Algorithm for Diagnosis of Metabolic Syndrome and Heart Failure Using CPET Biosignals via SVM and Wavelet Transforms

Rafael Fernandes Pinheiro[1(✉)] and Rui Fonseca-Pinto[1,2]

[1] Center for Innovative Care and Health Technology (ciTechCare), Polytechnic of Leiria, Rua de Santo André - 66-68, 2415-736 Leiria, Portugal
{rafael.f.pinheiro,rui.pinto}@ipleiria.pt
[2] School of Health Sciences (ESSLei), Polytechnic of Leiria, Campus 2 - Morro do Lena, Alto do Vieiro - Apartado 4137, 2411-901 Leiria, Portugal

Abstract. Early diagnosis of diseases is essential to avoid health complications and costs to the health system. For this purpose, algorithms have been widely used in the medical field to assist in the diagnosis of diseases. This work proposes an algorithm with a new approach to analyze biosignals from cardiopulmonary exercise testing (CPET) to identify metabolic syndrome (MS) and heart failure (HF). The algorithm uses the support vector machine (SVM) as a classification technique and wavelet transforms for extraction of the features. For training, CPET data from 30 volunteers were used, of which 15 are diagnosed with MS and 15 with HF. The SVM-L-W approach, which uses wavelet transforms, has been shown to have better accuracy (93%) compared to some other approaches found in the literature. In addition, the SVM-L-W algorithm can be applied to identify other diseases, and is also adaptable to modifications in order to obtain better performance, as suggested in future work to continue this research.

Keywords: Classification algorithms · Biosignals · CPET · Metabolic diseases · Heart diseases · Wavelet transforms

1 Introduction

Both metabolic syndrome (MS) and heart failure (HF) are public health issues of major global relevance. MS involves several metabolic risk factors, such as obesity and type 2 diabetes, increasing the risk of cardiovascular disease. More than 15 years ago, studies [1] already showed that the prevalence of MS was between 20% and 25% worldwide and was already considered one of the most

This work was funded by Portuguese national funds provided by Fundação para a Ciência e Tecnologia (FCT) in the scope of the research project 2 ARTs - Acessing Autonomic Control in Cardiac Rehabilitation (PTDC/EMD-EMD/6588/2020).

P. J. Coelho et al. (Eds.): GOODTECHS 2023, LNICST 556, pp. 159–171, 2024.
https://doi.org/10.1007/978-3-031-52524-7_12

common chronic diseases, ranking fourth in the list of leading causes of death worldwide. On the other hand, HF is when the heart does not pump enough blood to fulfil the body's needs, causing severe symptoms. Heart failure is increasing in prevalence, with approximately 26 million patients affected worldwide [5]. Prevention, early diagnosis and appropriate treatment are key to mitigating the impact of these conditions on society.

Cardiopulmonary Exercise Testing (CPET) is a test that assesses the body's response to exercise by combining cardiovascular, respiratory and metabolic analysis. It provides important information for diagnosis, prognosis and therapeutic planning in various medical conditions, including cardiac [17], metabolic [16] and pulmonary [14]. The interpretation of the data is done by health professionals and helps in the assessment of cardiorespiratory capacity, diagnosis of diseases and optimization of physical training.

In contrast, the interpretation of CPET data is based on a thorough analysis of the variables recorded during the test. Currently, interpretation follows guidelines and criteria established by medical and exercise physiology societies, as well as by scientific studies that provide references for understanding the results (basically the flowchart is used - see [13] and [10]). In this line, it is understood that the interpretation of CPET data for diagnosis is not yet a closed subject, and with the improvement of artificial intelligence techniques, new methods and algorithms have emerged to help doctors provide more accurate diagnoses and therapeutic plans.

In the field of artificial intelligence, classification algorithms have been used for the development of diagnostic algorithms, the most common are in the areas of machine learning and artificial neural networks. In the area of machine learning, for a training dataset of approximately 70 CPET files for each disease, [12] shows that the support vector machine (SVM) technique is an excellent tool for classifying diseases such as heart failure and chronic obstructive pulmonary disease, reaching accuracy levels of 100%. In the area of artificial neural networks, the work of Brown et al. [6] can be highlighted, which develops hybrid models using convolutional neural networks (CNN) and autoencoders (AE) with principal component analysis (PCA) and logistic regression (LR) for the classification of heart failure and metabolic syndrome with a set of 15 CPET files for each disease. The methodologies of these works are very different, however, both show very effective results, especially the one that uses a very small dataset.

In this paper, it is presented an algorithm for diagnosis of MS and HF, based on supervised learning, with two new approaches for analyzing CPET data. Diagnostic algorithms are of great value to medicine because they can prevent severe health damage and death. The training of this algorithm is performed using the same dataset of [6]. The simplest approach uses SVM classificator with kernel linear, the features being the means and variances of the CPET data (algorithm called SVM-L). The second version is a linear kernel SVM classifier algorithm that employs a different methodology to obtain the features and uses the means and variances of the coefficients of the Daubechies wavelets of order 2 with 3 levels (algorithm called SVM-L-W).

The main contributions of this paper are highlighted below:

- According to the literature search conducted by the authors, there are no previous works dealing with the development of algorithms for disease diagnosis from CPET data that combine the use of SVM with wavelets.
- The use of wavelet transforms to prepare the data and obtain the features from the CPET data is presented in this work as a very efficient alternative to reduce the computational cost. This technique allows to drastically reduce the dimension of the features used in the classification algorithms for analyzing CPET data. The authors believe that this approach is the greatest contribution of the work, presenting an efficient algorithm with low computational cost compared to [12] and [6].
- The work presents an algorithm with better accuracy compared to other algorithms that use CNN, PCA, LR and flowchart, proving to be able to compete with the AE+LR technique of [6]. In this sense, future work is proposed, which aims to train the SVM-L-W algorithm with artificial data for its improvement and with more features extracted from the CPET variables. This future work may show that SVM-L-W surpasses AE+LR in accuracy.
- Since the feature dimension reduction technique is essential for biosignals from the brain (e.g. electroencephalography data), due to the huge amount of information, it is understood that the SVM-L-W algorithm can be effectively used in the diagnosis of neurological and psychiatric diseases.

Figure 1 shows the illustrative design that covers all phases of the process, from data collection by the CPET to diagnosis after processing the patient data by the SVM-L-W algorithm. In the figure, the interrogation signs represent the authors' intention to continue with this investigation to design a more comprehensive algorithm that provides diagnostics in several other diseases from the CPET data collection.

The rest of the paper is developed as follows: in the Sect. 2, the main theoretical bases for the development of the algorithms are presented; in the Sect. 3, information about the data and the creation of the features are brought; the Sect. 4, deals with the construction of the algorithms; the results regarding the performance of the algorithm are presented in the Sect. 5; finally, the work ends with the conclusions, Sect. 6, where a summary of the findings is made and future work is proposed.

2 Theoretical Basis

This section presents the main theoretical basis for the development of the algorithm.

2.1 Support Vector Machine

Support Vector Machine (SVM), proposed by Vladimir Vapnik [4], is a supervised machine learning method of classification and regression that seeks to find

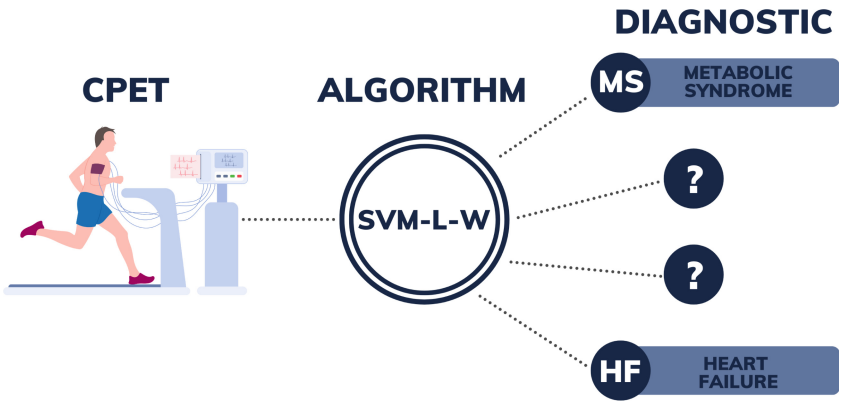

Fig. 1. Illustration of the diagnostic system from CPET data.

the optimal hyperplane in multidimensional spaces to separate two classes of data. Its effectiveness stems from the process of maximizing the margin between the support vectors, which are the closest points to the decision boundaries. This approach allows for robust generalization even in complex, high-dimensional datasets, making it a prominent choice in various data analysis and pattern recognition applications. Its variants, what is called the SVM kernel, include linear, polynomial, sigmoidal and RBF. Linear SVM is used to separate linearly separable classes, while polynomial and sigmoidal SVM are applied to non-linearly separable datasets using transformations. RBF SVM is highly powerful, mapping the data to a high-dimensional space to separate complex classes.

2.2 Wavelet Transforms

Wavelet transforms are a powerful mathematical tool widely used in signal and image analysis. Unlike traditional transforms, wavelets offer a multiresolution approach, allowing to capture both local and global signal information efficiently. Wavelets have stood out as a promising approach in classification algorithms [9,18]. By applying the wavelet technique in data analysis, it is possible to extract relevant features from different scales and frequencies, allowing a richer representation of the patterns present in the datasets. This ability to identify discriminative information at multiple resolutions has led to the development of more accurate and robust classification models in several areas, such as pattern recognition, image processing and medical diagnostics.

When applying a wavelet transform, the signal is decomposed into levels (d_1, d_2, d_3, etc.) representing details at different frequencies. The coefficients in these decompositions reflect the contributions of each level in the overall representation of the original signal. These wavelet coefficients allow a detailed analysis of the signal at various resolutions. In this work, it was used the Daubechies wavelet of order 2 with 3 levels (d_1, d_2, d_3,) and an approximation, both for the CPET variables under consideration presented in the next section.

2.3 Validation Process

For the validation of the algorithm, the cross validation k-fold was used (see more in [21]). The cross validation k-fold is a technique used in machine learning that is very useful when you have small datasets, which is the case of this work. The technique involves dividing the dataset into k subsets (folds) of approximately equal sizes. The model is trained k times, where, in each iteration, one of the subsets is used as a test set and the remaining $k-1$ subsets are used for training. At the end of the k iterations, the results are combined to produce a single metric for evaluating model performance. This allows for a more accurate and robust assessment of model performance, using all available data efficiently and avoiding biases in the assessment.

3 Dataset

In this section, it was presented the origin of the data used and the methodology for creating the features for use in the diagnostic algorithm, as well as the labels.

3.1 Data Source

The CPET dataset used to create the algorithm was obtained from a public database[1] which has been used in several other works [2,3,6–8,11,15]. The data on MS come from a study supported by the National Institute of Health/National Heart Lung and Blood Institute (NIH/NHLBI), "Exercise dose and metformin for vascular health in adults with metabolic syndrome" and the HF data came from patient studies supported by the American Heart Association, "Personalized Approach to Cardiac Resynchronization Therapy Using High Dimensional Immunophenotyping," as well as the NIH/NHLBI, "MRI of Mechanical Activation and Scar for Optimal Cardiac Resynchronization Therapy Implementation." In the data, a total of 30 individuals were sampled, 15 of them with a diagnosis of MS and another 15 with HF.

CPET provides a wide variety of information extracted from the patient during the test. In this work, for the creation of the features, the CPET variables were used according to Table 1.

3.2 Features and Labels

Features are the information taken from the data for training the classification algorithms. Labels are the classifications (names) given to a set of features. For example, consider the heart rate (HH) and respiratory rate (RR) data of a given patient. You can use the variables themselves (HH and RR) as features, this usually gives a large dataset. However, for computational gain, one can extract parameters from these variables, for example the mean. Thus, the features for this patient will be the mean HH and RR. On the other hand, the labels are the

[1] https://github.com/suchethassharma/CPET.

Table 1. Variables for the creation of features.

Description	Feature
Metabolic equivalents	$METS$
Heart Rate	HR
Peak oxygen consumption	$\dot{V}O_2(L/min)$
Volume of carbon dioxide released	$\dot{V}CO_2(L/min)$
Respiratory exchange ratio	RER
Ventilation	$VE(L/min)$
Expiratory tidal volume (expiratory time)	$Vtex(L)$
Inspiratory tidal volume (inhale time)	$Vtin(L)$

classifications given to the patient connected to their features, for example, if it is a non-diabetic patient, it is given the label 0 and if it is a diabetic patient it is given label 1. Therefore, a set of data from several patients is the information used to train the algorithm. The greater the number of patients for training, the better the algorithm's ability to determine a disease. More content on feature extraction can be found at [19,22].

For this work, the features were obtained from the CPET variables presented in the previous subsection. Two sets of features were used. The first set, called X, consists of the mean and variance of each variable in Table 1, with the first 15 lines corresponding to data from patients with HF and the other 15 lines with data from patients with MS. The X feature was constructed by organizing the data using Excel software functions where the means and variances were also extracted. Table 2 shows how the data are presented to the algorithm. The table shows that the matrix X has a dimension of 30 rows with 16 columns.

Table 2. Features with mean and variance.

	METS		HR		...	Vtin	
	Mean	Var	Mean	Var	...	Mean	Var
1	2.639	0.285	113.68	193.79	$\cdots$	1.079	0.187
2	4.075	0.704	102.613	81.259	$\cdots$	1.778	0.128
3	2.647	0.264	124.911	130.344	$\cdots$	1.716	0.125
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
30	5.074	1.708	131.645	439.237	$\cdots$	1.643	0.13

The second set of features, called W, contains the mean and variance of the wavelet transform coefficients at three levels (d_1, d_2 and d_3) obtained from the CPET variables. The algorithm presented in the next section is implemented to obtain the matrix of features W, with the first 15 rows corresponding to data

from patients with HF and the other 15 rows with data from patients with MS. Figure 2 shows the structure of the matrix W returned by the algorithm. It can be seen from the figure that the matrix W has a dimension of 30 rows with 64 columns.

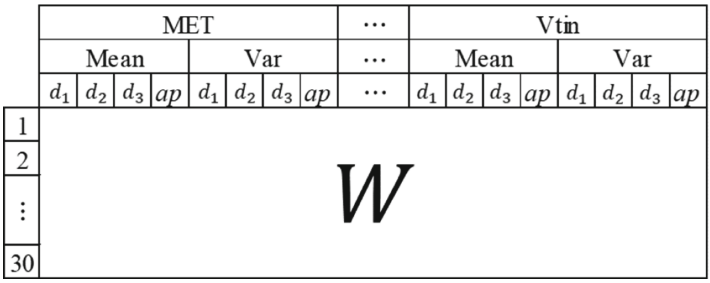

Fig. 2. Features using wavelets provided by the algorithm.

On the label, the value 0 (zero) was used to represent HF and the value 1 (one) to represent MS. These values were inserted in a vector, called Y, of dimension 30, with the first 15 elements of values 0 and the other 15 of values 1.

4 The Algorithm

Diagnostic algorithms have been widely used in the medical field to support the diagnosis of diseases. These types of algorithms generally use artificial intelligence techniques and are built using a database for training. Considering a supervised learning algorithm, it is trained with a database that provides the data of people with a certain disease, and this disease receives a type of classification. Thus, when the model is ready, i.e., after training, the algorithm is able to identify the type of disease when a new patient's data is presented.

The algorithm for diagnosis developed in this work uses Matlab functions specific to machine learning projects, and its training is done using the database presented in the previous section. Two approaches were used to obtain the results. In both cases, the Matlab reference code [20] was used to construct the confusion matrices.

The first development approach, called SVM-L, which is simpler, receives the feature X and labels Y, performs the classification via SVM, cross-validates and prints the confusion matrix. The second approach, called SVM-L-W, differs from the first in the creation of the feature. This version receives the data from the CPET variables (Table 1) and creates the W feature, i.e., it applies the wavelet transforms and obtains their coefficients, and then obtains the mean and variance creating the W matrix.

Figure 3 illustrates the methodology for the development of the algorithm with the two approaches and presents the Matlab functions used. It can be

observed that in the SVM-L-W approach, it is necessary to create only the SVM with Linear kernel, since the results of the first approach showed that the SVM classifier with linear kernel has better accuracy (see next section).

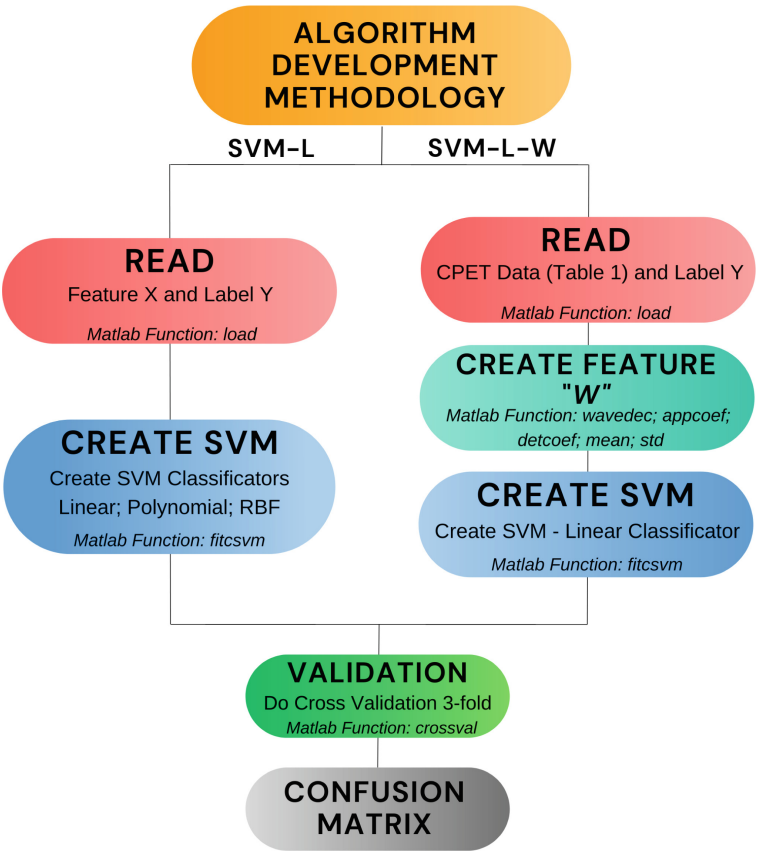

Fig. 3. Algorithm development methodology.

5 Results

In this section, the performance results of the algorithm for the SVM-L and SVM-L-W approaches are presented. Then, comparisons are made with some existing results in the literature.

In this type of work, some evaluation metrics are used to validate the algorithms, the most common are: accuracy, precision, recall and F1-score. The formula for each metric is obtained from the confusion matrix (see Fig. 4). In this

paper, only the accuracy metric is taken into account. Accuracy shows the overall performance of the model, indicating, among all diagnostics, how many the model indicated correctly.

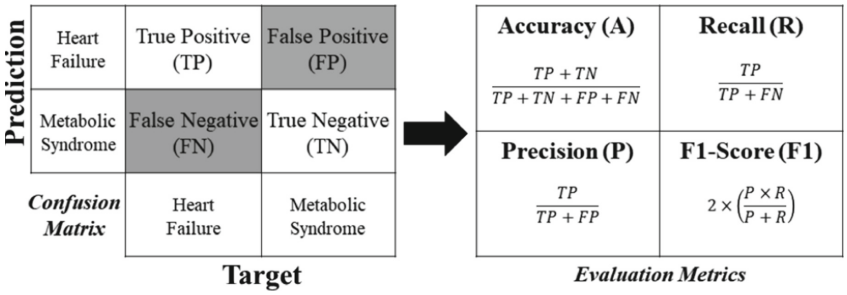

Fig. 4. Formulas of the evaluation metrics.

5.1 SVM-L Algorithm

This version of the algorithm uses the X feature with linear kernel SVM. Figure 5 shows the confusion matrix for each classifier. With emphasis on accuracy, it is observed that the linear classifier performs better.

5.2 SVM-L-W Algorithm

This second version of the algorithm uses the W feature with the linear kernel SVM model that showed better accuracy in the previous version. It should be noted that the W feature is constructed by the algorithm from the means and variances of the wavelet transform coefficients with the biosignal data provided by CPET (Table 1). Figure 6 shows the confusion matrix for the SVM-linear classifier with wavelet transforms.

5.3 Comparisons

The basic method used to interpret CPET results is a flowchart. According to [6], flowcharts have been used to interpret CPET results for more than 30 years. Here, the guidelines FRIEND [13] and of Hansen, et al. [10] were used. In addition to the flowchart method, the results of this work were compared with those obtained from Brown, et al. [6] who propose methods that utilize neuralnetwork

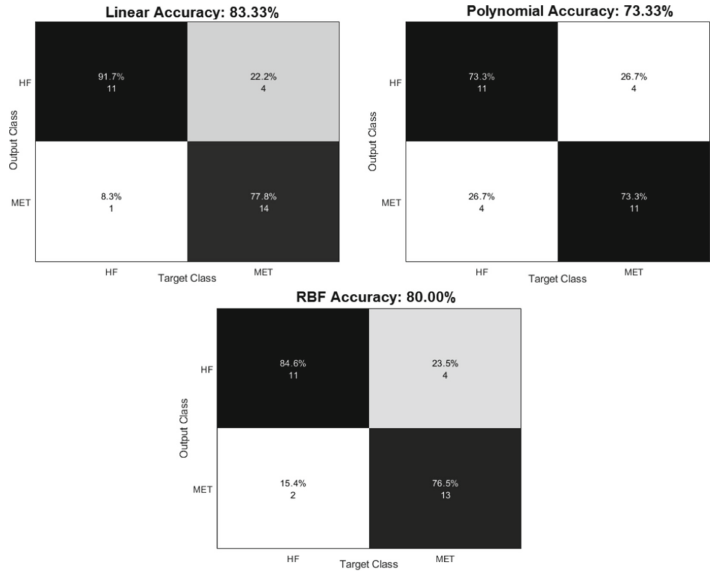

Fig. 5. Confusion matrices.

techniques (AE and CNN), PCA and LR. Table 3 shows the accuracy obtained in each of the methods.

Table 3. Comparisons with other methods.

Method	Accuracy (%)
AE + LR [6]	97
SVM-L-W	93
CNN [6]	90
PCA + LR [6]	90
SVM-L	83
Flowchart [13]	77
Flowchart [10]	70

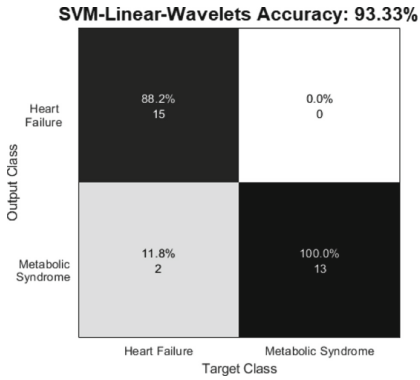

Fig. 6. Confusion matrix of the approach SVM-L-W.

6 Conclusion

This work presented a new approach to build an algorithm that uses CPET data for diagnosis of metabolic syndrome and heart failure. The algorithm, called SVM-L-W, presented an accuracy of 93%, ranking better than some other algorithms found in the literature. However, in this first investigation, SVM-L-W presented lower accuracy compared to AE+LR of [6], in this case, one believes that in the future, the SVM-L-W could overcome AE+LR with some adaptations in terms the use of more features and artificial data for training. It is also considered that the SVM-L-W algorithm can be applied to determine other types of pathologies that may be related to biosignals from CPET, given proper training with an adequate database.

For future works, it is suggested to use this algorithm to analyze electroencelography signals to assist in the diagnosis of neurological and psychiatric diseases. Also, one intends to continue the development of the algorithm by investigating the use of artificial data for training and a larger number of features, in order to improve its accuracy, as well as designing it to provide diagnostics of a larger number of diseases.

References

1. Alberti, K.G.M.M., Zimmet, P., Shaw, J.: Metabolic syndrome-a new world-wide definition: a consensus statement from the international diabetes federation. Diab. Med. **23**(5), 469–480 (2006)
2. Auger, D.A., et al.: Reproducibility of global and segmental myocardial strain using cine dense at 3t: a multicenter cardiovascular magnetic resonance study in healthy subjects and patients with heart disease. J. Cardiovasc. Magn. Reson. **24**(1), 1–12 (2022)
3. Bilchick, K.C., et al.: CMR dense and the seattle heart failure model inform survival and arrhythmia risk after CRT. Cardiovasc. Imaging **13**(4), 924–936 (2020)

4. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152 (1992)
5. Bowen, R.E., Graetz, T.J., Emmert, D.A., Avidan, M.S.: Statistics of heart failure and mechanical circulatory support in 2020. Ann. Transl. Med. **8**(13) (2020)
6. Brown, D.E., Sharma, S., Jablonski, J.A., Weltman, A.: Neural network methods for diagnosing patient conditions from cardiopulmonary exercise testing data. BioData Mining **15**(1), 16 (2022)
7. Gaitán, J.M., Eichner, N.Z., Gilbertson, N.M., Heiston, E.M., Weltman, A., Malin, S.K.: Two weeks of interval training enhances fat oxidation during exercise in obese adults with prediabetes. J. Sports Sci. Med. **18**(4), 636 (2019)
8. Gao, X., et al.: Cardiac magnetic resonance assessment of response to cardiac resynchronization therapy and programming strategies. Cardiovasc. Imaging **14**(12), 2369–2383 (2021)
9. Guo, T., Zhang, T., Lim, E., Lopez-Benitez, M., Ma, F., Yu, L.: A review of wavelet analysis and its applications: challenges and opportunities. IEEE Access **10**, 58869–58903 (2022)
10. Hansen, D., et al.: Exercise training intensity determination in cardiovascular rehabilitation: should the guidelines be reconsidered? Eur. J. Prev. Cardiol. **26**(18), 1921–1928 (2019)
11. Heiston, E.M., et al.: Two weeks of exercise training intensity on appetite regulation in obese adults with prediabetes. J. Appl. Physiol. **126**(3), 746–754 (2019)
12. Inbar, O., Inbar, O., Reuveny, R., Segel, M.J., Greenspan, H., Scheinowitz, M.: A machine learning approach to the interpretation of cardiopulmonary exercise tests: development and validation. Pulmonary Med. **2021**, 1–9 (2021)
13. Kaminsky, L.A., Imboden, M.T., Arena, R., Myers, J.: Reference standards for cardiorespiratory fitness measured with cardiopulmonary exercise testing using cycle ergometry: data from the fitness registry and the importance of exercise national database (friend) registry. In: Mayo Clinic Proceedings, vol. 92, pp. 228–233. Elsevier (2017)
14. Luo, Q., et al.: The value of cardiopulmonary exercise testing in the diagnosis of pulmonary hypertension. J. Thorac. Dis. **13**(1), 178 (2021)
15. Malin, S.K., Gilbertson, N.M., Eichner, N.Z., Heiston, E., Miller, S., Weltman, A., et al.: Impact of short-term continuous and interval exercise training on endothelial function and glucose metabolism in prediabetes. J. Diab. Res. **2019** (2019)
16. Rodriguez, J.C., Peterman, J.E., Fleenor, B.S., Whaley, M.H., Kaminsky, L.A., Harber, M.P.: Cardiopulmonary exercise responses in individuals with metabolic syndrome: the ball state adult fitness longitudinal lifestyle study. Metab. Syndr. Relat. Disord. **20**(7), 414–420 (2022)
17. Saito, Y., et al.: Diagnostic value of expired gas analysis in heart failure with preserved ejection fraction. Sci. Rep. **13**(1), 4355 (2023)
18. Serhal, H., Abdallah, N., Marion, J.M., Chauvet, P., Oueidat, M., Humeau-Heurtier, A.: Overview on prediction, detection, and classification of atrial fibrillation using wavelets and AI on ECG. Comput. Biol. Med. **142**, 105168 (2022)
19. Subasi, A.: EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst. Appl. **32**(4), 1084–1093 (2007)
20. Tshitoyan, V.: Plot confusion matrix (2023). https://github.com/vtshitoyan/plotConfMat. Accessed 27 July 2023

21. Wong, T.T., Yeh, P.Y.: Reliable accuracy estimates from k-fold cross validation. IEEE Trans. Knowl. Data Eng. **32**(8), 1586–1594 (2019)
22. Xing, Z., Pei, J., Yu, P.S., Wang, K.: Extracting interpretable features for early classification on time series. In: Proceedings of the 2011 SIAM International Conference on Data Mining, pp. 247–258. SIAM (2011)