators, but they are also more likely to falsely predict not having heart disease among patients with heart disease. Thus, machine learning algorithms with only self-measurable physical condition indicators should not be used until more indicators are measurable at home in the future.

5.2. Study Limitation

The findings of this study have to be seen in light of some limitations. It is noteworthy that the dataset used in this is a subset of the original database, which contained 76 attributes instead of 14, which is used in this study. Within the original 76 attributes, other attributes could be measured at home and thus improve the accuracy and reduce the false negative rate of the machine learning algorithms with only self-measurable physical condition indicators.

5.3. Future Work

The limitations of this study have indicated the following areas as recommenda-

tions for future work. First, include other health attributes from the original dataset to discover the machine learning algorithm with the highest accuracy and lowest false negative rate. Second, since every patient has different health conditions, it is recommended to group the patients with similar health conditions and ages to investigate each machine learning algorithm's accuracy and false negative rate.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Heron, M. (2012) Deaths: Leading Causes for 2008. *National Vital Statistics Reports*: *From the Centers for Disease Control and Prevention*, National Center for Health Statistics, National Vital Statistics System, **60**, 1-94.

[2] Makino, K., Lee, S., Bae, S., Chiba, I., Harada, K., Katayama, O., Shinkai, Y. and Shimada, H. (2021) Absolute Cardiovascular Disease Risk Assessed in Old Age Predicts Disability and Mortality: A Retrospective Cohort Study of Community-Dwelling Older Adults. *Journal of the American Heart Association*, **10**, e022004. https://doi.org/10.1161/JAHA.121.022004

[3] WHO (2017) Cardiovascular Diseases. http://www.who.int/mediacentre/factsheets/fs317/en/

[4] ABS (2009) Causes of Death, Australia. Australian Bureau of Statistics. http://abs.gov.au/ausstats/abs@.nsf/Products/696C1CF9601E4D8DCA25788400127BF0?opendocument

[5] AHA (2017) American Heart Association. http://www.heart.org

[6] Liu, X., Wang, X.L., Su, Q., Zhang, M., Zhu, Y.H., Wang, Q.G. and Wang, Q. (2017) A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. *Computational and Mathematical Methods in Medicine*, **2017**, 1-11. https://doi.org/10.1155/2017/8272091

[7] Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. and Kannel, W.B. (1998) Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, **97**, 1837-1847. https://doi.org/10.1161/01.CIR.97.18.1837

[8] Liu, W., Tang, Q., Jin, J., *et al.* (2021) Sex Differences in Cardiovascular Risk Factors for Myocardial Infarction. *Herz*, **46**, 115-122. https://doi.org/10.1007/s00059-020-04911-5

[9] Lee, H.G., Noh, K.Y. and Ryu, K.H. (2007) Mining Biosignal Data: Coronary Artery Disease Diagnosis Using Linear and Nonlinear Features of HRV. In: *Emerging Technologies in Knowledge Discovery and Data Mining*, *PAKDD* 2007, *Lecture Notes in Computer Science*, Vol. 4819, Springer, Berlin, Heidelberg.

[10] Nahar, J., Imam, T., Tickle, K.S. and Chen, Y.-P.P. (2013) Association Rule Mining to Detect Factors Which Contribute to Heart Disease in Males and Females. *Expert Systems with Applications*, **40**, 1086-1093. https://doi.org/10.1016/j.eswa.2012.08.028

[11] Desai, F., Chowdhury, D., Kaur, R., Peeters, M., Arya, R.C., Wander, G.S., Gill, S.S. and Buyya, R. (2022) HealthCloud: A System for Monitoring Health Status of Heart Patients Using Machine Learning and Cloud Computing. *Internet of Things*, **17**,

Article ID: 100485. https://doi.org/10.1016/j.iot.2021.100485

[12] Nahar, J., Imam, T., Tickle, K.S. and Chen, Y.P.P. (2013) Computational Intelligence for Heart Disease Diagnosis: A Medical Knowledge Driven Approach. *Expert Systems with Applications*, **40**, 96-104. https://doi.org/10.1016/j.eswa.2012.07.032

[13] Xing, Y.W., Wang, J., Zhao, Z.H. and Gao, Y.H. (2007) Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. *Convergence Information Technology*, Gwangju, 21-23 November 2007, 868-872. https://doi.org/10.1109/ICCIT.2007.204

[14] Akkaya, B., Sener, E. and Gursu, C. (2022) A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques. 2022 *International Congress on Human-Computer Interaction, Optimization and Robotic Applications* (*HORA*), Ankara, 9-11 June 2022, 1-8. https://doi.org/10.1109/HORA55278.2022.9799978