

RESEARCH

Open Access



Machine learning classification of polycystic ovary syndrome based on radial pulse wave analysis

Jiekee Lim¹, Jieyun Li¹, Xiao Feng², Lu Feng¹, Yumo Xia¹, Xinang Xiao¹, Yiqin Wang^{1,3} and Zhaoxia Xu^{1,3*}

Abstract

Background Patients with Polycystic ovary syndrome (PCOS) experienced endocrine disorders that may present vascular function changes. This study aimed to classify and predict PCOS by radial pulse wave parameters using machine learning (ML) methods and to provide evidence for objectifying pulse diagnosis in traditional Chinese medicine (TCM).

Methods A case-control study with 459 subjects divided into a PCOS group and a healthy (non-PCOS) group. The pulse wave parameters were measured and analyzed between the two groups. Seven supervised ML classification models were applied, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, Random Forest, Logistic Regression, Voting, and Long Short Term Memory networks (LSTM). Parameters that were significantly different were selected as input features and stratified k-fold cross-validations training was applied to the models.

Results There were 316 subjects in the PCOS group and 143 subjects in the healthy group. Compared to the healthy group, the pulse wave parameters h3/h1 and w/t from both left and right sides were increased while h4, t4, t, As, h4/h1 from both sides and right t1 were decreased in the PCOS group ($P < 0.01$). Among the ML models evaluated, both the Voting and LSTM with ensemble learning capabilities, demonstrated competitive performance. These models achieved the highest results across all evaluation metrics. Specifically, they both attained a testing accuracy of 72.174% and an F1 score of 0.818, their respective AUC values were 0.715 for the Voting and 0.722 for the LSTM.

Conclusion Radial pulse wave signal could identify most PCOS patients accurately (with a good F1 score) and is valuable for early detection and monitoring of PCOS with acceptable overall accuracy. This technique can stimulate the development of individualized PCOS risk assessment using mobile detection technology, furthermore, gives physicians an intuitive understanding of the objective pulse diagnosis of TCM.

Trial registration Not applicable.

Keywords Machine learning, Polycystic ovary syndrome, Pulse diagnostics, Radial pulse wave, Traditional Chinese medicine

*Correspondence:

Zhaoxia Xu
zhaoxia7001@shutcm.edu.cn

¹School of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, P. R. China

²The First Affiliated Hospital, Guangzhou University of Traditional Chinese Medicine, Guangzhou 510405, P. R. China

³Shanghai Key Laboratory of Health Identification and Assessment, Shanghai 201203, P. R. China



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Polycystic ovary syndrome (PCOS) is one of the most common gynaecological endocrine disorders in women of reproductive age, affecting 5–20% of women worldwide [1]. PCOS is characterized by ovulatory dysfunction, hyperandrogenism, and polycystic ovarian morphology (PCOM). PCOS increases the risk of metabolic complications, cardiovascular disease, endometrial cancer, and mental health disorders [2–4]. The economic burden of PCOS is estimated at USD 8 billion annually [5]. Therefore, early diagnosis of PCOS is essential to prevent the long-term complications of the disease and to reduce the medical burden.

Pulse diagnosis, which is a non-invasive, convenient, and simple method, is one of the most common diagnostic methods in TCM. It is done by palpating the radial artery pulse as shown in Fig. 1, each pulse position is a reflection point of a certain internal organ system [6]. For centuries, the practitioner gains insights into patients' physical conditions and constitutions by interpreting the characteristics of the pulse, the pulse can reflect the conditions of internal organs, Qi (vital energy), and blood of individuals [7].

However, traditional pulse diagnosis relies on the palpation sensitivity of the practitioners, the judgments are always limited due to subjective observations based on one's experience. Consequently, the importance of objectifying and quantifying radial pulse waves has been drawing attention [8, 9]. Previous studies suggested that, compared to healthy individuals, changes in radial pulse wave in PCOS are related to hemodynamic changes, ventricular systolic function, and aortic compliance in the cardiovascular system [10, 11], and the pulse wave parameters varied at different body mass index (BMI) levels in PCOS [12].

Merely understanding specific relevant factors is not enough in actual clinical diagnosis and treatment of PCOS, to better serve the clinical needs, the application of ML algorithms for diagnosis and predictions has been reported using different features, such as Raman spectroscopy of follicular fluid [13], PCOS genes [14, 15], ovary ultrasound images and reports [16, 17], or clinical data set [18–20]. However, the study about the ML prediction model of PCOS based on pulse wave parameters has not been found yet. Some evidence supported that ML models using pulse waves are of great significance in disease predictions [21–25]. Therefore, through the

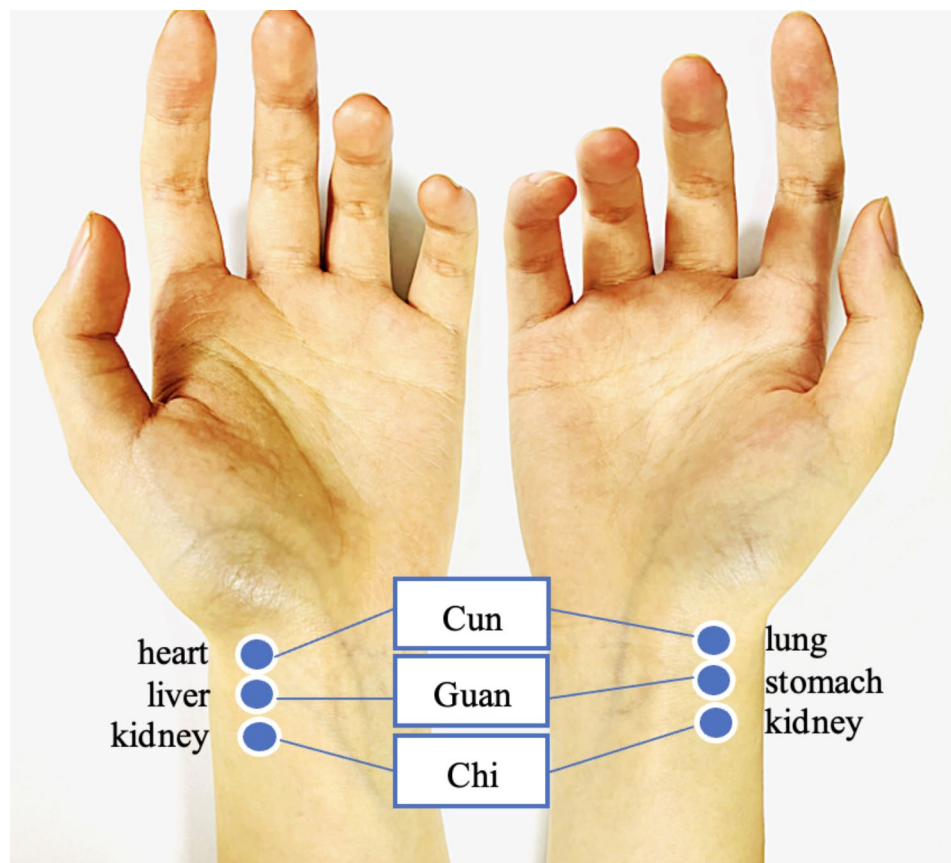


Fig. 1 Pulse diagnosis on the radial artery and the corresponding internal organ systems. Pulse diagnosis is done by palpating three adjacent regions of the wrist of both hands, namely Cun (distal), Guan (middle), and Chi (proximal) position

comparisons of ML prediction models, we intended to determine whether PCOS could be diagnosed and monitored by radial pulse wave.

The objective of this study is to classify and predict PCOS by radial pulse wave parameters using machine learning methods and to provide evidence for objectifying pulse diagnosis in TCM.

Methods

The study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.

Study subjects

This case-control study involved two groups. The PCOS group included 316 patients with PCOS while the healthy group included 143 normal subjects. All the participants were recruited from the Shanghai University of TCM (SHUTCM), Shanghai Municipal Hospital of Traditional Chinese Medicine, Shuguang Hospital, and Yueyang Hospital affiliated with SHUTCM, from August 2018 to January 2022.

Inclusion criteria

Female participants aged 18–40 years old were eligible for this study. Based on “Chinese guidelines for diagnosis and treatment of polycystic ovary syndrome (2018)” [26], the diagnosis of PCOS was as follows:

- a. oligomenorrhea or amenorrhea or irregular uterine bleeding is a necessary condition.
- b. 1 of the following 2 criteria must be met: clinical and/or biochemical hyperandrogenism (HA), polycystic ovarian morphology (PCOM).

Participants in the healthy group were required to show none of the PCOS criteria and were free from gynaecological and organic diseases. All participants must sign the written informed consent.

Exclusion criteria

The exclusion criteria were as follows:

- a. other diseases that may cause hyperandrogenism and abnormal ovulation;
- b. participants with other apparent gynaecological diseases and organic diseases such as liver or kidney disorders;
- c. participants with adenomyosis, Cushing’s syndrome, chromosomal abnormalities, congenital adrenal cortical hyperplasia, and chocolate cyst of the ovary;
- d. participants with serious primary diseases in internal medicine and surgery;
- e. patients with significant incomplete clinical data;
- f. patients who were unable to cooperate.

Radial pulse signal collection

The pulse signals were collected from the Guan position of participants’ left and right hands using the Z-BOX pulse meter, radial pulse signals can be detected most clearly and easily at the Guan position. The time of collection was from 9 AM to 11 AM or from 1 PM to 4:30 PM. The participants were required to keep calm and prohibited to eat and drink 30 min before the test, they also had to avoid violent mood swings. During the test, the participant was required to breathe calmly, sit upright, keep the left arm relaxed, and spread forward the left forearm naturally, the wrist was placed on a pulse pillow with the palm facing up and the fingers slightly bent. The Z-BOX pulse meter was attached to the wrist where the pressure sensor was placed on Guan position, in the meantime, the participant should avoid speaking or moving. A series of radial pulse signals within the pulse pressure range of 25–250 g were recorded continuously for 30s. The sampling process repeats with the right arm. The radial pulse signal with the highest main amplitude, apparent fluctuation of three peaks, and a steep ascending branch without incisure was selected for time-domain parameter analysis. As pulse signals can be affected by a variety of noise sources, including patient tremors, respiration, mechanical vibrations of instruments, and power frequency interference. We used the PulseSystem software [jointly developed by our research group and East China University of Science and Technology (Shanghai)] to de-noise pulse signals and extract the pulse wave parameters. Pulse signals are concentrated in the low-frequency range, so the software used a Butterworth filter to remove high- and low-frequency noise. The filter order was set to 3, and the passband range was set to 0.2–20 Hz. To prevent bias, the pulse meter used for data collection was consistent, and data collection was done by the same executors (XF and LF) with adequate training, double entry and verification are adopted for data entry by the same executors again.

Time-domain parameters of the radial pulse wave signal

Figure 2 shows the time-domain parameters which are commonly used in TCM radial pulse wave analysis. 30 time-domain parameters were extracted for comparisons, including h1, h3, h4, h5, t1, t4, t5, t, w, As, Ad, h3/h1, h4/h1, h5/h1, w/t, from both left and right Guan position. The interpretation of the meaning of pulse wave parameters was done by referring to “Pulse Diagnosis of Modern TCM” [27], each parameter corresponded to specify physiological significance. Parameters h1, h3, h4, and h5 are the main wave, tidal wave, dicrotic notch, and dicrotic wave amplitude accordingly. Parameter t is the time for a complete pulse cycle, t1 is the time between the starting point to the crest of the main wave, t4 is the time between the starting point to the dicrotic notch,

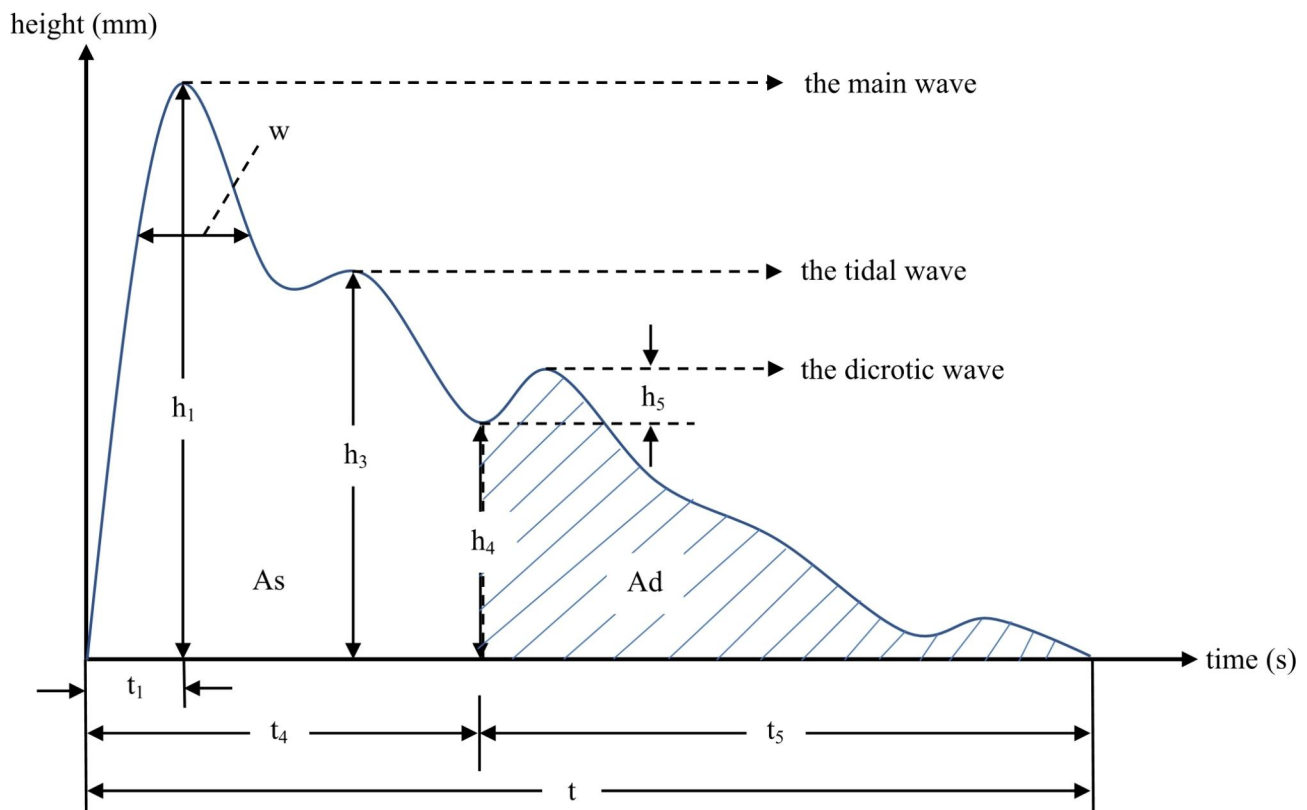


Fig. 2 The amplitude, time, and area parameters of the radial pulse diagram

t_5 is the time between the dicrotic notch to the ending point. Parameter w is the width of the main wave at its 1/3 height. A_s is the area of the systolic phase while A_d is the area of the diastolic phase. The ratio h_3/h_1 reflects vascular wall compliance and peripheral resistance, h_4/h_1 reflects the level of peripheral resistance, h_5/h_1 mainly reflects aortic compliance and aortic valve function, and w/t corresponds to the duration of elevated aortic pressure and is related to peripheral resistance. Figure 3 showed the flowchart of pulse wave parameters collection and analysis.

Statistical analysis of pulse wave parameters

Statistical analysis of radial pulse wave parameters was performed using IBM SPSS Statistics (version 26). All measurements from the two groups (PCOS patient group and healthy control group) did not conform to normal distribution. Therefore, independent samples Wilcoxon rank sum test (Mann-Whitney U test) was used to compare between groups. The results were presented by median, $M(P_{25}, P_{75})$. The level of statistical significance was set at $P < 0.05$ for all the analyses.

Machine learning classification method

Machine Learning (ML) classification is used to predict categories, which are the PCOS group and the healthy

group here. Figure 4 shows the flowchart of the machine learning classification process.

Supervised learning is used for the study, it is a type of machine learning where the predicted values are already known. Given the features and target variables, a model is built to accurately predict target values of unseen data, in our case, predict a subject's class based on pulse wave parameters ($P < 0.05$). Seven supervised machine learning classifiers were used to build models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), Soft Voting (SV), and Long Short Term Memory networks (LSTM).

The pulse wave parameters dataset was split into a training set (75%) and a testing set (25%). The training set was used for model training and the testing set was used to evaluate final model performance. Data splitting is a crucial step to prevent overfitting since ML classifiers can perform relatively well on trained data [28]. Data standardization of features was done by subtracting the mean and dividing by variance so that all features were centred around zero and had a variance of one.

Stratified k-fold cross-validation was applied to the training data, the data were further split into train/test sets for 10 folds, the folds are made by preserving the percentage of subjects for each class, this technique is

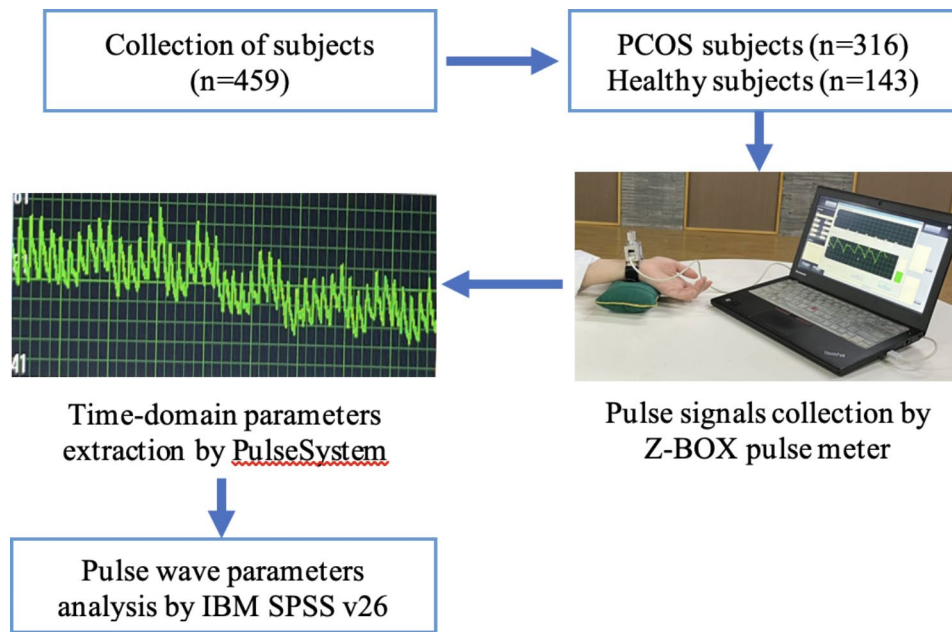


Fig. 3 Flowchart of pulse wave parameters collection and analysis

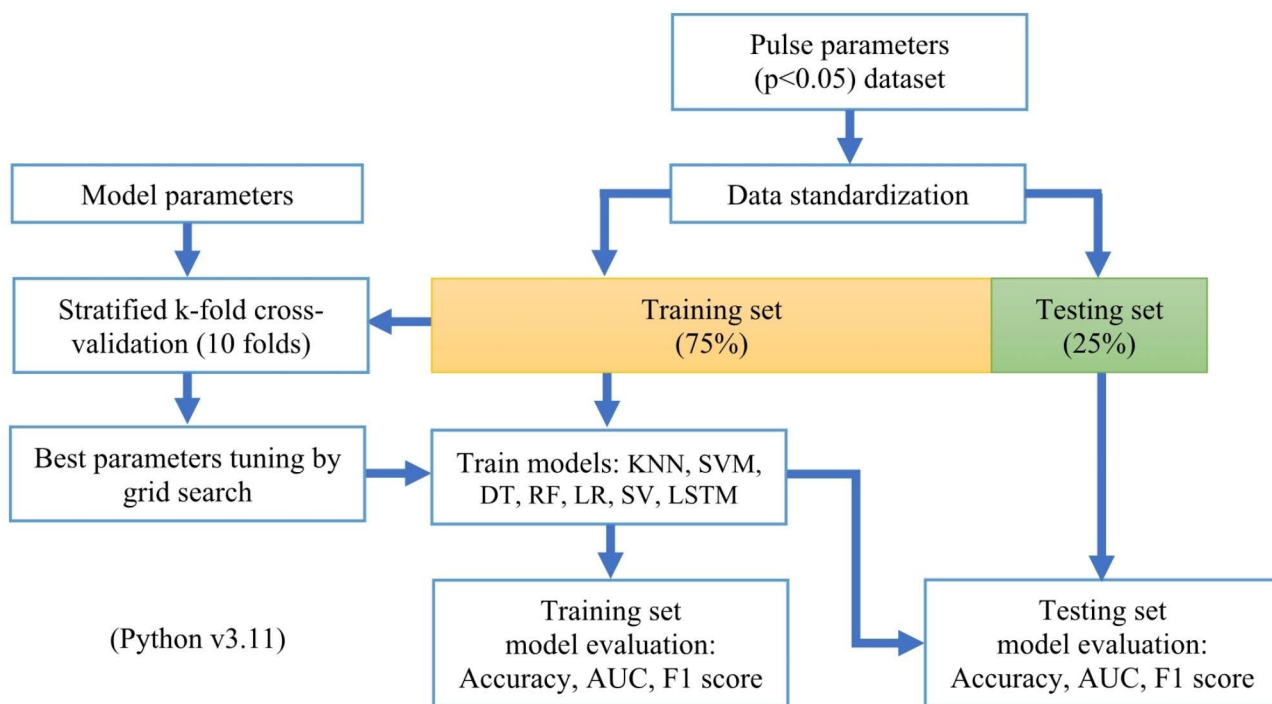


Fig. 4 Flowchart of PCOS and healthy group discrimination by machine learning classifiers. 15 parameters ($P < 0.05$) were used as the features for model training

good for an imbalanced class as in our case. Models were trained using the seven classifiers, the best model parameters were tuned by the grid search. The model performance based on the training data was evaluated by accuracy, area under the ROC curve (AUC), and F1 score, then, the evaluation was repeated on the testing

set. The final evaluation aimed to check the general ability of models to predict unseen data.

Machine learning classifiers

K-Nearest Neighbors (KNN) is a non-parametric algorithm that is popular for classification problems. KNN

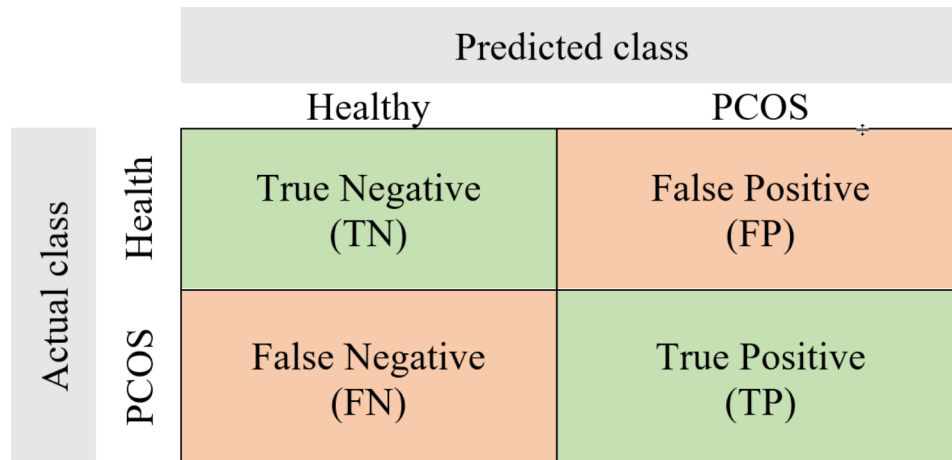


Fig. 5 The confusion matrix

uses “k” numbers of closest labelled data points to predict new data points, prediction is made based on the majority labels of the nearest neighbours. The determined value of k is 22 in this study.

Support Vectors Machines (SVM) work by finding the hyperplane that maximally separates the data points of different classes [29]. We applied the SVM radial basis function (RBF) kernel which is suitable for nonlinear data in this study.

Decision Trees (DT), also known as Classification and Regression Trees (CART)), can be used to predict categorical or continuous outcomes, which are commonly used in epidemiological and medical fields [30]. When a classification tree is trained, the tree learns a sequence of if-else questions about individual features to infer the class labels. The maximum depth determined by grid search was 2 here in this study.

Random Forest (RF) is one of the best ensemble learning methods of decision trees. Random Forest uses random subsamples of training data and randomizing the algorithm for base-level classifiers (decision trees), a subset of features is randomly selected by decision trees and the best is chosen among these at each step of tree construction [31]. Random Forest models are less prone to overfitting and can achieve higher accuracy in disease prediction. A Random Forest with 300 decision trees and a maximum depth of 4 was determined in this study.

Logistic Regression (LR) calculates the probability of an observation belonging to the binary class. The predicted

probability is compared to the default probability threshold to make the classification.

A voting classifier with soft voting (SV) is chosen, the model trains on the ensemble of the five models above (KNN, SVM, DT, RF, LR), and the class label is predicted based on the highest average probability given to that class. This ensemble-based Voting classifier is expected to improve model performance compared to a single classifier [32].

Long Short-Term Memory networks (LSTM) are a type of recurrent neural network (RNN) architecture specifically designed to capture and process sequential and time series data, which have been widely used in speech recognition, natural language processing, and time series prediction [33, 34]. While a simple RNN can use past predictions to infer new ones, LSTMs were introduced to overcome the limitations of RNN, which hard to manage long-range dependencies due to the vanishing gradient problem [35]. After parameters tuning, we built the LSTM model with 2 LSTM layers and 1 Dense layer, each followed by a Dropout layer, lastly an output layer. The model architecture was stated in the supplementary material 1.

Classification metrics

In the present study, the main metrics used for model performance evaluation are accuracy, F1 score, and AUC.

The confusion matrix can summarize the model performance as in Fig. 5. The true positives (TP) are the number of PCOS subjects correctly predicted; the true negatives (TN) are the number of healthy subjects correctly predicted; the false negatives (FN) are the number of healthy subjects incorrectly predicted; and the false positives are the number of PCOS subjects incorrectly predicted.

Table 1 Comparison of general information (n=459)

Group	Age	BMI (kg/m ²)
PCOS (n=316)	27.00 (24.00, 29.75)	22.03 (19.91, 25.08)
Healthy (n=143)	25.00 (21.00, 31.00)	20.44 (19.15, 21.77)
P	0.061	<0.01*

*P<0.01

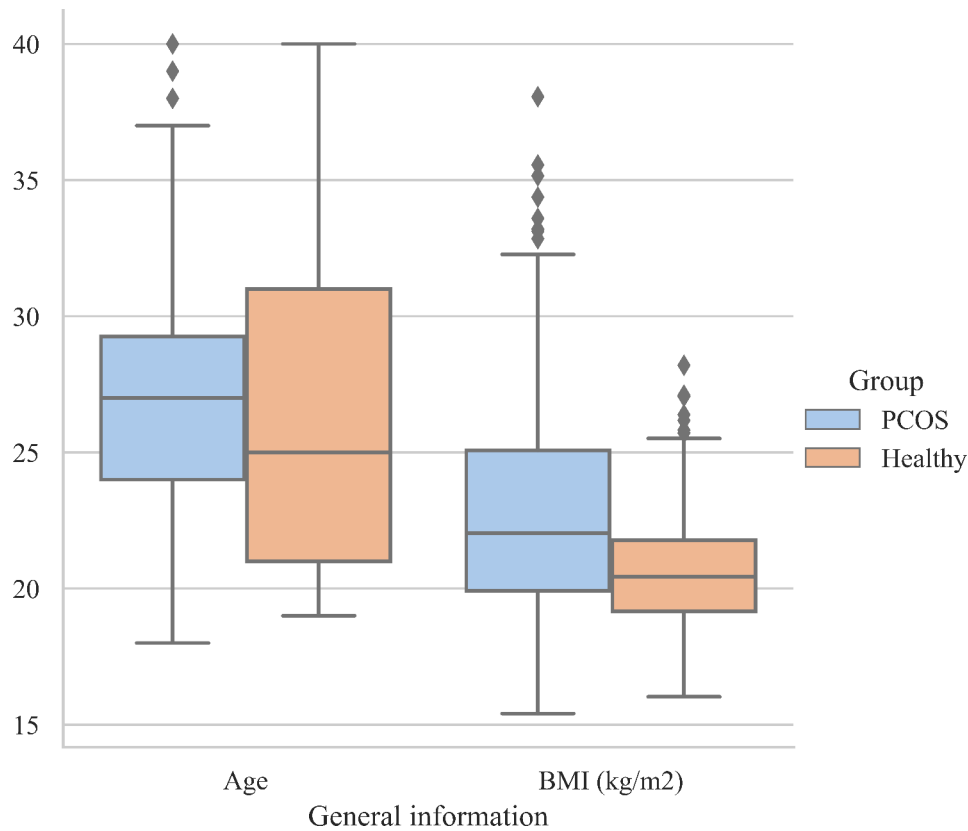


Fig. 6 Box plot of age and BMI distribution

Accuracy is the proportion of the correct predictions to the total number of observations. It is a commonly used metric to measure model performance in classification.

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

Precision is the positive predictive value. It is the ratio of true positives to the sum of all positive predictions. High precision means having a lower false positive rate.

$$Precision = \frac{TP}{TP + FP}$$

Table 2 Comparison of time-domain pulse wave parameters from left Guan position

Time domain parameters (left)	PCOS group (n=316)	Healthy group (n=143) (n=143)	MannWhitney U Test	P
h1	51.1463(36.1461,62.1430)	51.0814(37.7732,68.6276)	20663.00	0.14
h3	42.0940(29.3900,51.8107)	39.9151(29.1578,52.5250)	22553.50	0.98
h4	12.2748(4.9953,20.7704)	17.0547(7.1223,26.6643)	18800.50	<0.01**
h5	25.9800(16.5715,35.0660)	26.3911(16.6583,36.7083)	22180.00	0.75
t1	0.0436(0.0411,0.0478)	0.0439(0.0413,0.0473)	22432.00	0.9
t4	0.0665(0.0475,0.0849)	0.0810(0.0573,0.0960)	17108.50	<0.01**
t5	0.1560(0.1320,0.1791)	0.1575(0.1326,0.1865)	21420.50	0.37
t	0.2199(0.2036,0.2416)	0.2351(0.2158,0.2572)	16446.00	<0.01**
w	0.0616(0.0505,0.0706)	0.0577(0.0484,0.0688)	20269.50	0.08
As	1110.1580(562.0087,1820.6702)	1540.8545(726.5162,2432.4840)	18174.00	<0.01**
Ad	2051.2545(1322.9600,3009.1210)	2042.1479(1306.9180,2853.1460)	22224.00	0.78
h3/h1	0.8620(0.7703,0.9068)	0.8056(0.7113,0.8768)	17230.00	<0.01**
h4/h1	0.2716(0.1384,0.3915)	0.3533(0.1886,0.4463)	18570.00	<0.01**
h/5h1	0.5197(0.4133,0.7125)	0.5021(0.3787,0.6911)	21547.00	0.43
w/t	0.2781(0.2431,0.3124)	0.2478(0.2119,0.2867)	15750.00	<0.01**

*P<0.05 **P<0.01

Table 3 Comparison of time-domain pulse wave parameters from the right Guan position

Time domain parameters (right)	PCOS group (n = 316)	Healthy group (n = 143) (n = 143)	MannWhitney U Test	P
h1	53.6468(39.4475,64.9978)	56.3246(41.1000,71.4271)	20346.00	0.09
h3	41.0358(30.5258,51.8072)	40.3392(29.0920,53.4807)	22365.00	0.86
h4	14.8259(7.1823,23.1606)	20.3658(11.8300,29.3906)	17541.00	< 0.01**
h5	23.8942(14.4083,36.8861)	26.5248(17.3289,34.6542)	21196.00	0.29
t1	0.0426(0.0405,0.0458)	0.0441(0.0415,0.0472)	19140.50	< 0.01**
t4	0.0752(0.0570,0.0905)	0.0929(0.0730,0.1014)	15079.00	< 0.01**
t5	0.1469(0.1233,0.1702)	0.1480(0.1289,0.1708)	21916.50	0.61
t	0.2199(0.2036,0.2416)	0.2351(0.2158,0.2572)	16446.00	< 0.01**
w	0.0556(0.0459,0.0660)	0.0523(0.0416,0.0651)	20568.50	0.12
As	1413.7820(762.0707,2039.2582)	1892.5390(1175.0580,2725.4790)	16729.00	< 0.01**
Ad	1961.8885(1163.8118,2810.0853)	1729.3390(1201.8840,2604.1810)	22138.00	0.73
h3/h1	0.8117(0.7123,0.8798)	0.7249(0.6472,0.8534)	17282.00	< 0.01**
h4/h1	0.3224(0.1935,0.4141)	0.3866(0.2749,0.4459)	17573.00	< 0.01**
h5/h1	0.4689(0.3546,0.6285)	0.4862(0.3920,0.5685)	22082.00	0.7
w/t	0.2582(0.2212,0.2844)	0.2308(0.1792,0.2741)	16914.00	< 0.01**

*P<0.05 **P<0.01

Recall is the ratio of true positives to the sum of true positives and false negatives. It is also called sensitivity. High recall reflects a lower false negative rate. For the present study, high recall means predicted most PCOS subjects correctly.

$$Recall = \frac{TP}{TP + FN}$$

F1 score is the harmonic mean of precision and recall, thus, it evaluates a model's precision and recall ability. F1 score ranges from 0 (worst) to 1 (best).

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Area under the Receiver Operating Characteristic curve (AUC/ ROC AUC), is a useful metric to visualize and evaluate classification ability [36]. ROC graph reveals the relationship between true positive rate (TPR) and false positive rate (FPR). AUC ranges from 0 to 1.0, 0.5 means random guessing, the larger the AUC the better the model is.

Results

Baseline characteristics

Table 1; Fig. 6 show the comparisons of age and BMI of the subjects from the PCOS group and healthy group. Subjects from different groups did not show significant differences in age ($P > 0.05$) but showed significant differences in BMI ($P < 0.01$). The BMI of the PCOS group was significantly higher than the healthy group.

Comparisons of time-domain parameters

The time-domain parameters of pulse diagrams from both left and right Guan positions were compared between the PCOS group and healthy control group using the Wilcoxon rank sum test. Tables 2 and 3 showed the results of the comparisons. For both the left and right sides, compared to the healthy group, the parameters h4, t4, t, As, h4/h1 were significantly lower in the PCOS group ($P < 0.01$) while the parameters h3/h1 and w/t were significantly higher in PCOS group ($P < 0.01$). The right t1 was significantly lower in the PCOS group compared to the healthy group ($P < 0.01$). No significant differences were observed in the rest of the pulse wave parameters between the groups ($P > 0.05$).

Table 4 Performance evaluation of models

Models	Training results (Stratified k-fold cross-validations)			Testing results		
	Accuracy	AUC	F1 score	Accuracy	AUC	F1 score
KNN	72.672 ± 6.259	0.691 ± 0.105	0.823 ± 0.040	69.565	0.680	0.795
SVM	72.387 ± 4.543	0.696 ± 0.098	0.825 ± 0.027	72.174	0.694	0.818
Decision Trees	72.395 ± 4.307	0.624 ± 0.070	0.819 ± 0.032	71.304	0.646	0.811
Random Forest	72.689 ± 5.188	0.689 ± 0.118	0.825 ± 0.033	69.565	0.695	0.798
Logistic Regression	70.933 ± 3.874	0.698 ± 0.111	0.814 ± 0.028	72.174	0.704	0.814
Voting	73.546 ± 4.232	0.701 ± 0.115	0.831 ± 0.027	72.174	0.715	0.818
LSTM	74.135 ± 5.437	0.702 ± 0.115	0.828 ± 0.023	72.174	0.722	0.818

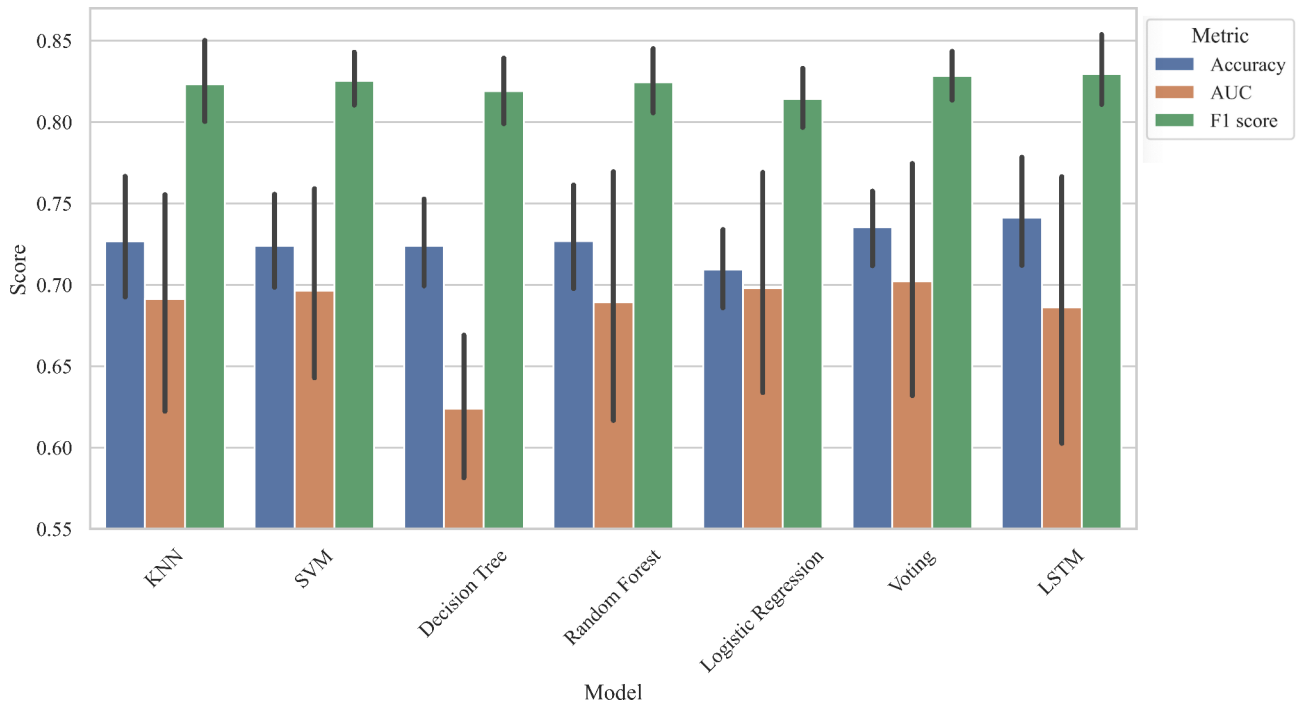


Fig. 7 Stratified k-fold cross-validation training results of machine learning models

Machine learning classification based on radial pulse wave parameters

15 parameters ($P < 0.05$) were selected as feature variables for model training including right t1 and h4, t4, t, As, h4/h1, h3/h1, w/t from both left and right sides. The target variables were the groups of the subjects.

The performance metrics including accuracy, AUC, and F1 score were calculated, the summary of results is presented in Table 4, Fig. 7 and 8.

For the stratified k-fold cross-validation training, the result showed that KNN, SVM, Decision Trees, and Random Forest performed similarly by gaining

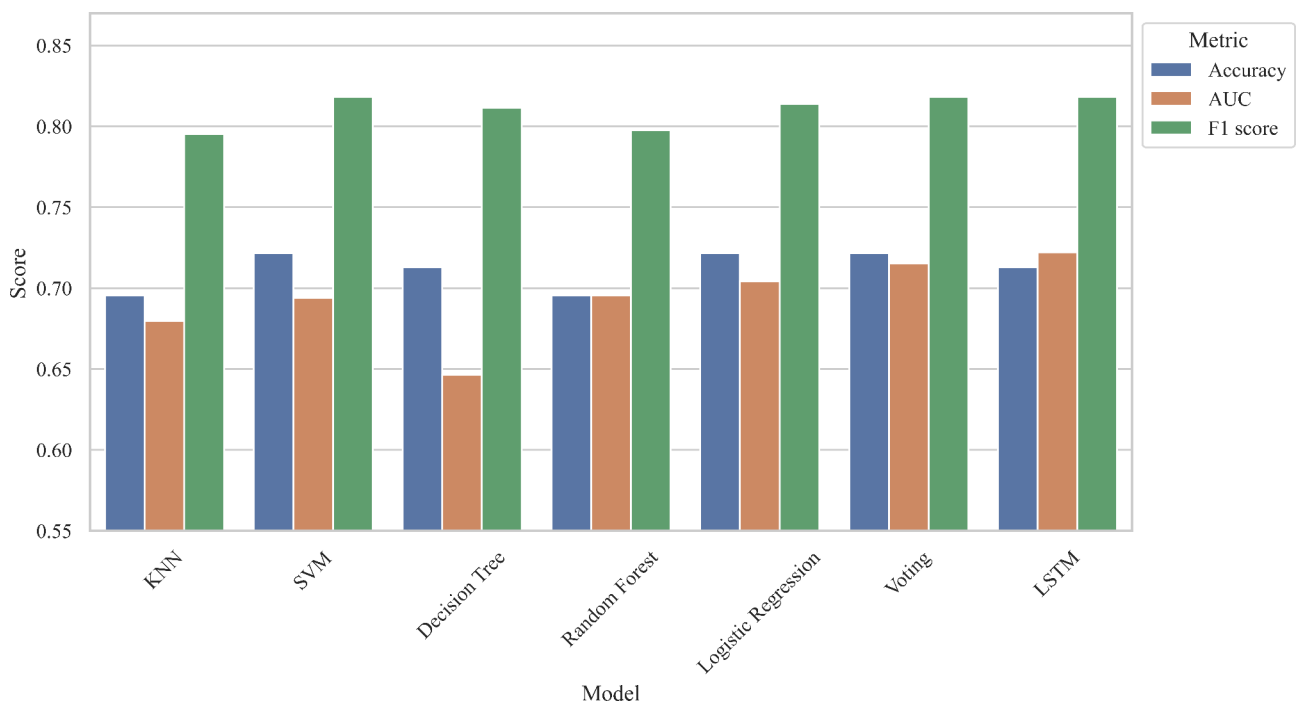


Fig. 8 Testing results of machine learning models

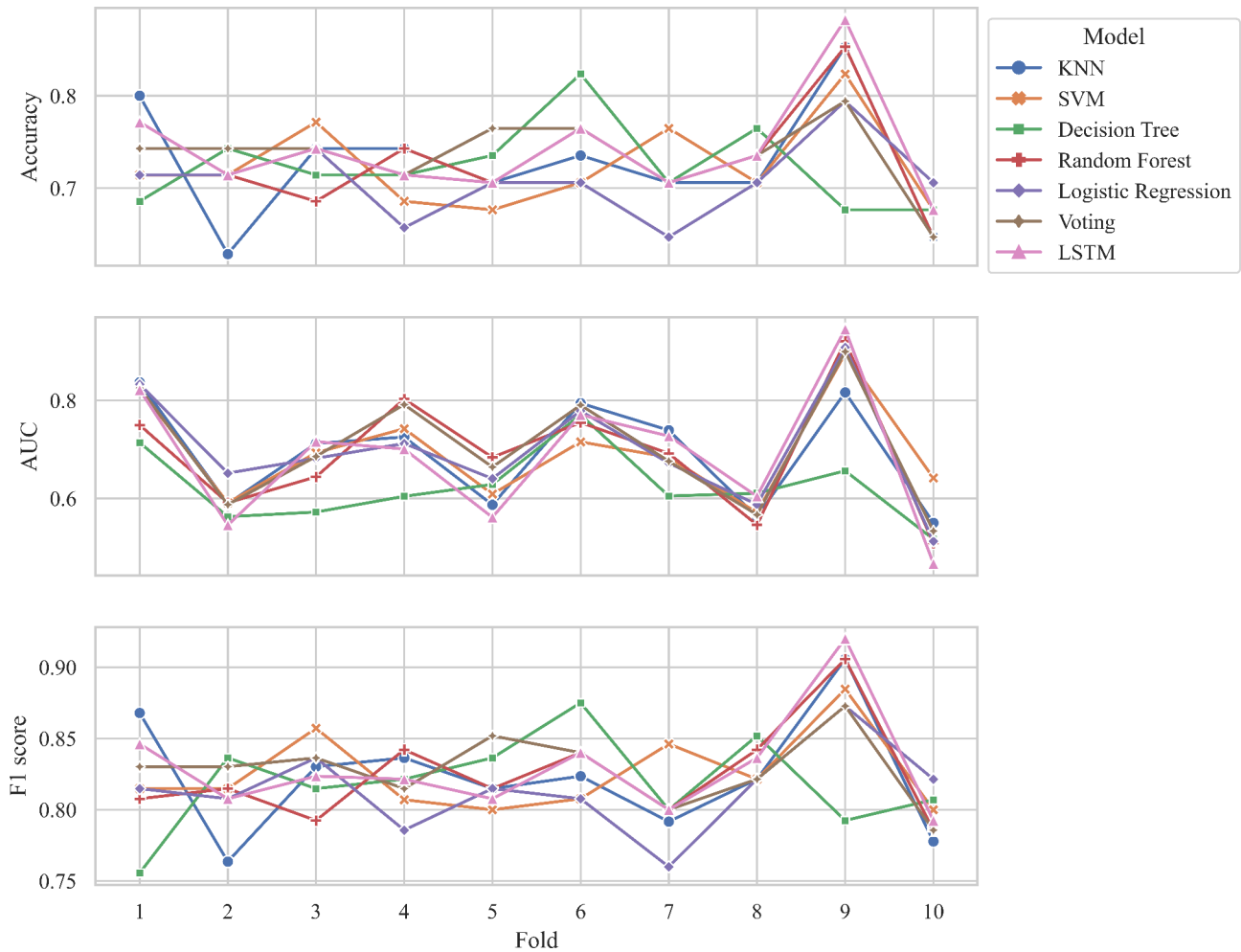


Fig. 9 Results of stratified k-fold cross-validation for each fold, accuracy, AUC, and F1 score are visualized respectively

similar accuracy scores. They have slightly lower performance compared to LSTM and Voting ensemble. LSTM achieved the highest accuracy and AUC at 0.74135 ± 0.05437 and 0.702 ± 0.115 respectively while Voting achieved the highest F1 score at 0.831 ± 0.027 . The cross-validation results for each fold are visualized in Fig. 9.

For the testing set evaluation, SVM, Logistic Regression, Voting, and LSTM achieved the highest accuracy (72.17%). SVM, Voting, and LSTM got the highest F1 score (0.818), LSTM again performed the best for AUC (0.722), thus, LSTM had the best testing performance. Figure 10 showed the AUC of the ROC graph among models and Fig. 11 compared the performance metrics between the training set and testing set. Both the Voting and LSTM exhibited similar levels of performance across all metrics.

Feature importance

The features contributing most to the Random Forest model were reported (Fig. 12). Random forest with training accuracy $72.689 \pm 5.188\%$, AUC 0.689 ± 0.118 , and F1

score 0.825 ± 0.033 . The top five features are right t4, left w/t, right t1, left As, and left t, these features contributed to 46.89% of importance over all features.

Discussion

The PCOS group and the healthy group were age-matched ($P > 0.05$) but the BMI of the PCOS group was higher ($P < 0.01$) due to its metabolism disorders. The pulse wave parameters h3/h1, w/t, h4, t4, t, As, h4/h1 from both sides and right t1 were significantly difference between groups ($P < 0.01$). Compared to healthy individuals, PCOS patients experienced poorer vascular compliance, faster heart rate, and reduced left ventricular systolic function. Results of ML classification proved that pulse signal analysis could be used to predict PCOS patients, among the seven models, LSTM achieved the best testing performance.

Radial pulse wave is affected by the pulsation of the heart and conditions of arteries, tissues, and organs; thus, it could contain rich physiological and pathological information about individuals [37]. Time-domain

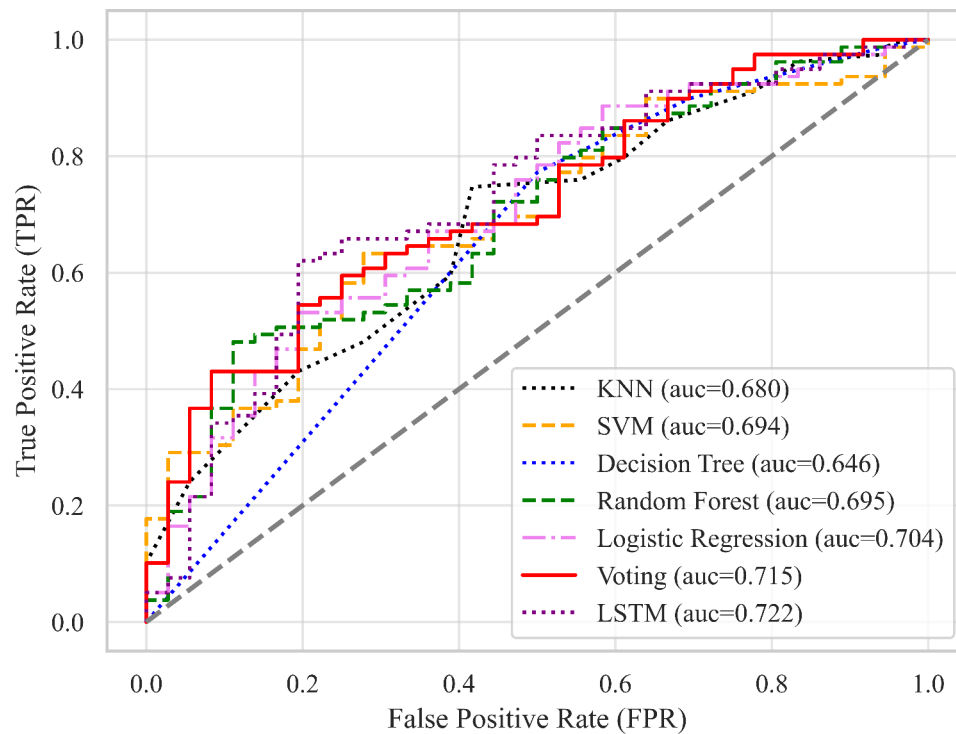


Fig. 10 Testing scores of Area under the Receiver Operating Characteristic curve (AUC).

analysis, which is one of the most widely used methods [9], was applied in this study to determine the correlation between PCOS and radial pulse wave parameters.

The metabolic disorder of PCOS causes vascular function changes. Studies reported that PCOS patients had evidence of an increased risk of hypertension [38], proven early atherosclerosis, and endothelial dysfunction, thus, increasing the risk of developing cardiovascular disease (CVD) compared to healthy women [39]. The relative risk for myocardial infarction is 7.4 for PCOS patients [40]. Radial pulse wave analysis can provide valuable information on the cardiovascular health of PCOS women.

From the results of this study, we found that the pulse wave parameters $h3/h1$ and w/t from both left and right sides were significantly higher in the PCOS group than in the healthy group, meanwhile, parameters right $t1$, and $h4$, $t4$, t , As , $h4/h1$ from both sides were significantly lower in PCOS group. Higher $h3/h1$ and w/t reflect poorer arterial elasticity or higher peripheral resistance in the PCOS group. Liu et al. found that $h3/h1$ was higher in the mild and severe occlusion group of coronary heart disease (CHD) patients compared to the non-occlusion group [41], which proved that high $h3/h1$ is related to poor vascular compliance, a risk factor of CVD.

Parameter $t1$ corresponds to the rapid ejection period, $t4$ reflects the systolic time of the left ventricle, and t is the time of a cardiac cycle of the left ventricle, therefore lower in $t1$, $t4$ and t indicates a shorter systolic phase and a faster heart rate of PCOS patient. As is the area of the

systolic phase, lower As in the PCOS group suggests a decrease in cardiac output, which may result in insufficient peripheral tissue perfusion. Our result agreed with the finding that suggested PCOS women had lower left ventricular ejection fraction [42]. Parameters $h4$ and $h4/h1$ reflect the peripheral resistance of the artery, lower $h4$, and $h4/h1$ in the PCOS group probably because of decreasing in peripheral blood volume, as a result, the heartbeat increased compensatory. In brief, PCOS patients experienced poorer vascular compliance, faster heart rate, and reduced left ventricular systolic function compared to healthy individuals.

From the feature importance analysis of Random Forest, we noted that the top five features are right $t4$, left w/t , right $t1$, left As , and left t , these features contributed to 46.89% of importance over all 15 features for classification, the features are related to left ventricular function directly or indirectly. This result revealed that the left ventricle function is the prominent factor to distinguish PCOS patients from healthy individuals. The correlation between PCOS patients and left ventricular function is worth further exploration, previous research had also shown that PCOS women are associated with a higher left ventricular mass index and larger left atrial diameter [43]. For the $t1$ differed significantly only on the right wrist, we assumed that this is because the right wrist is more sensitive than the left wrist upon pulse wave detection, a study demonstrated that the predictive power of

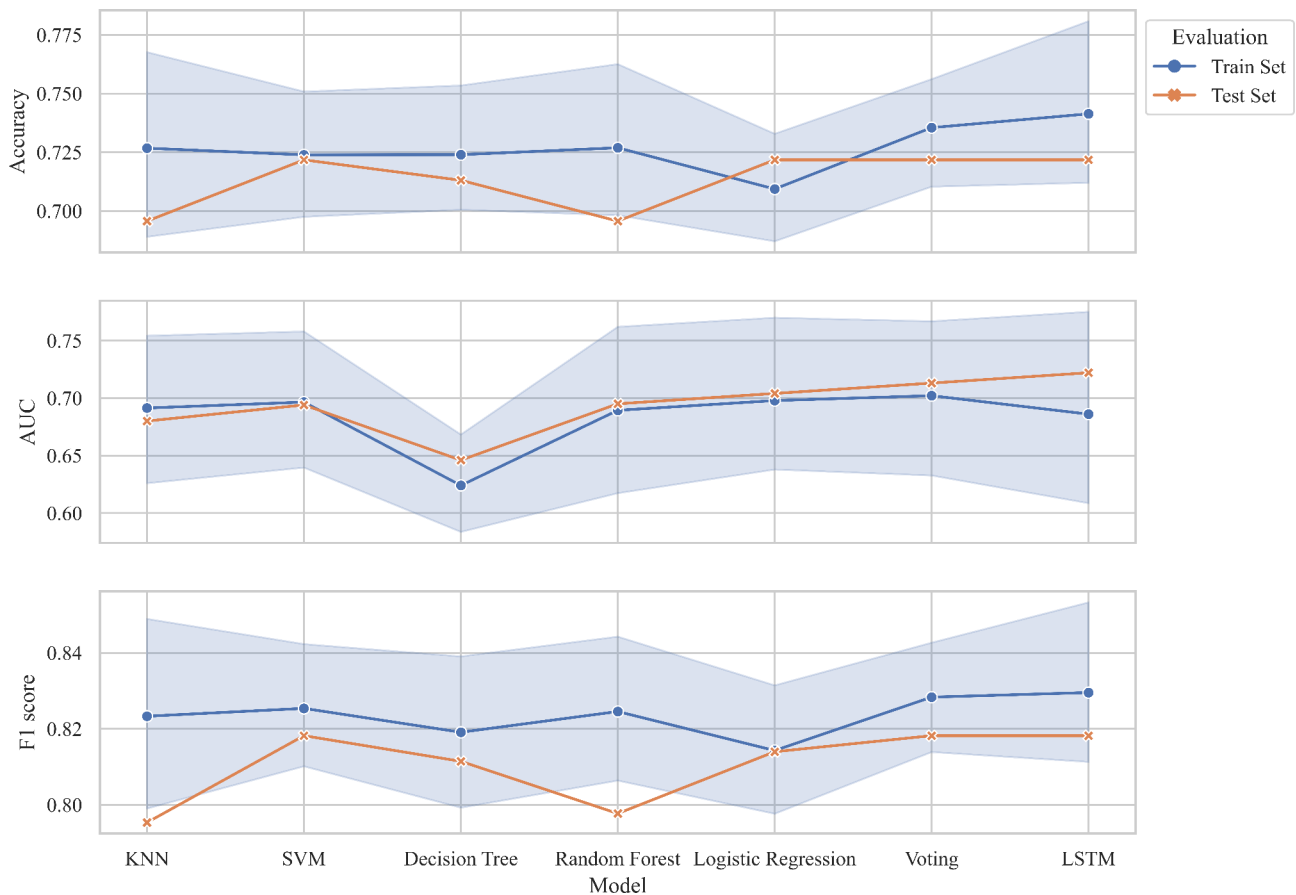


Fig. 11 Model performance comparisons of training results and testing results. Training results came from the training data that was used for both training and testing while testing results came from the ability of the trained model to identify unseen testing data that were not used in training

physical factors from the right wrist was higher than that of the left wrist [44].

During ML modelling, to reduce the data imbalanced effect, we applied a stratified k-fold cross-validation method. Then, the models were tuned by the grid search with cross-validation to get the best parameters. The cross-validations were repeated 10 times to ensure reliable results. Seven models were trained sequentially including KNN, SVM, Decision Trees, Random Forest, Logistic Regression, Voting, and LSTM. Rui Guo et al. found that SVM with Gaussian radial basis function is an effective tool for solving pattern recognition and function estimation problems and is suitable for pulse wave analysis [45]. On the other hand, Ding et al. found that Logistic Regression achieved the most satisfactory result among others in waveform classification [46], Logistic Regression classifier is consistent with the physiological process of the pulse wave. Su et al. proved that Random Forest could obtain higher accuracy in disease prediction due to its bootstrap aggregation and randomization of predictors, and it is less prone to overfitting [25]. To improve the model performance, we applied ensemble learning methods by using a Voting classifier, which

made decision based on the highest average probability given to certain class from the first five models. LSTMs are suitable for capturing the temporal patterns present in the pulse wave parameter data studied [47, 48].

Our findings showed that these prediction models indicated similar performance in classifying the PCOS group and healthy group overall, but the Voting and LSTM did surpass the others. LSTM, a deep learning model known for its ability to capture temporal dependencies, demonstrated competitive performance across all metrics. It was the best-performing model on testing data, which achieved an accuracy of 72.174%, an AUC of 0.722, and an F1 score of 0.818. The Voting ensemble, composed of base models KNN, SVM, Decision Trees, Random Forest, and Logistic Regression, also performed remarkably well. It achieved an accuracy of 72.174%, an AUC of 0.715, and an F1 score of 0.818. Notably, both have the best performance across all training and testing metrics, showcasing their potential for effective modelling. The F1 scores from LSTM and Voting are considered at a good level, which means that the models can predict most of the PCOS cases and be accurate with them. However, the values of accuracy and AUC are only moderate and far

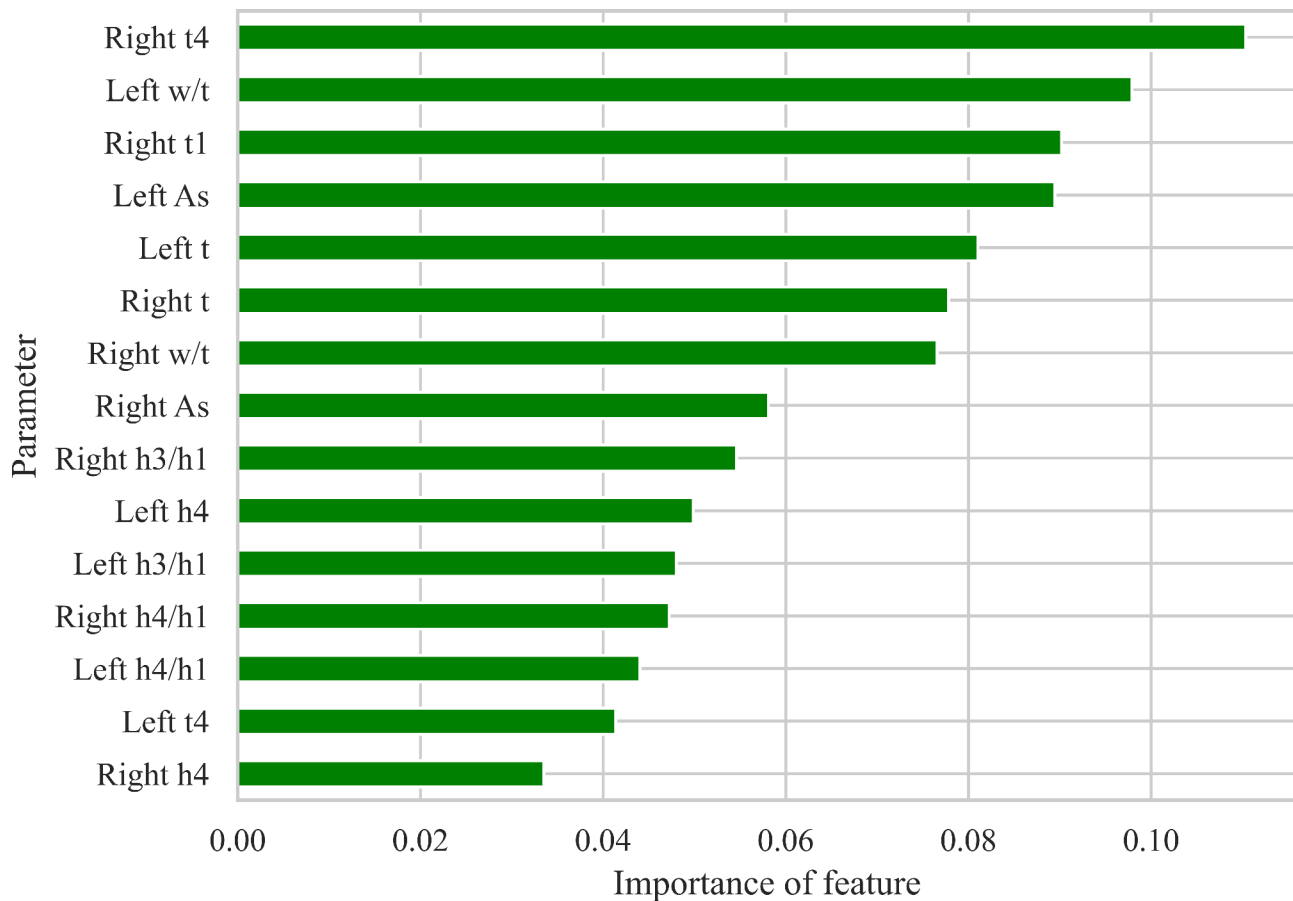


Fig. 12 Feature importance of 15 features based on Random Forest

from excellent compared to the other ML disease classifiers (featuring pulse parameters) as mentioned previously. The lower accuracy and AUC may be caused by the limitations including relatively small sample sizes ($n=459$), and an imbalanced dataset (70% PCOS cases: 30% healthy cases).

In overall, the performance metrics on the testing data are generally slightly lower than those on the training data. This is expected as models tend to perform slightly worse on unseen data. None of the models seems to show significant overfitting, as the drop in performance from training to testing data is relatively small, and overfitting would occur only when a model performs well on training data but poorly on testing data.

Considering computational efficiency and simplicity, Voting might be the preferred option for model selection. It is important to note that LSTM, as a deep learning model, is a more computationally expensive model compared to Voting. LSTM training involves complex backpropagation through time steps, and hyperparameter tuning can contribute to its computational cost. On the other hand, Voting comprising simpler base models generally requires less computation.

The correlation between pulse diagnosis and PCOS was again clarified in this study. By comparing the different prediction models, the results could provide a reference for other clinical research. This study is subject to several limitations, comprising (1) even though the participants were instructed to remain calm and refrain from eating or drinking 30 min prior to the test, it is still possible for biases to be introduced due to factors such as patient tremors, respiration, mechanical vibrations from instruments, and power frequency interference; (2) although we debugged the equipment to a certain extent and tried our best to ensure a noise-free environment during data collection, the current level of science and technology still has certain limitations to fully distinguish noise from pulse signals; (3) a relatively small sample size, we used ensemble methods, cross-validation, and hyperparameter tuning to overcome the limitation caused by small sample size to a certain extent, and the results were verified that there was no overfitting. However, we will expand the sample size in the future to provide more reliable assessment results. In the follow-up study, we will increase sample sizes, balance the dataset, integrate pulse wave analysis with frequency domain parameters, and investigate the relationship between pulse conditions

with different PCOS phenotypes. To increase the prediction ability in a way of TCM diagnostics, we suggest: (1) the diversifying of the pulse wave analysis methods; (2) the integration of the pulse data with tongue data and/or TCM symptoms.

In conclusion, there were significant differences in radial pulse waves between PCOS patients and healthy individuals. ML classification based on pulse wave analysis could identify most PCOS patients accurately with good F1 score and is valuable for early detection and monitoring of PCOS with acceptable overall accuracy. Voting classifier and LSTM with ensemble learning ability gave the best model performance among others. This radial pulse wave-based ML prediction method can stimulate the development of individualized PCOS risk assessment using mobile detection technology, with the advantages of being simple, convenient, non-invasive, and cost-effective. Nonetheless, this study gives physicians an intuitive understanding of the objective pulse diagnosis in TCM.

Abbreviations

AUC	Area under the Receiver Operating Characteristic curve
BMI	Body mass index
CART	Classification and Regression Trees
CHD	Coronary heart disease
CVD	Cardiovascular disease
DT	Decision Trees
FN	False negatives
FP	False positives
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long Short Term Memory networks
ML	Machine learning
PCOM	Polycystic ovarian morphology
PCOS	Polycystic ovary syndrome
RBF	Radial basis function
RF	Random Forest
RNN	Recurrent neural network
ROC	Receiver Operating Characteristic
SHUTCM	Shanghai University of Traditional Chinese Medicine
SV	Soft Voting
SVM	Support Vector Machine
TCM	Traditional Chinese Medicine
TN	True negatives
TP	True positives
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12906-023-04249-5>.

Supplementary Material 1

Acknowledgements

Not applicable.

Authors' contributions

Z.X. and J.Lim provided the research conceptualization. J.Li. provided research methodology. L.F. and X.F. acquired the data. J.Lim analyzed the data and perform machine learning, was a major contributor in writing the manuscript.

All authors edited and revised the manuscript and approved the final version. Y.W. in charged of funding acquisition.

Funding

This research was supported by the Shanghai Key Laboratory of Health Identification and Assessment (21DZ2271000).

Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to ethical concern but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All experimental protocols were approved by the institutional review board of Shuguang Hospital affiliated with SHUTCM (2018-623-52-01). All research methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 21 March 2023 / Accepted: 7 November 2023

References

- Azziz R. Reproductive endocrinology and infertility: clinical expert series polycystic ovary syndrome. *Obstet Gynecol.* 2018;132:321–36.
- Hoeger KM, Dokras A, Piltonen T. Update on PCOS: consequences, challenges, and Guiding Treatment. *J Clin Endocrinol Metab.* 2021;106:e1071–83.
- Zore T, Joshi Nv, Lizneva D, Azziz R. Polycystic ovarian syndrome: long-term Health consequences. *Semin Reprod Med.* 2017;35:271–81.
- Teede H, Deeks A, Moran L. Polycystic ovary syndrome: a complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan. *BMC Med.* 2010;8.
- Riesterberg C, Jagasia A, Markovic D, Buyalos RP, Azziz R. Health Care-Related Economic Burden of Polycystic Ovary Syndrome in the United States: pregnancy-related and long-term Health consequences. *J Clin Endocrinol Metab.* 2022;107:575–85.
- Haixia Y, Yiqin W. Research and clinical application of pulse diagnosis in traditional Chinese medicine. 1st ed. Shanghai: Shanghai Scientific & Technical Publishers; 2016.
- Yiqin W. Diagnostics of Traditional Chinese Medicine. 2nd ed. Beijing: Higher Education Press; 2016.
- Matos LC, Machado JP, Monteiro FJ, Greten HJ. Can Traditional Chinese Medicine diagnosis be parameterized and standardized? A narrative review. *Healthc (Basel).* 2021;9.
- Velik R. An objective review of the technological developments for radial pulse diagnosis in traditional Chinese medicine. *Eur J Integr Med.* 2015;7:321–31.
- Hsu CL, Hsiu H, Hsu WC, Chen CY. Characteristics of harmonic indexes of the arterial blood pressure waveform in polycystic ovary syndrome. *Blood Press Monit.* 2014;19:226–32.
- Wang W, Zeng W, Chen X, Tu L, Xu J, Yin X. Parameter study on characteristic pulse diagram of polycystic ovary syndrome based on logistic regression analysis. *J Obstet Gynaecol (Lahore).* 2022. <https://doi.org/10.1080/01443615.2022.2158317>.
- Feng X, Feng L, Gao H, Wang QS, Xia YM, Xu ZX et al. Characteristics of Pulse Parameters in Patients with Polycystic Ovary Syndrome Varied at Different Body Mass Index Levels. *Evid Based Complement Alternat Med.* 2022;2022.
- Zhang X, Liang B, Zhang J, Hao X, Xu X, Chang HM et al. Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening. *Mol Cell Endocrinol.* 2021;523.

14. Xie NN, Wang FF, Zhou J, Liu C, Qu F. Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network. *Biomed Res Int.* 2020;2020.
15. Ho CH, Chang CM, Li HY, Shen HY, Lieu FK, Wang PSG. Dysregulated immunological and metabolic functions discovered by a polygenic integrative analysis for PCOS. *Reprod Biomed Online.* 2020;40:160–7.
16. Suha SA, Islam MN. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. *Sci Rep.* 2022;12.
17. Cheng JJ, Mahalingaiah S. Data mining polycystic ovary morphology in electronic medical record ultrasound reports. *Fertil Res Pract.* 2019;5.
18. Nasim S, Almutairi MS, Munir K, Raza A, Younas F. A Novel Approach for Polycystic Ovary Syndrome Prediction using machine learning in Bioinformatics. *IEEE Access.* 2022;10:97610–24.
19. Denny A, Raj A, Ashok A, Ram CM, George R. I-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques. *IEEE Region 10 Annual International Conference, Proceedings/TENCON.* 2019;2019-October:673–8.
20. Tiwari S, Kane L, Koundal D, Jain A, Alhudaif A, Polat K et al. SPOSDS: a smart polycystic ovary syndrome diagnostic system using machine learning. *Expert Syst Appl.* 2022;203.
21. Shi YL, Liu JY, Hu XJ, Tu LP, Cui J, Li J et al. A New Method for Syndrome Classification of Non-Small-Cell Lung Cancer Based on Data of Tongue and Pulse with Machine Learning. *Biomed Res Int.* 2021;2021.
22. Zhang Z, Zhang Y, Yao L, Song H, Kos A. A sensor-based wrist pulse signal processing and Lung cancer recognition. *J Biomed Inform.* 2018;79:107–16.
23. Hao Y, Cheng F, Pham M, Rein H, Patel D, Fang Y et al. A noninvasive, economical, and instant-result method to diagnose and monitor type 2 Diabetes using pulse Wave: case-control study. *JMIR Mhealth Uhealth.* 2019;7.
24. Zhang CK, Liu L, Wu WJ, Wang YQ, Yan HX, Guo R et al. Identifying Coronary Artery Lesions by Feature Analysis of Radial Pulse Wave: A Case-Control Study. *Biomed Res Int.* 2021;2021.
25. Su X, Xu Y, Tan Z, Wang X, Yang P, Su Y et al. Prediction for Cardiovascular Diseases based on laboratory data: an analysis of random forest model. *J Clin Lab Anal.* 2020;34.
26. Endocrinology Subgroup and Expert Panel CS of O and GCMA. Chinese guideline for diagnosis and management of polycystic ovary syndrome. *Chin J Obstet Gynecol.* 2018;53:2–6.
27. Fei ZF. *Pulse Diagnosis of Modern Traditional Chinese Medicine*. 1st edition. Beijing: People's Medical Publishing House; 2003.
28. Biswas A, Saran I, Wilson FP. *Introduction to Supervised Machine Learning* *Kidney360.* 2021;2:878–80.
29. Zhi Hua Z. *Machine learning*. Beijing: Beijing Tsinghua University Press; 2016.
30. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A primer for the epidemiologist. *Am J Epidemiol.* 2019;188:2222–39.
31. Kotsiantis SB. Decision trees: a recent overview. *Artif Intell Rev.* 2013;39:261–83.
32. Chatterjee S, Byun YC. Voting Ensemble Approach for Enhancing Alzheimer's Disease classification. *Sens (Basel).* 2022;22.
33. Oruh J, Viriri S, Adegun A. Long short-term memory recurrent neural network for Automatic Speech Recognition. *IEEE Access.* 2022;10:30069–79.
34. Shuang K, Li R, Gu M, Loo J, Su S. Major-minor long short-term memory for Word-Level Language Model. *IEEE Trans Neural Netw Learn Syst.* 2020;31:3932–46.
35. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9:1735–80.
36. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* 2006;27:861–74.
37. Song XH, Wang Y. Pulse wave: the bridge connecting traditional Chinese medicine with western medicine. *J Chin Integr Med.* 2008;6:891–6.
38. Wu CH, Chiu LT, Chang YJ, Lee CI, Lee MS, Lee TH, et al. Hypertension risk in Young Women with Polycystic Ovary Syndrome: a Nationwide Population-based Cohort Study. *Front Med (Lausanne).* 2020;7:585.
39. Guleria AK, Syal SK, Kapoor A, Kumar S, Tiwari P, Dabadghao P. Cardiovascular Disease risk in young Indian women with polycystic ovary syndrome. *Gynecol Endocrinol.* 2014;30:26–9.
40. Dahlgren E, Janson PO, Johansson S, Lapidus L, Odén A. Polycystic ovary syndrome and risk for Myocardial Infarction. Evaluated from a risk factor model based on a prospective population study of women. *Acta Obstet Gynecol Scand.* 1992;71:599–604.
41. Lu L, Chunke Z, Jianjun Y, Rui G, Yiqin W, Haixia Y, et al. Analysis of characteristics of pulse-graph parameters in patients with different degree of coronary artery occlusion. *J Beijing Univ Traditional Chin Med.* 2022;45:835–41.
42. Orio F, Palomba S, Spinelli L, Cascella T, Tauchmanová L, Zullo F, et al. The cardiovascular risk of young women with polycystic ovary syndrome: an observational, analytical, prospective case-control study. *J Clin Endocrinol Metab.* 2004;89:3696–701.
43. Wang ET, Ku IA, Shah SJ, Daviglius ML, Schreiner PJ, Konety SH, et al. Polycystic ovary syndrome is Associated with higher left ventricular Mass Index: the CARDIA women's study. *J Clin Endocrinol Metab.* 2012;97:4656.
44. Lee BJ, Jeon YJ, Ku B, Kim JU, Bae JH, Kim JY. Association of Hypertension with physical factors of wrist pulse waves using a computational approach: a pilot study. *BMC Complement Altern Med.* 2015;15.
45. Guo R, Wang YQ, Yan HX, Li FF, Yan JJ, Xu ZX, et al. Feature extraction and recognition of traditional Chinese medicine pulse based on hemodynamic principles. *Zhong Xi Yi Jie He Xue Bao.* 2010;8:742–6.
46. Ding X, Cheng F, Morris R, Chen C, Wang Y. Machine learning-based Signal Quality evaluation of single-period radial artery pulse waves: Model Development and Validation. *JMIR Med Inform.* 2020;8.
47. Mou H, Yu J. Cnn-Lstm prediction method for blood pressure based on pulse wave. *Electron (Switzerland).* 2021;10.
48. Huang L, Yan J, Cai S, Guo R, Yan H, Wang Y. Automated Segmentation of the Systolic and Diastolic Phases in Wrist Pulse Signal Using Long Short-Term Memory Network. *Biomed Res Int.* 2022;2022.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.