Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230745

Predicting pulse wave age from cardiovascular characteristics using machine learning algorithms

Ruihan Wang

University of California, Los Angeles, Los Angeles CA 90024, USA

ruihanw3@g.ucla.edu

Abstract. Arterial pulse waves are an essential and informative source of data for measuring cardiovascular health and are currently available on various devices, such as smartwatches. In addition, most experiments have focused on pulse wave velocity (PWV) and arterial pulse waveform (APW). However, in recent years, due to many factors, such as high work pressure among young people, cardiovascular diseases have become more prevalent. As a result, we need more research on pulse age to help patients focus on the consistency of their biological age with their legal age. Therefore, this paper introduces a machine learning framework to predict the age of patients according to the HR and SV data in the pulse wave of existing patients. At the beginning of the experiment, we attempted to use K-nearest neighbours (KNN), logistic regression (LR), and random forest (RF) models. Then we compared the accuracy of patient age estimates; RF had the best performance and was used as the final model. In addition, HR and SV were selected as the main features to predict patients' age according to the feature's importance. The final experimental results indicate that the RF model can predict the results with up to 1 accuracy only with two features, i.e., HR and SV. In particular, this is the point that we need to continue to pay attention to in the future to determine whether the prediction accuracy of this model can reach 1. Is it because the model is overly optimistic or because the selected HR and SV features strongly correlate with age? In the future, we can introduce the concept of actual body age and attract young people's attention to their body age. For example, more people will be urged to take a quick and efficient pulse diagnosis, complete a simple physical examination, and give rapid feedback by placing devices in places where young people gather, such as shopping malls. In this case, users can immediately know what part of their body may have problems. In general, our experiment demonstrates that patients' pulse wave ages could be well predicted based on their HR and SV data with high accuracy through the ML model of RF.

Keywords: pulse wave age, machine learning algorithm, cardiovascular characteristics.

1. Introduction

The arterial pulse wave (PW) is the wave of rhythmic arterial pressure felt while palpating an artery that is produced when the left ventricle contracts and travels through the arterial tree. [1] It is affected by the heart, whose features like heart rate (HR) and stroke volume (SV) affect its length and shape, as well as the vasculature, whose features like arterial stiffness and wave reflections affect it. Therefore, the arterial pulse wave (PW) is a valuable data source on cardiovascular (CV) health, widely measured by both consumer and medical gadgets. Additionally, it is commonly seen in consumer electronics like

DOI: 10.54254/2755-2721/13/20230745

smartwatches and fitness wristbands. [2] The PW can therefore provide valuable information on CV function in both clinical settings and everyday life. [3] This motivates this study to predict the relationship between the patient's age and pulse wave based on the RF and ML models using the existing data that contains patients' age and pulse wave indicators.

So far, pulse waves' different features have been applied in various areas. Among them, pulse wave velocity (PWV) is the most common measurement index of the pulse wave present, and many diseases are predicted based on PWV. The PWV is directly correlated with wall stiffness, changes inversely with artery radius, and is connected to wall thickness and elasticity. The most prevalent indicator, carotid-femoral PWV, is oblivious to changes in the proximal aorta, where most of the systolic damping function occurs. As a result, modifications in the carotid-femoral PWV imply adaptations of more distal arteries, most likely representing late effects of hypertension. [4]

Another famous study in the pulse wave area is the arterial pulse waveform (APW), another standard index to detect CV health and assess CV risk. Due to their utterly non-contact nature and ability to monitor skin surface displacement, particularly at the carotid artery location, optical sensors are an appealing instrumental option for APW evaluation. [5] However, the operator's knowledge and experience affect how well APW detection may be performed. Since the subjective criterion used to choose the pulse section to be analysed for operator variability causes the systematic variances between the operators and variations of a specific operator's score on a specific patient, these are caused by the subjective effect. [5]

Previous studies indicate that most studies have focused on the pulse wave's speed, waveform, and other characteristics to predict disease. However, little attention has been paid to the correlation between age and pulse waves, but now many diseases are proven to be age-related. Especially in recent years, cardiovascular diseases have become more common in young people, and the average age of patients is decreasing. According to research being presented at the American College of Cardiology's 68th Annual Scientific Session, new statistics not only confirm this trend but also show that more heart attacks are happening to those under the age of 40. [6] Additionally, the percentage of very young people experiencing a heart attack increased across the 16-year research period (2000–2016), rising by 2% annually for the past ten years. [6] Therefore, in our work, we use these two features as indicators to predict the age of patients based on Random Forest ML models.

In this experiment, we input the patient's pulse wave according to a series of index values, use the ML model to infer that the patient's pulse wave's characteristics generally exist in patients of which age, and then get the "pulse wave age" of the patient. The patient's health is estimated by comparing the patient's actual age with the age at which the pulse wave is expected. Moreover, it predicted possible cardiovascular disease based on the age at which the pulse wave appears.

This experiment can be widely used in places with a large number of young people, such as shopping malls and bars, to attract experimental subjects by detecting the pulse wave age. A small machine can be set up to detect pulse waves and, through the ML model, give feedback to the user on their pulse wave age and identify potential cardiovascular disease. In addition, this model can help future researchers conduct experiments on age-related pulse wave disease.

In the present work, we introduce a random forest model to predict the patient's age with their PWV data accurately. The result shows that the predicted value obtained by the experiment reaches 1. In addition, in the future, the model can be applied to places outside hospitals with fewer required features, faster operation, and higher accuracy to strengthen the awareness of physical examination among young groups.

2. Method

2.1. Problem formulation

This research aims to test pulse waves to obtain certain features' values and then use machine learning models to speculate the actual "age" of the pulse wave to remind patients that they may have disease risk and need to adjust their sleep schedule as well as seek medical treatment in time. The input of the

DOI: 10.54254/2755-2721/13/20230745

model in this experiment is typical cardiovascular characteristics X with length (30, 4347), and it outputs an age range Y from 25 to 75 years old in 10-year increments with length (1,4347).

2.2. Data

The dataset used in this research comes from the "Pulse Wave Database (PWDB): A database of arterial pulse waves representative of healthy adults." [3] PWDB is a simulated arterial pulse wave database aiming to represent pulse wave samples from 4374 real healthy adults ranging in age from 25 to 75 years old in 10-year increments. Peter H. Charlton et al. used typical cardiovascular characteristics of healthy participants in each age group to develop the baseline sets of PWDB. The chart below is the feature we used in the next with their mean and standard deviation values.

Table 1. The mean and standard deviation value of used features.

	HR [bpm]		SV [ml]		CO [l/min]		LVET [ms]		RFV [ml]		AIx [%]		SVR [10^6 Pa s / m3]	
age	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
25	72.86	9.11	66.74	13.18	4.86	1.14	282.74	23.24	0.73	0.02	2.17	10.32	154.13	34.33
35	76.74	9.18	64.09	12.49	4.92	1.14	283.11	23.30	0.74	0.04	7.91	10.64	158.47	36.18
45	77.67	9.08	61.56	11.83	4.78	1.07	282.74	23.37	0.74	0.04	17.13	10.59	168.82	37.87
55	77.30	9.16	59.22	11.19	4.58	1.02	282.59	23.38	0.72	0.00	26.14	10.17	176.73	40.07
65	76.40	9.09	56.38	10.49	4.31	0.96	282.59	23.20	0.80	0.07	34.71	9.87	186.11	42.50
75	74.40	9.06	54.08	9.85	4.02	0.87	282.35	23.22	0.76	0.06	42.39	10.31	196.46	43.72
25	72.86	9.11	66.74	13.18	4.86	1.14	282.74	23.24	0.73	0.02	2.17	10.32	154.13	34.33

This table includes the mean and standard deviation values for the seven features needed in the experiment, which are separated into different age ranges.

2.2.1. Feature normalization. The gradient descent method was used to find the optimal solution, which resulted in many iterations to converge. Also, the different parameters' magnitudes negatively affect the classification and prediction results. Therefore, using the MinMaxScaler normalization method to change the number to a decimal between (0, 1) predicts a higher accuracy score. [2] The MinMaxScaler formula is

$$\frac{x_i - \min(x)}{\max(x) - \min(x)}$$

2.3. Training

2.3.1. Training Schema. Features such as age, HR, and SV of simulated healthy adults' pulses were input as the independent variables. The subjects were divided into groups based on their age range. Randomly pick 80% of the total samples for the training examples and 20% for the testing examples.

2.3.2. Bootstrap. One of the ensemble methods used in Random Forest is Bootstrap, a resampling technique in the statistics area that involves randomly sampling and replacing datasets. [7] In Bootstrap, each RF tree is trained on a subset of the data instead of all the observations during training. [8] Then, average the predictions based on these subsets to mitigate the high variance. According to previous study, if N is an independent and identically distributed (iid) number, then the variance of the mean of the observations $Z_1, \ldots, Z_N$ is $\frac{\sigma^2}{N}$. [9] Thus, bootstrap aggregation can reduce variance by dividing the variance by the number of observations. [9] The estimator model with low-variance $\hat{f}_{avg}$ can be written as

DOI: 10.54254/2755-2721/13/20230745

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x)$$

where B is the number of separate bootstrap samples and $\hat{f}^{b}(x)$ is corresponded bootstrap estimators. [7]

Random forest uses a very similar operation process to bagging and increases the average prediction accuracy by reducing each DT's correlation, a process called "feature bagging."

Feature bagging operates by choosing a subset randomly of feature dimensions in each split DT. This random selection process is a good way to avoid some strong features leading to the same split of DT. [7] In this process, the leftover samples are known as “out of bag samples,” while the selected subset is referred to as the “bag.” [9]

2.3.3. Gini importance. This experiment uses the Gini Index to calculate the impurity of every feature so as to determine the root node. The Gini Index formula can be written as:

$$Gini\ Index = 1 - \sum_{i=1}^{n}(P_i)^2$$

where P_i is the probability of data lying in the i^{th} set of this split. After computing the Gini Index of each split, calculate the weighted Gini Index which is the total Gini index of this split.

$$Total\ Gini\ Index = \sum_{j=1}^{m} w_j \cdot Gini\ Index_j$$

where w_j is the weight of j^{th} split. Therefore, taking the feature which has the lowest Gini index as the root node, which means having the lowest impurity. [10]

2.3.4. Training objective. The misclassification error defined in this experiment was calculated using bootstrapping based on the out-of-bag data instead of a separate test set. For any kth tree, use the out-of-bag data as a test set for the kth tree and compute the out-of-bag error. Then, use the out-of-bag error estimate to evaluate the performance of the model on the test data set.

3. Results

3.1. Model selection

In addition to RF, this experiment also uses two other models: K-nearest neighbors (KNN) and logistic regression (LR). The comparison of the three models shows that RF has a better performance based on its relatively high accuracy. Based on the accuracy of RF (1.0) and LR (0.6395), the RF model was finally chosen to be used in this experiment.

3.1.1. Parameters of prediction model. We used N_estimators and max_depth as two parameters that needed to be adjusted in this experiment. N_estimators is the number of trees you wish to construct before averaging the results of the most votes or forecasts. [11] We adopt the out-of-bag (OOB) error image below to determine the possible n_estimators for selecting the optimized combination between n_estimators and max_depth later. According to the OOB image, we chose the number of N-estimators from [60, 80] since when “log2” is selected as max_features, the error is 0 after 60, but it is still underfitting at 20 or 40.

DOI: 10.54254/2755-2721/13/20230745

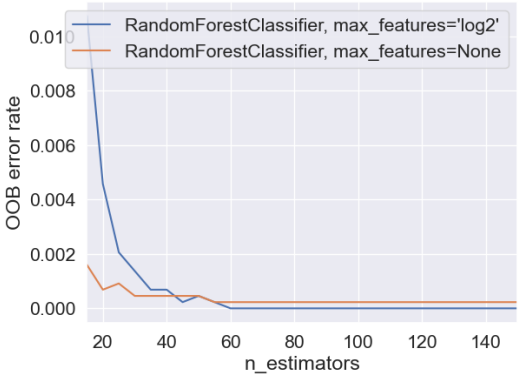

Figure 1. The Out-of-bag (OOB) error rate.

The OOB means the average error of each error derived using predictions from the trees that were not included in their corresponding bootstrap sample. In figure 1, the blue line means the trend of the OOB error rate corresponding to the n_estimators when max_feature is log2. Similarly, the orange line means the trend of the OOB error rate corresponding to the n_estimators when max_feature is none.

Max_depth is the maximum number of splits that a decision tree can have. The model underfits the data if there are too few splits and overfits the data if there are too many divides. [12] According to the max_depth image below, we can determine the maximum depth from [19, 20, 21, 22]:

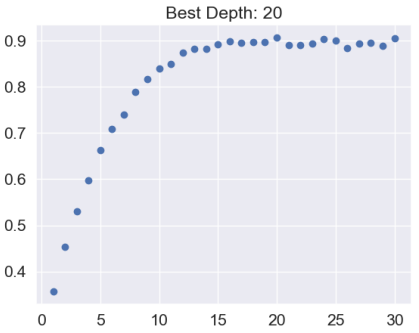

Figure 2. The Best Depth.

The scatter plot of the best depth value (x-axis) has the best cross-validation score (y-axis) and the highest y-value. In this experiment, we use the plot to determine the max_depth for the Random Forest model.

Then we tried different combinations of n_estimators and max_depth to see if optimized parameter combinations existed. Here is the table of series combinations corresponding to their RF accuracy score.

Table 2. Optimize Parameter Combinations.

N-estimators	max_depth	Accuracy (CO [l/min])
60	19	0.9632
60	20	0.9634
60	21	0.9604
60	22	0.9611
80	19	0.9618
80	20	0.9625
80	21	0.9639
80	22	0.9623

DOI: 10.54254/2755-2721/13/20230745

This table shows 8 combinations with N-estimators, max_depth, and the accuracy score of series combinations. N-estimators are selected from [60, 80], and max_depth is selected from [19, 20, 21, 22].

3.2. Feature selection

When all the features are predicted based on the RF model, the accuracy score is very high, even up to 1. Therefore, further feature selection is required. According to the figure of feature importance below, these features play a crucial role in influencing the test results: PWV_br [m/s], AIx [%], RFV [ml], LVET [ms], CO [l/min], SV [ml], and HR [bpm].

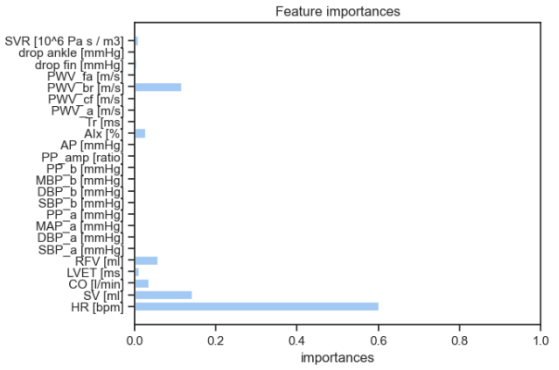

Figure 3. Feature Importance of all features.

All the features involved in the prediction are important in their own right. The longer the blue bar chart is, the more the specific feature can affect the prediction result.

Then, using the RF model, the accuracy score was calculated based on these features. Here is a table to show a series of features corresponding with their accuracy results in the RF model.

Table 3. High Accuracy Features.

Feature	Accuracy (RF)
HR [bpm]	1.0
SV [ml]	1.0
CO [l/min]	0.9652
LVET [ms]	0.3642
RFV [ml]	0.7387
AIx [%]	0.9762
SVR [10^6 Pa s / m3]	0.9262
HR [bpm] & SV [ml]	1.0

This table shows several features selected from Figure 3 as well as their accuracy score based on the RF model.

From the table above, we can see that an accuracy of 1 can be achieved even if only one of the HR or SV features is used, or a combination of HR and SV. Then, the next step is to fit the RF classifier based on its single feature value. Through this figure, we can conclude that the age range [65–75] will be displayed within [0,42.8] ml of the SV prediction value. And when the SV value is in the range (42.8-44.69), the age range will be [45-65].

3.3. Case study

We finally chose HR and SV as two features to predict the age range of the experimental subjects. Here is one decision tree for a random forest model with a max_depth of 6 and n_estimators of 100.

DOI: 10.54254/2755-2721/13/20230745

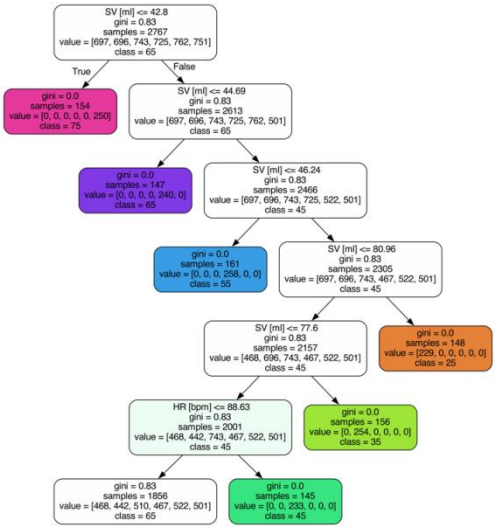

Figure 4. Random Forest of HR and SV.

This is a branch under the random forest that is predicted by HR and SV, which includes the Gini Index, the number of samples, and the range of SV values corresponding to their age range.

From the decision tree above, we can understand the judgment logic of the RF model. For example, we have a patient in our database who is 25 years old, but our model predicts that his SV value is 66 ml, which means that his pulse age should be in the mid-40s. This clearly indicates that the patient's physical age is much greater than his actual age. Therefore, if similar conditions are detected among our users, we need to remind the patients that they need to pay attention to their health and do further examinations, especially their cardiovascular health.

4. Discussion results

This paper introduces the current background on pulse wave diagnosis and a new concept, pulse wave age. After that, we selected 25 useful features by analysing various data attributes. Then we selected three ML models: KNN, LR, and RF. Calculate their prediction accuracy, respectively, and conclude that RF performs best. Later, the relatively optimal parameter combination is selected: N-estimators as 80 and max_depth as 21. After that, HR and SV, two features with the most significant influence on the experimental results, are selected by feature importance. And the prediction result with an accuracy of 1 is chosen. What's exciting about this is that we can use this model to successfully predict participants' biological ages, and by comparing it to their actual ages, we can get an idea of their general health.

However, the prediction results of the whole experiment are ideal, and the accuracy reaches 1. At the same time, it is worth thinking about why the accuracy is so high. One possibility is that the two features we use are SV and HR. However, according to Dr. Jana M Goldber, heart rate decreases linearly with age. [13] Thus, HR has a strong linear relationship with age, which is one possible reason we can predict the age of patients with very high accuracy through HR and SV. Therefore, after plotting the boxplot between age and several features in this research, we observed a robust linear relationship between age, HR, and SV. Thus, we decided to use HR and SV to avoid overfitting problems. Another possible reason is that the data used in our experiment was not clinical medical data but virtual data from Peter Charlton. PW data is simulated by the CV conditions of people aged 25–75. Therefore, one direction worth our research in the future is to fit a more feasible and comprehensive model through real clinical data collection.

In summary, the significance of this study is, first of all, that the age of the patient is clinically meaningful. We can predict the pulse wave age of patients with relative accuracy and convenience using easily measured HR and SV data such as electronic watches. Another application is to infer a

DOI: 10.54254/2755-2721/13/20230745

patient's biological age and make a corresponding diagnosis of possible conditions if not readily available. For example, specific demographic data cannot be obtained in many cases due to privacy issues. If we need to provide a detailed health care plan, we can calculate the age of patients according to the measured HR and SV data.

References

[1] H. Tomiyama Envelope, et al. "Ankle-Brachial Pressure Index and Pulse Wave Velocity in Cardiovascular Risk Assessment." Encyclopedia of Cardiovascular Research and Medicine, Elsevier, 30 Nov. (2017)

[2] Luo, Zhi-yu, et al. "A Study of Machine-Learning Classifiers for Hypertension Based on Radial Pulse Wave." BioMed Research International, Hindawi, 11 Nov. (2018).

[3] Charlton, Peter H., et al. "Modeling Arterial Pulse Waves in Healthy Aging: a Database for in Silico Evaluation of Hemodynamics and Pulse Wave Indexes." American Journal of Physiology (2019).

[4] J;, Pereira T;Paiva JS;Correia C;Cardoso. "An Automatic Method for Arterial Pulse Waveform Recognition Using KNN and SVM Classifiers." Medical & Biological Engineering & Computing, U.S. National Library of Medicine.

[5] L.IzzoJr., Joseph, et al. "Assessment of Hypertensive Target Organ Damage." Hypertension, W.B. Saunders, 15 May. (2009).

[6] "Heart Attacks Increasingly Common in Young Adults." American College of Cardiology, 7 Mar. (2019).

[7] "Bootstrap Aggregation, Random Forests and Boosted Trees." QuantStart, Retrieved November 22, (2022).

[8] Choudhary, Divya. "Bootstrapping and Oob Samples in Random Forests." Medium, Analytics Vidhya, 18 Apr. (2021).

[9] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) An Introduction to Statistical Learning, Springer.

[10] Saini, Anshul. "Random Forest Algorithm for Absolute Beginners in Data Science." Analytics Vidhya, 26 Aug. (2022).

[11] Srivastava, Tavish. "Random Forest Parameter Tuning: Tuning Random Forest." Analytics Vidhya, 26 June (2020).

[12] Ram, Sandeep. "Mastering Random Forests: A Comprehensive Guide." Medium, Towards Data Science, 18 Oct. (2020).

[13] "Relationship between Exercise Heart Rate and Age in Men vs. Women." American College of Cardiology.